NCVS Status and Progress Report
Volume 10/November 1996

The National Center for Voice and Speech is a consortium of institutions--The University of Iowa, The Denver Center for the Performing Arts, The University of Wisconsin-Madison and The University of Utah--whose investigators are dedicated to the rehabilitation, enhancement and protection of voice and speech.

Editorial and Distribution Information

Editor, Ingo Titze
Production Editors, Julie Lemke and Julie Ostrem
Technical Editor, Martin Milder

Distribution of this report is not restricted.
However, production was limited to 700 copies.

Correspondence should be addressed as follows:
Editor, NCVS Status and Progress Report
The University of Iowa
330 Wendell Johnson Building
Iowa City, Iowa 52242
(319) 335-6600
FAX (319) 335-8851
e-mail titze@shc.uiowa.edu

Primary Sponsorship

The National Institute on Deafness and Other Communication Disorders,
Grant Number P60 DC00976

Other Sponsorship

The University of Iowa
Department of Speech Pathology and Audiology
Department of Otolaryngology - Head and Neck Surgery

The Denver Center for the Performing Arts
Wilbur James Gould Voice Research Center
Department of Public Relations
Department of Public Affairs
Denver Center Media
Department of Development

The University of Wisconsin-Madison
Department of Communicative Disorders
Department of Surgery, Division of Otolaryngology
Waisman Center
Department of Electrical and Computer Engineering

The University of Utah
Department of Otolaryngology - Head and Neck Surgery
LDS Hospital

The University of Illinois
Department of Speech and Hearing Science

NCVS Personnel

Administration

Central Office
Ingo Titze, Director
Julie Ostrem, Program Associate
Julie Lemke, Secretary

Area Coordinators
Research - Ingo Titze
Training - Patricia Zebrowski
Continuing Education - Julie Ostrem
Information Dissemination - Cynthia Kintigh

Advisory Board

Katherine Harris, Ph.D.
Minoru Hirano, M.D.
Clarence Sasaki, M.D.
Johan Sundberg, Ph.D.

Investigators, Affiliates and Support Staff

Fariborz Alipour, Ph.D.
Kristin Baker, Ph.D.
Bridget Berning, B.S.
David Berry, Ph.D.
Florence Blager, Ph.D.
Diane Bless, Ph.D.
James Brandenburg, M.D.
Myrna Burt
John Butler, M.D.
John Canady, M.D.
Kelly Cavanaugh, B.A.
Geron Coale, M.A.
Stefanie Countryman, M.A.
Linda D'Antonio, Ph.D.
Charles Davis, Ph.D.
Wendy Edwards, B.A.
Jeffrey Fields, B.M.
John Folkins, Ph.D.
Charles Ford, M.D.
Curt Freed, M.D.
Amy Furness
Steven Gray, M.D.
Elizabeth Hammond, M.D.
Kelle Hasenberg, B.A.
Marilyn Hetzel, Ph.D.
Margaret Hoehn, M.D.
Henry Hoffman, M.D.
Bruce Jafek, M.D.
Darin Johnson, B.S.
Joel Kahane, Ph.D.
Michael Karnell, Ph.D.
Judith King, Ph.D.
Cynthia Kintigh, M.A.
David Kuehn, Ph.D.
Jennifer Lehnherr
Julie Lemke
Russel Long, M.S.
Erich Luschei, Ph.D.
Kathryn Maes, Ph.D.
Martin Milder, B.S.
Paul Milenkovic, Ph.D.
Jerald Moon, Ph.D.
Carrie Nachtwey, B.S.
John Nichols, B.A.
Chris O'Brien, M.D.
Lorraine Olson Ramig, Ph.D.
Julie Ostrem, B.S.
Namrata Patil, M.D.
Annette Pawlas, M.A.
Kathe Perez, M.A.
Donald Robin, Ph.D.
Robin Samlan, M.S.
Ronald Scherer, Ph.D.
Richard Schmidt, Ph.D.
Suzanne Segal, Ph.D.
Elaine Smith, Ph.D.
Marshall Smith, M.D.
Elaine Stathopoulos, Ph.D.
Brad Story, Ph.D.
Linda Suckow, B.A.
Edie Swift, M.S.
Laetitia Thompson, Ph.D.
Sue Thompson, Ph.D.
Ingo Titze, Ph.D.
Lou Tomes, Ph.D.
Vern Vail, B.S.
Katherine Verdolini, Ph.D.
Jennifer Waldron, B.S.
Patricia Ward
Barbara Williams
Darrell Wong, Ph.D.
Raymond Wood, M.D.
George Woodworth, Ph.D.
Patricia Zebrowski, Ph.D.
Jane Zernicke, B.S.
Lynn Zimba, Ph.D.

Doctoral Students

Todd Brennan, M.S.
Roger Chan, B.S.
Eileen Finnegan, M.A.
Elisa Mordue, M.S.
John Nelson, M. Aud.
Phyllis Palmer, M.A.
Annie Ramos, M.S.
Nelson Roy, M.S.
Helen Sharp, M.S.

Postdoctoral Fellows

Michael Edgerton, D.M.A.
Katsuhide Inagi, M.D.
Aliaa Khidr, Ph.D.

Visiting Scholars

Hanspeter Herzel, Ph.D., Germany
Patrick Mergell, M.S., Germany

Contents

Part I. Research papers submitted for peer review in archival journals

Part II. Tutorial reports and updates

Forward

We are extremely pleased to report that the Department of Speech Pathology and Audiology at the University of Iowa has embarked on a joint graduate student training and exchange program with Howard University in Washington, DC. Several graduate students and faculty members will exchange visits and work on joint research projects.

We are indebted to Professor Richard Hurtig, Chairman of the Department of Speech Pathology and Audiology, for initiating the exchange and writing the grant application. We are also indebted to the National Institute on Deafness and Other Communication Disorders for funding the project and appending it to our Center grant.

More will be reported about this new venture in our next progress report.

Ingo R. Titze, Director
November, 1996

Part I

Research papers submitted for peer review in archival journals

NCVS Status and Progress Report - 10
November 1996, 1-7

A Three-Dimensional Solution of the Acoustic Wave Equation in a Model of Vocal Tract

Fariborz Alipour, Ph.D.
Brad H. Story, Ph.D.
Chenwu Fan, M.S.
Department of Speech Pathology and Audiology, The University of Iowa

Abstract

The wave equation was solved in a cylindrical coordinate system for models of the vocal tract corresponding to vowels, /a/, /i/, and /u/. A straight cylindrical model of vocal tract was built upon area function data for each vowel. Using boundary fitted coordinates, the vocal tract shape and its boundary conditions were simplified to a straight tube. However, this simplification in geometry resulted in a complicated wave equation in the new coordinate system. The transformed wave equation was discretized in space over a 90x21 grid and solved in time using a finite difference method. The results indicate that pressure contours are typically planar in the narrow regions and nonplanar in the wider sections. The results of the model have been compared and validated against the open-open and closed-open uniform tubes with good accuracy. The predicted frequency response of this model was also compared with a wave-reflection type of model.

Introduction

The sound of human speech result from vocal tract enhancement and/or suppression of various regions of a source spectrum. The vocal tract shape is, of course, produced by the relative positioning of the articulators such as the tongue, lips, velum, etc., and for any given positioning of the articulators, the resultant vocal tract shape can be thought of as continuously changing three-dimensional duct. However, most theoretical representations of the vocal tract shape have assumed one-dimensional wave propagation and consequently reduce the three-dimensional acoustic duct to a one-dimensional representation, typically called an "area function". The one-dimensional approximation has been used with much success to simulate human speech sounds (Fant, 1960; Ishizaka & Flanagan, 1972; Liljencrants, 1985; Story, 1995). The combination of a one-dimensional model and electrical analogy has been a strong tool in simulating the time-varying vocal tract and digital simulation of speech (i.e. Portnoff, 1973; Maeda, 1982). However, a one-dimensional approximation to the acoustic wave propagation, by definition, prevents the propagation of any acoustic mode other than the plane wave mode. For most voiced sounds such an approximation is probably quite reasonable. But any frication noise such as that used for the production of fricative consonants as well as any high frequency (>3500 Hz) sound associated with the voiced sound may not be adequately represented by the one-dimensional approximation. Thus, for some analyses of the vocal tract acoustics, the restriction of one-dimensional wave propagation may need to be lifted.

Two- or three-dimensional wave propagation in the non-uniform or time-varying ducts has not been studied extensively and only a few studies can be named. For example Ling (1976) developed a Galerkin based finite element method to solve the Helmholtz equation in variable cross-sectional area ducts and reported some applications of his method such as the sound field in a bottle-like duct, in a exponential horn, and in a convergent-divergent duct with airflow. Also, Astley and Eversman (1978) developed a finite element method to study the transmission of sound in non-uniform ducts. Using an eight-node isoparametric elements, they solved the Helmholtz equation in Cartesian and Cylindrical coordinates. They reported reflection and transmission coefficients and compared they results with the method of weighted residuals.

The only detailed study that used modern computational methods to solve the wave equation in the vocal tract models is Lu (1993). He solved the Helmholtz equation with the 2-D and 3-D finite element method. His calculations of amplitude and phase of the sound field were compared with the measured data of Motoki et al. (1988) and showed some deviation from the plane wave propagation for the vowel /a/

. He also incorporated a yielding wall effect but noticed that its effect could be omitted at frequencies above 1 kHz.

In this article a new method is presented that is capable of solving the three-dimensional wave equation of the pressure for any vocal tract shape and can provide the time-varying sound field, wave propagation properties, and resonance frequencies. At this stage, the model does not include any type of energy loss. Thus, the walls are hard and sound is not radiated at the lips. Additionally, a 3-D, axisymmetric geometry of the vocal tract was constructed based on the one-dimensional area functions data of Story (1995). The results of a simplified vocal tract such as this can be directly compared to a 1-D model for validation and refinement.

Method

The wave equation that describes the acoustic wave propagation in the vocal tract is

$$\nabla^2 p - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0 \qquad (1)$$

where p is pressure, t is time, and c is speed of sound. Although this equation is linear, the geometry of the vocal tract is very complex, thus applying the boundary conditions are very difficult. For variable cross-sectional area ducts, an assumption of uniform pressure over the cross section could lead to a quasi one-dimensional version of wave equation which is known as Webster Horn equation (Pierce, 1981). The next logical improvement over the one-dimensional model would be a two-dimensional model. Although, a two-dimensional model (third dimension is assumed infinitely long) easily defines the geometry in the midsagittal plane, the effects of cross-sectional area can not be included with it. Thus a simplified three-dimensional model is needed. One typical simplification is straightening the vocal tract while preserving the area function. This straightening does not change the resonant frequencies more than few percent (Sondhi, 1986). With such an assumption and using circular cross sections, the vocal tract is replaced with a straight horn where area function is preserved. In this geometry, a cylindrical coordinate system will be suitable especially if a rotational symmetry is chosen, the wave equation then becomes:

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = \frac{\partial^2 p}{\partial r^2} + \frac{1}{r}\frac{\partial p}{\partial r} + \frac{\partial^2 p}{\partial z^2} \qquad (2)$$

Now that the wave equation has been simplified to be a two-dimensional equation, a coordinate transformation is introduced that simplifies the geometry. As shown in Figure 1, a body-fitted coordinate (BFC) system in a complicated physical domain (Fig. 1A) helps to solve the problem in a simple logical domain (Fig. 1B); i.e. the area function is mapped into an effective uniform tube. A grid generation is needed to discretize the problem for the curvilinear coordinate system. For simplicity we replace the (z,r) with (x,y) symbols.

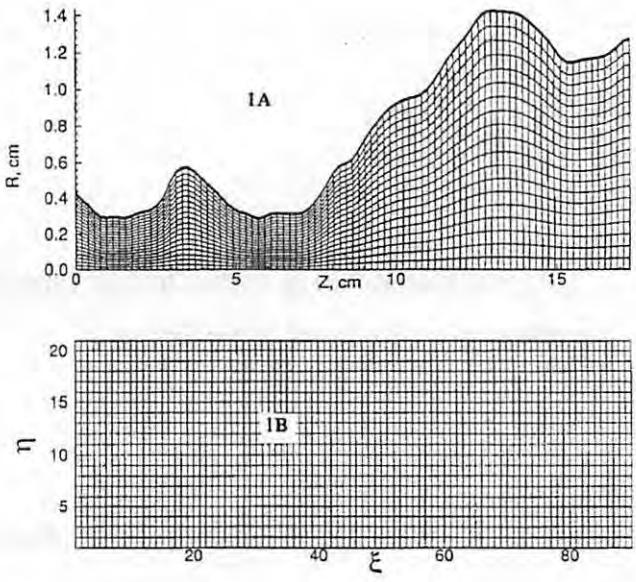

Figure 1. Computational grids in the physical domain (A) and logical space (B).

A one-to-one transformation from (x,y) space to (ξ, η) with non-zero Jacobian (Knupp and Steinberg, 1993) can be written as,

$$x = x(\xi, \eta),\ y = y(\xi, \eta) \qquad (3)$$

$$d\xi = \frac{y_\eta \cdot dx - x_\eta \cdot dy}{J}, \qquad d\eta = \frac{x_\xi \cdot dy - y_\xi \cdot dx}{J} \qquad (4)$$

where the subscripts stand for partial differentiation and $J = x_\xi \cdot y_\eta - y_\xi \cdot x_\eta$ is the Jacobian. The partial derivatives in the transformed coordinate system can be obtained using the above differential relations as:

$$\xi_x = \frac{y_\eta}{J}, \quad \xi_y = -\frac{x_\eta}{J}, \quad \eta_x = -\frac{y_\xi}{J}, \quad \eta_y = \frac{x_\xi}{J} \qquad (5)$$

then calculate the Laplacian in Cartesian coordinates as:

$$\begin{aligned}\nabla^2 f &= f_{xx} + f_{yy} \\ &= \frac{x_\eta^2 + y_\eta^2}{J^2}\cdot f_{\xi\xi} - 2\frac{x_\xi \cdot x_\eta + y_\xi \cdot y_\eta}{J^2}\cdot f_{\xi\eta} + \frac{x_\xi^2 + y_\xi^2}{J^2}\cdot f_{\eta\eta} \\ &+ \frac{1}{J^3}\left[(x_\eta \cdot y_{\xi\xi} - y_\eta \cdot x_{\xi\xi})(x_\eta^2 + y_\eta^2) + 2(y_\eta \cdot x_{\xi\eta} - x_\eta \cdot y_{\xi\eta})(x_\xi \cdot x_\eta + y_\xi \cdot y_\eta)\right. \\ &+ \left.(x_\eta \cdot y_{\eta\eta} - y_\eta \cdot x_{\eta\eta})(x_\xi^2 + y_\xi^2)\right]\cdot f_\xi \\ &+ \frac{1}{J^3}\left[(y_\xi \cdot x_{\xi\xi} - x_\xi \cdot y_{\xi\xi})(x_\eta^2 + y_\eta^2) + 2(x_\xi \cdot y_{\xi\eta} - y_\xi \cdot x_{\xi\eta})(x_\xi \cdot x_\eta + y_\xi \cdot y_\eta)\right. \\ &+ \left.(y_\xi \cdot x_{\eta\eta} - x_\xi \cdot y_{\eta\eta})(x_\xi^2 + y_\xi^2)\right]\cdot f_\eta\end{aligned} \qquad (6)$$

Using $\alpha = x^2_\eta + y^2_\eta$; $\beta = x_\xi \cdot x_\eta + y_\xi \cdot y_\eta$; $\gamma = x^2_\xi + y^2_\xi$; the Laplacian of f becomes:

$$\nabla^2 f = \frac{\alpha}{J^2} f_{\xi\xi} - 2\frac{\beta}{J^2} f_{\xi\eta} + \frac{\gamma}{J^2} f_{\eta\eta} + \frac{1}{J^3}(x_\eta \cdot L_y - y_\eta \cdot L_x) f_\xi + \frac{1}{J^3}(y_\xi \cdot L_x - x_\xi \cdot L_y) f_\eta \quad (7)$$

where: $L_x = \alpha \cdot x_{\xi\xi} - 2\beta \cdot x_{\xi\eta} + \gamma \cdot x_{\eta\eta}$,
$L_y = \alpha \cdot y_{\xi\xi} - 2\beta \cdot y_{\xi\eta} + \gamma \cdot y_{\eta\eta}$,

Finally, the Laplacian is written as:

$$\nabla^2 f = a_{11} \cdot f_{\xi\xi} + 2a_{12} \cdot f_{\xi\eta} + a_{22} \cdot f_{\eta\eta} + a_1 \cdot f_\xi + a_2 \cdot f_\eta \quad (8)$$

where:

$$a_{11} = \frac{\alpha}{J^2}; \quad a_{12} = -\frac{\beta}{J^2}; \quad a_{22} = \frac{\gamma}{J^2};$$

$$a_1 = \frac{1}{J}(x_\eta \cdot \psi_y - y_\eta \cdot \psi_x); \quad a_2 = \frac{1}{J}(y_\xi \cdot \psi_x - x_\xi \cdot \psi_y) \quad (9)$$

and

$$\psi_x = a_{11} \cdot x_{\xi\xi} + 2a_{12} \cdot x_{\xi\eta} + a_{22} \cdot x_{\eta\eta}; \quad \psi_y = a_{11} \cdot y_{\xi\xi} + 2a_{12} \cdot y_{\xi\eta} + a_{22} \cdot y_{\eta\eta}.$$

Now if we consider the wave equation in the new coordinate system we have:

$$\frac{\partial^2 P}{\partial^2} = c^2\left(a_{11} \cdot \frac{\partial^2 P}{\partial \xi^2} + 2a_{12} \cdot \frac{\partial^2 P}{\partial \xi \partial \eta} + a_{22} \cdot \frac{\partial^2 P}{\partial \eta^2} + a_1 \cdot \frac{\partial P}{\partial \xi} + a_2 \cdot \frac{\partial P}{\partial \eta}\right) \quad (10)$$

where the coefficients a_1 and a_2 are replaced later with a'_1 and a'_2 for the cylindrical coordinates as given by:

$$a'_1 = a_1 + y_\eta / yJ, \quad a'_2 = a_2 - y_\xi / yJ \quad (11)$$

Grid Generation

Grid generation is a common practice in modern computational fluid dynamics and usually is done numerically by an inverse transformation of coordinates. A detailed procedure may be found in Knupp and Steinberg (1993). Here a brief description of the technique is given. Since the logical space is rectangular, a uniform mesh in the ξ and η directions is assumed and by defining the range and the number of divisions in the ξ and η directions, the coordinates of the grid points in logical space are become known (Figure 1B). The task is now to find the coordinates of the corresponding grid points in the physical domain. This is done by a Poisson equation solver which uses the boundaries of the physical domain for its numerical calculations. As an example of grid generation, the area function for the vowel /a/ of a male subject from Story et al. (1996) was chosen and from it the "radius function" was calculated. This function is based on the assumption of the circular cross-sections and straightened vocal tract. Prior to grid generation, the digitized data were smoothed by a five-term linear filter (Hamming, 1973). After grid calculations, one fine grid was added near the wall to facilitate the application of the wall boundary condition. A typical calculated BFC grid is shown in Figure 1A for the vowel /a/ which has 90x21 nodes in x and y directions.

Numerical Solution

After the coordinate transformation, the solution domain is discretized into finite difference meshes, both in time and space as:

$$\frac{P^{n-1}_{i,j} - 2P^n_{i,j} + P^{n+1}_{i,j}}{c^2 \cdot \Delta t^2 / 2} = a_{11}(P^{n-1}_{i-1,j} - 2P^{n-1}_{i,j} + P^{n-1}_{i+1,j} + P^{n+1}_{i-1,j} - 2P^{n+1}_{i,j} + P^{n+1}_{i+1,j}) +$$
$$0.5a_{12}(P^{n-1}_{i-1,j-1} - P^{n-1}_{i-1,j+1} - P^{n-1}_{i+1,j-1} + P^{n-1}_{i+1,j+1} + P^{n+1}_{i-1,j+1} - P^{n+1}_{i-1,j-1} - P^{n+1}_{i+1,j+1} + P^{n+1}_{i+1,j-1})$$
$$+ a_{22}(P^{n-1}_{i,j-1} - 2P^{n-1}_{i,j} + P^{n-1}_{i,j+1} + P^{n+1}_{i,j-1} - 2P^{n+1}_{i,j} + P^{n+1}_{i,j+1}) \quad (12)$$
$$+ 0.5a_1(P^{n-1}_{i+1,j} - P^{n-1}_{i-1,j} + P^{n+1}_{i+1,j} - P^{n+1}_{i-1,j}) + 0.5a_2(P^{n-1}_{i,j+1} - P^{n-1}_{i,j-1} + P^{n-1}_{i,j+1} - P^{n-1}_{i,j-1})$$

These equations must be solved at every time step in space (for all the nodes). To avoid time consuming matrix inversion, an alternating direction implicit (ADI) iterative method was used (Hoffmann & Chiang, 1993). In each iteration, the resulting tridiagonal system was solved for all the rows and then followed by the columns.

Let:

$$s_{i,j} = a_{11}(P^{n-1}_{i-1,j} - 2P^{n-1}_{i,j} + P^{n-1}_{i+1,j}) + 0.5a_{12}(P^{n-1}_{i-1,j-1} - P^{n-1}_{i-1,j+1} - P^{n-1}_{i+1,j-1} + P^{n-1}_{i+1,j+1})$$
$$+ a_{22}(P^{n-1}_{i,j-1} - 2P^{n-1}_{i,j} + P^{n-1}_{i,j+1}) + 0.5a_1(P^{n-1}_{i+1,j} - P^{n-1}_{i-1,j}) + 0.5a_2(P^{n-1}_{i,j+1} - P^{n-1}_{i,j-1})$$
$$+ \frac{(4P^n_{i,j} - 2P^{n-1}_{i,j})}{c^2 \cdot \Delta t^2} \quad (13)$$

then in the ξ direction:

$$(0.5a_1 - a_{11})P^{n+1}_{i-1,j} + 2(a_{11} + a_{22} + \frac{1}{c^2 \cdot \Delta t^2})P^{n+1}_{i,j} - (0.5a_1 + a_{11})P^{n+1}_{i+1,j} = s_{i,j} + \quad (14)$$
$$0.5a_{12}(P^{n+1}_{i-1,j-1} - P^{n+1}_{i+1,j-1} - P^{n+1}_{i-1,j+1} + P^{n-1}_{i+1,j+1}) + a_{22}(P^{n+1}_{i,j-1} + P^{n+1}_{i,j+1}) + 0.5a_2(P^{n+1}_{i,j+1} - P^{n+1}_{i,j-1})$$

and in the η direction:

$$(0.5a_2 - a_{22})P^{n+1}_{i,j-1} + 2(a_{11} + a_{22} + \frac{1}{c^2 \cdot \Delta t^2})P^{n+1}_{i,j} - (0.5a_2 + a_{22})P^{n+1}_{i,j+1} = s_{i,j} + \quad (15)$$
$$0.5a_{12}(P^{n+1}_{i-1,j-1} - P^{n+1}_{i+1,j-1} - P^{n+1}_{i-1,j+1} + P^{n+1}_{i+1,j+1}) + a_{11}(P^{n+1}_{i-1,j} + P^{n+1}_{i+1,j}) + 0.5a_1(P^{n+1}_{i+1,j} - P^{n+1}_{i-1,j})$$

For these equation, (14) is solved implicitly for P in the ξ direction (varying i index) and (15) is solved implicitly in the η direction (varying j index). This process repeated alternatively until a convergence is achieved.

Initial and Boundary Conditions

The solution of the wave equation requires both initial and boundary conditions. The initial boundary conditions are usually specified at the first two time steps, especially when we apply an impulse as an excitation signal. For example, the solution requires specifying the initial pressure in the whole domain at time zero. The boundary conditions are dependent on the type of walls and end conditions of the vocal tract. Usually the vocal tract is assumed to be closed at the glottis and open at the mouth. Additionally, either the pressure at a boundary (Dirichlet type) or its derivative at a boundary (Neumann type) is specified. While Dirichlet type is easy to apply, the Neumann type involves some transformation from the (x,y) to (ξ, η) plane. That is, the $\frac{\partial P}{\partial n}$ in (x,y) plane needs to be calculated in (ξ, η) plane at specified conditions. To calculate the normal derivative, we need to have the expressions for unit vectors. The unit vector of ξ direction is:

$$\vec{n}(\xi) = \frac{\xi_x \vec{i} + \xi_y \vec{j}}{\sqrt{\xi_x^2 + \xi_y^2}} = \frac{y_\eta \vec{i} - x_\eta \vec{j}}{\sqrt{\alpha}}, \tag{16}$$

where $\vec{i}$ and $\vec{j}$ are unit vectors of x and y axes, then the unit vector of h direction is:

$$\vec{n}(\eta) = \frac{\eta_x \vec{i} + \eta_y \vec{j}}{\sqrt{\eta_x^2 + \eta_y^2}} = \frac{-y_\xi \vec{i} + x_\xi \vec{j}}{\sqrt{\gamma}} \tag{17}$$

The gradient of the pressure is:

$$\nabla P = \frac{\partial P}{\partial x}\vec{i} + \frac{\partial P}{\partial y}\vec{j} = \left(\frac{\partial P}{\partial \xi}\cdot\frac{\partial \xi}{\partial x} + \frac{\partial P}{\partial \eta}\cdot\frac{\partial \eta}{\partial x}\right)\vec{i} + \left(\frac{\partial P}{\partial \xi}\cdot\frac{\partial \xi}{\partial y} + \frac{\partial P}{\partial \eta}\cdot\frac{\partial \eta}{\partial y}\right)\vec{j} \tag{18}$$

The normal derivative of pressure can be calculated from the pressure gradient (equation 18) and the normal unit vector (equation 17). In the ξ direction we have:

$$\frac{\partial P}{\partial n(\xi)} = \nabla P \cdot \vec{n}(\xi) = \frac{1}{J\sqrt{\alpha}}\left(\alpha\frac{\partial P}{\partial \xi} - \beta\frac{\partial P}{\partial \eta}\right) = \frac{1}{\sqrt{a_{11}}}\left(a_{11}\frac{\partial P}{\partial \xi} + a_{12}\frac{\partial P}{\partial \eta}\right) \tag{19}$$

and in the η direction

$$\frac{\partial P}{\partial n(\eta)} = \nabla P \cdot \vec{n}(\eta) = \frac{1}{J\sqrt{\gamma}}\left(\gamma\frac{\partial P}{\partial \eta} - \beta\frac{\partial P}{\partial \xi}\right) = \frac{1}{\sqrt{a_{22}}}\left(a_{12}\frac{\partial P}{\partial \xi} + a_{22}\frac{\partial P}{\partial \eta}\right) \tag{20}$$

For example, the normal derivative at the glottis (west face, i=1) is

$$\frac{\partial P}{\partial n^*} = -\frac{\partial P}{\partial \xi} = \frac{1}{\sqrt{a_{11}}}\frac{\partial P}{\partial n} + \frac{a_{12}}{a_{11}}\frac{\partial P}{\partial \eta} = \frac{1}{\sqrt{a_{11}}}\frac{\partial P}{\partial n} + \frac{a_{12}}{2a_{11}}\left(P_{j+1,1} - P_{j-1,1}\right) \tag{21}$$

and at the exit (east face, i=nx)

$$\frac{\partial P}{\partial n^*} = \frac{\partial P}{\partial \xi} = \frac{1}{\sqrt{a_{11}}}\frac{\partial P}{\partial n} - \frac{a_{12}}{a_{11}}\frac{\partial P}{\partial \eta} = \frac{1}{\sqrt{a_{11}}}\frac{\partial P}{\partial n} - \frac{a_{12}}{2a_{11}}\left(P_{j+1,nx} - P_{j-1,nx}\right) \tag{22}$$

This derivative at the vocal tract wall, (the north face in logical domain, j=ny) is

$$\frac{\partial P}{\partial n^*} = \frac{\partial P}{\partial \eta} = \frac{1}{\sqrt{a_{22}}}\frac{\partial P}{\partial n} - \frac{a_{12}}{a_{22}}\frac{\partial P}{\partial \eta} = \frac{1}{\sqrt{a_{22}}}\frac{\partial P}{\partial n} - \frac{a_{12}}{2a_{22}}\left(P_{ny,i+1} - P_{ny,i-1}\right) \tag{23}$$

Using equations 21-23, the Neumann boundary conditions can be applied numerically at every time step.

Validation

To test the accuracy of our numerical method, a straight tube of 17.5 cm length and 4 cm^2 cross-sectional area was used as an initial validation case. The resonance frequency of this tube with open-open and closed-open ends can be calculated analytically. The wave propagation in the tube was solved numerically with this method. A single square pressure pulse of 10 microseconds in width and an amplitude of 20 Pascal was used to excite the tube. Figure 2 shows the frequency spectrum of the pressure signal within the tube with open-open (solid line) and closed-open (dashed line)ends. These spectrum are obtained by a fast Fourier transform of pressure waveform near the end of the tube. The first three resonance of open-open tube were

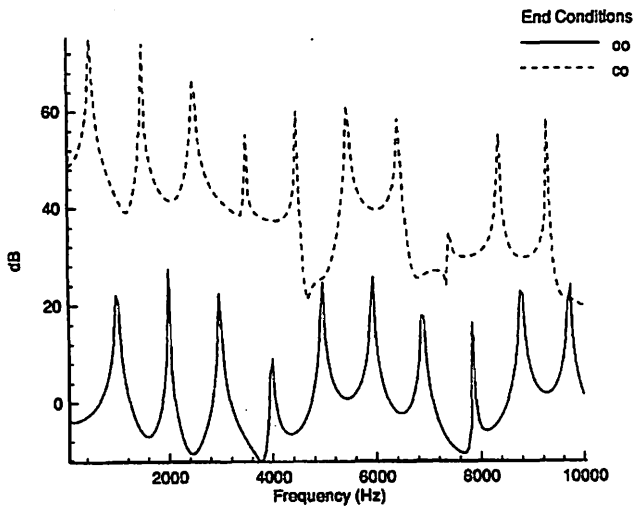

Figure 2. Frequency spectrum of a 17.5 cm straight tube with open-open (solid line) and closed-open (dashed line) end conditions excited by a single square pulse.

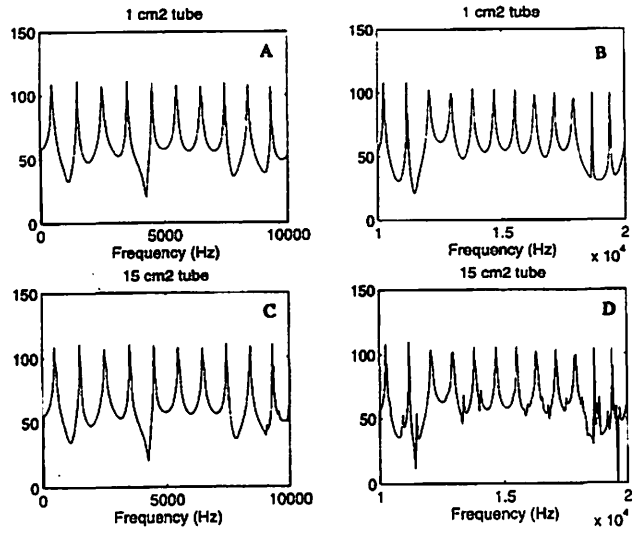

Figure 3. Frequency response of the 1-cm² and 15-cm² tubes due to a square impulse at different range of frequencies.

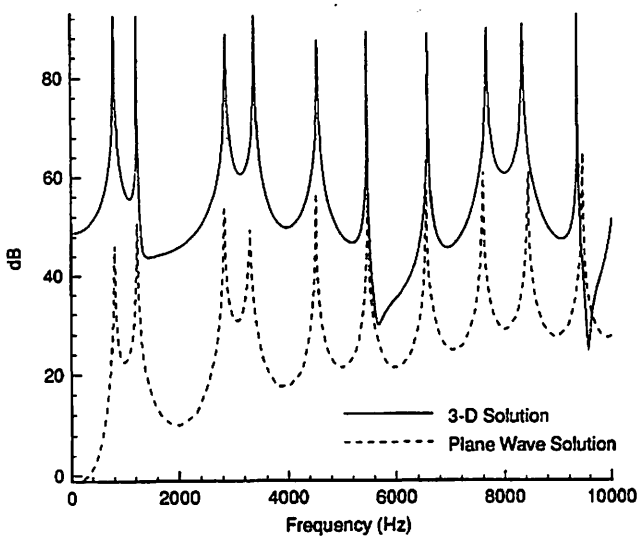

Figure 4. Frequency response functions of vowel /a/ obtained from this method (solid line) and from a one-dimensional wave reflection type model (dashed line).

located at 1000, 2000 and 3000 Hz with a frequency resolution of 40 Hz which indicate a possible uncertainty of better than 4% compared to analytical solution. Similarly in closed-open tube the first three resonance of 520, 1520 and 2520 Hz were obtained with the same resolution.

To test the capability of the model in resolving the nonplanar waves, the diameter of the straight tube was selected at 1 cm² and 15 cm², and impulse response these tubes were obtained with our 3-D BFC method. Figure 3 shows the frequency response of these models. The 1 cm² tube retains its shapes up to 15 kHz as one expect from its small diameter that makes it one-dimensional. But the 15 cm² tube shows some break up at about 9 kHz .

The next validation case is the comparison of formants of the vocal tract for the vowel /a/ with the grid shown in Fig. 1A obtained by this (3-D solution) method and by a one-dimensional wave-reflection type model (Story, 1995). Figure 4 shows the frequency response functions found from this method (solid line) and a wave reflection type (dashed line). Both models used a closed-open end and hard wall boundary conditions with no losses. The first three formants are about 800, 1240, and 2880 Hz from this BFC method and 817, 1231 and 2834 Hz from the other model showing a rather close match. However, some deviations may be observed in the higher formants that could be due to the limitation of the one-dimensional model. It appears that the present method is working properly.

Results

The preliminary results are obtained for a straight tube and vowels /a/, /i/, and /u/ with closed-open ends. At this stage, the model does not include any loss and walls are considered hard. The area functions of the three vowels were obtained from Story (1995) and converted to compu-

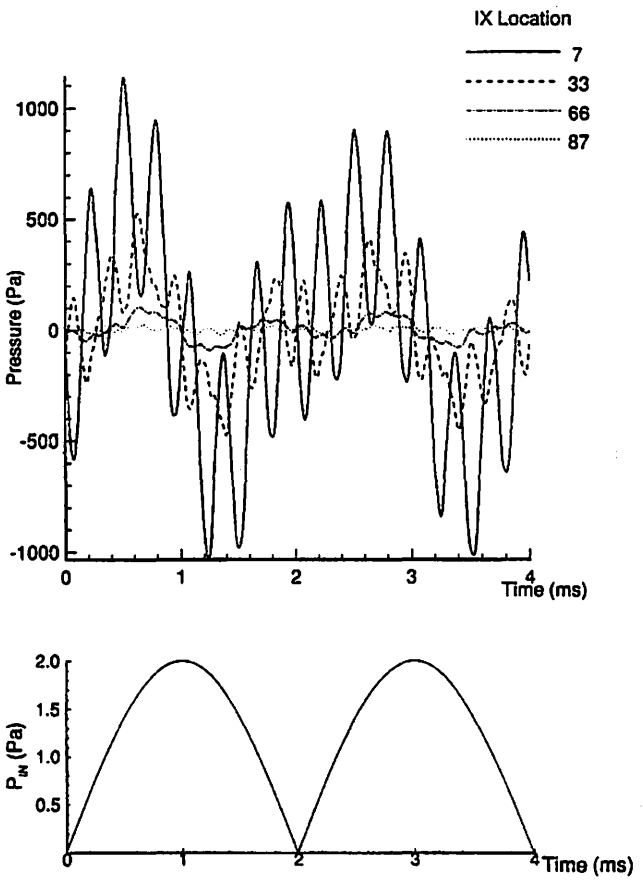

Figure 5. Pressure waveforms in a model of vocal tract for the vowel /a/ at four critical locations (upper graph), excited by a fully rectified sinewave of 500 Hz (lower graph).

tational grids. Each model was excited with a single square pulse as described before and a fully rectified sinewave of 500 Hz. During a 25 ms simulation period, time varying pressure at.a few points within the vocal tract were recorded

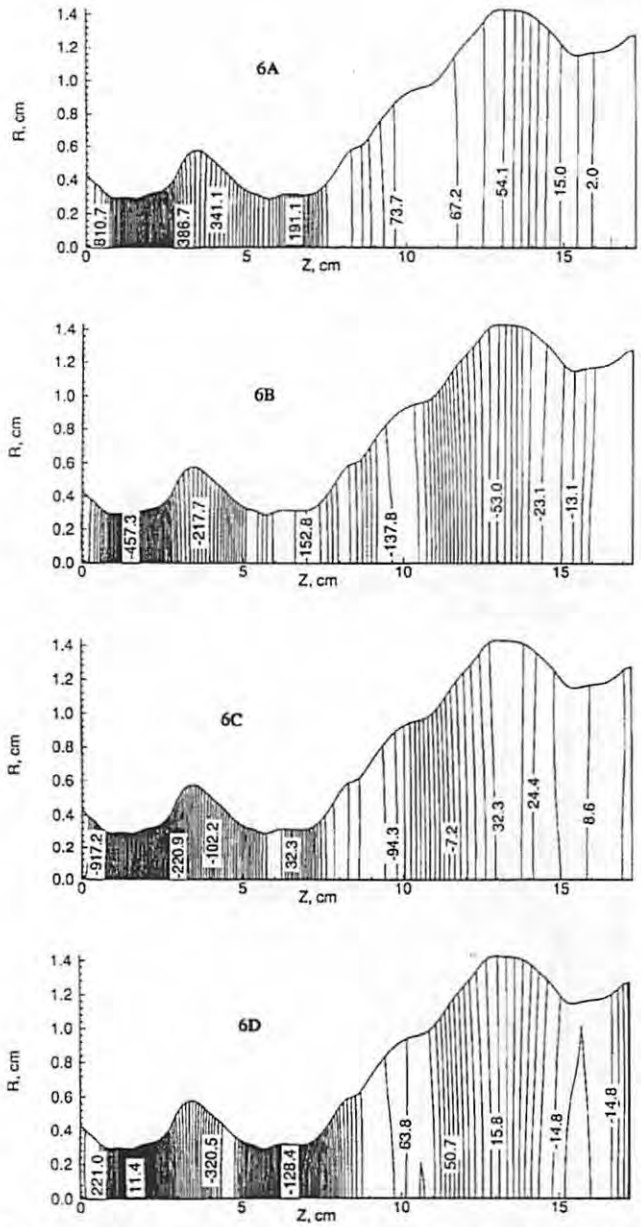

Figure 6. Sound pressure maps in the vocal tract model /a/ at four instants of time (A, B, C, and D). The numbered contour lines indicate the pressure values at that location.

for later analysis. Also, at certain intervals (such as 20 μs), the entire pressure field was retained for the study of wave pattern and propagation animation. To avoid instability and for a better accuracy, a time step of $\Delta t=5$ μs was used which satisfies the relation $\Delta t<\Delta x/c$, where Δx is the smallest grid spacing in the x direction and c is the speed of sound (Hoffmann & Chiang, 1993).

Figure 5 shows the excitation signal and pressure waveforms at the center of tract for the nodal positions, 7, 33, 66, and 87 for the vowel /a/, with low number being closer to the glottis and high number to the mouth. The level of maximum and minimum pressures reach higher values at

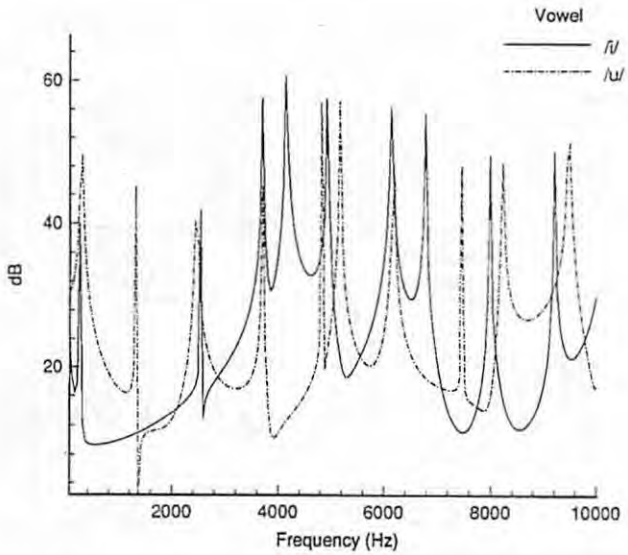

Figure 7. Frequency spectrum of the vowels /i/ (solid line) and /u/ (dash-dot line) obtained by this method with a single square pulse.

node 7 which is the narrowest location in the tract and the closest to the glottis.

Since the speed of sound is about 35000 cm/s and the length of the tract is about 17.5 cm, the acoustic wave should take about 0.5 ms to travel from one end to other. Thus the pressure field within the tract is changing very rapidly. Figures 6 shows the pressure fields at four different instants of time within the vocal tract model of /a/. These sound fields are obtained for the same fully rectified sinewave that gave waveforms of Figure 5. The numbers on the contour lines represent the pressure values in Pascals at that moment. When each line is rotated along the Z-axis a surface of the acoustic wave is generated. These surfaces are planar in some locations and at some points in time, but they are non-planar in other locations and at other times. For example, in Fig. 6A the pressure waves seem planar around 2 cm and 5 cm locations and nonplanar near 4 and 9 cm location. The non-planar waves are more noticeable in Fig. 6C and 6D. The curvature of non-planar surfaces may show considerable increases if both axes were plotted with the same scale. These non-planar surfaces indicate the existence of the transverse modes in the vocal tract.

Figure 7 shows the frequency spectrum of the vowels /i/ (solid line) and /u/ (dash-dot line) excited by the same square pulsed that was mentioned earlier. The first three formants for /i/ are 240, 1840 and 2560 Hz and for the vowel /u/ are 280, 1320, and 2480 Hz. There are some similarity between some modes that may be due to the some similarity of shape of the lower vocal tract for the two vowels. Because the large cavities in the vocal tract during the pronunciation of the vowels /i/ and /u/, the non-planar waves are expected to cause some error in the one-dimensional calculation.

Conclusions

The wave equation was solved with a new methodology and results were compared and validated. The method described in this paper can be combined with a airflow model based on the Navier-Stokes equations to include the effects of airflow turbulence and vortex shedding as a source of sound. The BFC method is capable of modeling the shape of the lips (actual shape) that improves the sound radiation. Although this method is used in the analysis of sound propagation in the vocal tract, it is capable of solving the sound fields in rotary machines such as turbofans and nozzles. The future studies should include the refinement and extension of this method by including a yielding wall boundary condition that contribute to the energy loss in the vocal tract. Also, the glottal flow pulse with different shaping parameters such as open and skewing quotients should be used to excite the tract. Another possible extension might be the use of non-circular vocal tract shapes in the future models.

Acknowledgments

This work was supported by research grant number 5 R01 DC00831-04 from the National Institute on Deafness and Other Communication Disorders, National Institute of Health.

References

Astley, R.J. and Eversman, W. A finite element method for transmission in non-uniform ducts without flow: comparison with the method of weighted residuals. *Journal of Sound and Vibration* . Vol. 57, No. 3, pp. 367-388, 1978.

Fant, G. *Acoustic theory of speech production.* 2nd Ed. The Hague: Mouton, 1960.

Hamming, R.W. *Numerical Methods for Scientists and Engineers* . 2nd Edition McGraw-Hill New York, 1973.

Hoffmann, K.A. and Chiang, S.T. *Computational Fluid Dynamics for Engineers- Volumes I & II.* Engineering Education Systems, Wichita, KS, 1993.

Ishizaka, K. and Flanagan, J. L. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell System Technical Journal*, 51(6), pp. 1233-1268, 1972.

Knupp, P. and Steinberg, S. *Fundamentals of Grid Generation* CRC Press, Inc. Boca Raton Florida, 1993.

Liljencrants, J. *Speech Synthesis with a Reflection-Type Line Analog.* Doctoral Dissertation, Department of Speech Communications and Musical Acoustics, Royal Institute of Technology, Stockholm, Sweden, 1985.

Ling, S-F. *A Finite Element Method for Duct Acoustic Problems.* Ph.D. Thesis, Department of Mechanical Engineering, Purdue University, W. Lafayet, IN 1976.

Lu, C.X. *A numerical simulation of sound production in the vocal tract.* Doctoral Thesis, Graduate School of Electronic Science and Technology, Shizuoka University, Japan, 1993.

Maeda, S. A digital simulation method of the vocal tract system. *Speech Communication* Vol. 1, pp. 199-229, 1982.

Motoki, K., Miki, N., and Nagai, N. Measurement of radiation characteristics using replicas of the lips. *Journal of the Acoustical Society of Japan* 9(3), pp. 123-130, 1988.

Pierce, A.D. *Acoustics: An Introduction to Its Physical Principles and Applications.* McGraw-Hill, New York, 1981.

Portnoff, M.R. *A quasi-one-dimensional digital simulation for the time-varying vocal.* M.S. Thesis, Department of Electrical engineering, MIT, 1973.

Sondhi, M.M. Resonance of a bent vocal tract. *Journal of the Acoustical Society of America* 79(4):1113-1116, 1986.

Story, B.H. *Physiologically-Based Speech Simulation Using An Enhanced wave-Reflection Model of the Vocal Tract.* Ph.D. thesis, The University of Iowa, Iowa City, Iowa 1995.

Story, B.H., Titze, I.R. & Hoffman, E. Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America* in press.

NCVS Status and Progress Report - 10
November 1996, 9-23

Parameterization of Vocal Tract Area Functions by Empirical Orthogonal Modes

Brad H. Story, Ph.D.
Ingo R. Titze, Ph.D.
Department of Speech Pathology and Audiology, The University of Iowa

Abstract

A set of ten vowel area functions [Story, Titze, and Hoffman, JASA, in press, 1996] has been parameterized by an "empirical orthogonal mode decomposition" which represents each area function as the sum of the mean area function and proportional amounts of a series of orthogonal basis functions (i.e. empirical orthogonal modes). Consequently, a compressed version of each area function can be represented by the amplitude coefficients used to multiply the mode shapes; each vowel has a unique set of modal amplitude coefficients. It is shown that four modes can explain up to 97% of the variance in the area function set, meaning that each vowel can be adequately represented by four modal amplitude coefficients. Since each area function was originally composed of 44 equal-length segments, the parameterization yields an 11:1 compression ratio. It is also found that pre-processing the area functions with either a logarithmic or square root operation avoids the possibility of reconstructing negative areas in the most constricted regions of the vocal tract. The mean area function was found to possess a formant structure similar to that of a uniform tube (i.e. nearly equally spaced formants) suggesting that empirical orthogonal modes are perturbations on the neutral vowel shape much like past vocal tract analyses have considered perturbations on a uniform tube (e.g. [Schroeder, M. R., JASA, 41(4), 1002-1010, 1966]). The acoustic characteristics of the two most significant empirical orthogonal modes were examined, showing that both modes tend to increase the first formant as the modal amplitude coefficients are both increased from negative to positive values. However, the second formant was found to decrease in frequency for increasing values of the first mode coefficient and increased for increasing values of the second mode coefficient. Finally, a grid of 2500 modal coefficient pairs for modes 1 and 2 was used to generate 2500 area functions. The frequency response of each area function was computed and the first two formant frequency location were extracted. Thus a mapping from coefficient pairs to the F2 vs F1 plane was generated. Within a large range, the mapping was one to one. This suggested the possibility of mapping speech waveforms to physiologically-constrained area functions and a simple example is presented.

Introduction

Articulatory models of the vocal tract have long been used to transform articulatory parameters, such as the positions of the tongue, lips, and velum, to an area function; i.e. the cross-sectional area of the vocal tract as a function of the distance from the glottis. These models have typically been defined with reference to the midsagittal plane which is a convenient reference because of the large body of x-ray films of speech production that are available for analysis and also the apparent physiological correlation between model parameters and human articulatory structures. Such parametric models provide a simple, compressed representation of the state of the vocal tract at a given point in time. Articulatory parameters typically are of lower dimension than a full area function representation, but they are dependent on an accurate transformation from midsagittal distance to cross-sectional area. Examples of such midsagittally-based models can be found in Lindblom and Sundberg (1971), Mermelstein (1973), Coker (1976), and Browman and Goldstein (1990).

Other highly compact articulatory models are those of Stevens and House (1955) and Fant (1960), both of which represented the vocal tract with only three parameters; the place and cross-sectional area of the main vocal tract constriction and a ratio of lip protrusion to lip open area. With these parameters, the entire area function from just above the glottis to the lips can be constructed by empirically-based rules.

Models such as these all depend heavily on the intuition and experience of the researcher to decide which features of the vocal tract shape are most significant and how to provide a numerical description of those features. This approach has produced valuable tools for synthesizing speech and for explaining many phenomena in both speech production and perception. However, it would be useful to have a more objective parameterization of the vocal tract shape. An example of such an approach is found in Liljencrants (1971) where he sought to explain the midsagittal profile of the tongue for 10 vowel shapes with a Fourier series representation. He made three key observations regarding his collection of tongue profiles: 1) the mean displacement of the tongue from a neutral position did not significantly change across vowels, implying a "conservation of mass" of the tongue body, 2) the fine structure of each profile was much smaller than that of the overall shape variation, and 3) many of the tongue profiles showed a strong resemblence to a sinusoid. These particular features suggested that the shape of the tongue for each vowel could be described by proportional amounts of a standard set of orthogonal basis functions; i.e. a Fourier series. Liljencrants found that the tongue shape could be reconstructed with small error using only a DC term and the first significant Fourier components. This representation produced about a 9 to 1 compression of the original data.

A similar study was performed by Harshman, Ladefoged, and Goldstein (1977) in which midsagittal tongue profiles for 10 English vowels were subjected to a factor analysis. In this case, a specialized 3-way factor analysis was developed to assist in explaining variations of tongue shapes from speaker to speaker. The factor analysis revealed two underlying displacement patterns, various proportions of which could be used to reconstruct the original tongue shapes. Much like the Fourier description proposed by Liljencrants (1971), the two displacement patterns uncovered by the factor analysis allow the tongue profile of all the vowel shapes analyzed to be represented by a set of basic features (or patterns) and a set of amplitude coefficients that define each individual vowel.

Using a factor analyis similar to that of Harshman et al., Jackson (1988) attempted to parameterize Icelandic vowels. Jackson found that three factors were needed to describe the Icelandic vowels, with the second factor having quite a different shape than the second factor given by Harshman et al. This led to the suggestion that factor shapes are not universal across languages but are language specific. However, Nix, Papcun, Hogden, and Zlokarnik (1996) have re-analyzed Jackson's data and re-compared the results to Harshman et al.'s results and found that two factors are actually adequate in describing the Icelandic vowels and shape of each factor in the re-analysis is remarkably similar to Harshman et al.'s original factor shapes.

Meyer, Wilhelms, and Strube (1989) have also used a similar approach to generate articulatory parameters for a speech synthesizer. Using the data from Harshman et al. (1977), they computed ten section vocal tract area functions based on midsagittal-to-area transformations. Each area function was assumed to have a length of 17.5 cm, giving a spatial resolution of 1.75 cm. Data from Fant (1960) was also used to supplement their collection. The area functions were then subjected to an eigenfunction decomposition that yielded three eigenvectors capable of explaining 93% of the variance in the area function set.

The quest for a vocal tract parameterization of this type is analogous to a Fourier-based spectral analysis of the acoustic speech waveform in which the sound wave is described by the amplitudes of the Fourier coefficients (Harshman et al., 1977). In Liljencrants (1971), Harshman et al. (1977) and Meyer et al. (1989), the vocal tract shape for each vowel is described by the amplitudes of a descriptive set of orthogonal basis functions; in Liljencrants a standard Fourier series formed the set of basis functions and in Harshman et al. and Meyer et al. the basis function set was empirically derived. A similar representation of the vocal tract area function was reported by Schroeder (1966) and Mermelstein (1967) based on purely acoustic considerations of perturbing the shape of a closed-end tube of constant cross-sectional area. They both showed that the area function could be represented as the sum of a Fourier series and constant area tube. This representation was developed in an effort to find a possible mapping from the vocal tract resonance peaks (poles) in a frequency spectrum to a specific vocal tract shape. The problem with this approach is that the formants can only be used to determine the odd components of the Fourier series; the even components were set to zero. However, various sets of non-zero even Fourier series components can produce widely varying area functions while maintaining the same formant (pole) locations in the frequency spectrum. This is the classic "many-to-one" mapping. The problem of unknown even Fourier components is equivalent to not knowing the location of the zeroes of a vocal tract configuration. This lack of information about the zeroes is also the primary cause of the ineffectiveness of extracting vocal tract area functions based on LPC analysis.

Recently, Story, Titze, and Hoffman (1996) have reported a set of area functions corresponding to 12 vowels, 3 plosives, and 3 nasals for one adult male speaker. The area functions were obtained from 3-D reconstructions of the vocal tract using magnetic resonance imaging. It is the purpose of this paper to develop a speaker-specific parameterization of a subset of those vocal tract shapes using a technique that decomposes the subset of area functions into empirical orthogonal modes. The primary goal of this analysis was to develop a compressed representation of the vocal tract area functions, but possible connections between

mode shape and articulation and mode shape and acoustic characteristics are also explored. It should be stressed that this empirical "modal" analysis considers the entire vocal tract. Thus, the shape of the tongue as well as jaw position, lip opening and lower pharyngeal structures are all included. The study is most similar to that done by Meyer et al. (1989) in which a set of area functions were subjected to an orthogonal decomposition. However, the area functions used in the present analysis are from only one adult male speaker, have been measured from 3-D reconstructions of each vowel shape, and have a spatial resolution of 0.396 cm (44 sections for a 17.5 cm vocal tract) as compared to Meyer et al.'s 1.75 cm resolution.

The specific aims of the paper are to: 1) use a covariance method to decompose ten vowels (i, ɪ, ɛ, æ, ʌ, ɑ, ɔ, ʊ, o, u) into a set of orthogonal basis vectors from which specific area functions can be reconstructed, 2) infer some articulatory meaning to the two most significant empirical modes, and 3) relate the two most significant modes to acoustic features in the formant spectrum.

Empirical Orthogonal Mode Decomposition

Statistical techniques that utilize a linear orthogonal transformation to extract prominent features from a high-dimensional input set to produce a low-dimensional set of features have been used in many fields, and as a result go by several different names. Principal components analysis, Karhunen-Loeve transform, empirical orthogonal functions, or singular value decomposition are a few of the names used to identify the method. In this paper, the term "empirical orthogonal modes" will be used since it implies that basis functions are derived purely from empirical data and the term "mode" is used to emphasize a similarity to a modal decomposition of a dynamical system into natural modes. Occasionally, the terms "orthogonal modes" or simply "modes" will be used; these should be assumed to mean the same as "empirical orthogonal mode".

Decomposition of a data set into orthogonal modes transforms a high-dimensional input into a low-dimensional output consisting of significant, uncorrelated features where a small number of the features contain most of variance of the original data set. The method used in this study to extract modes is given in general terms in Herzel et al. (1995) and specifically applied to vocal fold vibratory patterns in Berry et al. (1994). The formulation of the method will now be given in terms that are specific to the analysis of the area function set in Story et al. (1996). The notation will be similar to that used by Herzel et al. (1995) except that the temporal dimension will be replaced by a vowel dimension.

Each area function in Story et al. (1996) was reported as a set of cross-sectional areas with a inter-point spacing of 0.396 cm; the space between data points was assumed to represent a cylindrical tube section. The number of sections comprising each area function was chosen to most closely represent the measured length of the vocal tract during production of a given vowel. However, to extract empirical orthogonal functions, all of the area functions must be represented as equal length vectors. A length of 17.46 cm was chosen as a reasonable length compromise over the set of the ten vowels to be analyzed; this yielded 44 element area vectors. Thus, each area function was normalized to be 17.46 cm long and then was resampled with a cubic spline interpolation to generate a 44 element area vector. The normalization has the effect of slightly stretching the shorter vowels and compressing the longer ones. The normalized area functions are given in Table 1.

Table 1.
Area functions of ten vowels based on Story et al. (1996). Each area function has been normalized to a standard length of 17.46 cm (44 sections x 0.396 cm/section). The glottal end of the area function is at section 1 and thelip end at section 44.

section	i	ɪ	ɛ	æ	ʌ	ɑ	ɔ	ʊ	o	u
1	0.33	0.20	0.23	0.23	0.33	0.45	0.61	0.32	0.18	0.60
2	0.30	0.18	0.13	0.26	0.28	0.20	0.28	0.39	0.17	0.36
3	0.36	0.16	0.14	0.27	0.23	0.26	0.19	0.39	0.23	0.29
4	0.33	0.19	0.19	0.17	0.15	0.21	0.10	0.43	0.28	0.44
5	0.64	0.11	0.04	0.15	0.17	0.32	0.07	0.56	0.59	0.69
6	0.46	0.67	0.26	0.14	0.33	0.30	0.30	1.46	1.46	2.15
7	1.70	1.70	1.08	0.59	0.39	0.33	0.18	2.20	1.60	3.00
8	3.14	1.64	1.26	1.31	1.02	1.05	1.13	2.06	1.11	2.72
9	2.89	1.45	1.21	1.54	1.22	1.12	1.42	1.58	0.82	2.15
10	2.45	0.97	0.96	1.06	1.14	0.85	1.21	1.11	1.01	2.48
11	2.87	0.84	0.72	0.93	0.82	0.63	0.69	1.11	2.72	4.95
12	3.71	1.90	0.74	0.67	0.76	0.39	0.51	1.26	2.71	5.91
13	3.77	2.35	0.91	1.98	0.66	0.26	0.43	1.30	1.96	5.49
14	3.92	2.97	1.64	2.25	0.80	0.28	0.66	0.98	1.92	5.05
15	4.50	3.21	1.91	2.08	0.72	0.23	0.57	0.93	1.70	4.60
16	4.44	3.37	2.20	1.90	0.66	0.32	0.32	0.83	1.66	4.41
17	4.47	3.33	2.62	2.35	1.08	0.29	0.43	0.61	1.52	3.77
18	4.71	3.61	2.77	2.92	0.91	0.28	0.45	0.97	1.28	3.39
19	4.44	3.91	2.98	3.33	1.09	0.40	0.53	0.75	1.44	3.18
20	4.15	3.82	3.00	3.76	1.06	0.66	0.60	0.93	1.28	3.29
21	4.07	3.86	2.83	3.80	1.09	1.20	0.77	0.53	0.89	3.24
22	3.51	3.47	2.84	3.69	1.17	1.05	0.65	0.65	1.25	2.33
23	2.98	3.00	2.86	3.87	1.39	1.62	0.58	0.95	1.38	2.08
24	2.10	2.65	2.44	3.73	1.55	2.09	0.94	0.99	1.09	2.04
25	1.69	2.41	2.15	3.23	1.89	2.56	2.02	1.07	0.71	1.42
26	1.44	2.07	2.10	3.24	2.17	2.78	2.50	1.39	0.46	0.62
27	1.13	1.85	1.84	3.30	2.46	2.86	2.41	1.47	0.39	0.18
28	0.72	1.82	1.77	3.21	2.65	3.02	2.62	1.79	0.32	0.17
29	0.39	1.47	1.84	3.21	3.13	3.75	3.29	2.34	0.57	0.22
30	0.33	1.49	1.72	3.24	3.81	4.60	4.34	2.68	1.06	0.25
31	0.21	1.23	1.45	3.28	4.30	5.09	4.78	3.36	1.38	0.46
32	0.10	0.91	1.37	3.62	4.57	6.02	5.24	3.98	2.29	0.71
33	0.08	0.79	1.36	3.86	4.94	6.55	6.07	4.74	2.99	0.75
34	0.27	0.88	1.43	3.86	5.58	6.29	7.08	5.48	3.74	1.33
35	0.21	1.14	1.72	4.15	5.79	6.27	6.81	5.69	4.39	2.23
36	0.26	1.48	2.08	4.52	5.51	5.94	6.20	5.57	5.38	2.45
37	0.45	1.75	2.36	4.59	5.49	5.28	5.89	4.99	7.25	3.16
38	0.21	1.95	2.66	4.77	4.69	4.70	5.04	4.48	7.00	5.16
39	0.43	1.57	2.38	4.36	4.50	3.87	4.29	3.07	4.57	4.92
40	0.77	2.09	1.95	4.36	3.21	4.13	2.49	1.67	2.75	2.73
41	1.69	1.86	2.68	4.30	2.79	4.25	1.84	1.13	1.48	1.21
42	2.06	1.60	2.61	4.55	2.11	4.27	1.33	0.64	0.68	0.79
43	2.01	1.35	2.19	4.30	1.98	4.69	1.19	0.15	0.39	0.42
44	1.58	1.18	1.60	3.94	1.17	5.03	0.88	0.22	0.14	0.86

To perform the decomposition of the area vectors into empirical orthogonal modes, assume that any area function in the set can be represented by a mean and a variable part,

$$A(x,v) = A_0(x) + \alpha(x,v) \quad (1)$$

where $A(x,v)$ is the area function for a given vowel v, $A_o(x)$ is the mean area function across the data set, $\alpha(x,v)$ is the variation that is superimposed on $A_o(x)$ to produce a specific area function. The x denotes the index vector [1,2, ... 44] where 1 represents the first section above the glottis and section 44 is at the lips. The *vowel-dimension* is denoted by v and v = [1, 2, ...10] represents the vowels ordered as [i, ɪ, ɛ, æ, ʌ, ɔ, ʊ, o, u] (see Table 1). The mean area function $A_o(x)$ is computed by,

$$A_0(x) = \frac{1}{M}\sum_{v=1}^{M} A(x,v) \quad (2)$$

M is the number of area functions included in the analyzed set. The superimposed variations are computed by subtracting the mean area function from all the area functions in the input data set,

$$\alpha(x,v) = A(x,v) - A_0(x) \quad (3)$$

A covariance matrix can now be computed using the $\alpha(x,v)$'s,

$$R_{ij} = \frac{1}{M}\sum_{v=1}^{M} \alpha(x_i,v)\alpha(x_j,v) \qquad (i,j = 1,2,\ldots,N) \quad (4)$$

where M is again the number of area functions in the analyzed set and N is the number of elements in each area vector (N=44). Normalized eigenvectors $\varphi(x)$ can now be computed from the real, symmetric, covariance matrix (R_{ij}). The *empirical orthogonal modes* are defined to be the normalized eigenvectors. The eigenvalues λ_i of the covariance matrix indicate how much of the total variance in the input data set can be explained by the corresponding modes. Defining $c_i(v)$ as the amplitude coefficient of the i^{th} mode corresponding to the v^{th} area function, the superimposed variation defining each area vector can be computed by,

$$\alpha(x,v) = \sum_{i=1}^{N} c_i(v)\varphi_i(x) \quad . \quad (5)$$

The amplitude coefficients are obtained by projecting the original data set onto the set of modes,

$$c_i(v) = \sum_{j=1}^{N} \alpha(x_j,v)\varphi_i(x_j) \qquad (i=1,2,\ldots,N) \quad (6)$$

Once the coefficients have been computed, approximate area functions can be reconstructed using equations (5) and (1). A property of a decomposition into empirical orthogonal modes is that they are ordered such that the first ones capture the most prominent spatial variance. Thus, the reconstruction of the area functions can be performed by summing over less than N modes. This provides a compression of the original area functions from N cross-sectional areas to much less than N modal amplitude coefficients.

Reconstruction of Area Functions with Empirical Orthogonal Modes

A graphical representation of the first ten vowels of Table 1 is shown in Figure 1a while the mean area function is given in Figure 1b. The four most prominent empirical orthogonal modes are shown with solid lines in Figure 2; the dashed lines represent the reflection of each mode across the zero axis. Cumulative variances for the first ten modes are given in column 2 of Table 2 and the modal amplitude coefficients for each vowel are shown in Table 3 . Note that just four modes account for over 97% of the total variance in the set. This means that each 44-section area function can be closely approximated by only four parameters, providing a significant (~11:1) compression of the important information. Figure 3 shows reconstructions of the "corner" vowels /i/, /ɑ/,/æ/, and /u/ using the first four mode shapes and the coefficients given in Table 3. The original vowel area

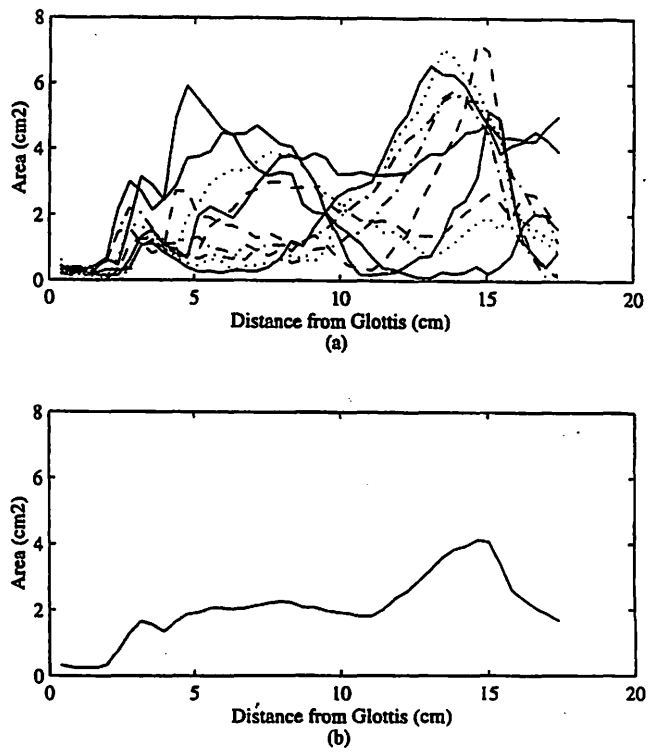

Figure 1. a) Ten area functions for the vowels (i, ɪ, ɛ, æ, ʌ, ɑ, ɔ, ʊ, o, u) and b) mean area function.

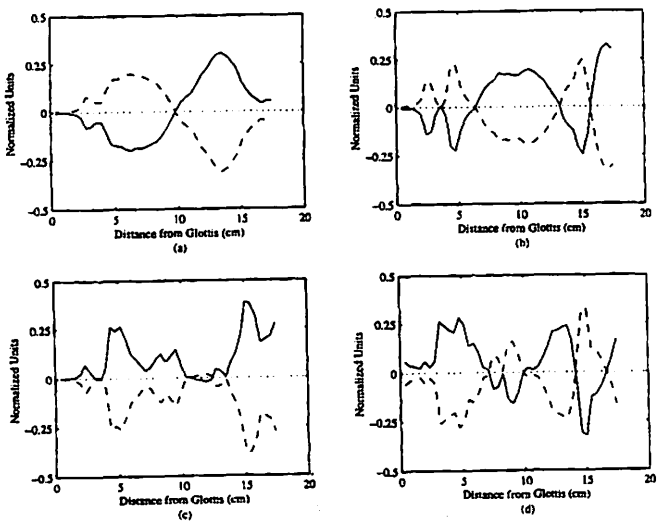

Figure 2. Four empirical orthogonal modes obtained without a pre-processing operation (the dashed lines are the reflection of each mode about the zero axis): a) mode 1, b) mode 2, c) mode 3, and d) mode 4.

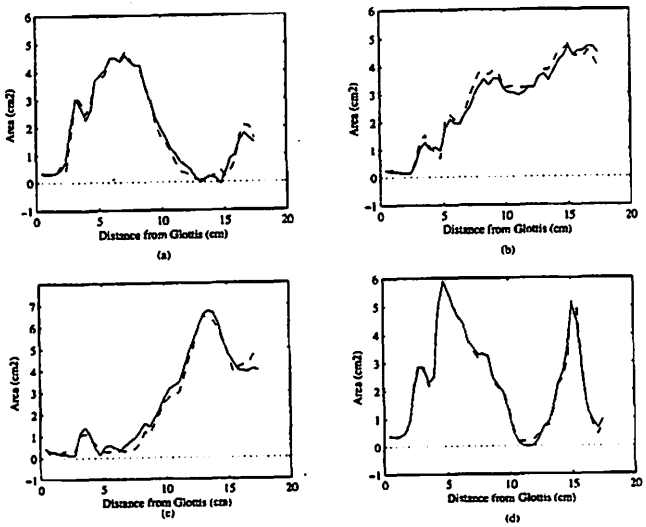

Figure 3. Reconstructions of the four "corner" vowels (no pre-processing operation): a) /i/, b) /æ/, c) /ɑ/, and d) u.

function is shown with the dashed line. Note that the reconstructed area functions appear to be closely matched to the originals except in the regions of tightest constriction. In these regions the area tends to go below zero. Negative area values are not possible in the vocal tract and this problem is a result of losing some of the information in these regions by using only four empirical orthogonal modes.

To avert the problem of producing negative areas in the area function reconstructions, a base ten logarithm was applied to each original area function prior to performing the modal decomposition. Taking the logarithm of the area functions has the effect of expanding the constricted regions and compressing the more open regions. The area functions are reconstructed just as they were before except that an antilogarithm must be applied to the final result. The first four mode shapes for this case are shown in Figure 4, cumulative variances for the first ten principal component vectors are given in column 3 of Table 2, and the amplitude coefficients for each vowel are shown in Table 4. In this case, the first four modes account for nearly 95% of the total variance, a few percent lower than the previous case. Figure 5 shows reconstructions of the vowels /i/, /ɑ/,/æ/, and /u/

Table 2.
Cumulative variances for the first ten empirical orthogonal modes.

Mode No.	Area	Log_{10}(Area)	√(Area)
1	68.46	60.03	66.90
2	87.13	83.02	87.90
3	94.27	90.72	94.31
4	97.22	94.88	96.81
5	98.72	97.18	98.39
6	99.21	98.62	99.23
7	99.73	99.60	99.71
8	99.92	99.90	99.81
9	100.00	100.00	100.00
10	100.00	100.00	100.00

Table 3.
Modal amplitude coefficients corresponding to ten vowels. No pre-processing operation was used.

Mode	i	ɪ	ɛ	æ	ʌ	ɑ	ɔ	ʊ	o	u
1	-12.4348	-7.4723	-3.9460	1.5183.	6.6356	9.8041	9.0407	4.6589	1.2210	-9.0256
2	1.8631	2.4365	2.8938	5.9207	-0.1328	3.8517	-1.5592	-3.7667	-6.3723	-5.1350
3	-1.7112	-1.9455	-2.2287	3.9988	-0.5215	1.5295	-1.5786	-2.7190	1.1167	4.0595
4	2.1190	-1.0360	-2.1260	-0.9825	-0.0430	1.4427	1.0373	0.8768	-2.6582	1.3701

Table 4.
Modal amplitude coefficients corresponding to ten vowels with the Log_{10}(area) pre-processing operation.

Mode	i	ɪ	ɛ	æ	ʌ	ɑ	ɔ	ʊ	o	u
1	-3.4818	-1.1912	-0.2926	0.5232	1.4916	2.4226	2.1813	0.8504	-0.2544	-2.2490
2	1.0087	0.7304	1.0341	1.2699	0.1866	0.6384	-0.0673	-1.5237	-1.8710	-1.4060
3	-1.1864	0.3317	0.3980	1.1273	0.0385	-0.5540	-0.5148	-0.5000	0.2792	0.5804
4	0.1348	-0.5819	-0.4542	0.1698	0.0048	0.8650	-0.2086	-0.6219	-0.0038	0.6961

Table 5.
Modal amplitude coefficients corresponding to ten vowels with the $\sqrt{}$(area) pre-processing operation.

Mode	i	ɪ	ɛ	æ	ʌ	ɑ	ɔ	ʊ	o	u
1	-4.5868	-2.2662	-0.9671	0.6019	2.2705	3.4088	3.0793	1.5314	0.0892	-3.1610
2	0.8701	0.9242	1.1817	2.0505	0.1111	1.2017	-0.4369	-1.6899	-2.3804	-1.8320
3	-1.0200	-0.3415	-0.3772	1.4780	-0.0643	0.1532	-0.6305	-0.8610	0.4413	1.2219
4	0.7094	-0.7028	-0.4845	-0.3044	-0.1079	0.9708	-0.0853	-0.2056	-0.1372	0.3477

using the first four modes and the coefficients given in Table 4. The dashed lines in the figure are the original area functions for each vowel. It is observed that the regions of the area functions that contain the most error (or are least well represented) are those with the largest areas. Thus, the logarithm operation has transferred the error from the constrictions to the expansions. Also note that the regions of constriction are very well represented and that no negative areas are produced.

Another operation that can be applied to each area function prior to the modal decomposition is a square root function. Similar to the logarithm, the square root has the effect of expanding constrictions and compressing expansions. The square root operation also has as aesthetic appeal in that the square root of an area essentially produces the radius of a circular cross-section, except without the scaling factor of pi. Thus, taking the square root of each area function transforms each area function into a "radii" function. To reconstruct the area functions after the modal decomposition, the final result needs only to be squared. The first four modes are shown in Figure 6, cumulative variances for the first ten modes are given in column 4 of Table 2, and the amplitude coefficients for each vowel are shown in Table 5. In the square root case, the first four modes account for almost 97% of the total variance which is the nearly the same as the first case where no pre-processing was performed on the area functions. Figure 7 shows the reconstructions of vowels /i/, /ɑ/,/æ/, and /u/ using four modes and the coefficients given in Table 5. It appears that the area function error is spread out across the entire area function. The reconstructions in the constricted areas are not as accurate as the logarithmic case but there are no negative areas (this is guaranteed by the squaring operation required for reconstruction). The expanded regions are more accurately represented than for the log area case but not quite as good as for the unprocessed areas. Thus, the square root pre-processing seems to be a compromise between using unprocessed areas and the base ten log operation.

Significance of the Mean Area Function

It is of interest to discuss the possible connections between empirical orthogonal modes and the physiology of articulation. However, before any association between mode shapes and articulatory configurations is considered, the significance of the mean area function shape will be examined. The empirical orthogonal modes resulting from the area function pre-processing schemes all showed the same general characteristics, but for discussion purposes the focus will be on the final case (Figure 6); i.e. using the square root pre-processor.

It has often been the case that theoretical analyses of the vocal tract formant structure begins by considering a uniform tube, closed at the glottis and open at the mouth, with formants spaced according to,

$$F_n = \frac{nc}{4L} \qquad n = 1,3,5,\ldots \qquad (7)$$

which for an ideal, lossless tube of length L=17.5 cm and a speed of sound c=350 m/s will generate formants located at

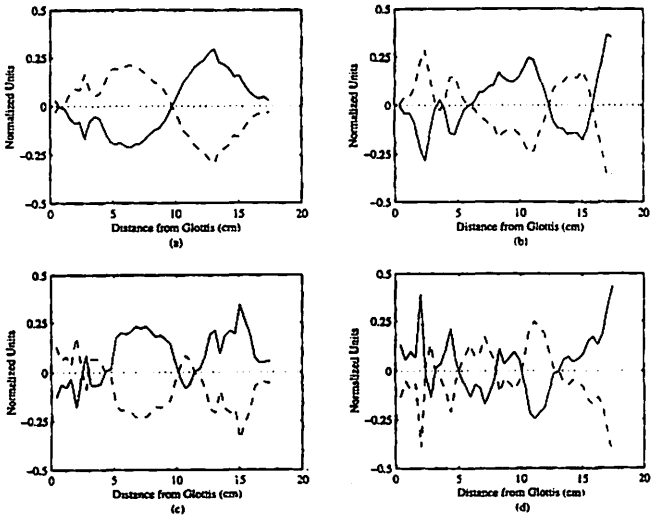

Figure 4. Four empirical orthogonal modes obtained with the $\log_{10}$(area) pre-processing operation (the dashed lines are the reflection of each mode about the zero axis): a) mode 1, b) mode 2, c) mode 3, and d) mode 4.

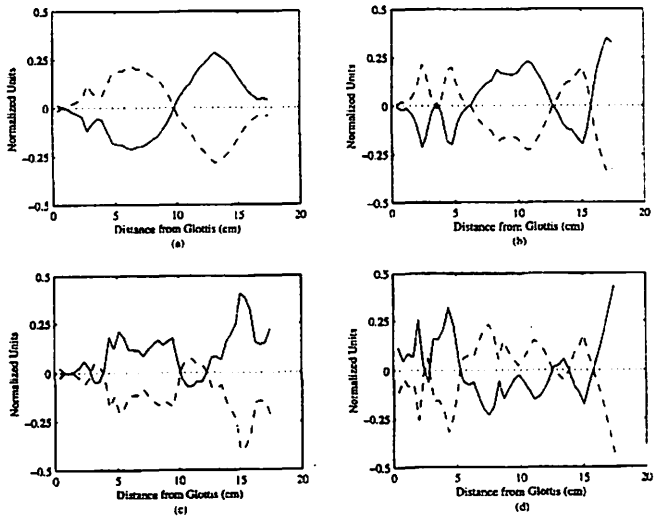

Figure 6. Four empirical orthogonal modes obtained with the √(area) pre-processing operation (the dashed lines are the reflection of each mode about the zero axis): a) mode 1, b) mode 2, c) mode 3, and d) mode 4.

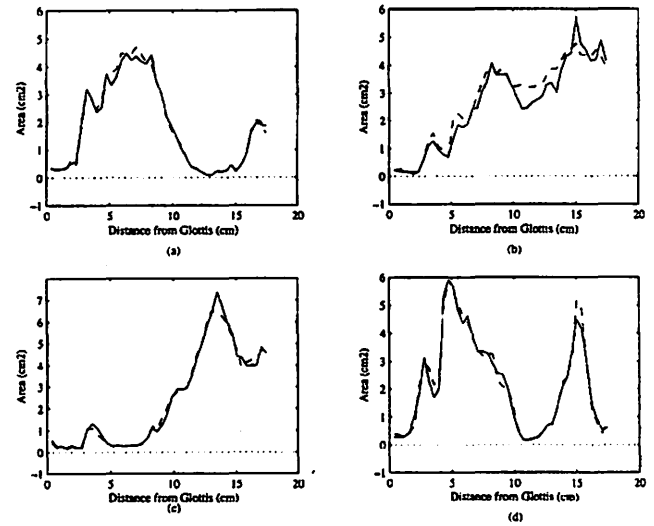

Figure 5. Reconstructions of the four "corner" vowels ($\log_{10}$(area) pre-processing operation): a) /i/, b) /æ/, c) /ɑ/, and d) u.

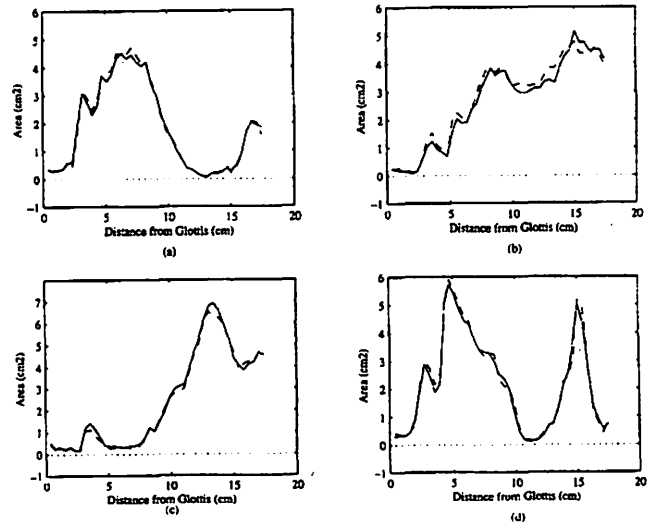

Figure 7. Reconstructions of the four "corner" vowels (√(area) pre-processing operation): a) /i/, b) /æ/, c) /ɑ/, and d) u.

500, 1500, 2500, ... Hz. The uniform tube is assumed to produce the acoustic characteristics of a neutral /ə/ vowel. The tube is then systematically perturbed to shift the formants up or down in frequency to produce other formant structures (Fant, 1960; Schroeder, 1966; Mermelstein, 1967; Mrayati et al., 1988). A 1.0 cm^2 uniform tube is shown with the mean area function in Figure 8a, while the frequency responses of both area functions, computed with a frequency domain transmission line technique (Sondhi and Schroeter, 1987), are shown in Figure 8b. The formant peaks for the mean area function occur at similar locations as those of the uniform tube. The first four formants generated by the mean area function show an upward shift in frequency relative to those of the uniform tube while the fifth formant is shifted down in frequency. Formant frequencies and percent differences are given in Table 6 for the two cases. Note that the uniform tube formants do not occur at exactly 500, 1500, 2500, ... etc., but are shifted because of the inclusion of frequency dependent losses (e.g. yielding walls and radiation impedance). The first formant is the worst match, with an 18.9 percent difference. Both F2 and F3 show about a 2 percent difference while F4 differs by 4.8 percent.

The characteristic of the mean area function to produce formants similar to a uniform tube is a demonstration of how two (and theoretically more) different tube shapes can produce essentially the same formant structure. Consider the mean area function to be a deformation of a uniform tube. Then the constriction of the mean area func-

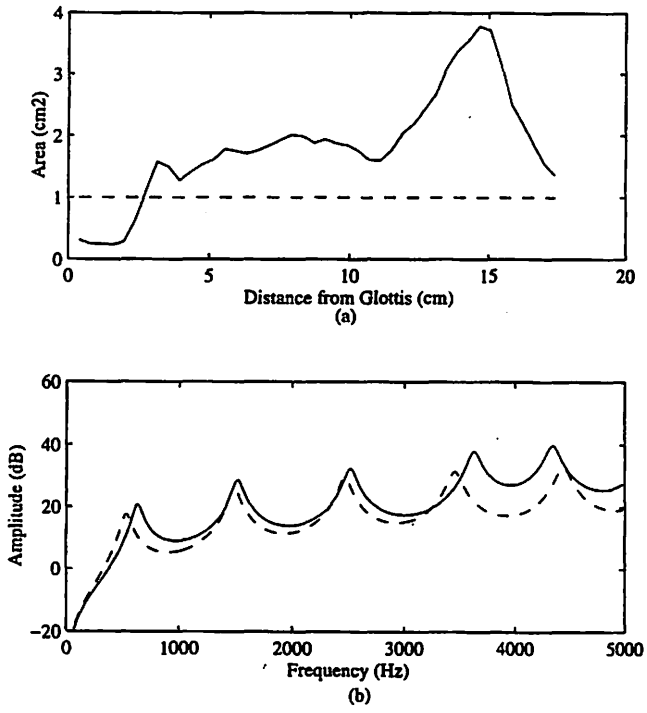

Figure 8. a) Mean area function (solid) and 1 cm² uniform tube (dashed), and b) frequency response of the mean area function (solid) and 1 cm² uniform tube (dashed).

Table 6.
First four formant frequencies of a 1 cm² uniform tube and the mean area function. Percent differences between the two set of formant frequencies are also shown.

	F1	F2	F3	F4
uniform tube	530	1486	2466	3452
mean area function	630	1516	2517	3624
%diff	18.9	2.0	2.0	4.8

tion in the short region above the larynx is countered with an expansion at the opposite end of the tract, keeping the formants in nearly the same locations. This suggests that vowels are produced by perturbing formants from uniformly spaced locations in the frequency spectrum to vowel formant locations. However, due to anatomical constraints, the physiological neutral vowel shape is not that of a uniform tube, but rather a deformed version that can produce an equivalent neutral vowel formant structure. This suggests that the mean area function is the neutral vowel configuration and other vowels are produced by imposing perturbations on it. The decomposition of the area function set into empirical orthogonal modes quantifies these perturbations. A comparison of the Fourier series representation of the area function suggested by Schroeder (1966) and the modal representation presented in this paper show an interesting similarity. From Schroeder (1966), the area function is represented by

$$A(x) = A_0 + \sum_{m=1}^{no.\ fmts} \upsilon_m \cos\left(\pi m x / L\right) \qquad (8)$$

(after Schroeder, 1966)

in which υ_m is the coefficient determining the magnitude of the m^{th} Fourier component and L is the vocal tract length. By combining equations (1) and (5), the empirical orthogonal mode representation for any given vowel is,

$$A(x) = A_0(x) + \sum_{i=1}^{n} c_i \varphi_i(x) \qquad (9)$$

In equation (8), the area function $A(x)$ is sum of the uniform tube cross-sectional area and the summation of the odd Fourier cosine series (a standard set of orthogonal basis functions) over the number of formants extracted from a speech waveform spectrum. In Schroeder (1966), A_0 is a scalar value since the area along the length of the tube was constant. Inclusion of even cosine terms in equation (8) will not significantly effect the locations of the formants but will greatly alter the resulting area function. Thus, in theory, an infinite number of area function shapes could generate the same formant structure. In equation (9), $A_0(x)$ is a spatially varying vector and $A(x)$ is the sum of $A_0(x)$ and the summation of proportional amounts of the empirical orthogonal modes (an empirically derived set of orthogonal basis functions) that contain the majority of the variance in the input data set. The equation from Schroeder (8) is based on the *theoretically-derived acoustical* possibilities of deforming a uniform tube, whereas equation (9) is based on the *empirically-derived physiological* possibilities of deforming the neutral vocal tract shape. Equation (9) is automatically constrained to generate physiologically realizable area functions as long as the coefficients c_i are not extended beyond the maximum and minimum values obtained in the modal decomposition.

A Possible Link Between Mode Shapes and Articulatory Configuration

It would be desirable, although not necessary, to attach some articulatory meaning to the modes shapes that were derived by the decomposition into empirical modes. Since the modal decomposition extracts the most prominent features or patterns from the input data set it seems likely that the most significant modes would contain significant articulatory information.

The first mode shape in Figure 6 accounts for 66.9% of the total variance in the area function set, which makes it, by far, the most prominent mode. It has a back-to-

front asymmetry that, for a negative modal amplitude coefficient, would largely replicate the forward and upward movement of the tongue and some upward jaw movement. Equivalently, a positive modal amplitude coefficient would signify a backward and downward tongue movement and a dropping of the jaw. Thus it is not surprising that the coefficients for /ɑ/ and /i/ (the most extreme front and back vowels) given in Table 5 have the largest positive and negative values for the first mode in the set, respectively. However, because each mode encompasses the entire length of the vocal tract, the structure of the tongue and jaw cannot account for the complete shape of the first mode. For example, the shape of the region between 0 and 5 cm is due to the lower pharyngeal structure such as the epilaryngeal tube and epiglottis.

The second mode, which accounts for 21% of the total variance, crosses the zero axis several times and allows for tract variations in areas in which the first mode has dimished amplitude, as would be expected for othogonal modes. Because it can affect a large region in the middle of the vocal tract it is plausible that this mode captures the up-down and possible arching motion of the tongue. Note that the coefficients for the second mode in Table 5 have large negative values for the vowels /o/, /u/, and /u/ all of which have a mid-tract constriction. The /ʌ/ has large positive cofficient for the second mode combined with a low-valued first mode coefficient to create a expansion that is slightly farther back than an /ɑ/. Additionally, the region of the second mode between 0 and 5 cm defines the shaping of the epilarynx and the lower pharynx with more detail than mode 1. At the lip end of the vocal tract, mode 2 can exert much more influence than can mode 1. Hence, mode 2 may also contain the shaping of the vocal tract corresponding to lip rounding and spreading. Again, note that the large negative coefficient values for the vowels /o/, /u/, and /u/, would generate lip rounding as well as a mid-tract constriction.

Any region in which the amplitude of the mode shape is close to zero represents a portion of the vocal tract that changes very little across vowels. With regard to modes 1 and 2, such a region exists from 0 cm to about 2 cm above the glottis. Both modes have amplitudes close to zero indicating that the epilaryngeal section does not change much across the vowels. This, of course, can be seen in Figure 1a where all ten vowels are plotted together. This stable region of the vocal tract is what forces the mean area function to have an expanded oral section. The expansion is required to maintain a neutral tract formant structure.

It is of interest to compare the first two modes in the present analysis to the two modes (or components) derived by Meyer et al. (1989). The first two components from Meyer et al. have been reconstructed from their Figure 1 (p. 524) and are shown with modes 1 and 2 from this study in Figures 9a and 9b. It should be noted that they subjected

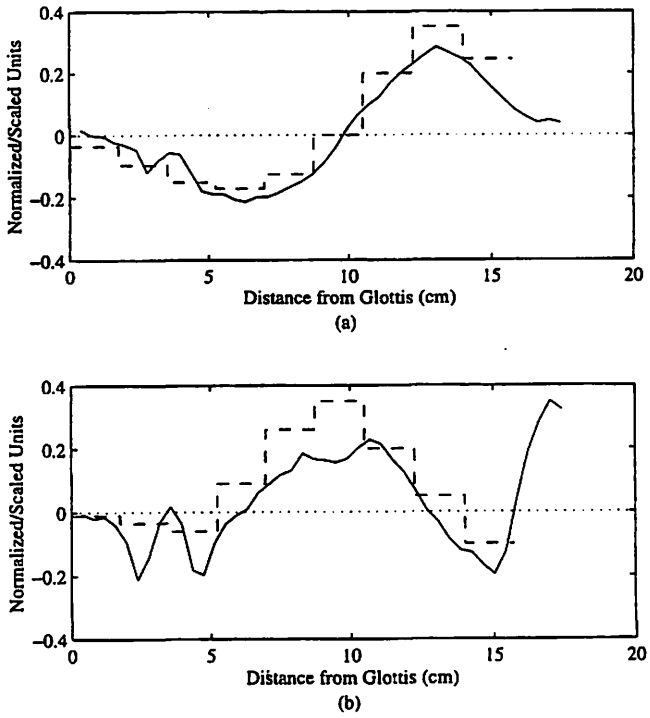

Figure 9. Comparison of modes 1 and 2 (solid lines) with those reported by Meyer et al. (1989) (these have been multiplied by -1 and scaled to fit on the same axis as the modes from this study) (dashed). a) mode 1, and b) mode 2.

only nine segments of each ten-segment area function to their eigenfunction decomposition. The lip section was defined as a separate, isolated parameter. Hence, only nine sections of their components can be shown in each figure. Also, the Meyer et al. components have been multiplied by -1 and scaled to have similar amplitudes to those from the present study. For both modes 1 and 2, the Meyer et al. study and the present study demonstrate very similar results. The first mode (or component) appears to correspond to the forward-upward and backward-downward movement of the tongue while mode 2 represents the up-down tongue motion required to create mid-tract constrictions. The zero crossings for each mode occur in nearly the same place across the two studies, but the finer spatial resolution of the modes in the present study is apparent. It is also quite apparent that the large positive portion at the lip end of mode 2 (app. 16 cm to 17.5 cm) from the present study is due to lip motion, since the Meyer et al. components did not include the lips.

The third and fourth modes correspond mainly to higher spatial frequency detail or the fine structure of the area function shape, which makes it difficult to speculate on the possible articulatory connection to these modes. Their respective coefficients in Table 5 indicate their dimished significance for reconstructing the vowel shapes.

Acoustic Characteristics of the Mode Shapes

In Section 4, the mean area function was shown to have a similar formant spectrum to that of a uniform tube. In this section the displacement of the mean vowel formant locations due to the superposition of varying amounts of modes 1 and 2 upon the mean area function will be examined. Formant spectra were computed with a frequency domain transmission line method (Sondhi and Schroeter, 1987) from which the first three peaks were extracted by peak-picking and the formant frequencies determined with a parabolic interpolation (Titze et al., 1987).

The first test investigated the effect of increasing or decreasing the amounts of mode 1 and 2 in isolation; mode 1 was varied while mode 2 was held at zero amplitude and vice versa. Additionally, all other modes were set to zero amplitude. Figure 10a shows the frequency locations of F1, F2, and F3 as a function of amplitude coefficient for mode 1. The value of mode 1 ranges from 10 percent below the most negative value in Table 5 up to 10 percent greater than the most positive. F1 and F2 change in a nearly monotonic fashion, but in different directions, as functions of the first modal coefficient. F1 begins at a value of 212 Hz and rises to 746 Hz, while F2 begins at 2267 Hz and decreases to 1104 Hz. The third formant remains reasonably flat between coefficient values of -5 to +1 and then shows a slight increase in frequency out to the final coefficient value. Results of varying the mode 2 coefficients, while holding mode 1 at zero, are shown in Figure 10b. Again, the coefficients range from 10% below their most negative value to 10% above there most positive value. The first formant as a function of the second modal coefficient shows a similar trend of increase from negative to positive coefficient values as in Figure 10a. However, the second formant has almost exactly the opposite trend for the varying mode 2 coefficient as for the mode 1 coefficient. The third formant also shows nearly an opposite trend to that seen in Figure 10a but the effect is much more subtle than for F2. What these tests show is that both modes 1 and 2 similarly affect the location of F1, but oppositely affect F2 (and to some degree F3). Thus mode 1 and mode 2 apparently engage in a "tug-of-war" to precisely position F2 for the production of a desired vowel quality.

To further investigate the effect of each mode on the resulting formant spectrum, it is helpful to compute the sensitivity functions for the first three formants of the mean area function. The sensitivity of a particular formant is defined as the difference between the kinetic energy (KE) and potential energy (PE) divided by the total energy (Fant and Pauli, 1975),

$$S_n = \frac{KE_n - PE_n}{TE_n} \tag{10}$$

where n is the section number; sensitivity can be computed at discrete sections throughout the tract. The sensitivity function can then be used to compute the change in a

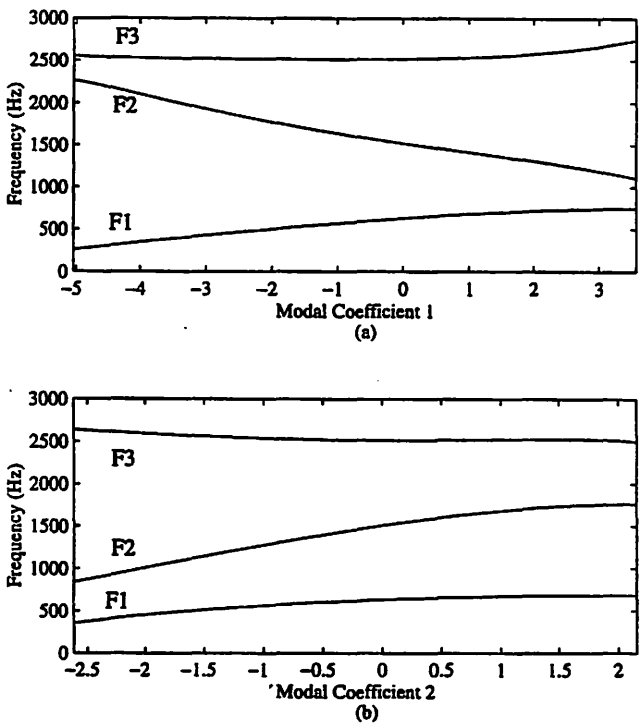

Figure 10. Formant frequencies (F1, F2, and F3) as a function of the modal coefficients: a) modal coefficient 1 was varied while modal coefficient 2 = 0, and b) modal coefficient 2 was varied while modal coefficient 1 = 0.

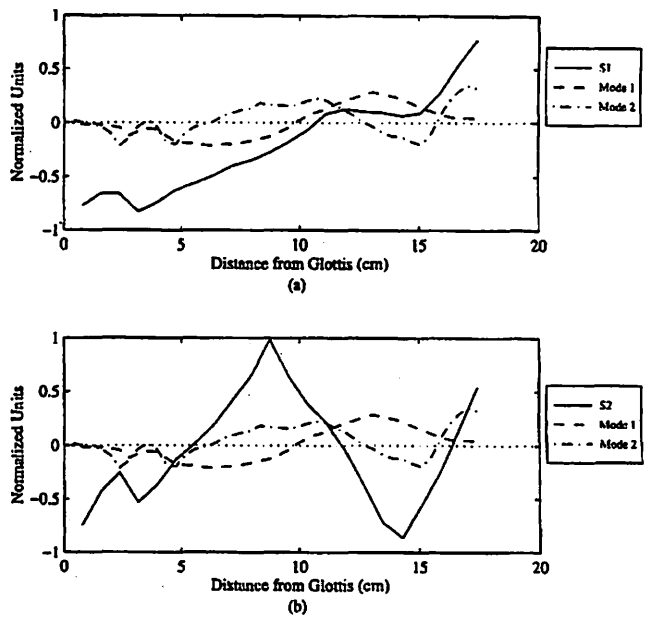

Figure 11. F1 and F2 sensitivity functions with modes 1 and 2: a) F1 sensitivity with modes 1 and 2, and b) F2 sensitivity with modes 1 and 2.

particular formant frequency (F_i) due to perturbation of the area function (ΔA) with the relation,

$$\frac{\Delta F_i}{F_i} = \sum_{n=1}^{N} S_{n,i} \frac{\Delta A_n}{A_n} \qquad (11)$$

where i is the formant number (F1, F2, . . .). This equation says that, if the sensitivity function is positive valued and the area perturbation is positive (i.e. area is increased) the change in formant frequency will be upward (positive). If the area change is negative (area decreased) the formant frequency will decrease. When the sensitivity function is negative, the opposite effect occurs for positive or negative area perturbations.

The sensitivity function for the first formant is shown superimposed onto the first two mode shapes in Figure 11a. From 0 cm to 11 cm, the sensitivity function is negative valued, meaning that expanding this portion of the mean area function would move F1 downward in frequency. Conversely, from 11 cm to the lips, an expansion in the mean area function will raise F1. In the same two portions of the vocal tract (i.e. 0 to 11 cm and 11 cm to the lips), mode 1 maintains nearly the same polarity as the F1 sensitivity function. This means that a positive valued modal amplitude coefficient multiplying mode 1 will decrease the cross-sectional area where the sensitivity is negative and increase it where the sensitivity is positive, thus efficiently raising F1. It is then no surprise that /ɑ/, which typically has a high F1 around 700 Hz, also has a large positive coefficient for mode 1 and /i/ has a large negative mode 1 coefficient to generate its characteristically low F1 of about 300 Hz. The second mode maintains the same polarity as the F1 sensitivity function from 0 cm to about 6.5 cm(except for a small positive value at 3.5 cm) and then has opposite polarity out to 10 cm. From 10 cm to the end of the vocal tract, mode 2's polarity with respect to the F1 sensitivity function oscillates. The net effect on F1 due to a positive valued modal amplitude coefficient multiplying mode 2 would be a raising of F1 but with less effectiveness as mode 1. This trend can be seen in Figure 10 where F1 rises with increasing values of the second modal coefficient.

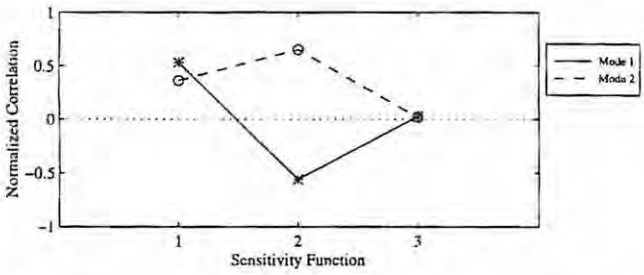

Figure 12. Normalized correlations of mode shapes with sensitivity functions: mode 1 (solid) and mode 2 (dashed).

The F2 sensitivity function is given in Figure 11b along with the shapes of the first and second modes. Except for the region from 0 cm to 5 cm, mode 1 has mostly opposite polarity to the sensitivity function while mode 2 is primarily of the same polarity. Thus, positive valued coefficients for mode 1 tend to lower F2 while positive coefficients for mode 2 will raise F2; the same trend was observed in Figures 10a and 10b.

The condense these observations, the normalized correlation (at zero lag) of each mode with each sensitivity function was computed and are shown graphically in Figure 12. Mode 1 has a positive correlation of 0.53 with the F1 sensitivity function and a negative correlation of -0.56 with the sensitivity function for F2. The F3 sensitivity was also computed but found to be nearly uncorrelated with either modes 1 or 2. Mode 2 is positively correlated, with a value of 0.36, to F1 sensitivity and also to F2 sensitivity with a value of 0.65. The largest correlation value for both modes occured for F2 even though they have opposite signs. Thus, with nearly equal valued modal amplitude coefficients, mode 1 and 2 can have almost the same effect on F2 except in opposite directions.

Mapping from Formants to Modal Coefficients

The results in the previous section have indicated a strong correlation between the empirically-derived orthogonal modes and the first two formants. It is of interest, then, to map many combinations of modal amplitude coefficients to the F2 vs F1 plane. A two-dimensional grid of mode 1 and 2 amplitude coefficients was generated by choosing ten percent beyond the maximum and minimum values of each and pairing 50 incremental values between them. The result is a 50x50 (2500 point) grid shown as a 2-D mesh in Figure 13a; the intersecting points of each horizontal and vertical line represent a modal amplitude coefficient pair. The coefficient pairs corresponding to each of the original ten vowels are shown with solid dots. Each pair of coefficients

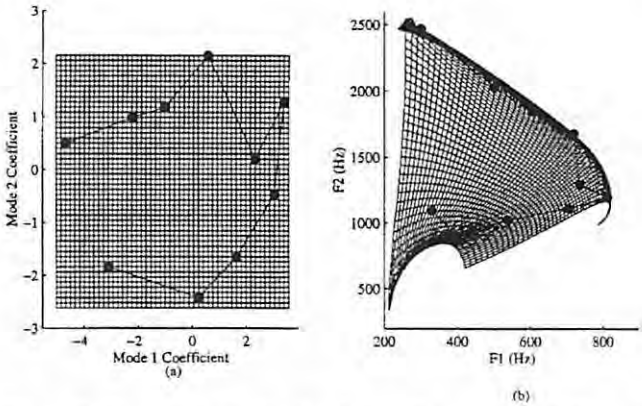

Figure 13. Mapping of mode 1 and mode 2 coefficient pairs to F1F2 pairs where the dots represent the coefficients and formant values of the original ten vowels, a) modal coefficient grid, and b) corresponding F1F2 grid.

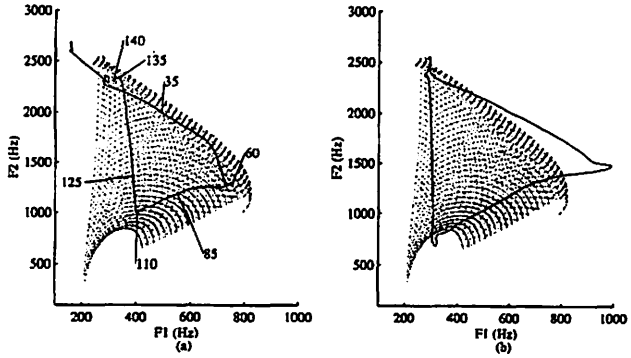

Figure 14. F1F2 trajectories for the utterance /iɑui/ superimposed on the F1F2 grid of Figure 13b: a) "conversational" manner, b) over-articulated manner. The numerical symbols represent specific analysis frames.

in the grid was then used to generate an area function by equation (9). The formant spectrum for each area function was computed by the same frequency domain method as was explained in Section 6 and the locations of F1 and F2 were similarly extracted from the spectrum. The F1-F2 pairs corresponding to each modal coefficient pair were plotted in the F2 vs F1 plane, generating the deformed grid shown in Figure 13b. As in the coefficient grid, the F1-F2 pairs corresponding to the ten vowels are shown with solid dots. Each line connecting formant pairs in this grid represents a constant value of either the first or second modal coefficient; i.e. each line is an "iso-coefficient" line.

The effect is that many formant combinations have been created that were not in the original set. Qualitatively, the F1-F2 plane is created by deforming the modal coefficient grid such that the upper right-hand corner of the coefficient grid is pulled down and to the right to stretch the upper portion of the grid and at the same time a compression pushes the lower right-hand corner upward and to the left. The upper boundary of the grid shows a saturation in the form of an apparent folding of the F1-F2 pairs, so that several pairs of coefficients corresponding to the boundary would produce nearly the same formant locations for F1 and F2. The coefficient pairs in the upper portion of the coefficient grid would correspond to large positive multipliers of mode 2 which constricts the area function at two points in the lower pharynx (at about 2.5 cm and 5 cm above the glottis), widens the area function between about 6.5 cm and 12.5 cm from the glottis, constricts a region between 12.5 cm and 16 cm, and increases the lip open area.

With the exception of the upper border where saturation occurs, the F1-F2 grid represents a one-to-one mapping between F1 and F2 formant locations and modal coefficients (or equivalently an area function created by the coefficients). This implies that an utterance consisting of connected vowels could be analyzed to extract F1 and F2 as functions of time and each of the time-dependent pairs of F1 and F2 could be mapped back to the modal coefficient grid,

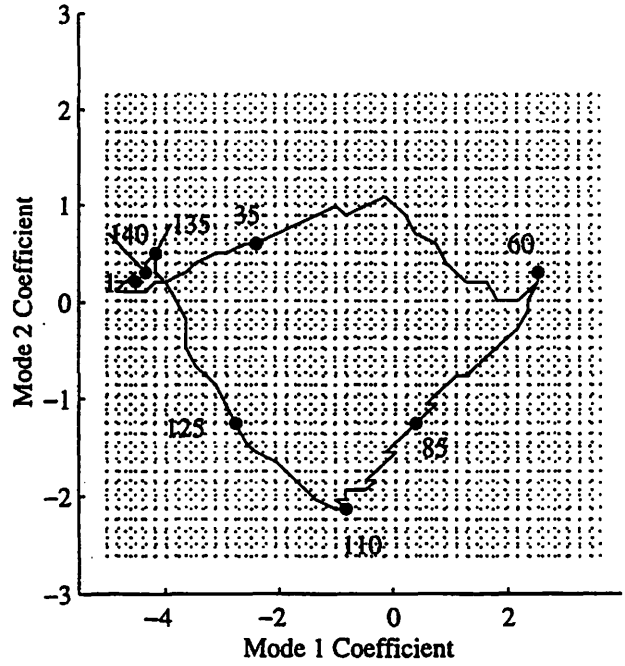

Figure 15. Modal coefficient trajectory corresponding to the F1F2 trajectory shown in Figure 14a. The numerical symbols correspond to the analysis frames indicated in Figure 14a.

and consequently to area functions. Hence, a time-dependent series of physiologically realizable area functions could be obtained from the acoustic speech waveform.

An Example of the Speech-to-Area Transform

To test the idea of mapping the locations of F1-F2 pairs back to the modal coefficient grid, two utterances recorded from the same subject who was imaged for the original MRI acquired area functions (see Story et al., 1996), were selected to be analyzed. The two utterances both consisted of the series of vowels /iɑui/ such that the standard F1F2 vowel chart will be maximally traversed. For the first utterance the subject was asked to produce the /iɑui/ in a natural "conversational" mode, while the second production of /iɑui/ was done in an over-articulated style. Each utterance was recorded at a sampling frequency of 44100 Hz in an anechoic chamber with a Panasonic SV-3700 DAT recorder and an AKG CK22 microphone. After recording, the recorded utterances were downloaded digitally via a Digidesign Audiomedia board installed in a MacIntosh Quadra 950. The audio files were later read into MATLAB where a 50 coefficient LPC analysis coupled with a peak-picking algorithm (see Section 6) extracted the first two formants at 5 millisecond intervals.

Figure 14a shows the trajectory of the conversational/iɑui/ superimposed on the F1F2 grid mesh (shown with dotted lines so that the F1-F2 trajectory can be better seen). The "1" and "140" symbols represent the beginning and end of the utterance, respectively, and the other numeri-

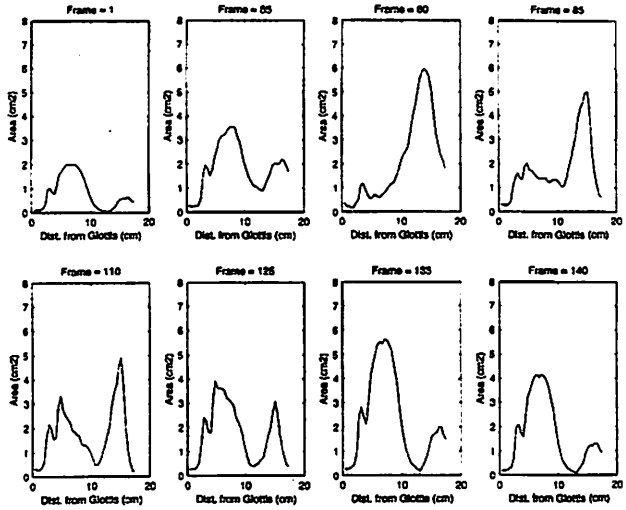

Figure 16. Series of eight area functions sampled from the 140 frame sequence of /iɑui/.

cal symbols represent specific analysis frames that are discussed later. The over-articulated version is similarly shown in Figure 14b. The F1F2 trajectory for the conversational version lies comfortably within the F1F2 mesh and thus can easily be mapped into corresponding mode 1 and 2 coefficients. However, the over-articulated version produced a F1F2 trajectory that moves outside the boundaries of the generated F1F2 mesh. For the portions of the trajectory outside the mesh, any mapping back to modal coefficients is necessarily deficient because no combination of the two modal coefficients would produce F1F2 pairs in those regions. This is, however, expected since the original area functions derived from MRI were suspected to be slightly centralized due to subject fatigue during the long image acquisition (Story et al., 1996). Additionally, the decomposition of the area functions into empirical orthogonal modes required that all vowels be normalized to one length. Thus any vocal tract length changes such as lip-rounding/spreading or larynx raising/lowering are not represented by the empirical orthogonal modes. Over-articulated speech would almost certainly use these articulatory manuevers to excessive degrees. Nonetheless, for conversational type speech the mapping from F1F2 pairs back to modal coefficient pairs may be useful.

The F1-F2 trajectory shown in Figure 14a was mapped to the modal amplitude coefficient mesh in a least squares sense. This F1-F2 pair was then matched to its corresponding modal coefficient pair. In Figure 15, the coefficient trajectory is shown for the corresponding F1-F2 trajectory in Figure 14a. Again the "1" represents the beginning of the utterance, "140" the end, and the numerical symbols and adjacent dots are the coefficient pairs that correspond the frame numbers in Figure 14a. The coefficient trajectory has a jagged characteristic due to forcing each F1-F2 pair extracted from the speech utterance to a discrete point in the F1-F2 mesh. A smoother mapping could be realized by interpolating within mesh cells, but for this preliminary study a simple minimum distance criterion was assumed to be adequate. The general shape of the coefficient trajectory is similar to that of the F1-F2 trajectory but rotated by approximately +45 degrees. The coefficient trajectory can now be used to generate area functions, using equation (9), with the same time interval that the first two formants were extracted from LPC spectra of the original speech. This series of area functions could be fed into a speech simulator (or articulatory synthesizer) coupled with a voice source to produce a simulated version of the original utterance.

Figure 16 shows a series of eight area functions sampled from the 140 frame sequence of /iɑui/. The numerical symbols shown in Figure 14a are the sampled frames and their frame number is indicated at the top of each area function plot. Frame 1 is the starting "i" -like vowel which evolves into an "ɑ" -like shape by frame 60. The shape then changes into a more of an "u" by frame 110. Frames 125 through 140 represent the transition back to the "i" -like vowel.

Discussion

Decomposition of a set of ten area functions (Story et al., 1996) into empirical orthogonal modes provides a parameterization that compresses each 44-section vowel area function into a set of four "modal" coefficients with minimal loss of information; the four empirical orthogonal modes explained approximately 97% of the variance. However, to avoid producing negative areas in highly constricted portions of the area function, either a square root or log base ten operation should be performed on the area function set prior to the orthogonal mode decomposition. Accordingly, the final step in reconstruction then must be the inverse of these two operations. While both pre-processing operations ensure the absence of negative areas, the logarithmic operation did not reconstruct the expanded portions of the area functions with the same fidelity as either the unprocessed areas or the square root areas. The log operation shifted the error from the smallest areas to the largest. The square root operation, which effectively turns each area function into "radii" (except for a scaling factor of π), was deemed to be the most useful since it guarantees no negative areas (because of the final squaring operation) and its reasonable reconstruction of the larger, more open area sections. Thus, each area function can be reconstructed from four modal coefficients, four orthogonal mode vectors, and the mean "radii" function.

The primary goal at the outset of the this study was to develop the parameterization of the vowel area function set. However, finding the mean area function to have a similar formant structure to that of a uniform tube suggested that the empirical orthogonal modes could be considered as

perturbations on a neutral vowel - that is a *physiologically-realistic* neutral vowel. The area function for this neutral vowel, along with the empirical orthogonal modes (a set of orthogonal basis functions) could be considered an empirically-based physiological equivalent to the Fourier series representation (also a set of orthogonal basis functions) of the area function proposed by Schroeder (1966) and used by many researchers since. It would also be useful if some physiological meaning could be attached to the shape of each empirical orthogonal mode. With the data presented in this paper there is no way to quantitatively determine which articulatory movements might be captured in each orthogonal mode shape. But since the empirical orthogonal mode decomposition extracts the most prominent features from the input data set, it is useful to at least speculate on the articulatory motions that might be captured in the most significant modes. The asymmetrical shape of the first mode is almost certainly due to front-back tongue movements while mode 2 appears to capture both the up-down tongue motion and lip opening and closing. The remaining higher and less significant modes fill in much of the fine detail in each area function but their shapes do not lend themselves easily to any articulatory interpretation.

Since the first two modes account for 88% of the total variance in the area function set, the articulatory movements captured in those modes would likely be the primary mechanisms of perturbing the formants of the neutral area function. Hence, the amount of each of the first two orthogonal modes imposed on the mean area function were varied in isolation to demonstrate the acoustical effect of each mode by itself. It was found that F1 increased in frequency with an increase in the amplitude of mode 1 or mode 2, and interestingly F2 was moved down by increasing mode 1 and up by increasing mode 2. A calculation of the F1 and F2 sensitivity functions for the mean area function showed that both modes 1 and 2 were positively correlated with the F1 sensitivity function and oppositely correlated with F2 sensitivity, but with nearly the same absolute value. Modes 1 and 2 seem to be shaped so that they efficiently exploit the most acoustically sensitive regions of the neutral vowel and the tendency for modes 1 and 2 to act cooperatively for F1 and in opposition for F2 allows an efficient coding of the first two formants by unique combinations of modal coefficients. A mapping of a 2-dimensional grid of modal amplitude coefficient pairs to a deformed grid of F1F2 pairs showed that each coefficient pair, within a reasonable range of values, could be mapped to a unique F1F2 pair. Thus, the selection of any combination of coefficient pairs is also a selection of a unique F1F2 pair. This property leads to the possibility of mapping formants extracted from a speech signal to physiologically realizable area functions and was shown to be moderately successful in doing that with the vowel-only utterances /iɑui/.

By understanding the acoustical consequences of imposing the empirical mode shapes upon the neutral vowel, a parsimonious model of vowel production (at least for ten vowels) can be suggested. The articulatory movements are largely captured in the first two empirical orthogonal mode shapes which have an effect along the entire length of the vocal tract. The degree to which they are used can be specified by their respective coefficient values. Thus, a very simple model for simulation of speech production could be depicted as a time-dependent voice source specified by the desired fundamental frequency (Fo) and amplitude (Ao); glottal flow pulse shaping parameters such as skewing and open quotients could also be specified to control the voice quality. The time-dependent articulation could be governed by the choice of two coefficient values (c_1 and c_2) by which the orthogonal modes will be multiplied. The selection of a given pair of coefficient values is also a selection of a unique pair of F1 and F2. This model lumps the typical articulatory specifiers such as tongue tip, tongue body, lip positions, etc. into a modal representation such that the empirical orthogonal modes "package" a specific orchestration of the articulatory musculature. This is reminiscent of a statement made by Coker in his 1976 paper:

> *"Linguistic control appears to be organized around the modes of articulatory response. The reason is probably nothing more than the physical separation of articulators. But there would be tendencies for languages to align themselves around modes, even without physically separate articulators. A mode-oriented control strategy is simplest to learn in this domain, cause and effect are most directly associated."*

The modes to which Coker was referring were based more on the natural response of individual articulators but the idea that linguistic control could be organized around some "natural" articulatory modes is attractive. Are there "natural" modes that serve as building blocks of linguistic gestures? Only a speculative result can be given at this point but in studying other vibratory and natural phenomena it is often the case that movement and/or vibration is composed of fundamental modes. Classical modal analyses of strings, membranes, bars, air columns, etc. always show fundamental modes of vibrations. Berry et al. (1994) have found, using a similar decomposition of input data into empirical orthogonal modes, that the vocal folds, even though they possess complicated tissue layers and boundary conditions, typically vibrate with combinations of just a few fundamental vibratory modes. It would seem parsimonious that vowel production (and possibly speech production in general) would be organized around a few fundamental *articulatory* modes, especially if those modes efficiently exploit the

acoustic modes (i.e. sensitivity functions) of the vocal tract. This idea is not new of course. Based on the factor analysis of tongue shapes, Harshman et al. (1977) suggested that vowel production might be organized around two prominent factors extracted from the input data. Jackson (1988) also suggested that a factor analysis of Icelandic tongue shapes yielded three fundamental factors (although this conclusion has recently been refuted by Nix et al. (1996)) that could be used as building blocks for the vowels. He referred to the factors as "core articulatory primes". It must always be remembered that any statistically-based pattern extraction has the characteristic that the patterns or modes found are those that have been *excited* within the system. There may be other natural modes that exist within the system but may not be excited or utilized for the purpose at hand.

To restate the limitations of this study, the results obtained were based on ten vowel area functions obtained by magnetic resonance imaging of one adult male speaker (native of the midwestern United States) making the results speaker dependent. However, the similarity of the mode shapes to those derived by Meyer et al. (1989) (see Figure 10) suggests that similarly shaped modes might be expected for other speakers. All of the vowels were normalized to one standard length which destroyed any information about vocal tract length changes such as lip rounding/spreading or larynx raising/lowering. Additionally, since only ten vowels were subjected to the analysis, the effect of consonantal area functions on the results and subsequent conclusions are unclear.

Even with the limitations cited, the parameterization of the vowel area functions provides a compact system by which to specify commands that can be used to simulate vowel production with some type of the speech simulation model (e.g. Liljencrants, 1985; Sondhi and Schroeter, 1987; Story, 1996). The quality of the simulation can be enhanced by directly mapping from F1F2 pairs extracted from natural speech to physiologically realistic area functions and using those area functions to simulate the original speech.

Acknowledgements

This work was jointly supported by Grants P60 00976 and R01 DC02532 from the National Institute on Deafness and other Communication Disorders. The authors would like to thank Dr. David Berry for fruitful discussions on empirical orthogonal mode decomposition.

References

Berry, D. A., Herzel, H., Titze, I. R., and Krischer, K., "Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions," JASA, 95(6), 3595-3604, 1994.

Browman, C., and Goldstein, L., "Gestural specification using dynamically-defined articulatory structures," Haskins Lab. Stat. Rep. on Speech Res., SR-103/104, 95-110, 1990.

Coker, C. H., "A model of articulatory dynamics and control," Proc. IEEE, 64(4), 452-460, 1976.

Fant, G., The Acoustic Theory of Speech Production, Mouton, The Hague, 1960.

Fant, G., and Pauli, S., "Spatial characteristics of vocal tract resonance modes," in Fant, Proc. Speech Comm. Sem. 74. Speech Communication, vol. 2, 121-132, 1975.

Harshman, R., Ladefoged, P., and Goldstein, L., "Factor analysis of tongue shapes," JASA, 62(3), 693-707, 1977.

Herzel, H., Krischer, K., Berry, D. A., and Titze, I. R., "Analysis of spatio-temporal patterns by means of empirical orthogonal functions," In Spatio-Temporal Patterns in Nonequilibrium Complex Systems, P. E. Cladis & P. Palffy-Muhoray (Eds.), Reading MA: Addison-Wesley, 505-518, 1995.

Jackson, M.T.T., "Analysis of tongue positions: Language-specific and cross-linguistic models," JASA, 84(1), 124-143, 1988.

Liljencrants, J., "Speech Synthesis with a Reflection-Type Line Analog," DS Dissertation, Dept. of Speech Comm. and Music Acous., Royal Inst. of Tech., Stockholm, Sweden, 1985.

Liljencrants, J., "Fourier series description of the tongue profile," STL-QPSR, 4, 10-18, 1971.

Lindblom, B., and Sundberg, J., "Acoustical consequences of lip, tongue, jaw, and larynx movement," JASA, 4(2), 1166-1179, 1971.

Meyer, P., Wilhelms, R., & Strube, H. , "A quasiarticulatory speech synthesizer for German language running in real time", JASA,86(2),523-539, 1989.

Mermelstein, P., "Articulatory model for the study of speech production," JASA, 53(4), 1070-1082, 1973.

Mermelstein, P.," Determination of the vocal-tract shape from measured formant frequencies," JASA, 41(5), 1283-1294,1966.

Mrayati, M., Carre, R., and Guerin, B., "Distinctive regions and modes: A new theory of speech production," Speech Comm., 7, 257-286, 1988.

Nix, D. A., Papcun, G., Hogden, J., and Zlokarnik, I., "Two cross-linguistic factors underlying tongue shapes for vowels," JASA, 99(6), 3707-3717, 1996.

Schroeder, M. R., "Determination of the geometry of the human vocal tract by acoustic measurements," JASA, 41(4), 1002-1010, 1966.

Sondhi, M. M., and Schroeter, J., "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. ASSP, ASSP-35(7), 1987.

Stevens, K. N., and House, A. S., "Development of a quantitative description of vowel articulation," JASA, 27(3), 484-493, 1955.

Story, B. H., Titze, I. R., and Hoffman, E. A., "Vocal tract area functions from magnetic resonance imaging," JASA, 100(1), 537-554, 1996.

Story, B. H., Speech Simulation with an Enhanced Wave-Reflection Model of the Vocal Tract, Ph. D. Dissertation, University of Iowa, 1995.

Titze, I. R., Horii, Y., and Scherer, R. C., "Some technical considerations in voice perturbation measurements," JSHR, 30, 252-260, 1987.

NCVS Status and Progress Report - 10
November 1996, 25-34

Acoustic Interactions of the Voice Source With the Lower Vocal Tract

Ingo R. Titze, Ph.D.
Brad H. Story, Ph.D.
Department of Speech Pathology and Audiology, The University of Iowa

Introduction

Adjustments in the lower vocal tract (the epilarynx tube, the piriform sinuses, the pharynx) and the nasal port seem to play an important role in regulating voice quality (Laver, 1980). Fiberoptic observations of the supraglottal structures of voice impersonators (Feder, 1988) suggest that the epiglottis, the pharyngeal walls, the opening to the piriform sinuses, the false vocal folds, and the aryepiglottic folds can all be adjusted to create special vocal effects. Colton and Estill (1981) observed differences in these supralaryngeal configurations between normal *speech* quality, *opera* quality, *twang*, and *sob*. Later, Estill and coworkers added *belt*, which is a quality characteristic of modern musical theatre singing (Yanagisawa et al., 1990; Estill et al., 1994). *Twang* and *belt* qualities generally have both a narrow epilarynx tube and a narrow pharynx, whereas *sob* exhibits a widening of both structures. *Opera* quality demonstrates a relatively wide pharynx with a relatively narrow epilarynx tube. Part of this quality is vocal *ring*, the percept of a bell-like constant background tone, which spectrally shows up as a prominence of acoustic energy in the 2500-3500 Hz range, independent of vowel (Bartholomew, 1934).

Quantitative analysis of vocal *ring* was conducted by Sundberg (1974). By modeling the epilarynx tube as a separate small resonator inside and at the base of a larger tube representing the vocal tract, Sundberg was able to show that a large concentration of acoustic energy was produced around 3000 Hz. He called this the singer's formant because it spectrally appears like a new formant, although it is generally a clustering of the third, fourth, and fifth formant. The singer's formant was particularly prominent when the open end of the small tube (the aryepiglottic collar) expanded abruptly into the wider tube (the lower pharynx). An area expansion of 6:1 or greater produced large amounts of *ring*. In such a sudden expansion, the small resonator is relatively independent of the large resonator and has its own formant structure. In particular, for an approximate 3 cm length, the epilarynx tube is about 1/6 the length of a typical 17.5 cm male vocal tract. Given the same boundary condition (open at the top and closed at the bottom), the first formant frequency should be 6 times higher than the 500 Hz first formant frequency of a uniform vocal tract, or about 3000 Hz.

Figures 1 and 2 show sagittal and anterio-lateral views, respectively, of a vocal tract airway as imaged by electron beam computed tomography (EBCT). The epilarynx tube is the narrow portion above the glottis in each figure. In Figure 1, the epilarynx tube is slanted upward to the left and a little more of the tracheal airway is shown below the glottis

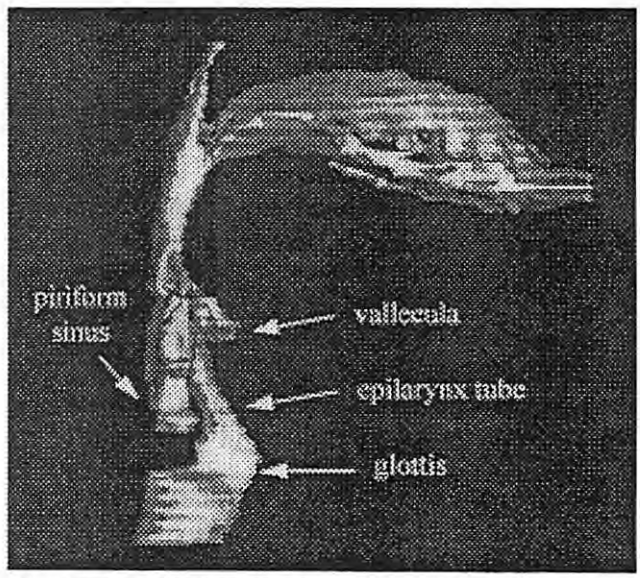

Figure 1. Sagittal section of an electron beam CT scan of the vocal tract airway of a normal 30-year old male subject phonating an /ə/ vowel (from Story, Titze, & Hoffman, 1996; by permission).

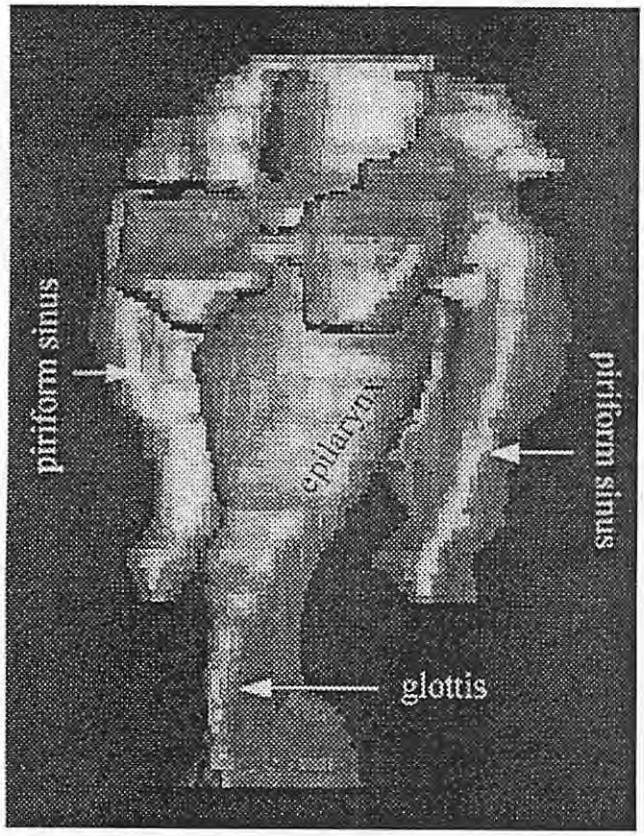

Figure 2. Anterio-lateral view of an approximate 2:1 enlargement of the laryngeal-pharyngeal portion of Figure 1 (from Story, Titze, & Hoffman, 1996; by permission).

than in Figure 2. Note that at the top of the epilarynx tube, the airway widens and becomes complicated. The lower pharynx, near the vallecula, is seen as a sudden expansion and the two piriform sinuses protrude downward from this expansion. The piriform sinuses are acoustic side branches that can have a profound effect on the resonance of the entire system. Dang and Honda (1996; also in press) have recently studied this effect in detail. Using both a human subject and a physical model, they filled the piriform sinuses with water to eliminate the acoustic side branches. In each case, the epilarynx tube resonance was enhanced with the elimination of the side branches. Thus, the piriform sinuses acted as energy absorbers (spectral zeros). It would seem, then, that a vocalist intending to produce *ring* in the voice would find a strategy to block or re-orient the side branches in a way that would minimize their absorption effect.

The current study is a theoretical investigation of how the entire back portion of the vocal tract can be adjusted geometrically to produce the most favorable conditions for vocal fold oscillation. The basic question is: As relative areas or lengths of the epilarynx, the pharynx, the piriform sinuses, and the nasal tract are varied, what is the effect of the spectral balance, the glottal waveform, and the phonation threshold pressure? The first step in the analysis is to compute the input impedance and transfer function of the vocal tract for a variety of geometries. This computation is in the form of a sensitivity analysis to determine the critical parameters. The second step is to simulate the glottal waveform and the oral radiated pressure interactively with the vocal tract for critical parameter changes.

Vocal Tract Input Impedance Calculation

To determine the relative importance of a variety of areas and lengths in the airway structures, it is instructive to begin with a sensitivity analysis of the input impedance and the transfer function of the vocal tract. The input impedance, defined as the ratio of supraglottal pressure to glottal flow (both as complex numbers), can have a profound effect on the conditions of the vocal fold oscillation, as will be seen later. The transfer function, defined as the ratio of oral radiated pressure to glottal flow (likewise as complex numbers), determines the output for a given input.

Consider the vocal tract to be a transmission line with side branches as shown in Figure 3. The following parameters are defined:

A_m, L_m = cross-sectional area and length of the mouth section

A_n, L_n = cross-sectional area and length of the nasal section

A_p, L_p = cross-sectional area and length of the pharynx section

A_s, L_s = cross-sectional area and length of the piriform sinus section

A_e, L_e = cross-sectional area and length of the epilarynx section

The impedance and transfer function calculations follow the chain matrix approach used by Sondhi and Schroeter (1987). We will not restate all the mathematical steps in this calculation, but give an equivalent abbreviated derivation for the impedance only. The calculation begins by defining a frequency f and a corresponding wave number for one-dimensional wave propagation in a tube,

$$\beta = 2\pi f/c \quad , \qquad (1)$$

where c is the sound propagation velocity (350 m/s in a warm, humid vocal tract). Next, an attenuation factor α is defined (or more detailed frequency-dependent losses), which together with the wave number gives a complex propagation constant

$$\gamma = \alpha + j\beta \quad , \qquad (2)$$

where $j = (-1)^{1/2}$.

Using standard transmission line theory (see, for example Flanagan, 1972; p. 52, equation 3.53), the input impedance to the mouth section can be written as

$$Z_m = \frac{\rho c}{A_m} \frac{Z_{rm}\cosh\gamma L_m + \frac{\rho c}{A_m}\sinh\gamma L_m}{\frac{\rho c}{A_m}\cosh\gamma L_m + Z_{rm}\sinh\gamma L_m} \quad , \qquad (3)$$

where ρ is the density of air and z_{rm} is the radiation impedance at the mouth end. The radiation impedance is also given by Flanagan and can be written as

$$Z_{rm} = \rho f\left[\beta + j\frac{16}{3}(\pi A_m)^{-1/2}\right] \quad . \qquad (4)$$

In an identical fashion, the input impedance to the nasal section is

$$Z_n = \frac{\rho c}{A_n} \frac{Z_{rn}\cosh\gamma L_n + \frac{\rho c}{A_n}\sinh\gamma L_n}{\frac{\rho c}{A_n}\cosh\gamma L_n + Z_{rn}\sinh\gamma L_n} \quad , \qquad (5)$$

where

$$Z_{rn} = \rho f\left[\beta + j\frac{16}{3}(\pi A_n)^{-1/2}\right] \quad . \qquad (6)$$

Continuing downward in the impedance calculation of Figure 3, the impedance at the end of the pharynx section is now the parallel combination of the mouth and nose input impedance,

$$Z_{mn} = \frac{Z_m Z_n}{Z_m + Z_n} \quad . \qquad (7)$$

Using the transmission-line equation again, the input impedance to the pharynx is

$$Z_p = \frac{\rho c}{A_p} \frac{Z_{mn}\cosh\gamma L_p + \frac{\rho c}{A_p}\sinh\gamma L_p}{\frac{\rho c}{A_p}\cosh\gamma L_p + Z_{mn}\sinh\gamma L_p} \quad . \qquad (8)$$

The input impedance to the piriform sinuses is simpler because the termination impedance of the sinuses is infinite at the closed end. If we assume that the two sinuses are identical, then a single cross-sectional area A_s can represent the sum of the two cross-sectional areas, and

$$Z_s = \frac{\rho c}{A_s}\frac{\cosh\gamma L_s}{\sinh\gamma L_s} \quad . \qquad (9)$$

Once again, the parallel combination of Z_s and Z_p can be computed as

$$Z_{ps} = \frac{Z_p Z_s}{Z_p + Z_s} \quad , \qquad (10)$$

and the final expression for the input impedance to the vocal tract becomes

$$Z_i = \frac{\rho c}{A_e} \frac{Z_{ps}\cosh\gamma L_e + \frac{\rho c}{A_e}\sinh\gamma L_e}{\frac{\rho c}{A_e}\cosh\gamma L_e + Z_{ps}\sinh\gamma L_e} \quad . \qquad (11)$$

We now choose a *nominal configuration* around which the parameter variation will be carried out

A_e =	0.5 cm²	L_e =	3.1744 cm
A_s =	1.0 cm²	L_s =	2.3808 cm
A_p =	3.0 cm²	L_p =	4.7616 cm
A_n =	0.0 cm²	L_n =	11.1104 cm
A_m =	3.0 cm²	L_m =	9.5232 cm

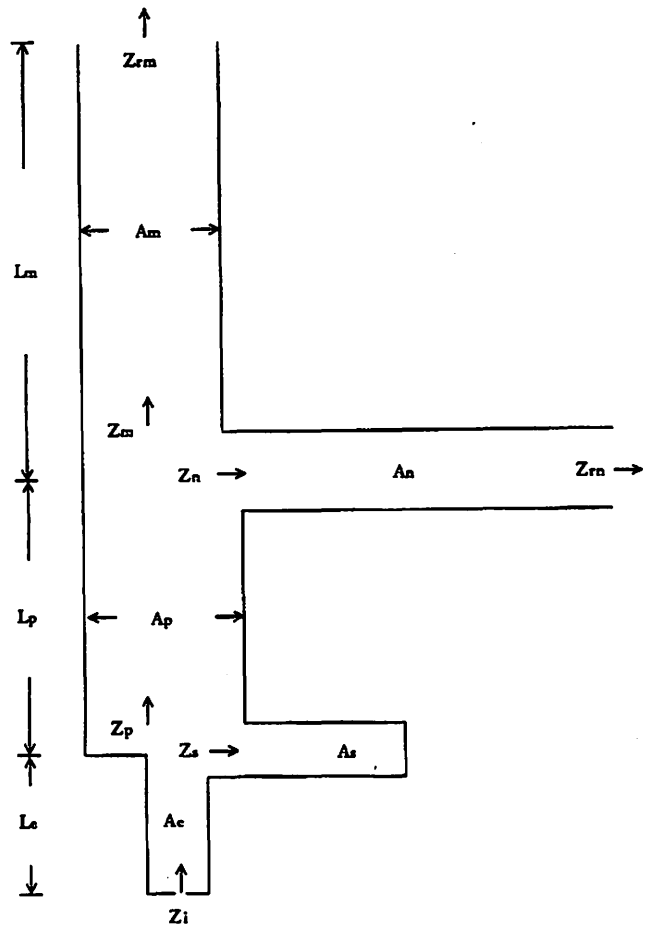

Figure 3. Schematic diagram of an acoustic transmission line with a narrowed epilarynx at the bottom and two side branches.

None of the areas are constant as a function of length in a human vocal tract (Story, 1995; Story, Titze & Hoffman, 1996; Dang & Honda, 1996) but in the calculation they are chosen constant to limit the number of parameters. The lengths are chosen to correspond to integral numbers of discrete sections of the vocal tract to be discussed later. Note that the nominal configuration has a closed nasal port, but the piriform sinuses are open. In various plots to follow, the nominal configuration will always be represented by dashed lines.

Figure 4 shows the magnitude of the transfer function (in dB), the magnitude of the input impedance (in dB), the real part of the input impedance (in dyn-s/cm^5), and the imaginary part of the input impedance (in dyn-s/cm^5). The solid curve on each plot illustrates the case without a narrowed epilarynx tube. With this uniform tube ($A_e = A_p = 3.0$ cm^2), there are roughly equally-spaced formant frequencies near (2n-1) (501) Hz, the quarter-wave resonance frequencies for a 17.46 cm tube, but the formants are detuned slightly because of radiation impedance and the piriform sinuses. This detuning will be discussed later. For the moment, the most important observation is that the reactance curve (bottom right) crosses the zero line at the formant frequencies, alternating between positive (inertive) reactance and negative (compliant) reactance. Also, both resistance and reactance are generally low.

With a narrowed epilarynx tube ($A_e = 0.5$ cm^2, dashed lines) the magnitudes of all the functions are generally higher throughout the frequency range, particularly in the 2500-3500 Hz region. The resistance (real part of input impedance) rises sharply for the third and fourth formant, while the reactance is generally more positive (inertive) below 2500 Hz and generally more negative above 2500 Hz. The increased inertive reactance

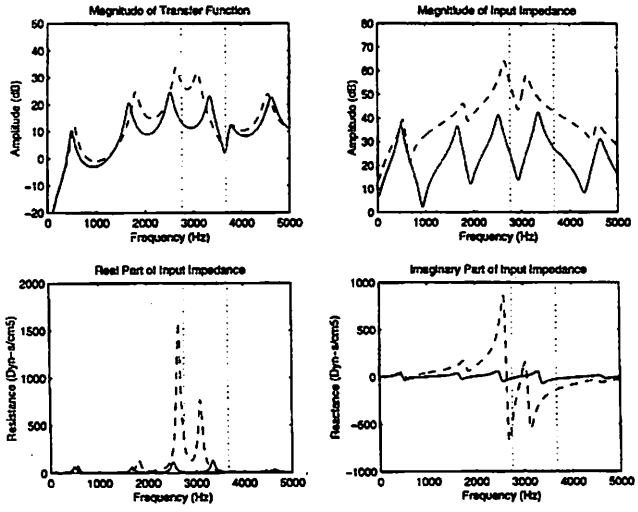

Figure 4. Magnitude of the transfer function, magnitude of input impedance, resistance (real part) and reactanc (imaginary part) of the input impedance for the idealized transmission system shown in Figure 3. Solid lines are for $A_e = 3.0$ cm^2 (a wide epilarynx tube) and dashed lines are for $A_e = 0.5$ cm^2 (a narrow epilarynx tube).

below 2500 Hz is highly significant because inertive reactance facilitates vocal fold oscillation, as will be shown later. Earlier analysis (Titze, 1988) has shown that inertive reactance effectively becomes negative resistance for small oscillation of glottal airflow and vocal fold tissue.

Figure 4 also shows that the first three formant frequencies for the narrowed epilarynx are raised in comparison to the uniform tract. But the fourth and fifth formant frequencies are lowered. The narrowed epilarynx tube therefore "attracts" all formant frequencies toward the 2500-3000 Hz region. This attraction of all the formant frequencies toward a single frequency focus was noted by Sundberg (1974) in his analysis.

If we consider the epilarynx tube to be a separate quarter-wave resonator, its uncoupled formant frequencies are

$$F_{en} = (2n-1)\frac{c}{4L_e} \qquad (12a)$$

$$= (2n-1)\frac{35{,}000\ cm/s}{(4)(3.1744)\ cm} \qquad (12b)$$

$$= (2n-1)\,2756\ Hz \quad . \qquad (12c)$$

We can say, then, that the first five formant frequencies are attracted toward F_{e1} (2756 Hz, represented by the left-most vertical dotted line in Figure 4). In the limit, as A_e becomes very small, the epilarynx tube has its own series of formants as given in equation 12c, with the formants of the larger tube (the vocal tract) becoming insignificant.

Note that the transfer function (upper left) has a spectral zero at 3675 Hz, the resonance frequency of the piriform sinuses (second vertical dotted line). Using again the quarter-wave resonator formula, a collection of resonance frequencies is predicted at

$$F_{sn} = (2n-1)\frac{c}{4L_s} \quad . \qquad (13)$$

The first resonance frequency F_{s1} is at 3675 Hz when L_s = 2.3808 cm, the nominal value selected above. At this piriform sinus frequency, there is a large energy exchange between the epilarynx tube and the sinuses, but little of this energy gets into the upper vocal tract and out of the mouth. The energy is dissipated within the vocal tract, as the transfer function shows. But the input impedance is not affected by this energy exchange, showing nothing remarkable in the 3675 Hz region, neither in terms of resistance nor reactance.

Dang and Honda's spectral zero was slightly higher (about 4000-4200 Hz) because they measured a piriform sinus length of slightly less than 2.0 cm on a human subject.

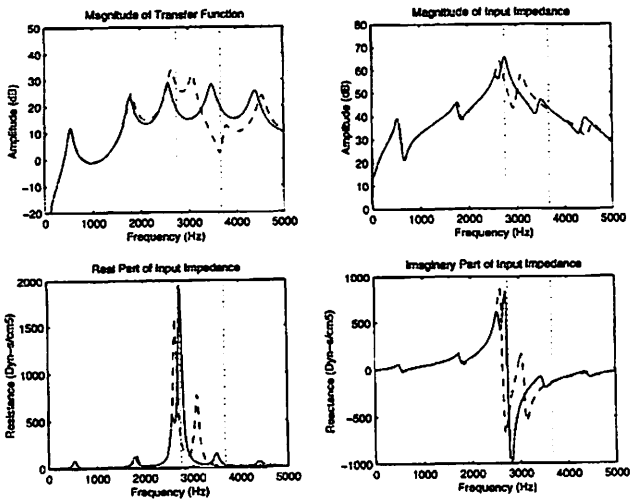

Figure 5. Magnitude of the transfer function, magnitude of input impedance, resistance (real part) and reactance (imaginary part) of the input impedance for the idealized transmission system shown in Figure 3. Solid lines are for $L_4 = L_e$ (length of piriform sinuses equals length of epilarynx) and dashed lines are for the nominal configuration.

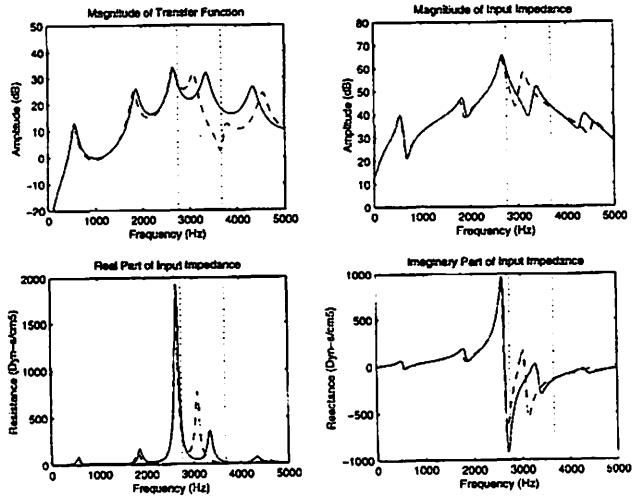

Figure 6. Magnitude of the transfer function, magnitude of input impedance, resistance (real part) and reactance (imaginary part) of the input impedance for the idealized transmission system shown in Figure 3. Solid lines are for no piriform sinuses and dashed lines are for the nominal configuration.

The length chosen here was dictated by a need for an even number of cylindrical sections of a wave reflection model (to be discussed later). The difference is not of fundamental importance, given the variability in laryngeal anatomy among humans.

An interesting situation arises when the length of the piriform sinuses is the same as the length of the epilarynx tube. A resonance and an anti-resonance (a pole and a zero) are then juxtaposed. This situation is shown in Figure 5 (solid lines). Note that the transfer function shows no effect of the zero anymore, but the input impedance rises to a maximum at the combined resonance frequency ($F_{el} = F_{sl} =$ 2756 Hz), which is very close to F_3 of the vocal tract. The low frequency region is basically unaffected, except for a small reduction in F_2.

Total elimination of the piriform sinuses has a similar effect. As shown in Figure 6 (solid lines), the zero is obviously removed from the transfer function, and the input impedance is again raised slightly; another effect is that F_2 and F_4 are raised (relative to the nominal configuration). We conclude, therefore, that the piriform sinuses have the effect of "repelling" the formants on both sides of their resonance frequency. They push F_1 to F_4 lower and F_5 higher. F_3 is not affected by this push because it is highly "attracted" by the epilarynx resonance.

Consider now a widened pharynx in addition to the epilarynx tube and the piriform sinuses, as in the *opera* quality. In Figure 7 (solid lines), A_p was increased from 3 cm² to 6 cm². Given that A_e was still 0.5 cm², the A_p/A_e ratio was now 12 instead of 6. Note that this widened pharynx lowers F_1 and raises F_2. This is well known from basic acoustic theory of an /i/ vowel, which has a widened pharynx

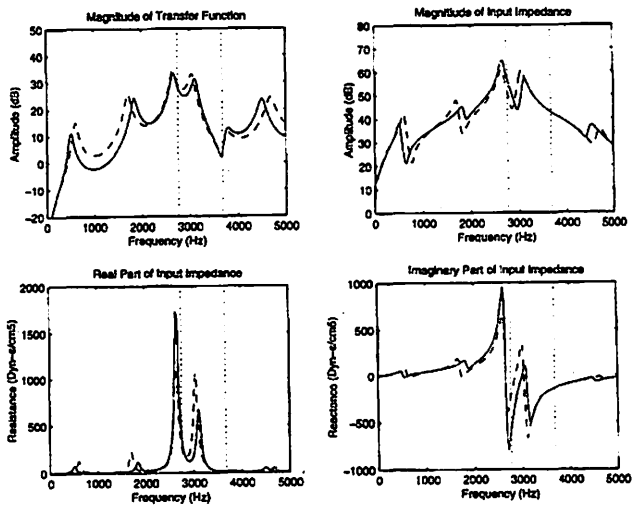

Figure 7. Magnitude of the transfer function, magnitude of input impedance, resistance (real part) and reactance (imaginary part) of the input impedance for the idealized transmission system shown in Figure 3. Solid lines are for a widened pharynx ($A_p = 6.0$ cm²) and dashed lines are for the nominal configuration.

relative to the oral tract. A more interesting result is that the resistance is lowered at F_1 and F_2 but raised at F_3. Furthermore, the reactance curve is smoothed out (less ripple) in the F_1 and F_2 regions and also rises to a large peak at F_3 (near 2500 Hz). F_4 on the other hand, is reduced in both resistance and reactance. Thus, the combination of a narrow epilarynx tube and a wide pharynx is ideal for maintaining a positive and steadily rising inertive reactance from 0-F_{el}, but at the expense of a more compliant reactance above F_{el}. This is

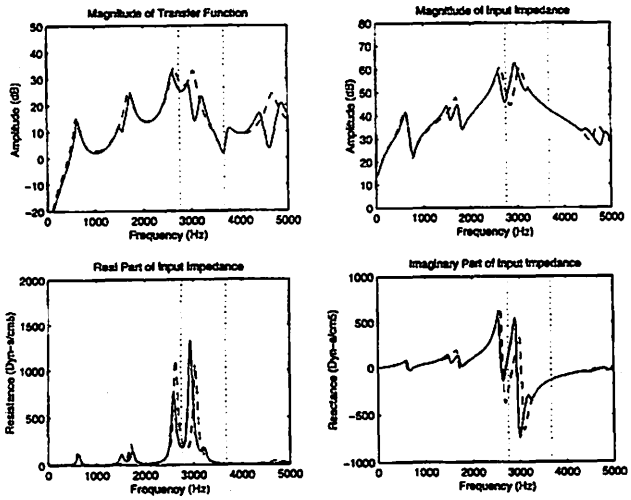

Figure 8. Magnitude of the transfer function, magnitude of input impedance, resistance (real part) and reactance (imaginary part) of the input impedance for the idealized transmission system shown in Figure 3. Solid lines are for an open nasal port (A_n = 10 cm²) and dashed lines are for the nominal configuration.

important for self-oscillation conditions. A positive reactance, combined with a low resistance in the low frequency range, reduces the oscillation threshold pressure. This will be demonstrated in Section III.

If the pharynx is narrowed (to say 1.0 cm²), the opposite effects occur. Low frequency resistance increases, reactance is more fluctuating (even going negative), and F_3 exchanges its level of prominence with F_4. This is the condition for *belt* quality. The increased resistance probably accounts (at least in part), for the greater lung pressures used in this vocal quality.

As a final impedance calculation, consider the effect of the nasal tract. As shown in Figure 8 (solid lines), opening the nasal port to 1.0 cm² and maintaining a uniform nasal tract at this value adds some predictable zeros to the transfer function. Since the tube is open at both ends, the resonance frequencies are at

$$nc/2L_n = n(1575)Hz \quad . \qquad (14)$$

Note that the first three of these zeros are seen in Figure 8 at 1575 Hz, 3150 Hz, and 4725 Hz.

The input impedance shows no profound changes with nasality, but there are a few minor observations. An extra ripple is seen in the F_2 region (both in resistance and reactance) because the nasal zero is close to F_2. Also, the zero at 3150 Hz pushes F_3 and F_4 downward a bit (recall that a zero repels a nearby formant).

Effect of Vocal Tract Impedance on Oscillatory Conditions

The effect of vocal tract impedance on vocal fold oscillation was discussed previously in analytical terms (Titze, 1988). The results are restated here to relate them to the new findings. Neglecting the subglottal system, the pressure-flow equation was written as

$$P_L = \frac{1}{2}k_t \rho U^2/A_g^2 + RU + I\dot{U} \quad , \qquad (15)$$

where P_L is the lung pressure, k_t is a transglottal pressure coefficient (determined empirically to be about 1.1), A_g is the glottal area, U is the glottal flow, $\dot{U}$ is the flow derivative, and R and I are vocal tract input resistance and inertance, respectively. This equation is a nonlinear differential equation in U that can, under certain conditions, produce an oscillatory solution. By relating the glottal area A_g to vocal fold displacement, mass, stiffness, and damping of a simple one-mass oscillator, it was shown that the oscillation threshold pressure is

$$P_{th} = \frac{1}{2}k_t \rho \left(\frac{B}{2LI}\right)^2 \qquad (16)$$

where B is the viscous damping coefficient of vocal fold tissue and L is the vocal fold length. The assumption was made in this derivation that vocal tract resistance R was small in comparison to vocal tract reactance $\omega_o I$ at the fundamental frequency of oscillation ($\omega_o = 2\pi F_o$).

This assumption of small resistance is valid only if the resonance qualities (Qs) of the formants (and epilaryngeal and piriform sinuses) are greater than about 10, which is evident from the impedance and transfer function curves in Figures 4-8. (A resonance Q is defined as the ratio of the resonance frequency to the resonance bandwidth, ± 3 dB down from the peak.)

Consider now a low-frequency approximation to the input impedance Z_i as given in equation 11. If the epilarynx is narrow (0.5 cm² or less) and the expansion into the pharynx is wide (3.0 cm² or more), then $Z_{ps} \approx Z_p$ is negligible in relation to $\rho c / A_e$. Furthermore, for low frequencies, $\sinh(\gamma L_e) \approx \gamma L_e$, and equation 11 reduces to

$$Z_i = \frac{\rho c}{A_e}\gamma L_e \quad . \qquad (17)$$

Substituting γ from equation 2 and β from equation 1, we get

$$Z_i = \frac{\rho c}{A_e}\left(\alpha + j\frac{\omega_o}{c}\right)L_e \quad . \qquad (18)$$

The resistance in this expression is

$$R = \rho c \alpha L_e / A_e \quad , \tag{19}$$

and the inertance is

$$I = \rho L_e / A_e \quad . \tag{20}$$

It is clear that the assumptions of low loss, low frequency, and narrow epilarynx are rather drastic, but they are interesting in that they reduce the vocal tract impedance to nothing but the epilarynx tube impedance. This is a bit like saying that the impedance of a trumpet is the same as the impedance of its mouthpiece. Although it cannot be justified in general, it helps to understand the nature of the source-resonator coupling when the entry to the resonator is narrow.

If we now substitute I from equation 20 into equation 16, the oscillation threshold pressure becomes

$$P_{th} = \frac{1}{8}(k_t / \rho)\left(\frac{BA_e}{LL_e}\right)^2 \quad . \tag{21}$$

Focusing on the A_e/L_e ratio in this expression, it is clear that a long, narrow epilarynx tube can have a substantial effect on lowering the oscillation threshold pressure. This will now be shown with a more sophisticated self-oscillating model.

Simulation of Glottal Flow and Output Spectra

Having discussed the input impedance to the vocal tract in detail, it is now appropriate to examine how this input impedance can affect self-sustained oscillations of the vocal folds. This is best done with numerical simulation. The vocal tract is spatially discretized into 44 cylindrical sections from the glottis to the lips (Figure 9a) and represents the neutral vowel /ə/ of a 30-year old male subject as measured with magnetic resonance imaging (MRI). The length of each section is set to 3.968 mm, as dictated by a chosen sampling frequency of 44.1 kHz and a wave propagation velocity of 350 m/s. A 28 section nasal tract and an 8 section piriform sinus are included. (We are assuming that the two piriform sinuses are identical and that their combined effect can be modeled by a single sinus of twice the cross-sectional area). A subglottal system is also included, which is 36 sections long from the glottis to the bronchial bifurcation.

One-dimensional acoustic wave propagation is simulated in all the tubes according to the algorithms described by Liljencrantz (1985) and Story (1995). Basically, at each junction between two adjacent cylindrical sections,

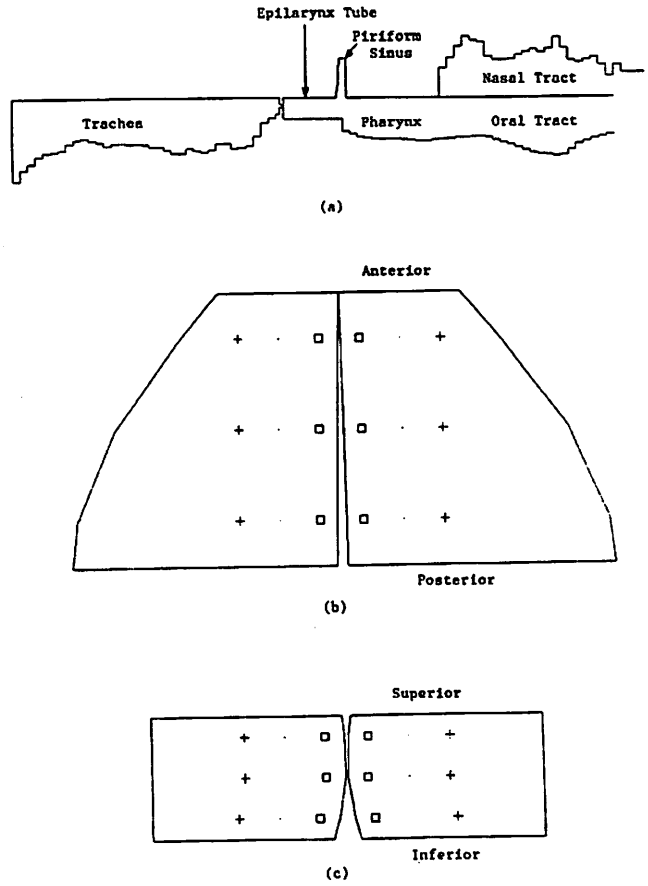

Figure 9. Sketches illustrating the components of a simulation model. (a) Vocal tract outline with the bend removed, showing equivalent tube diameters of all sections, (b) top view of vocal folds, showing tissue-points of the cover (open squares), ligament (dots) and muscle (plus signs) to form a 3 x 3 matrix, (c) coronal section through vocal folds, showing a similar matrix of tissue points.

two reflected waves are computed from two incident waves known from the previous time step. The solution is synchronized in time with the wave velocity through the tubes. Due to the nature of the wave reflection algorithm, the vocal tract is effectively sampled at 88.2 kHz while the glottal source and the pressure output from the vocal tract are sampled at 44.1 kHz (Liljencrants, 1985).

The vocal folds are represented by a three-dimensional array of point masses and their respective nearest-neighbor coupling springs. The array is 3 x 3 x 3, with three masses along the length of the folds, three masses in the vertical dimension and three masses in the lateral dimension (Figure 9 b, c). In the programming, which follows the multi-mass approach by Titze (1973, 1974), Wong et al. (1991), and Story and Titze (1995), the open squares, dots, and plus symbols represent the flesh points of the mucosa, the vocal ligament, and the thyroarytenoid muscle, respectively. The use of three masses in all directions approximates the low-order modes of the vocal fold tissues (Titze &

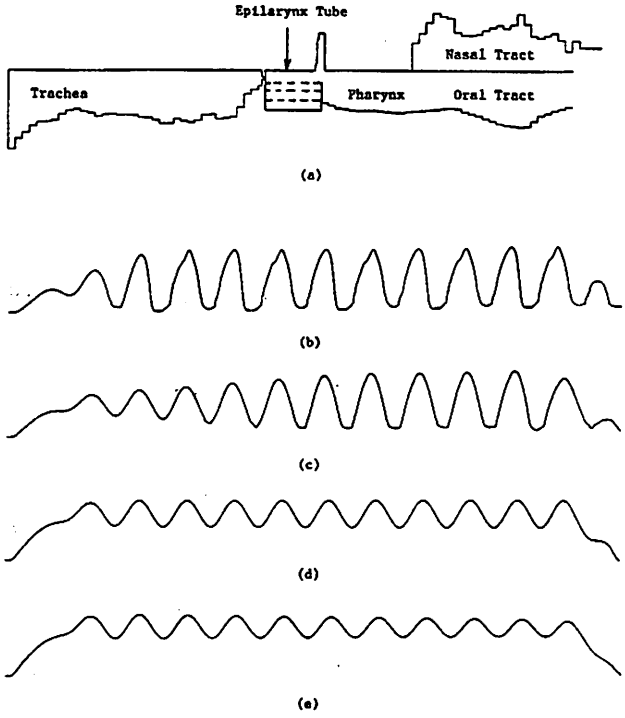

Figure 10. The effect of a widening the epilarynx tube. Dashed lines in part (a) show the equivalent diameters for A_e = 0.2, 0.5, 1.1, and 2.0 cm² and correspond to the glottal flow waveforms shown in (b)-(e). The waveform amplitudes are normalized to show the shape differences only.

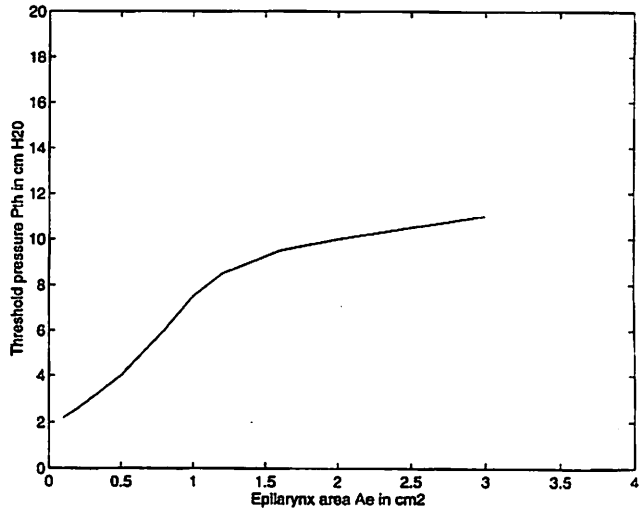

Figure 11. Oscillation threshold pressure (P_{th}) as a function of epilarynx tube area A_e.

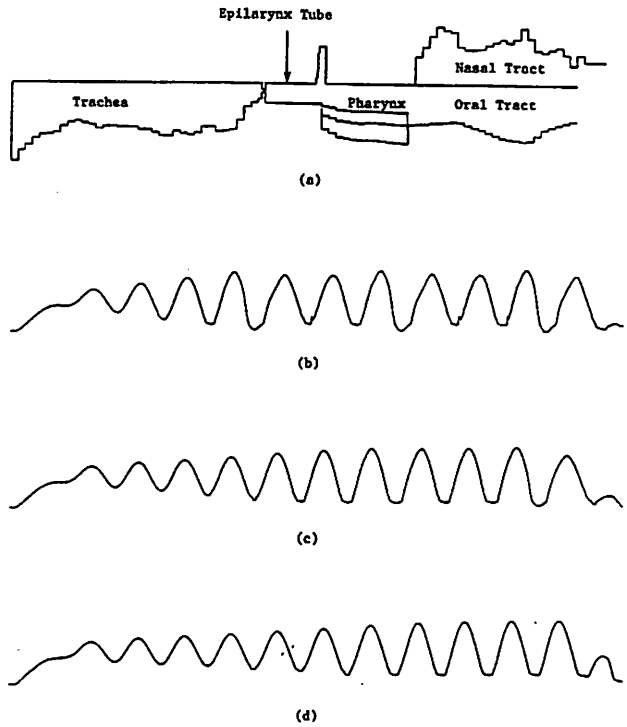

Figure 12. The effect of increasing and decreasing the pharyngeal area by a factor of 2. (a) Vocal tract outline, (b)-(d), glottal flow.

Strong, 1975; Berry, Herzel, Titze, & Kirscher, 1994). It should be kept in mind that increasing the number of masses in each direction not only increases the degrees of freedom of the tissue, but is also allows for a more accurate adjustment of the pre-phonatory glottis. In all simulations to follow, the pre-phonatory glottis was slightly open posteriorly (at the vocal process), closed anteriorly (at the anterior commissure), and curved in the coronal plane as shown in Figure 9. Oscillations were then computed around this configuration.

Figure 10 shows the effect of a widening the epilarynx tube. Part (a) shows the change in vocal tract configuration and Parts (b)-(c) show the corresponding glottal flow waveforms. All parameters of the vocal fold model were kept identical; only A_e was systematically changed from 0.2 cm² to 0.5 cm², then to 1.1 cm², and finally to 2.0 cm². The lung pressure was always maintained 8 cm H_2O (about 0.8 kPa). Note that this pressure is the oscillation threshold pressure for A_e = 1.1 cm² because the amplitude is neither growing nor decaying and vocal fold collision has not been reached.

The oscillation threshold pressure was probed for many values of A_e by moving gradually from damped oscillation to growing oscillation without collision. The results are shown in Figure 11. Note the strong dependence of P_{th} on epilaryngeal tube area in the lower range of A_e. The curve appears to follow a parabolic path (up to about 1.0 cm²) as predicted by equation 21. Above about 1.0 cm², however, the parabolic path is abandoned and the threshold pressure flattens out. This is because the resistive part of the vocal tract becomes more significant and the epilarynx tube merges with the vocal tract as a single tube of larger dimensions. The vocal folds then vibrate more independently of the vocal tract, their threshold pressure being determined by the tissue properties (Titze, 1988).

Of all the other parameters of the vocal tract probed, none had the dramatic effect on oscillation that A_e had. Figure 12 shows the effect of increasing and decreasing the pharyngeal area by a factor of 2. The narrowed pharynx (typical of *twang* and *belt* quality), produced some roughness in the waveform (a period-3 subharmonic). This is probably because the epilarynx tube is now acoustically coupled to the entire vocal tract. It no longer serves as an independent resonator, but mixes with all the formants of the vocal tract. Hence, the source-tract interactions are more complicated.

Conclusions

A number of interesting conclusions can be drawn from this study. First and foremost, a narrow epilarynx acts a bit like the mouthpiece of a brass instrument, matching the high internal impedance of the glottis to the lower impedance of the vocal tract and free space. In terms of an impedance matching concept, the small-amplitude acoustic impedance of the glottis can be approximated as

$$Z_g = \frac{dP_g}{dU} = \frac{d}{dU}\left[\frac{1}{2}k_t \rho v^2\right] = k_t \rho v \frac{dv}{dU} = k_t \rho v / A_g \quad , \tag{22}$$

where P_g is the pressure across the glottis and v is the glottal air particle velocity. Glottal inertance and several shape-dependent factors are neglected. This relationship was introduced earlier in equation 15, with U being the glottal flow, k_t being an empirical pressure coefficient (about 1.1), ρ being the air density, and A_g being the glottal area. Z_g is a time-varying impedance, of course, both v and A_g changing throughout the glottal cycle. But we are only interested in order-of-magnitude effects here.

If we compare this impedance with the intrinsic acoustic impedance of the epilarynx section of the transmission line,

$$Z_e = \rho c / A_e \quad , \tag{23}$$

which is also only a rough estimate because it neglects all the standing waves in the line, there is a clear symmetry between *particle velocity* in the glottis and *wave velocity* in the vocal tract. Likewise, there is a symmetry between *glottal area* and *epilarynx tube area*. The ratios v/c and A_g / A_e are apparently significant in establishing general impedance matching conditions. Typically, v is on the order of 20-50 m/s during glottal opening and c = 350 m/s. Thus, the v/c ratio is on the order of 0.1. Also, A_g / A_e is typically on the order of 0.1 if the time-average glottal area is about 0.1 cm^2 (a 1 cm glottal length times a 1 mm glottal width) and A_e is on the order of 1.0 cm^2. To maintain similar impedance conditions across pitches and loudnesses, a vocalist may want to decrease the epilaryngeal area when the glottal area decreases. This may occur at high pitches or low intensity. Also, the air particle velocity v can be adjusted by subglottal pressure to bring the v/c ratio in line.

It is not yet clear, however, whether the larynx works best as a constant flow source (with $Z_g >> Z_e$) or under nearly matched conditions ($Z_g \approx Z_e$). In normal speech, and more so in the *sob* and *falsetto* quality, the epilarynx tube can be fairly wide. There appears to be no great difficulty in maintaining vocal fold oscillation in these cases. In high-pitched operatic singing, *belting*, and *twang* quality, on the other hand, the narrow epilarynx seems to be the configuration of choice. It may also be the preferred configuration in the so-called *resonant speaking voice* (Verdolini et al., 1994). This configuration provides the desirable inertive reactance to facilitate vocal fold oscillation. When the narrowed epilarynx is combined with a wide pharynx, the reactance never goes negative below about 3000 Hz, which means that the acoustic load is inertive for all possible values of F_o.

Our findings confirm the earlier results of Sundberg (1974) that the epilarynx tube clusters the third, fourth, and fifth formants to generate the vocal *ring* (singer's formant). The focal point in the spectrum is the uncoupled (free) resonance frequency of the epilarynx tube, which can be computed simply on the basis of tube length. The epilarynx resonance frequency "attracts" all formant frequencies of the vocal tract. Generally, F_1 and F_2 are pulled upward, whereas F_4 and F_5 are pulled downward; F_3 is often entrained by the epilarynx resonance, thereby not moving much. But a short epilarynx tube can also entrain F_4, in which case F_3 will be pulled upwards toward it.

The piriform sinuses introduce a zero into the transfer function, but have no profound effect on the magnitude of the input impedance. The formant frequencies are shifted slightly, however. In contrast to the epilarynx tube, the piriform sinuses "repel" the formants, generally pushing F_1, F_2, F_3, and F_4 lower and pushing F_5 higher. If the length of the piriform sinuses is the same as the length of the epilarynx tube, the pole-zero pair produces a complete cancellation in the transfer function, but the impedance remains high.

An open nasal port introduces zeros into the spectrum (at the resonance frequencies of the vocal tract), but in this study, it showed no measurable effect on oscillation threshold pressure or glottal flow. We suspect, therefore, that the highly touted benefit of nasalization in singing is less acoustic than biomechanical. As Estill (1995) has pointed

out, the palatopharyngeal muscles probably help to stabilize the larynx to maintain a constant epilaryngeal configuration. This may open the nasal port or close it in the process. Acoustically, it may not matter much, but our investigation was not extensive enough throughout the F_0 and intensity ranges. Further investigation of nasality as a pedagogical tool is therefore recommended.

Acknowledgement

This work was supported by a grant from the National Institute on Deafness and Other Communication Disorders, No. P60 DC00976.

References

Bartholomew, W.T. (19834). A physical definition of "good voice-quality" in the male voice. *Journal of the Acoustical Society of America, VI*, 1, 24-33.

Berry, D., Herzel, H., Titze, I., & Krischer, K. (1994). Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions. *Journal of the Acoustical Society of America, 95*(6), 3595-3604.

Colton, R.H., & Estill, J. (1981). Elements of voice quality: Perceptual, acoustic and physiologic aspects. In N.J. Lass (Ed.), *Speech and Language: Advances in Basic Research and Practice*, Vol. V. New York: Academic Press, pp. 311-403.

Dang, J., & Honda, K. (1996). Local and global effects of the pyriform fossa on speech spectra. *Journal of the Acoustical Society of America, 98*(No. 5, Pt. 2), pg. 2931.

Dang, J., & Honda, K. (in press). Acoustic characteristics of the piriform fossa in models and humans. *Journal of the Acoustical Society of America.*

Estill, J. (1995). Voice Craft: With Compulsory Figures. Estill Voice Training Systems, New York.

Feder, F. (1988). Personal communication, based on videostroboscopic examination of professional voice impersonators.

Flanagan, J.L (1972). *Speech Analysis: Synthesis and Perception.* Berlin: Springer-Verlag.

Laver, J. (1980). *The Phonetic Description of Voice Quality.* Cambridge: Cambridge University Press.

Liljencrants, J. (1985). Dynamic line analogs for speech synthesis. *Quarterly Progress and Status Report, STL-QPSR* 1/1985. Speech Transmission Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden, pp. 1-14.

Sondhi, M., & Schroeter, J. (1987). A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-35*, 7, 955-967.

Story, B., & Titze, I.R. (1995). Voice simulation with a body cover model of the vocal folds. *Journal of the Acoustical Society of America, 97*(2), 1249-1260.

Story, B. (1995). *Physiologically-based Speech Simulation Using an Enhanced Wave-Reflection Model of the Vocal Tract.* Doctoral dissertation, University of Iowa, Iowa City.

Story, B., Titze, I., & Hoffman, E. (1996). Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America, 100*(1), 537-554.

Sundberg, J. (1974). Articulatory interpretation of the singing formants. *Journal of the Acoustical Society of America, 55*, 838-844.

Titze, I.R., & Strong, W.J. (1975). Normal modes in vocal cord tissues. *Journal of the Acoustical Society of America, 57*(3), 736-744.

Titze, I.R. (1973). The human vocal cords: A mathematical model. Part I. *Phonetica, 28*(3-4), 129-170.

Titze, I.R. (1974). The human vocal cords: A mathematical model. Part II. *Phonetica, 28*(3-4), 129-170.

Verdolini, K., Druker, D., Palmer, P., & Samawi, H. (1994). Psychological study of "resonant voice". *National Center for Voice and Speech Status and Progress Report*, Vol. 6, 147-153.

Yanagisawa, E., Estill, J., Kmucha, T., & Leder, S.B. (1990). The contribution of aryepiglottic constriction to "ringing" voice quality - A videolaryngoscopic study with acoustic analysis. *Journal of Voice, 3*, 342-350.

Wong, D., Ito, M., Cox, N., & Titze, I. (1991). Observation of perturbations in a lumped-element model of the vocal folds with applications to some pathological cases. *Journal of the Acoustical Society of America, 89*(1), 383-394.

NCVS Status and Progress Report - 10
November 1996. 35-44

A Numerical Simulation of Laryngeal Flow in a Forced-Oscillation Glottal Model

Fariborz Alipour, Ph.D.
Chenwu Fan, M.S.
Department of Speech Pathology and Audiology, The University of Iowa
Ronald C. Scherer, Ph.D.
Department of Speech Pathology and Audiology, The University of Iowa
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts

Abstract

A numerical simulation of laryngeal flow was developed to study flow patterns and pressure and velocity waveforms in a model of the oscillating glottis. The unsteady Navier-Stokes equations were solved with a finite volume method using a nonuniform staggered grid. The numerical method was tested against published experimental data. In this study of glottal aerodynamics, the vocal folds independently and sinusoidally were moved from a converging to a diverging and back to a converging shape, and the input airflow sinusoidally varied from zero to a maximum and back to zero. The typical results were obtained for a Reynolds number of 2000 and for an oscillation frequency of 100 Hz. Results indicate that with this simulation of the entire flow field, periodic velocity and pressure fields exist throughout the laryngeal duct. The airflow separates within the glottis, creating intraglottal (and downstream) asymmetric flow throughout the glottal cycle, with formation of eddies downstream of the glottis. The observed maximum velocity delays due to the glottal wall movement would contribute to the well-known glottal volume velocity skewing during phonation.

Introduction

The study of laryngeal aerodynamics is essential to the development of a more accurate and complete theory of phonation. Air pressure distributions, friction factors, and airflow are required in the assessment of the aerodynamic forces on the vocal fold surfaces, forces which contribute significantly to vocal fold oscillations and the creation of sound (Alipour & Titze, 1988). First order estimates of pressure and velocity distributions within the glottis were obtained from the data gathered either from empirical studies of steady laryngeal flows (Ishizaka & Matsudaira, 1972; Scherer, 1981; Scherer, Titze & Curtis, 1983; Scherer & Titze, 1983; Binh & Gauffin, 1983; Scherer & Guo, 1990, 1991) or from theoretical studies of steady flow in laryngeal models (Liljencrants 1991; Iijima, Miki & Nagai, 1992; Alipour & Patel, 1991, 1994; Guo & Scherer, 1993). These types of models have been extrapolated to oscillating conditions (pulsatile flow) under the quasi-steady assumption (Alipour & Titze, 1988).

Due to advances in computer and measurement technology, computational and empirical research of pulsatile laryngeal flow can augment steady flow studies. The experimental work on laryngeal pulsatile flow using excised larynges is promising (Berke et al., 1989; Alipour & Scherer, 1995), despite spatial limitations for transducers and lack of easy access to certain critical locations such as the intraglottal space during phonation. Numerical simulation of pulsatile flow in the glottis, the topic of this study, benefits from the validation data of empirical studies, and provides extensive aerodynamic predictive data throughout the laryngeal duct.

The purpose of this study was to develop a computational model that predicts pressure and velocity distributions within a model of the laryngeal airway during oscillations at typical ranges of fundamental frequency and flow rate. At the current stage of development, a forced oscillation model is employed to demonstrate the capability of this numerical technique in laryngeal aerodynamics. The model, described below, independently "forces" the vocal folds to move while applying an input flow. In this way the model

avoids complications of the more complex biomechanic-aerodynamic modeling (Alipour & Titze, 1988), while providing a means to study aerodynamic details of pulsatile flow. The advantages of such an approach are to study the effects of the moving glottal walls and the contributions of the various terms of the Navier-Stokes equations. The model developed here can then be applied to the more complex biomechanical modeling of speech production.

Oscillating Wall Model

Since the focus of this study was to apply a two-dimensional computational model of air flow through the glottis, the vocal fold tissue mechanics were simplified to a two-dimensional forced oscillation model wherein the vocal folds were independently "forced" to move in a prescribed manner. The model here is therefore not a self-oscillation model of the vocal folds. The forced oscillation model was built with a three-piece wall as shown in Fig. 1. The first piece of the glottal wall is the glottal entry section which is a fixed sinusoidal curve with height of $H1$ extending from point (1) on the flat wall to point (2) as shown on Figure 1. The glottis is created by a straight segment (2-3) approximately 0.95 cm in length. The oscillating glottal wall is attached to a cosine shaped wall (3-4) at the glottal exit, the height of which is $H2$, which varies according to:

$$H2 = H1 + A\cos\omega t \tag{1}$$

where A is amplitude, $\omega=2\pi f$, and f is the frequency of oscillation. The tangent line (2-3) forming the medial glottal surface and its points of contact were calculated analytically at every simulation time step. Sinusoidal motion of the glottal wall has been assumed in other useful phonatory models (e.g., Titze, 1988).

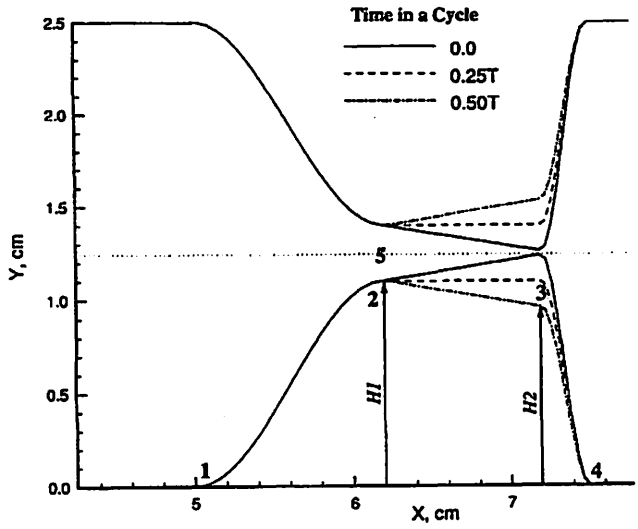

Figure 1. Schematic of moving glottal wall model.

The position of the glottal wall was updated at every time step and glottal constriction was simulated using a method similar to the so-called 'shadow method' used in heat transfer (Patankar, 1980). In this method a large source term is assigned to the region under the wall to ensure zero air velocity within the body of the vocal folds. Use of this method has three advantages. The first advantage is that the grid points are fixed in the domain and boundary motion has no effect on them, and thus they do not need to be calculated and updated at every time step. The second advantage is that this permits a more convenient form of the governing equations for a fixed grid configuration. The third advantage is the saving of computation time that would have been used for grid generation, calculation of metric coefficients, and the solution of more complicated equations. One of the disadvantages of this method is the difficulty in applying boundary conditions on the moving surface. Another disadvantage is the difficulty in estimating wall friction. The wall friction is useful in the prediction of the flow separation point on the moving boundaries and calculation of shear forces on the walls. These disadvantages will be overcome in time as code refinement continues and computation time decreases.

The glottal configuration had a constant glottal entry diameter of 0.3 cm. The glottal exit diameter varied during the phonatory cycle from a maximum of 0.58 cm to a minimum of 0.02 cm. Note that the inlet airflow was modulated sinusoidally with a minimum of zero flow, and the tissue modeling did not permit full glottal closure. The flow field of this simulation included a portion of the trachea, the glottis, and a portion of the pharynx without the ventricular folds. The pharyngeal duct was approximately half a meter in length (to reduce the effect of glottal wall motion on the exit boundary conditions).

Numerical Method

The governing equations were continuity and unsteady Navier-Stokes equations for two-dimensional incompressible laminar flow. In primitive dimensional form, These appear as:

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0 \tag{2}$$

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=-\frac{1}{\rho}\frac{\partial p}{\partial x}+\frac{\mu}{\rho}\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right) \tag{3}$$

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}=-\frac{1}{\rho}\frac{\partial p}{\partial y}+\frac{\mu}{\rho}\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}\right) \tag{4}$$

where u and v are velocity components (in m/s), p is pressure (in Pascal), x and y are coordinates (in meter), and t is time (in seconds). The mean flow rate is expressed in terms of

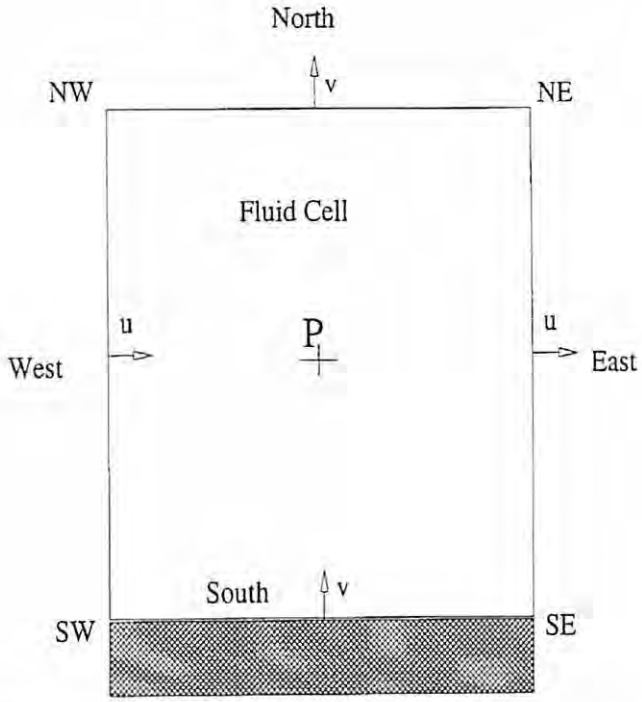

Figure 2. Schematic of finite volume cell. The shaded area represents the solid boundary.

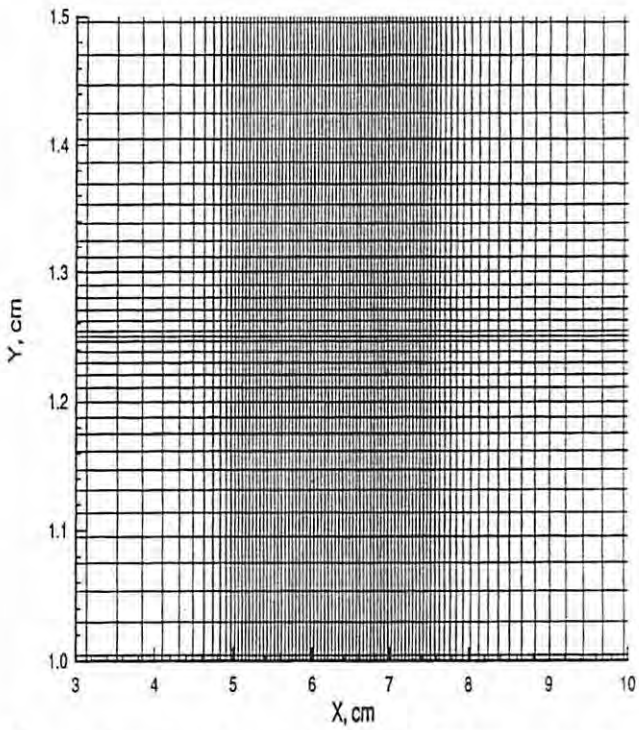

Figure 3. A portion of the computational grid. The x-axis is in the axial flow direction.

Reynolds number (Re), an important control parameter that indicates the flow regime, which was defined as $Re = \rho U_0 L/\mu$, where U_0 is the inlet cross-sectional average velocity, L is a reference length (channel height), μ is air viscosity and ρ is the air density.

These equations were solved with a cell-centered finite volume method with staggered grids (Patankar, 1980). In this method the governing equations (2-4) are integrated over an element such as a quadrilateral with sides denoted by "west-east" perpendicular to the x axis, and "south-north" perpendicular to the y axis (Figure 2). Applying the Green's Theorem and approximating the integration over the quadrilateral domain results in a set of algebraic equations to be solved. As shown in Figure 2, the grid is called staggered because the pressure and velocity components are not calculated at the same location. The pressure is calculated at the center of the cell (finite volume), velocity component u is calculated at the east and west faces, and velocity component v is calculated at the north and south faces. These calculations are required for the balance of momentum within each cell. The corner velocities and pressures can be calculated by interpolation.

Recent experimental results on pulsatile laryngeal flow have indicated that the inlet flow may resemble a laminar parabolic flow with a maximum velocity that changes with time in a periodic fashion (Alipour & Scherer, 1995). Thus the flow was modeled according to a sinusoidal varying rate with the same frequency as the glottal wall movement. The phase of the inflow was set such that as the glottis was closing, the flow diminished according to

$$u = 3y(1-y)(1-\cos \omega t) \qquad (5)$$

This is realistic in that real glottal flow essentially changes with glottal area. This method, however, does not impose a constant subglottal (or lung) pressure, and thus the pressure field may be different from that found in a real larynx.

The choice of the grid is crucial in any computational fluid dynamic analysis. Here a nonuniform rectangular grid was selected such that regions of higher velocity and larger pressure gradient contain more grid points. Figure 3 shows a portion of the grid that was used. Since the glottal gap changed during each cycle of oscillation, a logarithmic distribution of grids was designed to ensure the presence of a relatively large number of grid points in the region near closure. The typical grid had a total of 120 divisions in the x-direction with 15 divisions in the tracheal region, 55 divisions in the glottal region, and 50 divisions in the pharyngeal region. In the y-direction, 50 divisions were distributed from the wall to the centerline.

The Navier-Stokes equations were solved with the appropriate pressure and velocity boundary conditions at every time step. These included inlet and outlet conditions and the no-slip condition on the walls. The pressure gradient was treated as a mass source in the momentum equations and

its solution was obtained from continuity by an iteration method. The pressure boundary condition was thus enforced through the continuity of mass or the specification of the inlet and outlet mass flux. For the inlet boundary condition, the inflow was set to be a sinusoidally varying flow with the same frequency as the wall movement. For the outlet boundary, velocity was extrapolated linearly from the inner nodes. Also, the outlet pressure (at the end of the vocal tract) was set to zero. The outlet location was far enough from the glottis for these conditions to apply. The outlet mass flow could be different from the inlet mass flow, the difference just equaling the mass flux caused by the wall movement. The displacement flow of the moving wall (wall flux) was calculated from the wall velocity and was included in the boundary conditions.

The momentum equations (equations 3 and 4) were discretized for finite volume in space and for finite difference in time. The resulting equations were solved at every time step, and the velocity components were obtained by iterations and successive overrelaxation. The time integration was performed with a forward difference scheme. The velocity and pressure fields were updated at every time step of 50 microseconds. An exponential scheme was employed in discretizing the convective term to achieve better convergence (Patankar, 1980). Then pressure was calculated from continuity using the SIMPLER algorithm on the staggered grids (Patankar, 1980). Computations were performed initially on a DEC-Station 5000 and later on a Silicon Graphics Indy2 workstation. The solution to the N-S equations were obtained either for the lower half of the channel (forced symmetry) or for the whole domain. The half domain solution was used for grid resolution and static wall tests, the time variation of the N-S equations terms, and for some of the velocity and pressure profiles along the channel. The full domain solution was used to present the time dependent velocity profiles, flow patterns, and potential asymmetries in the channel.

Grid Resolution Test

Calculations were performed with three different grids in the solution domain to examine the sensitivity of the numerical solutions to grid refinement. Figure 4 shows the center pressure P_c and center velocity Uc for grid resolutions of 100 x 40, 120 x 50, and 150 x 60 for the Reynolds number 1000 and frequency of 100 Hz. The coarsest grid required 41 minutes CPU time per cycle on the Silicon Graphics machine and the finest grid used 4.5 hours of CPU per cycle. There were 200 steps in each cycle for these simulations. The two fine grids yield similar results; pressure are within 3% of each other, and velocity is within 1% except for the peak velocity which shows a difference of 10%. It is encouraging to note that even the coarser grid captures the main features of the pressures and velocities in the model.

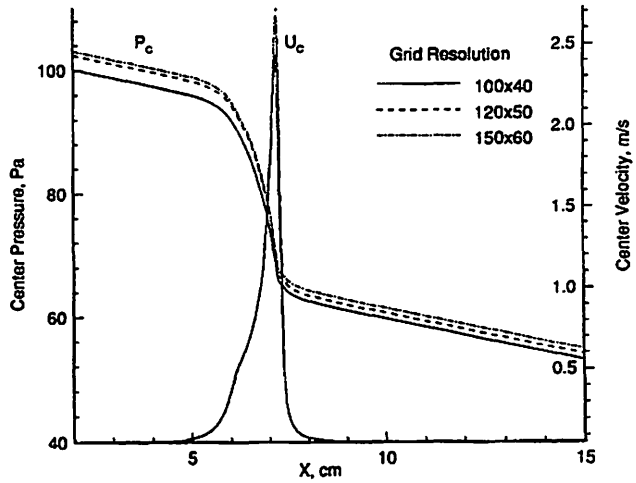

Figure 4. Effect of grid resolution on the center pressure and center velocity.

Simulation Tests

To test the simulation algorithm, results are obtained and compared with other methods for (1) a classical unsteady problem, (2) a body-fitted steady solution to a static model of the glottis, and (3) steady flow over a backward-facing step.

In case 1, Stokes' second problem, fluid oscillation above an infinite plate is considered. In this case, the lower plate is set to oscillate with $u(0,t)=U_0\cos\omega t$, with fluid in the farfield at rest. The analytic solution to this problem is known (White, 1991) to be

$$u=U_0\exp(-\eta)\cos(\omega t-\eta), \qquad \eta=y(0{\cdot}5\omega/\nu)^{0{\cdot}5} \qquad (6)$$

where U_0 is the amplitude of the plate oscillation, η is the normalized distance from the wall, y is the distance from the wall and ν is the kinematic viscosity of the fluid. Figure 5 shows the analytic and numerical solutions from this model for every 45 degrees of the oscillation cycle for a frequency of 10 Hz. As seen in Fig. 5, except for the first two angles, the model predicts the time varying solutions with reasonable accuracy.

The next case was intended to see if shadowed constriction has a reasonable pressure and velocity variation. For this purpose, steady flow with a Reynolds number of 900 was solved with this model for a static glottal model and was compared with results for a similar flow (the same Reynolds number) in the same channel based on a body-fitted coordinated (BFC) model (Alipour and Patel, 1994). Figure 6 demonstrates the variation of center pressure (P_C) and center velocity (U_C) along the channel for the two models. The solid line represents the BFC results and the dashed line is for the model proposed here. Although the

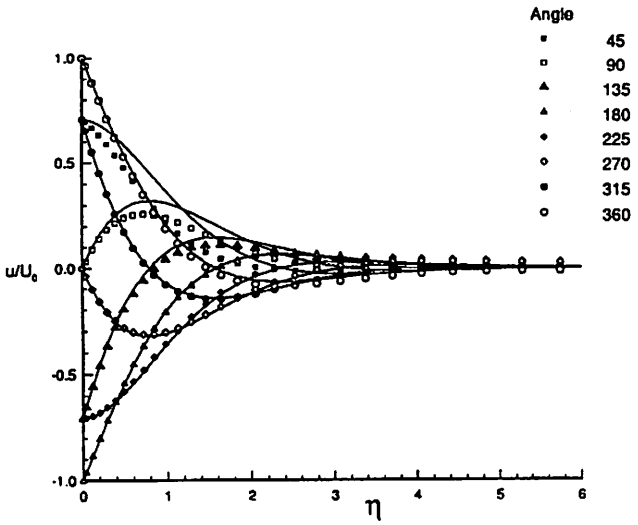

Figure 5. Analytical (lines) and solution from this model (symbols) of Stokes' second problem.

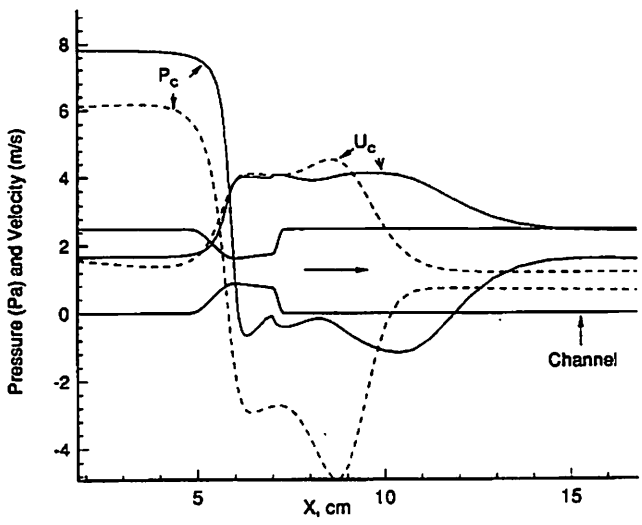

Figure 6. Steady flow in a divergent glottal model at a Reynolds number of 900 solved with body-fitted coordinates (solid lines) and the model proposed in this study (dashed lines).

pressure results do not overlap significantly, the amount of transglottal pressure drop differs by only 0.55 cm-H_2O and both models show the sharp pressure drop with minor recovery in the divergent glottal section. The BFC method predicted lower pressure within the glottis. The center velocity predicted by the two models are nearly identical in the glottis. However, downstream of the glottis, where vortex shedding may be strong, the velocities and pressures predicted by the two models have some discrepancies. The pressure predictions within and downstream of the glottis for the BFC model appear significantly low compared with results using physical models with a diverging glottis, whereas the method developed here appears to predict more realistic pressures (Scherer, 1981, 1983; Scherer and Titze, 1983).

The third case was flow in a backward facing step that was calculated and compared with the experiments by Armaly et al. (1983). They measured the velocity distribution downstream of a single backward-facing step using laser-Doppler anemometry for steady two-dimensional flow. The velocity profiles they reported for a Reynolds number of 1095 were used to test our simulation model. Figure 7 shows the comparison of velocity profiles downstream of the backward-facing step from the data of Armaly et al. and simulated profiles from the model of this report. In Fig. 7, profiles are shown at locations (x) relative to the edge of the step, normalized to the step's height (S). A reasonable accuracy can be observed in the magnitude of peak velocities. The simulation develops the flow faster than what is indicated by Armaly. That is, the outlet velocity distribution is preset by us, which forces a faster development of the flow, and the corresponding faster reduction of the duct vortices, eddies and other phenomena, than shown by Armaly's data. Otherwise, the prediction is qualitatively similar to the data.

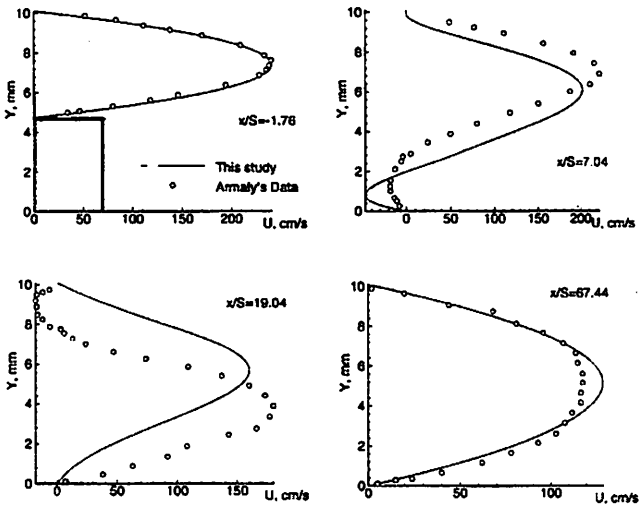

Figure 7. Velocity profiles in a backward-facing step solved in this study and compared with experimental data of Armaly et al. (1983). The parameter (x/S) refers to the location of the cross-section from the edge of the step (x) and the step height (S).

The grid and simulation tests suggest that the method proposed here both qualitatively and quantitatively should perform reasonably well in the current study of glottal flow dynamics.

Results and Discussion

Although the tracheal flow and glottal motion may be different from that expected during normal phonation, sufficient similarity exists that makes this model a suitable tool to study unsteady flow relevant to phonation. Aerodynamic characteristics such as flow separation and vortex

shedding during the modeled glottal wall oscillation should be predictive for human phonation. In this section, some cases of oscillations at Reynolds number of 2000 and frequency of 100 Hz will be discussed. The presentation of the results of the simulation experiments will give mean glottal velocities and the centerline air pressures for typical conditions. The time variation of various terms in the Navier-Stokes equations will be shown at typical locations. In addition, streamlines throughout the flow field will be shown and discussed. These data will demonstrate the potential usefulness of the simulation.

Figure 8 shows the time dependent center pressures for a simulation with Reynolds number of 2000 and frequency of 100 Hz. Except within the glottal region, the pressure lines are linear. They all meet at the end of the channel where the outlet pressure was set to zero. The linearity is predictable due to the uniformity of the channel downstream of the glottis. The pressure gradient or slope of these lines varied periodically due to the imposed sinusoidal (always positive) inlet flow. Because of the absence of the positive lung pressure, at some portion of the cycle the pressure gradient becomes positive and will cause flow reversal in this model. This model allows close examination of these reversal trends due to both the sinusoidal flow input and the glottal wall motion. One major feature of the flow is that the location of maximum velocity moves from the center towards the wall, and the velocity profile may have a notch in the center (White, 1991). The symmetric or half-domain solution was obtained for this case.

In order to better understand and interpret the results of this study, we will show and discuss the contributions of the four terms of the x-momentum (N-S) equation related to the results in Figure 9. The Navier-Stokes equation used here is comprised of the unsteady term, $(\partial u/\partial t)$, the convective term, $(u\,\partial u/\partial x + v\,\partial u/\partial y)$, the pressure gradient term, $(-1/\rho)\,\partial p/\partial x$, and the diffusion term, $(n\nabla^2 u)$. The contributions of these terms for a Reynolds number of 2000 and the oscillation frequency of 100 Hz at location 5 will be discussed. Location 5 is at the center of the glottal entrance where the fixed gap is 0.3 cm (ref. Figure 1). Figure 9 shows that in this location, the dominant terms are the unsteady (solid line) and pressure gradient (dash-dot line) terms. Essentially the pressure gradient term is balanced by the unsteady and convective terms (dashed line), with the diffusion term (dotted line) being negligible. In this modeling it is important to re-emphasize that the velocity is an input, and is sinusoidal beginning with zero value. The unsteady term is therefore essentially a derived input term, giving rise to the negative values during the second half of the cycle, and the pressure gradient is a dependent term, also becoming negative within the second half of the cycle. The convective term is a derived input term, dependent primarily on the increase in the velocity input.

Figure 10 shows the cross sectional average of velocity for five cycles of glottal wall motion. Data are given for the three locations, upstream of the glottis, glottal entry, and glottal exit (as shown in Fig. 1). The solid lines refer to the upstream section, the dashed line to glottal entry, and the dot-dash line to glottal exit. It is noted again that the upstream velocity waveform and the glottal motion are predetermined. The velocities elsewhere and all pressures are dependent results. Figure 10 shows that the velocity waveform at glottal entry is also symmetric and follows the upstream velocity well, whereas the average glottal exit velocity is skewed to the right. The cycle begins with the

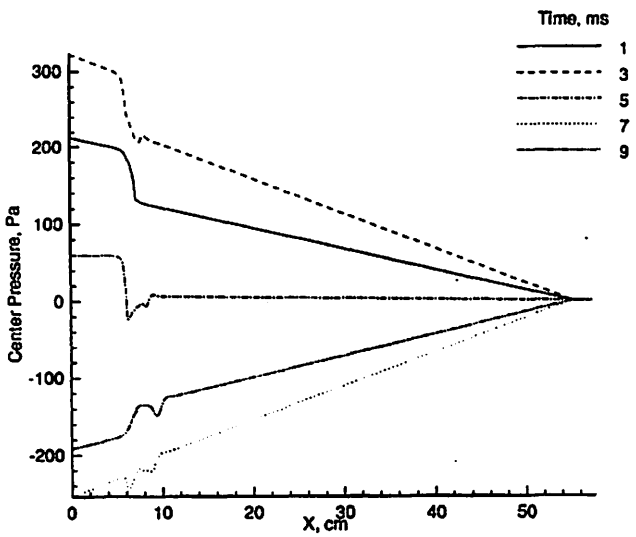

Figure 8. Time dependent pressure profiles along the channel within a cycle.

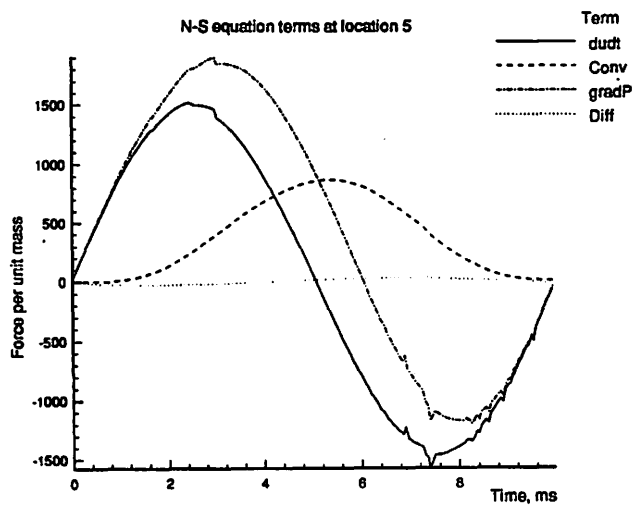

Figure 9. The time variation of the terms in the Navier-Stokes equations.

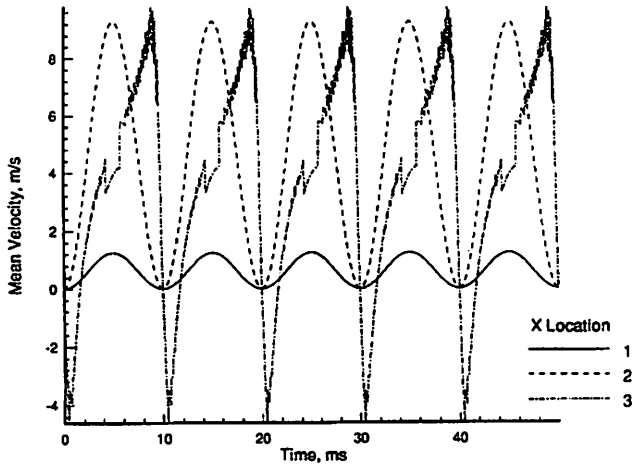

Figure 10. Waveforms of mean velocity profiles at three cross 1, 2 and 3.

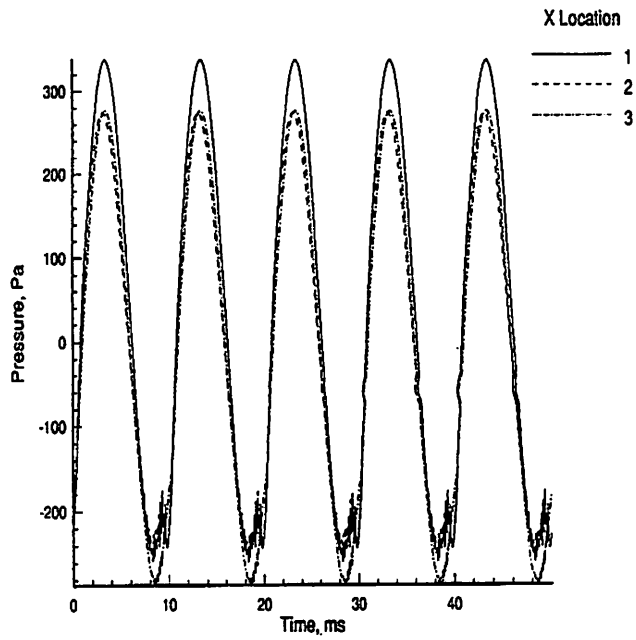

Figure 11. Waveforms of centerline pressure at cross-sections similar to Fig. 10.

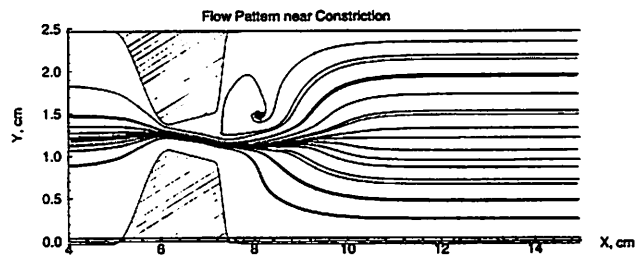

Figure 12. Flow pattern near glottal model at Re=2000, Fo=100.

glottal exit diameter being minimal. At mid-cycle, the walls have separated maximally (keeping the glottal entrance diameter constant), and then at the end of the cycle, the glottal exit is again at its minimum diameter. The skewing of the velocity signal is due to the glottal motion. As the walls separate, the glottal area increases, retarding the average velocity (note the negative velocities at the beginning of the cycle), and as the walls come together, the velocity is increased, resulting in the skewing. The maximum average velocity occurs while the walls still form a divergent glottis. The implication is that the motion per se of the glottal walls accounts for some of the skewing of the glottal volume velocity commonly seen in human phonation.

The waveforms of the average cross sectional pressure (Figure 11) are essentially in phase with each other for the three spatial locations. If the displacement flow due to wall motion were not included, the pressure waveforms would be zero at time zero (i.e., at the beginning of the cycle), as we have verified without showing a figure here. The inclusion of the displacement flow creates a phase shift of not only the average glottal exit pressure, but also of the pressures at the other two locations. Thus, whereas the average glottal entry velocity appears independent of the average glottal exit velocity, the pressures everywhere are highly dependent upon the wall displacement and the skewed exit velocity, but the pressures throughout the tract essentially stay in phase with each other.

Figure 12 displays a typical flow pattern within the glottal model, including a portion of the upstream and downstream duct, at an instant of divergent glottal shape. Despite the symmetric geometry of the walls, the flow pattern is asymmetric. This asymmetric solution is obtained when the whole channel height is solved and Reynolds number is large. This asymmetry of the flow in a plane symmetric channel has been reported by other investigators (Durst, Melling, and Whitelaw,1974; Durst, Pereria, and Tropea, 1993; Sobey and Drazin, 1986; Kim and Patel, 1992). For example, Durst et al. (1974) reported that at a Reynolds number of 114, the separation regions behind the sudden expansion (upper and lower) were of different length, and the flow became asymmetric. They reported that at higher Reynolds numbers multiple separation points may appear and there may be a few stable solutions to each asymmetric flow.

The converging streamlines in Figure 12 indicate an increase in the particle velocity within the glottis, forming a jet. The jet appears to bend toward one of the walls even with wall motion. The location of the maximum velocity in the jet may be obtained from the velocity profiles at each cross section.

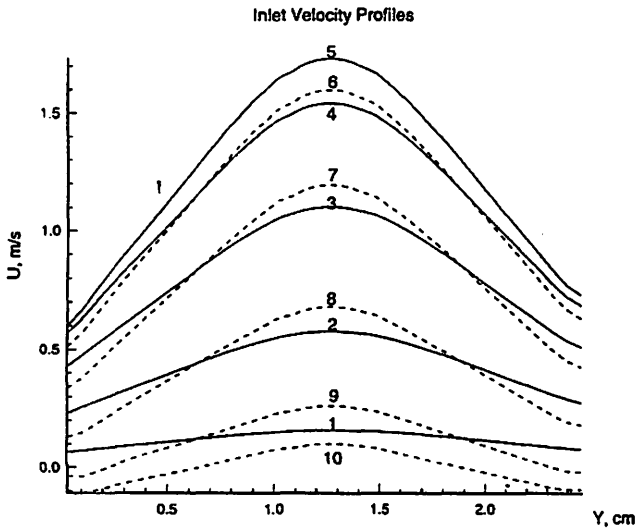

Figure 13. Velocity profiles at the inlet to the glottis (X=5 cm).

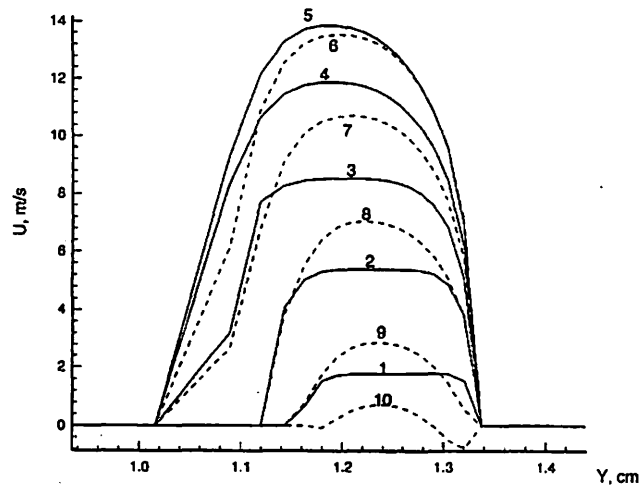

Figure 14. Velocity profiles within the glottis (x=6.7 cm).

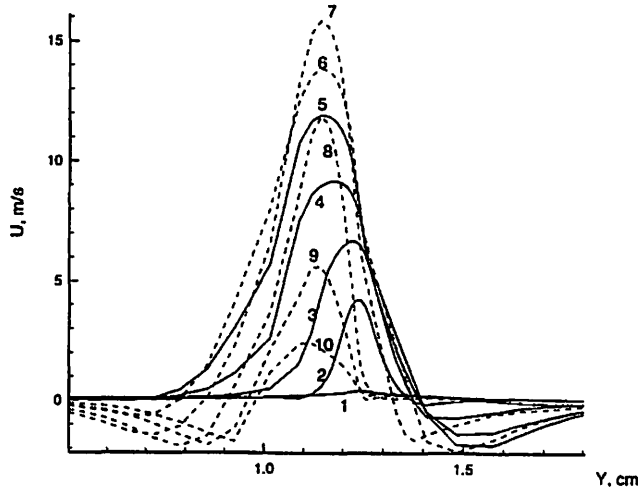

Figure 15. Velocity profiles at the glottal exit (x=7.5 cm).

Figure 13 gives the velocity profiles at the inlet of the glottis (location 1, x=5 cm) at ten equal time steps throughout the cycle. The symmetric inlet flow function controlled the mean flow but not the specific velocity profiles. Profile number 5 reflects the maximum mean inlet flow, and also has the greatest peak velocity, located in the center of the duct. Notice that times 4 and 6, 3 and 7, 2 and 8, and 1 and 9 correspond to time pairs of equal inlet mean flow, but differ in velocity profiles. This is due to the influence of the moving glottal walls on the inlet flow velocities. The walls are closest at time zero and 10, and are maximally separated at time 5. The motion of the walls is outward from time zero to 5 and inward from time 5 to 10, resulting in flatter profiles during glottal opening and profiles with greater maximum centerline flows during glottal closing. Note that at time increment 10 at the end of the cycle, the mean flow is zero, but there is flow reversal (negative flow) at the sides of the profile, with positive flow near the center. Some flow reversal is also seen at the profile tails at time frame 9.

Figure 14 shows corresponding velocity profiles within the glottis at x= 6.7 cm, halfway between locations 2 and 3 (see Fig. 1). The profiles grow in value as the inlet flow increases, so that step increment 5 yields the largest maximum particle velocity. What is of importance is that these intraglottal velocity profiles are asymmetric in time and space (contrary to the relatively symmetric profiles at location 1). First note that the center line of the glottis is at Y = 1.25 cm, the approximate center of the velocity profile number 1. The glottal walls start close together (with a minimum glottal diameter of 0.02 cm at glottal exit) and initially move outward. The initial convergence shape produces the expected relatively flat velocity profile seen in step increments 1 and 2. If the velocity profiles were symmetric, they would move left and right an equal amount with each time increment. The profile asymmetry of Fig. 14 indicates, however, that even by increment 2, the velocity remains zero at 1.34 cm in the glottal duct, suggesting flow separation from the upper wall of the airway shown in Fig. 1. Not only is there flow separation from the upper wall, but the velocity between about 1.34 cm and the actual glottal wall remains virtually zero throughout the cycle. Figure 14 also suggests that the midglottal velocity profiles are significantly different in shape between the opening (solid lines) and closing phases (dashed lines). The primary difference is that the profiles during closing are substantially more rounded with greater peaks than during glottal opening. Inward moving walls produce the effect of moving more air toward the center, and outward moving walls move air more toward the sides. There is flow reversal in the glottis at step 10 near the wall surfaces.

Figure 15 shows the velocity profiles at the section approximately 0.5 cm past the glottis exit at location 4 on Fig. 1 (7.5 cm). These profiles differ significantly from those

found in the midglottis of Fig. 14. The flow asymmetry follows the asymmetry direction of the midglottal section, that is, with greater flow toward the lower portion of the duct (Fig. 1). The positive flow is narrower, however, than in the midglottis, with flow reversal beginning by time increment 2 and continuing throughout the cycle. Of special significance is that the peak velocity occurs at time increment 7 rather than 5 as was found for the midglottal section. That is, the fastest particle velocities exiting the glottis occur after the inlet flow begins to reduce in mean flow. The cause of this delayed maximum particle velocity may be the movement of the glottal walls inward after the midway time point. The decrease in maximum velocity is relatively fast after time increment 7. This delay in reaching the maximum velocity, due to glottal motion, may relate directly to the skewness of the normal volume velocity of glottal exit flow. Figure 15 would indicate an "ac" portion of the exit velocities that help contribute to the glottal volume velocity skewness. It is noted that the reversed flows found lateral to the glottal exit would impinge upon the upper vocal fold surfaces, potentially creating positive forces on those surfaces.

Conclusions

A finite volume computational method was applied to the simulation of glottal flow. Both the motion of the vocal folds and the inlet flow function were prescribed. The results emphasize ac flow and pressure behavior within the glottis during phonation. Particle velocities were found to be more placed in the center during inward movement of the glottal wall. Intraglottal airflow, as well as flow downstream of the glottis, were asymmetric, with maximum velocity delays due to the motion of the glottal walls. These delays may contribute to the volume velocity and particle velocity skewing (Alipour & Scherer, 1995) found in normal phonation.

Acknowledgments

This work was supported by research grant number 5 R01 DC00831-04 from the National Institute on Deafness and Other Communication Disorders, National Institute of Health.

References

Alipour, F. & Titze, I.R. (1988). A Finite element simulation of vocal fold vibration. In *Proceedings of Fourteenth Annual Northeast Bioengineering Conference* Durham, N.H., IEEE publication #88-CH2666-6, 186-189.

Alipour, F. & Patel, V.C. (1991). Numerical simulation of laryngeal flow. In *Proceedings of ASME 1991 Winter Annual Meeting*, Edited by R. Vanderby, **BED-20**, 111-114.

Alipour, F. & Patel, V.C. (1994). Steady flow through modeled glottal constriction. *Journal of Engineering, Islamic Republic of Iran*, 7(1), 13-18.

Alipour, F. & Scherer, R.C. (1995). Pulsatile airflow during phonation: an excised larynx model. *Journal of the Acoustical Society of America.* 97(2), 1241-1248.

Armaly, B.F., Durst, F., Preira, J.C.F. & Schonung, B. (1983). Experimental and theoretical investigation of backward-facing step flow. *Journal of Fluid Mechanics*, **127**, 473-496.

Berke, G.S., Moore, D.M., Monkewitz, P.A., Hanson, D.G. & Gerratt, B.R. (1989). A preliminary study of particle velocity during phonation in an in vivo canine model. *Journal of Voice*, 3(4), 306-313.

Durst, F., Melling, A., and Whitelaw, J.H. (1974). Low Reynolds number flow over a plane symmetric sudden expansion. *Journal of Fluid Mechanics* 64(part 1):111-128.

Durst, F., Pereira, J.C.F., and Tropea, C. (1993). The plane symmetric sudden-expansion flow at low Reynolds numbers. *Journal of Fluid Mechanics* 248:567-581.

Guo, C.G. & Scherer, R.C. (1993). Finite element simulation of glottal flow and pressure *Journal of the Acoustical Society of America.* 94(2), Pt. 1, 688-700.

Iijima, H., Miki, N. & Nagai, N. (1992). Glottal impedance based on a finite element analysis of two-dimensional unsteady viscous flow in a static glottis. *IEEE Transaction on Signal Processing*, 40(9), 2125-2135.

Ishizaka, K. & Matsudaira, M. (1972). Fluid mechanical considerations of vocal cord vibration. *SCRL-Monograph* **8**, Speech Communication Research laboratory, Santa Barbara.

Kim, W.J. and Patel, V.C. (1992). Numerical Solution for Two-Dimensional Incompressible Flows. *Iowa Institute of Hydraulic Research*, University of Iowa, IIHR Report No 361.

Liljencrants, J. (1991). Numerical simulation of glottal flow. In J. Gauffin & B. Hammarberg (Eds.) *Vocal Fold Physiology: Acoustics, Perception, and Physiological Aspects of Voice Mechanisms.* Singular Publishing Group Inc., San Diego, 99-104.

Menon, A.S., Weber, M.E., and Chang, H.K. (1984). Model study of flow dynamics in human central airways. Part III: Oscillatory velocity profiles. *Respiration Physiology* 55(2):255-275.

Patankar, S.V. (1980). *Numerical Heat Transfer and Fluid Flow*, Hemisphere Publishing Corporation, McGraw-Hill, New York.

Scherer, R.C. (1981). *Laryngeal Fluid Mechanics: Steady Flow Considerations Using Static Models.* Ph.D. thesis, University of Iowa, Iowa City.

Scherer, R.C. (1983). Pressure-flow relationships in a laryngeal airway model having a diverging glottal duct. *Journal of the Acoustical Society of America*, 73(S1), S46(A).

Scherer, R.C. & Titze, I.R. (1983). Pressure-flow relationships in a model of the laryngeal airway with a diverging glottis. In D.M. Bless and J.H. Abbs (Eds.) *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, College-Hill Press, San Diego, 179-193.

Scherer, R.C., Titze, I.R. & Curtis, J.F. (1983). Pressure-flow relationships in two models of the larynx having rectangular glottal shapes. *Journal of the Acoustical Society of America*, 73(2), 668-676.

Scherer, R.C. & Guo, C.G. (1990). Laryngeal modeling: translaryngeal pressure for a model with many glottal shapes. In *Proceedings of 1990 International Conference on Spoken Language Processing*, 1, The Acoustical Society of Japan, Japan, 3.1.1 - 3.1.4.

Scherer, R.C. & Guo, C.G. (1991). Generalized translaryngeal pressure coefficient for a wide range of laryngeal configurations. In J. Gauffin and B. Hammarberg (eds.), *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, Singular Publishing Group, San Diego, 83-90.

Sobey, I.J. and Drazin, P.G. (1986). Bifurcation of two-dimensional channel flows. *Journal of Fluid Mechanics* 171:263-287.

Titze, I.R. (1988). Regulation of vocal power and efficiency by subglottal pressure and glottal width. In O. Fujimura (ed) *Vocal Fold Physiology: Voice Production, Mechanisms and Function* Raven Press, Ltd. New York, 227-238.

White, F.M (1991). *Viscous Fluid Flow*, 2nd edition, McGraw-Hill, New York.

NCVS Status and Progress Report - 10
November 1996, 45-50

Further Studies of Phonation Threshold Pressure in a Physical Model of the Vocal Fold Mucosa

Roger W. Chan, B.S.
Department of Speech Pathology and Audiology, The University of Iowa
Ingo R. Titze, Ph.D.
Department of Speech Pathology and Audiology, The University of Iowa
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts
Michael R. Titze
Department of Speech Pathology and Audiology, The University of Iowa

Abstract

This paper reports results of further experimentation on a previously developed physical model of the vocal fold mucosa [I.R. Titze, S.S. Schmidt, and M.R. Titze, J. Acoust. Soc. Am. 97, 3080-3084 (1995)]. The effects of vocal fold thickness, epithelial membrane thickness and prephonatory glottal geometry on phonation threshold pressure were studied. Phonation threshold pressures in the range of 0.13 to 0.34 kPa were observed for an 11-mm thick vocal fold with a 70 μm thick epithelial membrane for different mucosal fluid viscosities. Higher threshold pressure was always obtained for thinner vocal folds and thicker membranes. In another set of experiments, lowest offset threshold pressure was obtained for a rectangular or a near-rectangular prephonatory glottis (with a glottal convergence angle within about ± 5°). It ranged from 0.07 to 0.23 kPa for different glottal half-widths between 2.0 and 6.0 mm. The threshold for more convergent or divergent glottal geometries was consistently higher. This finding only partially agrees with previous analytical work which predicts a lowest threshold for a divergent glottis. The discrepancy between theory and data is likely to be associated with flow separation from a divergent glottis.

Introduction

Many empirical findings on vocal fold oscillation have been obtained through the use of physical models and excised larynges (human or animal). Because it is often difficult to isolate and control the parameters of theoretical interest in a human subject, this *in vitro* approach continues to be viable. A physical model of the larynx, for example, constructed with well-defined geometry and biomechanical features, allows for better control of variables and eliminates some variability inherent in human control. Such a model was built earlier for the purpose of measuring phonation threshold pressure (Titze *et al.*, 1995). Phonation threshold pressure (P_{th}) has been defined as the minimum lung pressure required to produce vocal fold oscillation (Titze, 1988, 1992). It was found to be consistently higher for oscillation onset than for oscillation offset both experimentally (Baer, 1975; Titze *et al.*, 1995) and analytically (Lucero, 1995). Thus *onset* P_{th} can be defined as the minimum lung pressure that initiates vocal fold oscillation from rest, while *offset* P_{th} ("minimum sustaining pressure" in Lucero, 1995) can be defined as the minimum lung pressure that sustains vocal fold oscillation after it has begun.

This study represents further experimentation on the previously developed physical model (Titze*et al.*, 1995), in order to explore some untested aspects of a small-amplitude oscillation theory (Titze, 1988). Specifically, the effect of vocal fold thickness, epithelial membrane thickness and prephonatory glottal convergence angle on P_{th} were explored. The apparent trade-off relation between epithelial membrane thickness and mucosal fluid viscosity was also studied.

Previous Analytical and Experimental Results

Assuming small-amplitude oscillation conditions and a surface wave propagating on the cover of a body-cover model of the vocal folds, Titze (1988) derived some analytical expressions for phonation threshold pressure. For a

rectangular prephonatory glottal geometry, the expression was

$$P_{th} = (k_t / T) B c \xi_0 \quad (1)$$

where k_t is the transglottal pressure coefficient (about 1.1 as determined by Scherer, 1981), T is the vertical vocal fold thickness, B is the viscous damping coefficient of vocal fold tissues, c is the mucosal wave propagation velocity, and ξ_0 is the prephonatory glottal half-width. Equation (1) predicts that P_{th} varies linearly with viscous damping in vocal fold tissues, mucosal wave velocity, and prephonatory glottal half-width. It also predicts that P_{th} varies inversely with vocal fold thickness.

Previous experimental results on the physical model showed that P_{th} indeed increases with mucosal fluid viscosity and prephonatory glottal halfwidth in a roughly linear fashion (Titze *et al.*, 1995), supporting equation (1), but linearity was not preserved to zero glottal width. Rather, P_{th} was actually lowest for a small positive value of glottal halfwidth (between 0.0 and 1.0mm), and rose when ξ_0 approached zero. This nonlinearity might be the consequence of viscous pressure losses which become dominant as ξ_0 gets small (Lucero, 1996). Another unpredicted result was the aforementioned hysteresis effect, with offset P_{th} being consistently lower than onset P_{th}. In response to this finding, Lucero (1995) extended the analytical approach to include large-amplitude oscillations. He came up with the same expression for onset P_{th} [equation (36) in Lucero's analysis reduces to equation (1) here when static vocal fold displacement $\bar{\xi} = 0$ for a rectangular prephonatory glottis]. Lucero then showed that offset P_{th} was equal to half the onset P_{th} for a rectangular glottis, as a consequence of nonlinear changes in effective aerodynamic damping on the vocal folds.

Returning to our former experimental results, there was an apparent trade-off between epithelial membrane thickness and mucosal fluid viscosity (Titze *et al.*, 1995). For the physical model with a 200 μm thick membrane and at a physiological range of fluid viscosity (on the order of 500cP according to Zhu and Mow, 1990), P_{th} was excessively high in relation to P_{th} values for human phonation reported in the literature (on the order of 0.3 kPa at 100 Hz; e.g., Verdolini-Marston *et al.*, 1990, 1994). Since the experimental membrane thickness was four times that of the epithelial layer in human vocal folds (50 μm according to Hirano, 1975), it was concluded that realistic P_{th} values could be obtained only with thinner membranes. On the other hand, lower fluid viscosity could be used to compensate for thicker membranes.

Now let us turn our attention to the more general case of non-rectangular (convergent and divergent) prephonatory glottal geometries. The analytical expression derived by Titze (1988) was

$$P_{th} = (2 k_t / T)(Bc)\, \xi_{01}^2 / (\xi_{01} + \xi_{02}) \quad (2)$$

where ξ_{01} and ξ_{02} are inferior and superior prephonatory glottal half-widths, respectively. Note that equation (2) reduces to equation (1) for a rectangular glottis, where $\xi_{01} = \xi_{02}$. For a divergent glottis ($\xi_{01} < \xi_{02}$), P_{th} is lower than that for a rectangular glottis, while P_{th} is higher for a convergent glottal geometry ($\xi_{01} > \xi_{02}$) because of the square power of ξ_{01}.

Lucero's (1995) large-amplitude analysis of this more general case yielded the following expression:

$$P_{th} = (k_t / 2T)(Bc)(\xi_{01} + \bar{\xi})(1 + \sqrt{1 - a^2}) \quad (3)$$

where $\bar{\xi}$ is the static vocal fold displacement and a is a normalized oscillation amplitude. Whether threshold for onset or offset is represented by this equation depends on the normalized amplitude of oscillation (e.g., $a = 0$ for oscillation onset).

The static displacement $\bar{\xi}$ is also a function of prephonatory glottal geometry and is given by

$$\bar{\xi} = (Bc / 2KT)(\xi_{01} - \xi_{02})(1 + \sqrt{1 - a^2}) / \sqrt{1 - a^2} \quad (4)$$

where K is the lumped effective stiffness per unit area of the vocal fold mucosa. The second term in parenthesis reveals that $\bar{\xi}$ is positive for a convergent glottis and negative for a divergent glottis. With this static displacement, equations (3) and (4) predict that P_{th} is lowest for the divergent glottal geometry and highest for the convergent glottis, a result similar to that of equation (2). Lucero's (1993) dynamic analysis of the two-mass model of vocal folds (Ishizaka and Flanagan, 1972) also predicted the same result.

Method

The physical model of the vocal fold mucosa used in this study was the same one described by Titze *et al.* (1995), where construction details can be found. Briefly, it consisted of a stainless-steel trapezoidal vocal fold "body" and a "mucosa" made of a silicone membrane (the "epithelium") encapsulating a fluid of varying viscosities (the "superficial layer of lamina propria"). Vocal fold thickness was adjustable by using bodies of different sizes. Epithelial membrane thickness was varied by controlled dipping of a silicone dispersion fluid and measured by a digital caliper. Thickness of membranes used in the previous study were on the order of 200 μm (as mentioned), but it was possible to manufacture intact membranes as thin as 70 μm in this study. This was much closer to the 50 μm epithelial thickness in human vocal folds (Hirano, 1975).

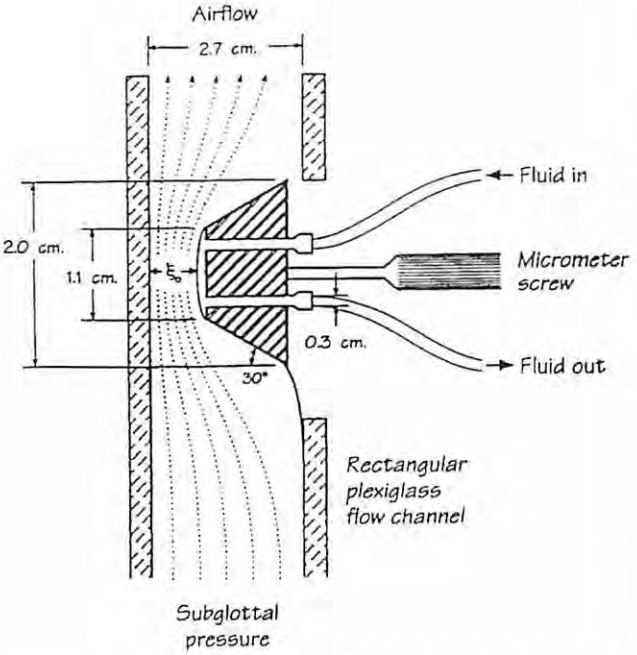

Figure 1. Sketch of the physical model of the vocal fold and the glottal airway with a rectangular prephonatory glottal geometry. From Titze et al., *1995.*

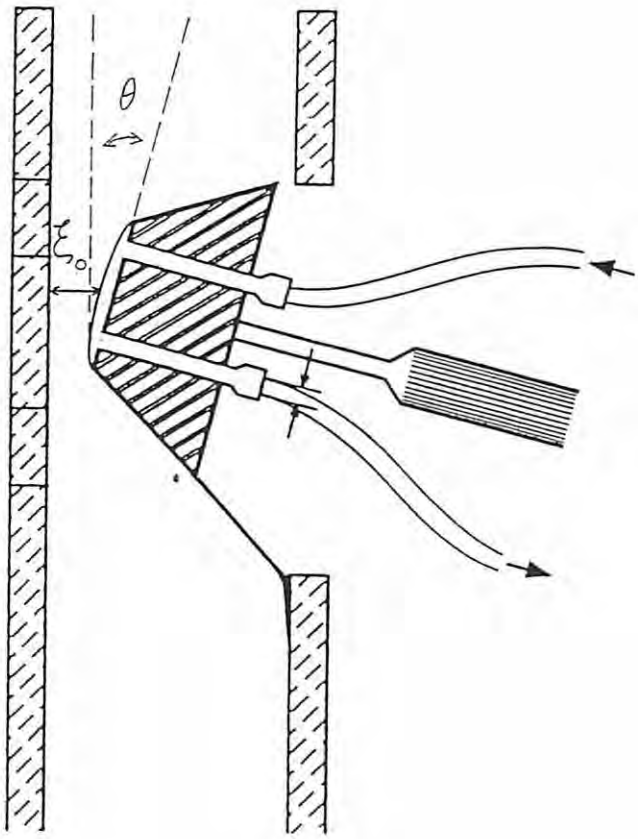

Figure 2. Schematic of the physical model with a divergent prephonatory glottal geometry. Note that the glottal convergence angle (θ) is negative in a divergent glottis by definition.

The vocal fold position within a rectangular Plexiglas airflow channel was controlled by a micrometer (Figure 1), "adducting" the fold against one wall of the channel, as in a hemilarynx set-up. The glottal geometry was further adjusted by tilting the micrometer while maintaining the mid-point prephonatory glottal half-width (ξ_0) constant (Figure 2). This geometry was expressed as a glottal convergence angle (θ). The entire model was then mounted onto a pipe that supplied compressed air to the flow channel. The pressure was regulated by a Fairchild Model 10 regulator (0 to 2 psi) and measured with an open-ended water manometer (Dwyer, 60 cm) with a resolution of 0.2 cm H_20 (0.02 kPa). A video camera was mounted directly above the model to magnify and record the mucosal oscillation.

Two sets of experiments were done. In the first set, a rectangular glottis was used with four variations of epithelial membrane thickness andmucosal fluid viscosity (see Table 1). For each variation, onset and offset P_{th} were measured for two different vocal fold thicknesses (7.5 mm and 11.0 mm). A glottal half-width of 2.0 mm was maintained throughout the experiments. This 2 × 2 × 2 design allowed the effect of vocal fold thickness on P_{th} to be studied, as well as the trade-off relation between membrane thickness and fluid viscosity.

Table 1.
Four variations of membrane thickness and fluid viscosity used in the first set of experiments. Pure water with a viscosity of 1.0 cP was used for sets A and B, and a 0.5% mixture (by weight) of sodium carboxymethyl cellulose (CMC) powder with H_2O was used for sets C and D.

	Set A	Set B	Set C	Set D
Membrane thickness (μm)	70	210	70	210
Fluid viscosity (cP)	1.0	1.0	471	471

In the second set of experiments, prephonatory glottal half-widths of 2.0, 2.5, 3.0, 3.5, 4.0, 5.0 and 6.0 mm were adopted and prephonatory glottal convergence angle was varied from about -10° to +10° for each of these widths. Other variables were maintained at constant values: vocal fold thickness at 15 mm, membrane thickness at 130 μm and tissue viscosity at 1.0 cP (pure water). Note that the membrane thickness was not the thinnest obtainable at the time of experimentation, but was less prone to rupture when angle was varied and repeated trials were taken. Onset and offset P_{th} were measured for each of the above conditions whenever stable oscillation was achieved (it was not always possible to achieve stable oscillation with smaller glottal half-widths and bigger angles because the mucosa collided partially with the opposing Plexiglas wall).

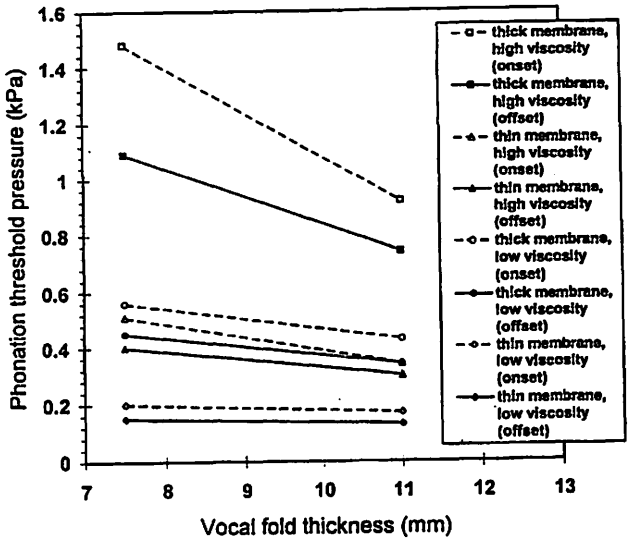

Figure 3. Phonation threshold pressure as a function of vocal fold thickness, epithelial membrane thickness and mucosal fluid viscosity. Each data point represents an average of measurements from three to four successive experimental trials. Prephonatory glottis was rectangular with a half-width of 2.0 mm. Epithelial membrane thickness was 70 μm (thin) or 210 μm (thick). Mucosal fluid was pure water with a viscosity of 1.0 cP (low) or sodium CMC solution with a viscosity of 471 cP (high).

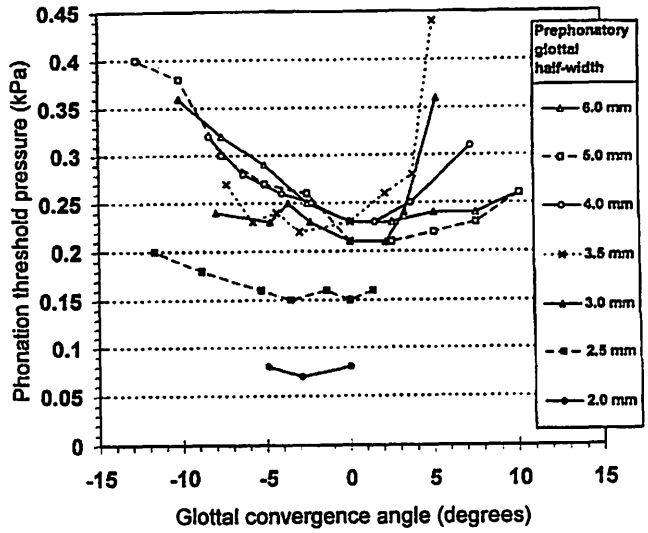

Figure 4. Offset phonation threshold pressure as a function of prephonatory glottal half-width and glottal convergence angle.

Results and Discussion

Figure 3 shows results of the first set of experiments, where onset and offset P_{th} are plotted against vocal fold thickness with epithelial membrane thickness and mucosal fluid viscosity as parameters. Each data point represents an averaged measurement of three to four successive trials. As reported in the previous study, experimental error or repeatability of the data was within about ± 0.02 kPa (Titze *et al.*, 1995).

There was again a hysteresis effect as observed before, with offset P_{th} (solid lines) consistently lower than corresponding onset P_{th} (dotted lines). The difference was smallest (about 0.04 kPa) for a thin membrane and a low viscosity and largest (about 0.2 to 0.4 kPa) for a thick membrane and a high viscosity. Interestingly, the ratios between offset and onset P_{th} are all very similar (about 0.75 to 0.80).

In every case, the thicker vocal fold had a lower P_{th} than the thinner one, across all four variations of membrane thickness and fluid viscosity. This finding agrees with the general prediction from equation (1). Like the hysteresis effect, the difference was smallest (about 0.03 kPa) for a thin membrane and a low viscosity and largest (about 0.4 to 0.6 kPa) for a thick membrane and a high viscosity. Unlike the hysteresis effect, however the ratios between P_{th} for the two vocal fold thicknesses are less constant, ranging from 0.65 to 0.85. This may suggest that the relation between P_{th} and vocal fold thickness is more than simply inverse, and that previous analytical work might be inadequate in describing its interaction with other parameters.

Figure 3 also clearly shows the trade-off relation between membrane thickness and fluid viscosity. For a non-physiological 210 μm thick membrane and a physiological fluid viscosity of 471 cP, P_{th} was at around 1.0 kPa, considerably above measured P_{th} values for human phonation (about 0.3 kPa at 100 Hz). However, when a near-physiological 70 μm thick membrane was used, realistic P_{th} at about 0.3 to 0.5 kPa were obtained with the physiological fluid viscosity of 471 cP. Comparably realistic P_{th} values were also obtained for a directly opposite and yet totally non-physiological condition, with a 210 μm thick membrane and pure water. A practical implication of this finding is that hydration treatments on benign vocal fold lesions (Verdolini-Marston *et al.*, 1994) might facilitate the ease of phonation by compensating for the unfavorable increase of P_{th} when a lesion causes an increase in epithelial thickness (and mass).

Results of the second set of experiments are shown in Figure 4, where offset P_{th} is plotted against prephonatory glottal convergence angle, with glottal half-width as the parameter. Only data on offset P_{th} are presented because data on onset P_{th} show a similar pattern. Again, each data point represents an average of measurements from three to four successive trials.

It can be seen that P_{th} generally increases with prephonatory glottal half-width, (vertical separation of the curves) a finding already discussed in the previous study (Titze *et al.*, 1995). P_{th} also changes considerably with prephonatory glottal convergence, the major variable of interest in this set of experiments, except for the case of smallest glottal half-width (2.0 mm) where vocal fold colli-

sion occurred at all but the smallest convergence and divergence angles. Note that P_{th} ranges only between 0.07 and 0.08 kPa for this condition and should be regarded as constant, given an estimated experimental error of ± 0.02 kPa.

With experimental error being considered, Figure 4 shows that the lowest P_{th} was obtained for zero glottal convergence angle (within about ± 2°), i.e., a rectangular prephonatory glottis. P_{th} for convergent or divergent glottal geometries were consistently higher, except for cases where stable oscillation was not achieved.

This finding only partially agrees with equations (2), (3) and (4), which predict that P_{th} is lowest for the divergent glottis and highest for the convergent one. The discrepancy between theory and data might be interpreted in terms of flow separation effects in a divergent glottis. Pelorson *et al.* (1994, 1995) described glottal fluid mechanics with a theoretical quasi-steady flow model and predicted flow separation for a divergent glottis, in particular a moving flow-separation point during closing phase of the vibratory cycle. Guo and Scherer (1993) used the finite element method to simulate two-dimensional steady air flow and air pressure through the glottis. Flow separation was found to occur immediately downstream of the point of minimal constriction in a divergent glottis. They simulated one-dimensional pressure profiles along the glottal airway for different prephonatory glottal geometries (Figure 8 in Guo and Scherer, 1993). Results showed that more divergent glottal angles are associated with less negative air pressure along the glottis, presumably because the point of flow separation moves upstream with an increase in glottal divergence, a result implicated in Pelorson *et al.*'s (1994) model of moving flow-separation point. A similar pattern of pressure change was also found in their later experimental study (Pelorson *et al.*, 1995). Hence, a smaller asymmetric driving force is available for the more divergent glottis in each vibratory cycle, given the same subglottal pressure. An increase of P_{th} with glottal divergence is therefore likely.

Indeed, it could be possible that the prephonatory glottis diverges to a point that oscillation cannot be established at all, because of the failure to build up sufficient alternate positive and negative pressures or asymmetric driving force necessary for self-sustained oscillation (cf. Scherer, 1981; Titze, 1988). A "critical divergent angle" might therefore exist and should be explored in future studies.

Summary and Conclusion

Experiments were conducted on a previously described physical model of the vocal fold mucosa. The effects of vocal fold thickness, epithelial membrane thickness and prephonatory glottal geometry on phonation threshold pressure were quantified. As predicted by previous analytical result, higher P_{th} was always obtained for a thinner vocal fold, although the relation between vocal fold thickness and P_{th} might not be exactly inverse.

An increase in epithelial membrane thickness was also found to raise P_{th}, which could be compensated by a decrease in mucosal fluid viscosity. As far as the prephonatory glottal geometry is concerned, lowest P_{th} was always observed for a rectangular prephonatory glottis, while it was predicted from previous analytical work that a divergent glottis should have the lowest threshold pressure. The discrepancy between theory and data could be interpreted in terms of flow separation effects in a divergent glottis. More analytical work has to be done to further explore this topic in terms of its relationship to phonation threshold pressure.

Clinically, these findings imply that prephonatory glottal geometry should be considered and carefully monitored in vocal-fold medialization types of phonosurgery (e.g., vocal-fold augmentation with injectable biomaterials, or thyroplasty), in order to obtain a nearly rectangular glottal geometry. This geometry likely facilitates the ease of phonation. Highly convergent or divergent glottal geometries are associated with higher phonation threshold pressures and should be avoided.

Acknowledgment

This study was supported by Grant No. P60 DC00976 from the National Institutes on Deafness and Other Communication Disorders, for which the authors are grateful.

References

Baer, T. (1975). "Investigation of phonation using excised larynges," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Guo, C.G., and Scherer, R.C. (1993). "Finite element simulation of glottal flow and pressure," JASA, 94, 688-700.

Hirano, M. (1975). "Phonosurgery: Basic and clinical investigations," Otologia (Fukuoka), 21, 239-240.

Ishizaka, K., and Flanagan, J.L. (1972). "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell Syst. Tech. J., 51, 1233-1268.

Lucero, J.C. (1993). "Dynamics of the two-mass model of the vocal folds: Equilibria, bifurcations, and oscillation region," JASA, 94, 3104-3111.

Lucero, J.C. (1995). "The minimum lung pressure to sustain vocal fold oscillation," JASA, 98, 779-784.

Lucero, J.C. (1996). "Relation between the phonation threshold pressure and the prephonatory glottal width in a rectangular glottis," JASA (in press).

Pelorson, X., Hirschberg, A., van Hassel, R.R., Wijnands, A.P.J., and Auregan, Y. (1994). "Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model," JASA, 96, 3416-3431.

Pelorson, X., Hirschberg, A., Wijnands, A.P.J., and Bailliet, H. (1995). "Description of the flow through in-vitro models of the glottis during phonation," Acta Acustica, 3, 191-202.

Scherer, R.C. (1981). "Laryngeal fluid mechanics: Steady flow considerations using static models," Ph.D. dissertation, The University of Iowa, Iowa City, IA.

Titze, I.R. (1988). "The physics of small-amplitude oscillation of the vocal folds," JASA, 83, 1536-1552.

Titze, I.R. (1992). "Phonation threshold pressure: A missing link in glottal aerodynamics," JASA, 91, 2926-2935.

Titze, I.R., Schmidt, S.S., and Titze, M.R. (1995). "Phonation threshold pressure in a physical model of the vocal fold mucosa," JASA, 97, 3080-3084.

Verdolini-Marston, K., Titze, I.R., and Druker, D.G. (1990). "Changes in phonation threshold pressure with induced conditions of hydration," J. Voice, 4, 142-151.

Verdolini-Marston, K., Sandage, M., and Titze, I.R. (1994). "Effect of hydration treatments on laryngeal nodules and polyps and related voice measures," J. Voice, 8, 30-47.

Zhu, W.B., and Mow, V.C. (1990). "Viscometric properties of proteoglycan solutions at physiological concentrations," *Biomechanics of Diarthrodial Joints*, edited by V.C. Mow, A. Ratcliffe, and S. Woo (Springer-Verlag, New York), pp. 313-344.

NCVS Status and Progress Report - 10
November 1996, 51-57

The Dynamics of Length Change in Canine Vocal Folds

Ingo R. Titze, Ph.D.
Department of Speech Pathology and Audiology, The University of Iowa
Jack J. Jiang, M.D., Ph.D.
Department of Speech Pathology and Audiology, The University of Iowa
Department of Otolaryngology - Head and Neck Surgery, Northwestern University School of Medicine
Emily Lin, Ph.D.
Department of Otolaryngology - Head and Neck Surgery, Northwestern University School of Medicine

Abstract

The time courses of vocal fold elongation and contraction have been measured as a function of intrinsic laryngeal muscle activity. The superior and recurrent laryngeal nerves of anesthetized canines were stimulated supramaximally (on-off in all combinations) while the vocal folds were surgically exposed and illuminated for conventional and higher-speed (300 frames per second) video recording. Microsutures were placed on various points on the vocal folds to measure elongation and contraction. Vocal fold strain, defined as elongation divided by rest length, ranged from -17% to +45%. The typical time constant for exponential increase or decrease in strain was about 30 ms. This reflects primarily the intrinsic muscle activation times rather than a passive (inertial or viscoelastic) response of cricothyroid joint rotation or translation.

Introduction

It is well known that the length of the vocal folds changes with fundamental frequency (F_0). Although in theory it is possible to change F_0 isometrically (at constant length), laryngoscopic observations have shown that this is generally not the control strategy used by humans.[1,2] Length usually increases directly with F_0, but more in some subjects than others. There are two basic questions to which only partial answers have been given to date: (1) what are the maximum possible length changes that can occur and (2) how rapidly can these changes be executed? With regard to the first question of range of elongation, it is known that the vocal folds shorten upon adduction, become even shorter at low pitches, and attain their greatest length at very high pitches (particularly in the falsetto register). But the practical limits in terms of the movements of the cricoid, thyroid, and arytenoid cartilages are not well known.

Maximum speed of pitch changes in human subjects has also been examined.[3] Eight subjects, all adult males between 20 and 31 years of age, performed ascending and descending pitch jumps as rapidly as they could. The subjects were not vocally trained, but were briefly trained in the exercise. The pitch jumps were on the order of an octave. Response time was about 100 ms for ascent and about 80 ms for descent. Sundberg[4] repeated the study and found similar results for untrained subjects. For trained singers, however, rise time was 70-80 ms for males and 57-65 ms for females; fall time was between 60-80 ms, except for trained females, who lowered their pitch in about 60 ms. Sundberg concluded that, in general, females can change pitch faster than males, trained vocalists faster then untrained vocalists, and speed of pitch change is reduced slightly (about 10-20 ms) if the pitch interval increases from 4 semitones to 12 semitones. It is likely that the difference in direction of F_0 change can be accounted for by passive relaxation of vocal fold tissues under stress,[5] but a solid proof has not been offered.

In this paper, we address only the *maximum* range of length change and the *maximum* speed of length change under supramaximal stimulation of laryngeal muscles. This is the type of experiment that is too risky to perform on human subjects; hence, canines were used as a model. The use of canines always raises questions of validity, but the assumption is that anatomical and physiological differences accounted for in other experiments may bridge the gap between this experiment and observations on humans. For

example, since contraction times of isolated laryngeal muscles are known to differ slightly between humans and canines[6] perhaps the difference in the speed of length change can be inferred from these contraction times. Also, any difference in the range of elongation between the species may be accounted for (in later modeling) by the inclusion of a vocal ligament, which is known to be present in humans, but absent in canines.

Methods

The experiments were conducted in two parts, several years apart. In Part I, seven dogs were used to obtain a small statistical sample of maximum length change. Unfortunately, at the time the experiment was conducted we had no access to high speed video imaging. Thus, the time course of elongation could only be obtained with conventional videography (about 30 frames per second). When high speed video imaging was obtained several years later, one additional dog was used to get better detail of the time course. This will be called Part II of the experiment. Given that the results were qualitatively the same, we felt no need to sacrifice additional animals.

Subjects

Mongrel dogs with total body mass between 20-30 kg were obtained from the University of Iowa Animal Care Facility. The first three columns of Table 1 indicate the identification number, the sex, and the approximate mass for each dog used in Part I of the experiment. In Part II, one additional female dog (22 kg) was used.

The dogs were anesthetized with Ketamine (8.25 mg/kg) and Rompun (2.75 mg/kg) and placed in the supine position. Additional Nembutal was added later at a rate of approximately 1 cc per hour to maintain a constant level of anesthesia. The room temperature was controlled at 25°C and a blanket was used to stabilize body temperature.

Table 1.
Maximum vocal fold strain for 7 canine larynges when superior and recurrent laryngeal nerves were stimulated supramaximally and bilaterally.

No. of larynges	sex	Body Weight (kg)	Stimulation Condition Recurrent	Superior	Both on
#1	?	20-30	—	+26.3%	+18.4%
#2	F	20-30	-12.3%	+37.4%	+29.9%
#3	M	20-30	-14.6%	+48.2%	+33.2%
#4	F	25	-18.9%	+29.1%	+13.2%
#5	M	28	-25.4%	+56.6%	+16.2%
#6	M	26	-15.1%	+44.2%	—
#7	F	25	—	+71.0%	+47.1%
Mean			-17.3%	+44.7%	+26.3%
S.D.			5.13%	15.71%	12.88%

With the dog's neck shaved from the mandible to the clavicle, a vertical midline incision was made from the level of the hyoid bone to the superior border of the manubrium of the sternum. The trachea was then severed, and an endotracheal cannula inserted into the caudal portion to maintain a free air passage. The cuff of the cannula was inflated with air to prevent blood leakage into the trachea. The strap muscles were severed. The external branch of the superior laryngeal nerve and recurrent nerve were dissected and exposed. Once identified, the bilateral superior and recurrent laryngeal nerves were disconnected from the central nervous system to avoid possible reflexes. A simplified supraglottic horizontal laryngectomy was then performed to expose the glottal area. The upper margin of the thyroid cartilage was sutured with the mucosa of the laryngeal ventricle to stop bleeding and to keep the area clean. An electrical cauterizer (set at high intensity) was also used during the surgery to minimize bleeding. Hooked-wire electrodes were placed on both the superior and recurrent laryngeal nerves bilaterally and fixed with stitches. Normal saline was applied to any exposed areas throughout the procedure to keep the exposed tissues moisturized.

Identification Marks

Black 8-0 ophthalmic nylon sutures with tape needles were used to stitch nodes at three positions of interest (Figure 1). Stitches were made on the upper surface of the membranous portion of the right vocal fold (points 1,

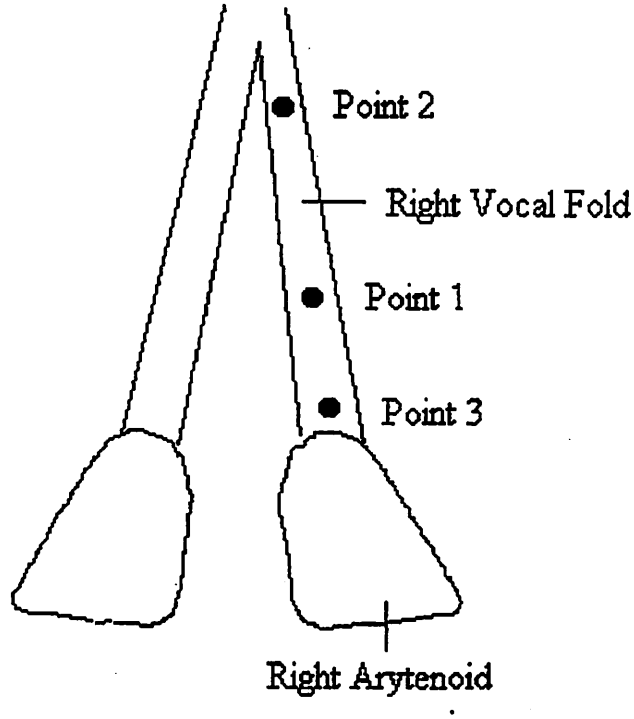

Figure 1. Schematic diagram of points where sutures were placed on the right vocal fold.

2 and 3). The needle penetrated through the epithelium into the superficial layer so that each stitch could be secured to show the movement of that particular surface point.

Instrumentation

A Grass S-88 stimulator was used for electrical stimulation. Each output channel of the stimulator was connected to a Grass PSIU6D isolation and current control unit. After distribution via a 10 kiloohm potentiometer, the output of each isolation unit was connected to the hooked-wire electrodes, which were attached to the nerve or the muscle of interest.

A Kodak Ektapro TR High Speed Video Imaging System (maximum 1000 frames per second full screen, black-and-white) was used to record the glottal image onto videotape. The camera was mounted approximately 20 cm from the larynx with the view angle normal to the glottis. Close-up views were obtained by using a Tamron 90 mm F2.5 telephoto lens.

Images were transferred from high speed tape onto three Maxell P/I VHS video tapes (T-30). A Panasonic Video Cassette Recorder/Player (AG-7500) was used with a 12 inch SONY Trinitron Color Monitor (CVM-1271) to play back the video tapes for frame-by-frame viewing. Fine point markers and transparencies were used for tracing the sutured nodes from the monitor screen. SigmaPlot installed on a PC-AT was used for data reduction and graphics.

Stimulation and Length Calibration

Electrical stimulation was applied to the superior and recurrent laryngeal nerves (separate stimulators). The stimulation was a 0.5 second pulse train at a frequency of 1 train per second, with the pulse frequency being 100 Hz and the pulse duration being 1 ms. With this frequency and a current amplitude of 0.5 mA, the stimulation was supramaximal in all cases and was applied in the following manner: (1) SLN alone, (2) RLN alone, and (3) both SLN and RLN simultaneously. For length calibration, glottal images were videotaped with a piece of metrically ruled paper resting on one vocal fold for the purpose of determining the ratio of actual size to screen size.

Data Analysis

The tapes were viewed frame by frame, based on the frame number superimposed on the glottal image. The duration of each frame was approximately 33 ms for the conventional video and 3.3 ms for the high speed video (analysis was done at a 300 Hz frame rate). From a segment without stimulation, 10 frames were analyzed for a measurement of an inter-suture distance d_o at rest (points 1-2 in Figure 1, which were always in view). This distance was 6.2 mm in Part II, with a standard deviation of 0.01 mm, based on the 10 observations. For each of the three stimulation conditions, 300 frames were analyzed to determine the time-course of elongation. The ratio of real size to screen size was calibrated to be 1:5. All the measured length values were adjusted using this correction factor.

Table 2.

The No. of the Measurement	Stimulation Conditions None	RLN + SLN
No. 1	16.7 cm	22.7 cm
No. 2	16.8 cm	22.8 cm
No. 3	16.6 cm	22.6 cm
No. 4	16.7 cm	22.6 cm
No. 5	16.6 cm	22.5 cm
Mean	16.68 cm	22.64 cm
S.D. (n-1)	0.084 cm	0.114 cm

To estimate the repeatability of stimulation conditions and length measurement in Part I, five trials were taken during a 10 second duration in one of the animals (one contraction per second with one second intervals among trials). The results obtained from the video screen (in cm) are shown in the Table 2. In this case, a longer inter-suture distance was used. The standard deviation (S.D.) of the measurement is about 0.1 cm on the screen. The actual distance measurement are accurate to 0.1 cm.

Results

When the superior laryngeal nerve was stimulated supramaximally and bilaterally, the cricothyroid muscles contracted and the vocal fold was thinned and elongated. When the recurrent laryngeal nerve was stimulated bilaterally and supramaximally, the vocal folds shortened and thickened. Stimulation of both nerves resulted in an intermediate posture. By measuring the distance $d(t)$ between the marks on the vocal fold frame by frame, the strain over time was obtained as

$$\epsilon(t) = \frac{d(t) - d_o}{d_o} \quad . \qquad (1)$$

We report the results in two parts, one for the maximum strain, and one for the typical time course of strain.

Maximal Length Changes

Vocal fold length reached a steady state after 100-200 ms of stimulation. Since the pulse train lasted 500 ms, the last 300 ms were used to determine this steady (maximum) strain. Results are shown in the right three columns of Table 1. Note that the average maximum strain over seven animals is +44.7% when the SLN is stimulated alone, +26.3% when both nerves are stimulated, and -17.3% when

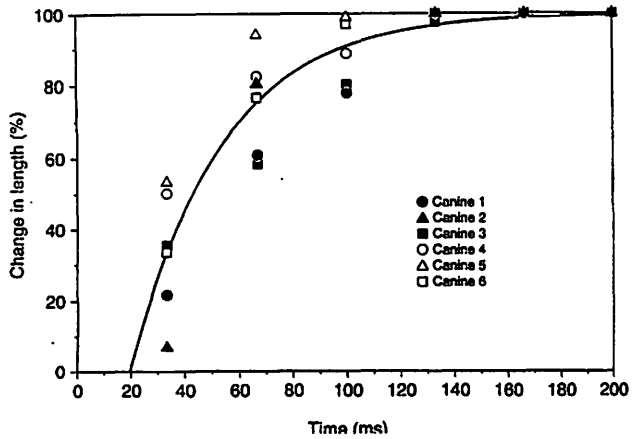

Figure 2. The time course of vocal fold elongation (relative to maximum elongation, 100%) from 6 larynges when the superior laryngeal nerve was stimulated alone. The solid curve is an exponential fit.

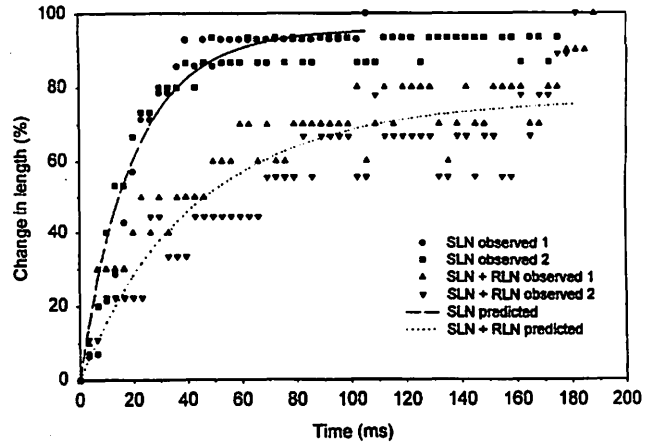

Figure 3. The time course of vocal fold elongation for one animal under two stimulation conditions. Upper curve is for SLN alone and lower curve is for SLN + RLN.

the RLN is stimulated alone. Larynx No. 7 had an extremely wide range of elongation. The maximum negative elongation could not be measured because the stitch mark was blocked; the larynx shifted during the stimulation. Similar methodological problems occurred for two other length measurements, as seen in the table.

The single larynx examined in Part II yielded similar results (not shown in Table), with a maximum strain of +46.3% when the SLN was stimulated alone, +30.4% when both nerves were stimulated, and -11.27% when the RLN was stimulated alone. Thus, the later single subject data fit well into the range of the earlier data, suggesting that our methodology had not changed remarkably.

It should be pointed out parenthetically that Table 1 is an extension of a smaller table reported previously.[7] From the smaller table we made some F_o predictions. Since the smaller data set is within one standard deviation of the present set, the F_o predictions can still be considered valid.

Change of Vocal Fold Length Over Time

The time-course of elongation was divided into an activation phase, a steady-state phase, and a relaxation phase. The steady-state phase was used to determine the maximum strains, as described above. Given that the maximum strains differed for different larynges, each time course was normalized to the maximum strain to obtain an average time course for all larynges. The result for stimulation of the SLN alone is shown in Figure 2 for Part I of the experiment. Recall that this is for the low video frame rate. An exponential fit to the average data for six larynges,

$$e(t) = e_m(1 - e^{-t/\tau}) \qquad (2)$$

yielded a time constant τ of 33.7 ms, with a standard deviation of 12 ms. (Note that no attempt was made to line up the t = 0 event because there was no synchronization of the video frame rate with the onset of stimulation.)

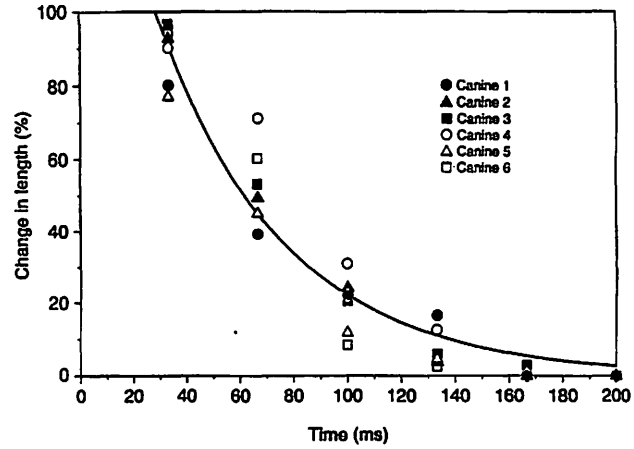

Figure 4. The average time course of relaxation from 6 larynges after superior laryngeal stimulation has been turned off.

Figure 3 shows the same time course of elongation (upper curve) for one additional larynx examined at the higher frame rate in Part II of the experiment. An exponential fit to this data set yielded a time constant of 23 ms. This is considerably less than the average of the six larynges of Figure 2, but within the standard deviation.

Figure 3 also shows the time course of elongation when both the SLN and RLN were stimulated. Note that the time constant here is much longer. The exponential model suggests about 89 ms. The internal mechanisms for this longer response time under combined activation are not yet clear to us.

The relaxation phase after SLN stimulation is shown in Figure 4. This represents the group data in Part I again, in which the strains are normalized to maximum and the video frame rate is 30 Hz. An exponential decay model suggests a time constant of 47 ms.

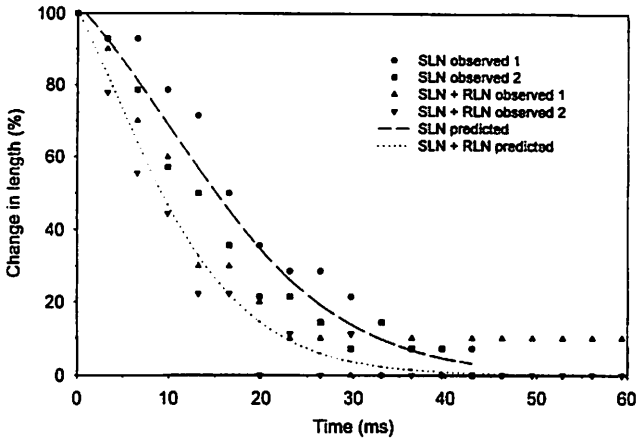

Figure 5. The time course of relaxation from one larynx after stimulation is turned off. The upper curve is for SLN alone and lower curve is for SLN + RLN.

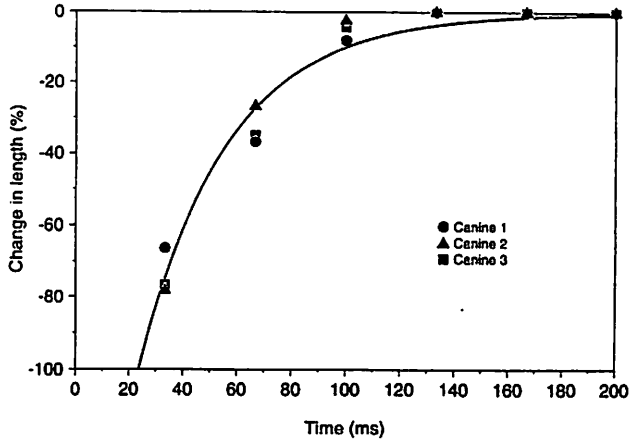

Figure 7. The time course of relaxation of vocal fold length from three larynges after removing the stimulation of RLN only.

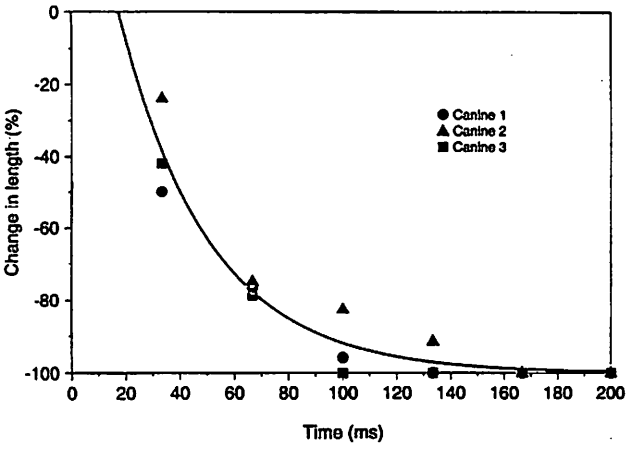

Figure 6. The time course of contraction of the vocal fold when the RLN was stimulated alone. Data are for three larynges.

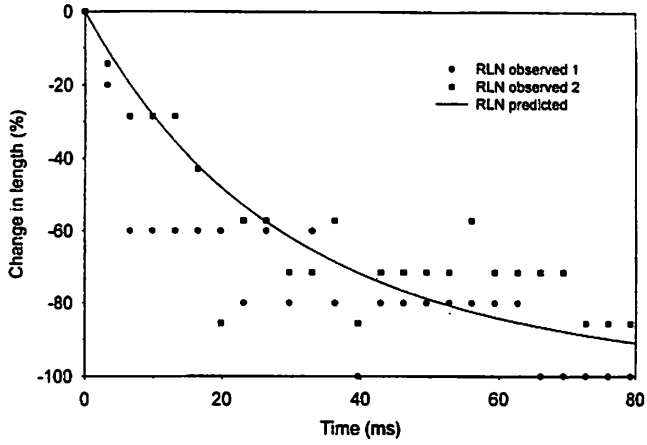

Figure 8. The time course of contraction from one larynx when the RLN was stimulated alone. The curve represents the average of two observations.

The higher frame rate single subject data are shown in Figure 5. The upper curve is for SLN alone and the lower curve for SLN plus RLN. Exponential time constants for relaxation are estimated to be on the order of 20-30 ms, but an exponential curve is not a good fit for this relaxation. There is a gradual release at the top, making the curves appear more like an ogive than an exponential.

When RLN was stimulated alone, the strain was negative. The contraction and relaxation curves, respectively, for the group data are shown in Figures 6 and 7. Only three larynges are represented because the stitch marks could not be seen well enough in the other larynges when the vocal folds adducted forcefully. Time constants were 33 ms for both contraction and for relaxation.

Data from the single larynx observed in Part II with higher frame rate also yielded a time constant of 33 ms for contraction, but a time constant of 36 ms for relaxation. Figures 8 and 9 show the contraction and relaxation curves for this case.

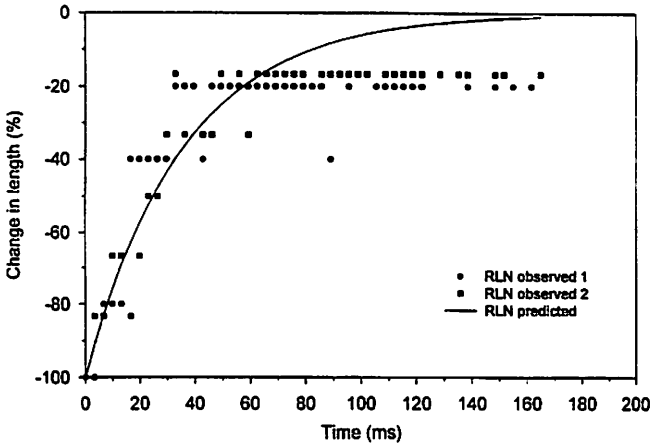

Figure 9. The time course of relaxation from one larynx after removing the stimulation of RLN only. The curve represents the average of two observations.

Discussion

It appears that the vocal fold elongation and contraction times measured here compare to the isometric contraction times of intrinsic laryngeal muscles. Alipour, Titze & Perlman[8] measured tetanic contraction times of 20-30 ms for the thyroarytenoid muscle. Cooper and colleagues[9] recorded about 40 ms for tetanic contraction of the posterior cricoarytenoid muscle and about 20 ms for a twitch response of the same muscle. We conclude, therefore, that the 20-40 ms response times measured here are mainly muscle response latencies. Mechanical delay due to cartilage rotation and translation is probably negligible. Specifically, there is little inertial or viscous delay in the cricothyroid joint and in the relative movement between the cricoid and thyroid cartilages.

This experiment did not address vocal fold lengths that are typical for phonation. During several experiments, we tried to get phonation with supramaximal RLN stimulation and a simulated lung system to derive the vocal folds. Glottal closure was too tight, however, to produce normal phonation. This result is noteworthy in relation to the canine in vivo studies by Berke and his coworkers. [10,11,12,13] Unless special care is taken to excite the laryngeal nerves in a submaximal manner (either by stimulating the distal branches to individual muscles or by pairing out separate sections of the nerve), the larynx will always be hyperadducted in phonation will require excessive subglottal pressure. This conclusion is also in agreement with our earlier findings,[14] in which we observed that normal phonation did not occur when the thyroarytenoid (TA) muscle was highly contracted without concomitant cricothyroid (CT) activity. At very high subglottal pressures (about 5 kPa), pressed phonation could be obtained. The phonation threshold pressure was much higher, however, than the 0.4-0.6 kPa pressure we see in the brainstem-evoked phonation[15] or the threshold pressure in excised larynges without any nerve stimulation (0.5 kPa to 2.0 kPa).

A comment about the current delivery is in order. To get supramaximal contraction the recurrent and superior nerves were stimulated with 0.5 mA. This experimentally determined current was larger than the current used in a previous report by Sato and Hisa,[16] who used 0.1 mA to get the tetanus contraction in the recurrent laryngeal nerve, but it is in the same range as the 0.5-1.2 mA used by Bielamonicz et al.[13] The difference in current delivery may be attributable to the difference in electrode structure. In our experiment, hooked-wire electrodes were used; Sato and Hisa used cuff electrodes.

Conclusions

Going back to the questions we asked at the beginning of this paper, (1) what is the maximum range of length change, and (2) what is the speed of length change, we can give the following answers on the basis of experiments with canines. The maximum positive strain of the membranous vocal folds is 26.3% to 71%, with a mean of 44.7% (n=7). This occurs under SLN stimulation only. The maximum negative strain of the vocal folds is -12.3 to -25.4%, with mean of -17.3%. This occurs under RLN stimulation only. When both SLN and RLN muscles are maximally contracted, the vocal folds elongate 18.4% to 47.1%, with a mean value of 26.3% (n=7).

The exponential time constant of elongation is 20-40 ms, with a mean of about 30 ms. The shape of the elongation curve appears to be truly exponential, but the relaxation curve appears more like an ogive, matching the response curves of individual laryngeal muscles under tetanic stimulation. A single time constant is therefore not as meaningful for relaxation as for contraction.

In relation to the maximum pitch change measured on humans (60-100 ms), the data are reasonable. First of all, humans don't contract muscles with supramaximal stimulation. Hence, one would expect a somewhat slower response due to motor unit recruitment and changing firing rates. Second, the time constants measured here describe the response up to 67% (1/e) of the target. We would expect full pitch changes to take two to three times that long if steady states are to be achieved. Hence, the 60-100 ms range for pitch change measured by Ohala and Ewan[3] and Sundberg[4] is quite consistent with our data.

Acknowledgments

This work has been funded by the National Institute on Deafness and Other Communication Disorders, Grant No: P60 DC00976.

References

1. Hollien, H. (1960). Vocal pitch variation related to changes in vocal fold length. *Journal of Speech and Hearing Research, 3*(2), 150-156.

2. Nishizawa, N., Sawashima, M., & Yonemoto, K. (1988). Vocal fold length in vocal pitch change. In: Fujimura, O. (Ed). *Vocal Physiology: Voice Production, Mechanisms and Functions*, 75-82. New York: Raven Press.

3. Ohala, J., & Ewan, W. (1973). Speed of pitch change. *Journal of the Acoustical Society of America*, 53, 345(A).

4. Sundberg, J. (1979). Maximum speed of pitch changes in singers and untrained subjects. *Journal of Phonetics, 7*, 71-79.

5. Alipour-Haghighi, F., & Titze, I. (1990). Stress relaxation in vocal fold tissues. *Advances in Bioengineering BED-17*, Amer. Soc. Mech. Eng., New York, N.Y., 21-24.

6. Cooper, D., Partridge, L., and Alipour-Haghighi, F. (1993). Muscle energetics, vocal efficiency, and laryngeal biomechanics. In: Titze, I. *Vocal Fold Physiology: Frontiers in Basic Science*, 37-92. San Diego, CA: Singular Publishing Group, Inc.

7. Titze, I., Jiang, J., & Druker, D. (1988). Preliminaries to the body-cover theory of pitch control. *Journal of Voice, 1*(4), 314-319.

8. Alipour-Haghighi, F., Titze, I.R., & Perlman, A. (1989). Tetanic contraction in vocal fold muscle. *Journal of Speech and Hearing Research, 32*, 226-231.

9. Cooper, D.S., Shindo, M., Sinka, U., Hast, M.H. & Rice, D.H. (1994). Dynamic properties of the posterior cricoarytenoid muscle. *Annals of Otology, Rhinology and Laryngology 103*(12), 937-944.

10. Berke, G., Moor, D., Hanson, D., Hantke, D., Gerratt, B., & Burstein, F.(1987). Laryngeal modeling: Theoretical, in vitro, in vivo. *Laryngoscope, 97*, 871-881.

11. Berke, G., Moore, D., Gerratt, B., Hanson, D., Bell, T., & Natividad, M. (1989). The effect of recurrent laryngeal nerve stimulation on phonation in an in vivo canine model. *Laryngoscope, 99*, 977-982.

12. Berke, G., Green, D., Smith, M., Arnstein, D., Honrubia, V., & Natividad, M. (1991). Experimental evidence in the in vivo canine for the collapsible tube model of phonation. *Journal of the Acoustical Society of America, 89*(3), 1358-1363.

13. Bielamowicz, S., Berke, G., Watson, D., Gerratt, B., Kreiman, J. (1994). Effects of RLN and SLN stimulation on glottal area. *Otolaryngology - Head & Neck Surgery, 110*(4), 370-380.

14. Titze, I.R., Luschei, E.S., & Hirano, M. (1989). The role of the thyroarytenoid muscle in regulation of fundamental frequency. *Journal of Voice, 3*(3), 213-224.

15. Solomon, N.P., Luschei, E.S., & Liu, K. (1995). Fundamental frequency and tracheal pressure during three types of vocalization elicited from anesthetized dogs. *Journal of Voice 9*(4), 403-412.

16. Sato, F., & Hisa, Y. (1987). Mechanical properties of the intrinsic laryngeal muscles and biomechanics of the glottis in dogs. In M. Hirano, J.A. Kirchner and D.M. Bless (Eds.). *Neurolaryngology: Recent Advances*. College Hill Press, pg. 97.

NCVS Status and Progress Report - 10
November 1996, 59-68

The Effect of Lung Volume Level on Selected Phonatory and Articulatory Variables

Christopher Dromey, Ph.D.
Lorraine Olson Ramig, Ph.D.
Department of Communication Disorders and Speech Science, The University of Colorado at Boulder
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts

Abstract

The purpose of the study was to examine the effects of manipulating lung volume level on phonatory and articulatory kinematic behavior during sentence production. Ten healthy adults repeated the sentence "I sell a sapapple again" under five respiratory conditions. These included a) a normal or reference condition, and conditions where the sentence was spoken b) after exhaling most of the air from the lungs, c) at end expiratory level, d) after a maximal inhalation, and e) after a maximal inhalation while attempting to maintain as normal a mode of speech as possible. From a multi-channel recording, measures were made of lung volume level, sound pressure level (SPL), fundamental frequency (F_0) and semitone standard deviation (STSD), and upper and lower lip displacements and peak velocities. When compared with the reference condition, the sentence was spoken more quickly at the lowest lung volume level. SPL increased for both of the higher lung volume conditions, as did females' fundamental frequency. STSD increased for the highest lung volume condition and decreased for speech at end expiratory level (EEL). Upper lip displacements and peak velocities generally decreased for lung volumes other than the reference condition. Lower lip movements showed inconsistent changes as a function of lung volume level. The data suggest that changes to the lung volume level for speech led to vocal intensity and fundamental frequency changes consistent with anticipated changes in subglottal pressure. However, less consistent effects were observed in the articulatory kinematic measures, possibly because of a less direct biomechanical linkage between respiratory and articulatory structures.

Introduction

Previous studies have shown that speakers often initiate speech at elevated lung volume levels to achieve the goal of increased vocal intensity (Hixon, 1973; Hixon, Goldman & Mead, 1973; Russell & Stathopoulos, 1988; Stathopoulos & Sapienza, 1993) or longer utterance durations (Winkworth, Davis, Ellis & Adams, 1994). The increased recoil forces at higher lung volume levels lead to passive increases in subglottal pressure (P_s). These forces can augment or substitute for the expiratory muscular effort required to achieve a desired speech intensity (Hixon, 1973). This suggests that the neural planning for speech can include the specification of higher lung volume levels in order to take advantage of passive biomechanical driving forces (McFarland & Smith, 1992). However, Winkworth, Davis, Adams and Ellis (1995) found the relationship between lung volume level and sound pressure level (SPL) to be inconsistent in spontaneous speech.

While researchers have examined the impact of speaking condition (e.g., loud versus soft--Russell & Stathopoulos, 1988; Stathopoulos & Sapienza, 1993), and various disorders (Hixon, Putnam & Sharp, 1983; Putnam & Hixon, 1983; Solomon & Hixon, 1993) on speech breathing, few studies have manipulated lung volume level as an independent variable to examine the effects on speech production. One study by Hoit, Solomon and Hixon (1993), examined voice onset time (VOT) as a function of lung volume level. They discovered that VOT generally decreased as lung volume decreased from the highest to the lowest levels their speakers could attain, and suggested that this might be due to changes in the position and configuration of the larynx secondary to movements of the diaphragm, lungs and trachea. In light of the previous studies that showed that speakers inhaled to higher lung volume levels to

speak loudly, it seems reasonable to assume that vocal intensity would also have increased at higher lung volumes in the Hoit et al. (1993) study. Since the authors did not report this variable, their findings could possibly have been confounded by changes in SPL. The paucity of published accounts of studies using lung volume as an independent variable lend justification to the present investigation into the effects of lung volume changes on speech production variables, including SPL. It might be hypothesized that individuals would increase their vocal intensity after inhaling to higher lung volume levels. On the other hand, it is possible that speakers would maintain a comfortable intensity level even when inspiring more deeply, implying that they would counteract the increased expiratory recoil forces by activating inspiratory muscles to check the descent of the ribcage (Hixon, 1973).

Some authors have noted that respiratory support can be inadequate in individuals who have voice disorders (Aronson, 1990; Hixon & Putnam, 1983; Wilson, 1987). Sperry, Hillman and Perkell (1994) found erratic respiratory patterns in a patient with vocal nodules, who frequently initiated utterances at low lung volumes. Sapienza and Stathopoulos (1994) reported increases in lung volume excursions for women and children with vocal nodules. Clinicians often encourage their patients to modify their respiratory patterns in order to provide the prerequisite pulmonary support for more normal voice production. This often entails inhaling more deeply prior to speaking, and using more appropriate breath groups (Cooper & Cooper, 1977; Greene, 1972). It would therefore be valuable to know how deliberate changes in respiratory function can influence phonatory as well as other speech production variables. McFarland and Smith (1992) suggested that neural planning processes take the prevailing lung volume into account for speech starting at different points in the respiratory cycle. When the recoil forces at a given lung volume level provide an adequate air pressure for speech, fewer preparatory movements of the chest wall are required than where pressure is insufficient. It therefore seems reasonable that deliberately manipulating lung volume level could lead to modifications in the way speech is produced.

If vocal intensity were to increase because of greater recoil forces at higher lung volume levels, it could be anticipated that various phonatory changes would occur. Increases in fundamental frequency (F_0) have been associated with higher subglottal pressure and elevated vocal intensity by several researchers (Dromey, Stathopoulos & Sapienza, 1992; Hixon, Klatt, & Mead, 1971; Lieberman, Knudson, & Mead, 1969; Scherer, 1991; Titze, 1989). Fundamental frequency variability, expressed as semitone standard deviation (STSD), has also been found to increase with vocal intensity (Dromey et al., 1995; Ramig, Countryman, Thompson & Horii, 1995). Increased vocal intensity could also contribute to improved phonatory stability (Orlikoff & Kahane, 1991).

While tentative predictions can be made about the effects of respiratory changes on phonatory behavior, there is a lack of published data to allow similar predictions to be made regarding articulation. Some authors have addressed the relationship between increases in vocal intensity and articulatory excursions and velocities. Schulman (1989) found that excursions and peak velocities of the lips and jaw increased for louder than normal speech. He suggested that the firmer bilabial closure associated with stop production in loud speech might be a means of compensating for increased intraoral pressures. He also speculated that the upward shifting of formant frequencies might help preserve vowel identities where F_0 rises with vocal intensity. Dromey, Ramig and Johnson (1995) reported increases in second formant transition duration and extent when vocal intensity increased, suggesting that the articulatory movements were larger. If intensity were to change as a function of lung volume level, articulatory excursions might increase concomitantly with this rise in intensity. On the other hand, it might be found that when concentrating on deliberate adjustments to respiratory activity, a speaker no longer maintains an association between increased vocal intensity and larger articulatory movements. It is unclear whether the increases in articulatory excursions and velocities found by Schulman (1989) occur only when individuals volitionally increase speech intensity, or whether similar increases might be anticipated when intensity increases as a biomechanical consequence of lung volume changes.

Based on their findings with cleft palate speakers, Warren, Dalston, Morr, Hairfield and Smith (1989) hypothesized that "activities of the respiratory and articulatory systems are coordinated for the purpose of regulating speech pressures or some correlate of pressure" (p. 571). This statement was made in the context of speakers increasing their respiratory drive to compensate for reduced vocal tract resistances. It could be speculated that a similar type of coordination exists, whereby changes in respiratory activity are associated with modifications to articulatory behavior, possibly with the goal of compensating for a rise in P_s (Schulman, 1989).

The purpose of the present study was to evaluate the effects of lung volume changes on phonatory and articulatory kinematic behavior. Preliminary data from a patient with Parkinson disease (Dromey et al., 1995) suggest that therapy aimed at increasing vocal intensity can have an impact on both respiratory and articulatory activity. The present study sought to determine whether changes targeting the respiratory system alone would affect the articulatory and phonatory aspects of speech. Specifically, it was hypothesized that: a) speaking at higher than normal lung volume levels will lead to increases in SPL, F_0 and STSD, b) this increase in SPL will be associated with increases in articulatory displacements and velocities, and c) speech produced at lower than normal lung volume levels will show phonatory and articulatory changes in the opposite direction.

Method

Subjects

The participants were 10 native speakers of English, who were non-smokers with no history of asthma, and no professional singing or acting experience. On the day of the study, they were free from respiratory infection. Two of the five female subjects had limited singing experience in the past, but had not at any time earned a living in performance. The five males ranged in age from 26 to 34 (mean 31.0) and the five females from 23 to 39 (mean 32.0). All experimental participants passed a hearing screening at 20 dB HL, and had no history of disordered communication.

Instrumentation

All data were collected while participants sat in a medical examining chair in an IAC sound-treated booth. Variable inductance plethysmograph bands (Respigraph--Non Invasive Monitoring Systems, Inc., PN SY03) were placed around the ribcage at the level of the nipples, and around the abdomen at the level of the umbilicus, below the level of the lowest rib to avoid sensing ribcage movements. The bands were secured to the participants' clothing with micropore tape. Signals from the ribcage (RC) and abdominal (AB) transducers were recorded.

Participants wore a head-mounted microphone (AKG C410), which was adjusted to maintain a constant 4 cm distance from the mouth. A sound level meter (Bruel and Kjaer Type 2230) was positioned 30 cm from the speaker's mouth. These distances had previously been found (in the sound booth where the study was conducted) to yield optimal signal to noise ratios without signal distortion as well as accurate SPL measurements without interference from ambient noise. The distance of the sound level meter from the speaker's mouth was periodically re-checked during the recording session.

A lightweight, head-mounted strain gauge cantilever system (Barlow, Cole & Abbs, 1983) was used to track the movements of the upper and lower lip during speech. This equipment was calibrated prior to the experiment to allow the analog voltage output to be interpreted as a displacement in mm. The cantilever beams were guided through small beads attached with an adhesive tab to the speaker's upper and lower lip at the midline.

Once all of the transducers were positioned appropriately, participants were asked to inhale and then exhale maximally. This task was performed twice, and was used to derive a measure of vital capacity, against which the lung volumes during the speech conditions were measured as percent vital capacity (%VC--see Russell & Stathopoulos, 1988). Experimental participants performed an isovolume maneuver at end expiratory level (EEL) to allow a sum channel to be generated which equally weighted the contributions of the rib cage and abdominal transducers. Measures of %VC were made from this sum channel.

All signals from the transducers were stored on an 8 channel digital audio tape (DAT) recorder (Sony PCM 108) for subsequent off-line digitization. Signals were digitized using a WINDAQ DI-200 hardware/software data acquisition system on a 486/66 PC. This equipment allowed the on-line monitoring of signals during data collection by displaying all 8 channels on a computer screen while they were being stored on the DAT recorder. The microphone signal was also recorded onto a 2 channel DAT recorder (Panasonic SV-3700), which provided higher bandwidth storage than the 8 channel device.

Subsequent analysis of the digitized signals was performed with WINDAQ EX software to extract calibrated measures of duration and amplitude from the sampled signals, as well as derived measures (e.g., velocity) following additional processing. The sampling rate for the movement, SPL, respiration and microphone signals was 500 Hz, with a low-pass filter cut-off at 200 Hz. The microphone signal in this data set served as a temporal marker. The microphone signal from the 2 channel DAT recorder was low-pass filtered at 5 kHz and digitized at 10 kHz for fundamental frequency analysis with CSpeech software.

Speech Tasks

Following the respiratory maneuvers, the participants were instructed to repeat the sentence "I sell a sapapple again" in various ways. The stimulus sentence was selected for several reasons. It allowed an examination of lip opening and closing during the /paep/ syllable of the word "sapapple" and also permitted relatively natural production because of its normal syntactic form. It was easy to say without placing stress on any particular part of the sentence. Another important reason is that there is a considerable precedent in the motor speech literature for the word "sapapple" (Abbs & Connor, 1991; Gracco, 1988; Gracco & Abbs, 1986, 1988, 1989). The speakers were instructed to say the sentence 10 times, with enough time between each token to relax and take a breath.

The specific instructions for each condition were as follows: "Say the sentence...

1. reference condition: "as you normally would"
2. maximum lung volume condition: "immediately after taking a very deep breath"
3. end expiratory level condition: "after a sigh, without taking in any air first"
4. low lung volume condition: "after breathing out most of your air first"
5. maximum lung volume while speaking normally: "immediately after taking a very deep breath, but concentrating on saying the sentence as normally as possible."

The sequence of instructions was the same for each participant. They were not required to reach specific lung volume targets set by the experimenter, but rather to follow

the verbal instructions as closely as possible. These five conditions were selected to allow comparisons to be made between speech produced normally, and at a series of points representing low, EEL and high lung volume levels. When the individuals were asked to speak as normally as possible after a maximal inhalation, the goal was to determine whether they would be able to compensate for the substantial recoil forces anticipated at this lung volume level. The five conditions were chosen because the participants could achieve them by following simple instructions, without reference to respiratory instrumentation. This offered the advantage of having speakers avoid precise target-matching paradigms, which have been found by previous researchers to have an impact on certain dependent measures (Hanson, Gerratt & Berke, 1990).

For the different respiratory conditions, the experimenter monitored the signals on-line from the Respigraph system to ensure that participants were achieving the goal of different lung volume levels. At no time were the tasks modeled by the experimenter, since it was felt that this could influence the speakers' performance if they were to imitate aspects of the experimenter's speech other than the requested respiratory manipulation.

A speech-language pathologist, who was monitoring the signals during recording, served as a second judge of the adequacy of the participants' performance of the requested tasks. Before experimental tokens were produced for each set of 10 sentences, speakers were encouraged to practice saying the sentence several times under each new speech condition until they felt comfortable with the new mode of production.

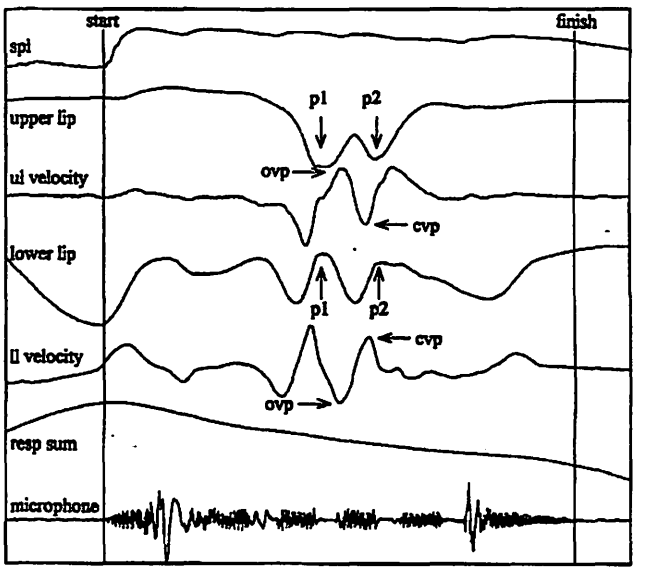

Figure 1. Measures derived from the simultaneously acquired signals. Start and finish represent the onset and offset of phonation for the sentence "I sell a sappaple again." P1 and P2 are the lip displacements associated with the first and second bilabial closures in the word "sappaple." OVP and CVP are opening and closing gesture velocity peaks for the bilabial gestures.

Data Analysis

Respiratory Activity

The signals from which the dependent measures were made are diagrammed in Figure 1. From the derived respiratory sum channel, which represents lung volume, measures were made of the lung volume at the start of the acoustic signal ("start"), and the lung volume at its end ("finish"). The difference in lung volume between these two points ("start, finish") was taken to represent the respiratory excursion for the speech task (Russell & Stathopoulos, 1988). All volumes were expressed as percent vital capacity (%VC), based on the vital capacity maneuver prior to the speech tokens.

Since the experimental conditions involved a wide range of lung volume levels, there was a possibility that the bands of the Respigraph system might not track lung volume linearly, particularly at the extremely high or low ends of the range. In order to ascertain whether the data might become contaminated through nonlinearities in the measurement system, sets of test calibrations were performed. The voltage of the Respigraph output was measured as a function of lung volume level as measured with the spirometer across the vital capacity range. A linear relationship was found between the Respigraph output and the spirometer volumes (0 to 100% VC in 20% increments). The R^2 values ranged from .968 to .994 for these linearity tests, and therefore the experimental data were considered valid.

Phonatory Activity

The utterance duration was measured from the microphone signal using cursors in the WINDAQ program (Figure 1 "start" and "finish"). This corresponds to the measure of mean speech rate over the entire utterance. Previous researchers have used duration as the inverse of mean speech rate for the utterance (Adams, Weismer & Kent, 1993).

The mean SPL during the utterance was determined from the calibrated sound level meter channel. The arithmetic mean of the data points between the cursors was taken as an index of overall intensity, expressed as dB SPL at 30 cm.

The mean and standard deviation of the fundamental frequency (F_0) were derived from the microphone signal from the 2 channel DAT recorder to determine the effects of lung volume level on the fundamental frequency contour of the voice during the production of the sentence. This analysis was performed with CSpeech 4.0. The F_0 contour was statistically analyzed to derive the mean and standard deviation in Hz, and the latter was converted into semitones (STSD) with a spreadsheet macro.

Articulatory Activity

The DC voltage signal from the strain gauge cantilever transducers represented a calibrated analog of the upward and downward movements of the lips. From this record, measures were made for the upper and the lower lip of the displacement (in mm) from /p/ to /ae/ and from /ae/ to the subsequent /p/ in the /paep/ of "sapapple" (Figure 1, points p1, p2 and the vowel displacement between them). It is recognized that the lower lip signal represented the sum of lower lip and jaw movements. This combined displacement was selected because it allowed data to be gathered regarding the degree of oral opening during speech. Additional instrumentation to gather a separate jaw movement signal was not available. A smoothed derivative (10 point moving average) of each lip's channel was produced with the WINDAQ software to calculate the peak opening (ovp) and closing velocities (cvp) for these speech gestures.

Measurement Reliability

Twenty percent of the data for each dependent variable were randomly selected and reanalyzed by the same experimenter for the purpose of assessing measurement reliability. The mean Pearson correlation coefficient across all dependent measures for the original and the re-checked data was $\underline{r} = 0.999$, and ranged from $\underline{r} = .995$ to $\underline{r} = 1.00$ ($\underline{p} < .001$) for the individual variables. The mean % difference between the original and re-checked measures was 0.05%, and ranged from 0% to a maximum of 1.18% for the individual variables. The largest measurement error in the data corresponded to a difference of 0.06% VC in the measure of lung volume excursion.

Intrasubject Variability

The statistical data reported below reflect the mean values and standard deviations for the 10 subjects together. The subjects differed in the degree to which they were consistent in their performance over the 10 repetitions under each speaking condition. For each subject, a coefficient of variation (CV) was calculated by dividing the standard deviation by the mean for that subject's tokens in each condition. The mean CV value across all subjects and conditions was 0.117, and ranged from 0.002 to a high of 0.931. However, this rather extreme value reflected only a few %VC variability around a mean which was close to zero for the end of sentence lung volume for one subject.

Statistical Analysis

Statistical analysis of the data was performed using the Statistical Package for the Social Sciences (SPSS) for Windows 6.1. The initial statistical analysis used was a repeated measures analysis of variance for each dependent measure across the lung volume conditions. Sex was included as a factor in the ANOVAs, but except for the few instances noted below, there were no interactions of sex with the levels of the independent variable. In each section below, the reported values represent the mean for all 10 subjects, except in the case of fundamental frequency, which is reported as the mean for the 5 subjects of each sex. Separate analyses were conducted for males and females for fundamental frequency, because of the substantial male/female differences in F_0. Fundamental frequency variability, on the other hand, was analyzed with both sexes together, since STSD already accounts for the male/female differences in mean F_0. Nonorthogonal contrasts were performed between the reference condition and each of the other levels of the independent variable. This allowed comparisons to be made between "normal" or reference condition speech, and sentences which were produced at the other lung volume levels. Because of the relatively small n and the large number of tests, an alpha level of .01 was selected to assess the significance of the results.

Results

Respiratory Activity

The changes in the dependent measures associated with the subjects' manipulation of lung volume level are summarized in Table 1. The F-ratios and p-values for these repeated measures ANOVAs are shown in Table 2. As would be expected, the lung volume level at which speech

Table 1.
Mean (and standard deviation) effects of lung volume level changes on respiratory, phonatory and articulatory measures.

Measure	Unit	low LV	sd	EEL	sd	ref	sd	mxlv	sd	mxno	sd
LV initiation	%VC	26.0	(12.9)	43.4	(9.7)	59.0	(8.1)	89.4	(5.2)	87.1	(6.7)
LV termination	%VC	19.7	(12.7)	35.7	(10.9)	51.6	(7.6)	83.8	(6.1)	79.3	(7.4)
LV excursion	%VC	6.4	(4.1)	7.7	(3.3)	7.4	(2.1)	5.6	(2.3)	7.8	(2.5)
Utterance duration	sec	1.54	(0.12)	1.56	(0.14)	1.65	(0.16)	1.61	(0.15)	1.58	(0.11)
Mean SPL	dB	64.3	(4.0)	64.6	(4.3)	66.6	(3.1)	71.9	(3.6)	68.3	(3.9)
Mean F0 (m)	Hz	100.6	(13.5)	101.1	(12.2)	103.5	(8.9)	109.9	(7.2)	103.8	(9.7)
Mean F0 (f)	Hz	199.0	(17.3)	197.3	(16.2)	200.9	(19.3)	228.7	(20.7)	225.3	(18.4)
STSD	ST	1.68	(0.52)	1.60	(0.60)	1.83	(0.71)	2.30	(0.58)	2.21	(0.77)
UL PAE displacement	mm	1.53	(0.61)	1.47	(0.87)	2.02	(0.93)	1.85	(0.88)	1.59	(0.80)
UL AEP displacement	mm	1.59	(0.80)	1.62	(0.89)	2.19	(0.94)	1.93	(0.90)	1.70	(0.93)
UL OP peak velocity	mm/s	33.5	(12.7)	35.6	(23.5)	44.2	(22.6)	39.6	(21.9)	34.9	(18.6)
UL CL peak velocity	mm/s	31.7	(12.8)	32.5	(16.6)	43.5	(19.3)	36.8	(16.3)	34.6	(18.4)
LL PAE displacement	mm	9.89	(3.12)	9.65	(3.47)	10.76	(3.42)	10.95	(3.66)	10.80	(3.76)
LL AEP displacement	mm	9.03	(2.90)	8.88	(3.23)	10.04	(3.34)	10.28	(3.65)	10.01	(3.61)
LL OP peak velocity	mm/s	149.9	(51.8)	152.4	(61.4)	164.6	(59.0)	164.3	(65.1)	171.5	(67.0)
LL CL peak velocity	mm/s	187.2	(63.4)	174.3	(65.3)	193.0	(64.9)	205.3	(71.7)	202.1	(78.6)

low LV = speech after exhaling most of the air in the lungs
EEL = speech after a sigh, without an inhalation
ref = reference or control condition
mxlv = speech after a deep inhalation
mxno = speech produced as normally as possible after a deep inhalation

Table 2.
F-ratios and probability values for repeated measures ANOVAs on all dependent measures.

Measure	F-ratio	p-value	df
LV initiation	136.84	< .001	4,32
LV termination	157.66	< .001	4,32
LV excursion	2.19	.093	4,32
Utterance duration	4.21	.008	4,32
Mean SPL	25.78	< .001	4,32
Mean F0 (m)	2.55	.080	4,16
Mean F0 (f)	15.65	< .001	4,16
STSD	18.97	< .001	4,32
UL PAE displacement	6.73	< .001	4,32
UL AEP displacement	4.77	.004	4,32
UL OP peak velocity	3.84	.012	4,32
UL CL peak velocity	5.72	.001	4,32
LL PAE displacement	2.44	.067	4,32
LL AEP displacement	3.28	.023	4,32
LL OP peak velocity	2.11	.103	4,32
LL CL peak velocity	2.66	.050	4,32

was initiated changed significantly, since this was the independent variable (F [4,32] = 136.84, $p < .001$). The results of the nonorthogonal contrast analyses are reported in Table 3. They indicate significant differences between the reference set (59.0% VC) and each of the other speaking conditions (low lung volume (26.0% VC), EEL (43.0% VC), maximum lung volume (89.4% VC) and maximum lung volume while speaking normally (87.1% VC)) for the measure of lung volume at the start of the sentence.

The end of sentence lung volume level also changed significantly (F [4,32] = 157.66, $p < .001$). The contrasts between the reference set (51.6% VC) and each of the other speaking conditions (low lung volume level (19.7% VC), EEL (35.7% VC), maximum lung volume (83.8% VC) and maximum volume while speaking normally (79.3% VC)) were all statistically significant.

There were inconsistent changes in lung volume excursion, which did not reach significance in the ANOVA

Table 3.
Nonorthogonal contrast analysis results for lung volume level conditions, comparing each against the reference or control condition.

Measure	low LV		EEL		mxlv		mxno	
	F-ratio	p-value	F-ratio	p-value	F-ratio	p-value	F-ratio	p-value
LV initiation	45.73	< .001	36.65	< .001	84.51	< .001	152.55	< .001
LV termination	51.88	< .001	27.26	.001	95.24	< .001	174.77	< .001
LV excursion	0.56	.472	0.14	.719	3.32	.102	1.10	.322
Utterance duration	10.17	.011	3.92	.079	1.47	256	4.83	.056
Mean SPL	4.47	.064	4.44	.064	38.95	< .001	5.83	.039
Mean F0 (m)	1.57	.278	2.26	.207	2.94	.162	1.88	.243
Mean F0 (f)	0.13	.735	1.27	322	37.64	.004	63.19	.001
STSD	1.84	.209	11.14	.009	22.41	.001	3.76	.085
UL PAE displacement	11.25	.008	20.56	.001	1.62	.235	5.37	.046
UL AEP displacement	10.69	.010	13.06	.006	2.78	.130	6.04	.036
UL OP peak velocity	6.74	.029	12.07	.007	2.35	.160	5.41	.045
UL CL peak velocity	10.19	.011	16.83	.003	4.88	.054	6.07	.036
LL PAE displacement	2.41	.155	4.00	.077	0.12	.733	0.01	.934
LL AEP displacement	4.27	.069	4.00	.077	0.25	.630	0.01	.944
LL OP peak velocity	3.74	.035	2.18	.174	0.00	.981	0.67	.434
LL CL peak velocity	0.29	.602	3.53	.093	1.44	.261	1.05	.332

low LV = speech after exhaling most of the air in the lungs
EEL = speech after a sigh, without an inhalation
mxlv = speech after a deep inhalation
mxno = speech produced as normally as possible after a deep inhalation

(F [4,32] = 2.19, $p = .093$) or in any contrast. There was considerable variability in this measure across the subjects.

Phonatory Activity

The utterance duration changed significantly when lung volume level was adjusted (F [4,32] = 4.21, $p = .008$). Only in the low lung volume level condition (1.54 s) did the contrast with the reference set (1.65 s) approach statistical significance ($p = .011$) for this measure. These data are shown in Figure 2.

SPL changed significantly as a function of lung volume level (F [4,32] = 25.78, $p < .001$). Contrasts were statistically significant on this measure between the reference set (66.6 dB) and for speech after a maximal inhalation (71.9 dB).

Mean fundamental frequency for females changed significantly with lung volume level (F [4,16] = 15.65, $p < .001$). Contrasts were significant between the reference set (201 Hz) and for speech after a maximal inhalation (229 Hz) and for speech produced as normally as possible after a maximal inhalation (225 Hz).

Mean fundamental frequency for males increased for the higher lung volume conditions, but the changes were not statistically significant for the ANOVA (F [4,16] = 2.55, $p = .080$) or any contrasts. The slightly greater increases for

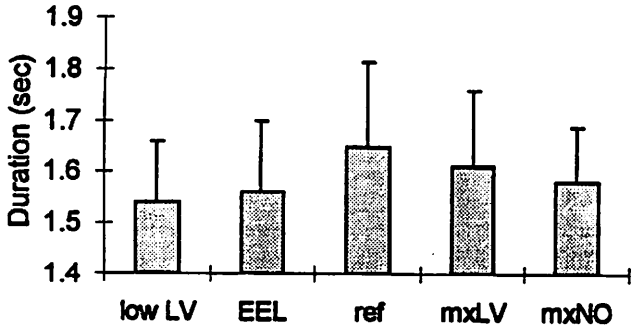

Figure 2. Mean and standard deviation sentence duration as a function of lung volume level. Low LV = speech produced after exhaling most of the air in the lungs. REL = speech following a sigh without an inhalation. Ref = control or reference condition. mxLV = speech after a deep inhalation. mxNO = speech produced as normally as possible after a maximal inhalation.

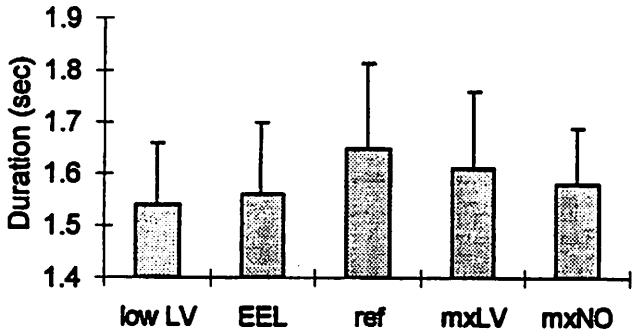

Figure 3. Mean Fo for male and female subjects as a function of lung volume level.

females led to a level by sex interaction when F_0 data for both sexes were analyzed together ($\underline{F}$ [4,32] = 5.56, $\underline{p}$ = .002; see Figure 3).

Fundamental frequency variability (STSD) changed significantly as a function of lung volume level condition ($\underline{F}$ [4,32] = 18.97, $\underline{p} < .001$). The significant contrasts were between the reference condition (1.8 ST) and speech initiated at EEL (1.6 ST) or after a maximal inhalation (2.3 ST - see Figure 4).

Some of the males showed a difference in STSD between the EEL and reference conditions, whereas the females did not generally show as great a difference. Some of the female subjects also showed more dramatic STSD increases for the higher lung volume level conditions. These patterns resulted in a level by sex interaction for this measure ($\underline{F}$ [4,32] = 5.85, $\underline{p}$ = .001).

Articulatory Activity

Upper Lip: The opening displacement of the upper lip for the /pae/ gesture (see Figure 5) changed significantly across the lung volume level range ($\underline{F}$ [4,32] = 6.73, $\underline{p} <$ .001). Significant contrasts were found between the reference set (2.0 mm) and the low lung volume (1.5 mm), and EEL (1.5 mm) conditions.

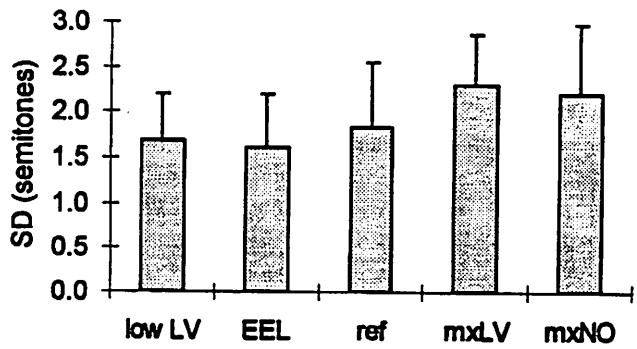

Figure 4. Semitone standard deviation as a function of lung volume level.

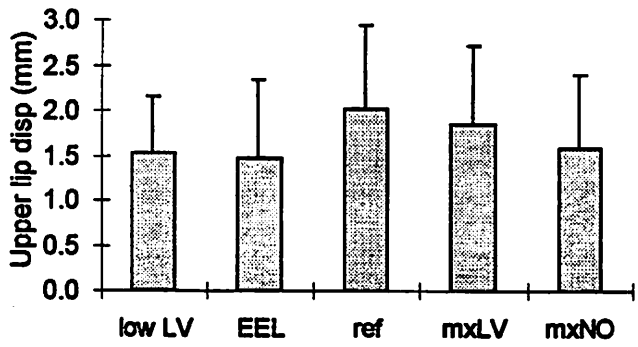

Figure 5. Upper lip opening displacement as a function of lung volume level.

The magnitude of the /aep/ gesture (for closing) changed significantly with lung volume level ($\underline{F}$ [4,32] = 4.77, $\underline{p}$ = .004). As was the case for the opening gesture, significant contrasts existed between the reference set (2.2 mm) and the low lung volume (1.6 mm), and EEL (1.6 mm) conditions.

The magnitude of the peak velocity for upper lip opening changed with lung volume level ($\underline{F}$ [4,32] = 3.84, $\underline{p}$ = .012). There was a significant contrast between the reference set (44 mm/s) and speech at EEL (36 mm/s).

The size of the closing velocity peak also changed across the respiratory conditions ($\underline{F}$ [4,32] = 5.72, $\underline{p}$ = .001). Again the significant contrast was between the reference set (44 mm/s) and speech at EEL (33 mm/s).

Lower Lip: The displacement for the lower lip closing gesture changed ($\underline{F}$ [4,32] = 3.28, $\underline{p}$ = .023) across the range of lung volumes, but none of the individual contrasts were found to be significant. For the lower lip opening gesture, changes in displacement were not significant, even at the $\underline{p} < .05$ level.

The peak closing velocity changed across the lung volume level continuum ($\underline{F}$ [4,32] = 2.66, $\underline{p}$ = .050), but none of the contrasts were significant. The peak velocity for opening did not change significantly ($\underline{F}$ [4,32] = 2.11, $\underline{p}$ = .103) across the respiratory conditions.

Discussion

Few studies have manipulated lung volume level as an independent variable (Hoit, Solomon & Hixon, 1993). Rather, the focus has been on the respiratory characteristics of speech produced under a variety of conditions (Russell & Stathopoulos, 1988; Smith & Denny, 1990; Stathopoulos, Hoit, Hixon, Watson & Solomon, 1991; Stathopoulos & Sapienza, 1993). The present study, therefore, represents a departure from the direction taken by most investigators of respiratory function, and augments previous research by examining changes that occur in phonation and articulation as respiratory patterns are deliberately manipulated.

The data support the view that there may be a fairly simple relationship between respiratory effort, sound pressure level and fundamental frequency. In other words, changes in lung volume level have a reasonably predicable impact on phonatory behavior. This could be due to aerodynamic, biomechanical factors, since the increased recoil forces found at high lung volume levels would be expected to elevate subglottal pressure, and thus SPL and F_0 (Hixon et al., 1971; Isshiki, 1964; Lieberman et al., 1969; Titze, 1989, 1994). It seems reasonable that the neural control of respiration and phonation should be closely coordinated, since laryngeal reflexes play such an important role in airway protection (Widdicombe, 1974), and because brain stem motor neurons which are active in respiratory control have also been associated with laryngeal activity (Newsom Davis, 1970).

In contrast, there does not seem to be a straightforward link between lung volume level and articulatory excursions and velocities. The changes in sound pressure level which occurred as a function of lung volume level led to phonatory changes, and might also be expected to lead to changes in articulatory activity, since previous work has shown that articulatory displacements and velocities increase when speakers talk more loudly (Dromey et al., 1995; Schulman, 1989). However, the SPL increases at higher lung volume levels were associated with slightly larger excursions and velocities for the lower, but not for the upper lip, when the mean data for 10 subjects are considered. When the individual subjects are examined, 6 of the 10 had the largest upper lip displacements and velocities for the reference condition. Thus, there may be a predictable association between respiratory effort and phonatory variables, but the impact on articulatory measures is not comparable. This may be because there is a less direct biomechanical relationship between respiratory effort and articulatory excursions, or because the neural control mechanisms for articulation are not as closely tied to respiratory regulation as are those for the larynx. It might be speculated that when intensity is deliberately increased (e.g., Schulman, 1989), the motor control strategies involved are different than when intensity increases as a consequence of modifications to lung volume level. It is also possible that the SPL increases at high lung volume levels in the present study (5 dB) were not sufficiently large to elicit a consistent articulatory kinematic effect.

It is unclear why there would be a shorter utterance duration (a faster mean rate of speech) in the non-reference respiratory conditions. While in individual contrasts the reference tokens only differed significantly from those initiated at low lung volume levels, there is nevertheless a pattern in the data, with 6 of the 10 subjects having the longest sentence duration in the reference condition. It is plausible that in the EEL and low lung volume conditions, speech became faster in many instances because of uncertainty on the part of the subjects as to whether there would be sufficient air for speech. Alternatively, the subjects might have had a sensation of increased effort in attempting to maintain adequate pressures and flows for speech when speaking at very low lung volume levels, since significant muscular effort is required to counteract the inspiratory recoil forces (Hixon, 1973). This might have prompted the subjects to speak more quickly to avoid a prolonged expenditure of expiratory effort. But the reason for more rapid speech for some subjects at the very high lung volume levels is unclear, unless perhaps, the extreme recoil forces activated pressure-sensitive upper airway afferents (Sant'Ambrogio, Mathew, Fisher & Sant'Ambrogio, 1983). This could have led to a sensation of excess pressure, and thus an urgency to speak the sentence because of the difficulty in counteracting such high recoil forces. Godfrey (1974) reported that the discomfort associated with breath holding at high lung volume levels could be relieved by movements of the chest wall. It is possible that speakers in the present study felt uncomfortable speaking at very high lung volume levels, and rushed to speak as a result.

The decreases in upper lip displacements and velocities in speaking conditions where the rate of speech increased could represent an articulatory undershoot of spatial targets (Ostry & Munhall, 1985). However, previous research has shown that an increase in rate can have very different effects on articulatory displacements and velocities across subjects (Adams et al., 1993). Some speakers increase the velocities of their articulatory movements while maintaining similar displacements; others reduce movement excursions but do not change their velocity; still others decrease both the amplitudes and the velocities of their articulatory movements (Flege, 1988). Thus, the present findings of decreases in upper lip displacement for non-natural respiratory conditions might not necessarily be a rate-induced phenomenon.

The condition in which subjects were instructed to speak as normally as possible at a high lung volume level was included to examine whether subjects would be capable of compensating for the increased recoil forces which would be encountered in this lung volume range. As the respiratory data show, part of the compensation involved inhaling to a slightly lower level, although this was only about 2% differ-

ent from the previous condition where the instruction was simply to speak after inhaling maximally. It is also possible that this slightly reduced lung volume level occurred as a sequencing effect, since subjects had spoken at their lowest lung volume level prior to this condition. SPL decreased 4 dB from the simple maximum lung volume condition, but remained 2 dB above the reference condition, suggesting an incomplete compensation for the elevated subglottal pressure at the maximum lung volume level.

If intraoral pressure increased substantially at high lung volume levels, the speakers in the study might have altered their articulatory activity in reaction to the pressure. Previous work (Netsell, 1969; Shipp, 1973) has shown that during voiceless stops, intraoral air pressure is essentially equal to tracheal pressure. Williams, Brown and Turner (1987) have shown that speakers can detect changes in intraoral pressure as small as 1 cmH2O. It is possible that the pressure changes in the present study were larger than this. It might be speculated that in the voiceless plosive context of the present study, intraoral air pressure increases disturbed the subjects' ability to achieve normal patterns of lip closure. Future studies where intraoral pressure is measured could help clarify this potential effect.

Speaking at elevated lung volume levels in the present study resulted in intensity increases, even though subjects were not instructed to speak more loudly. Many clinicians encourage their voice disordered patients to improve respiratory support for speech, especially where speech is initiated without adequate inhalation before an utterance. The present results suggest that such a strategy would be valuable in achieving improved vocal intensity, if this were one of the therapy goals. The present study did not examine vocal function in detail, and it might be found that there were improvements in the quality of phonation as speakers progressed from a lower than normal to a normal or slightly elevated lung volume level. Future studies which investigate the acoustics and aerodynamics of phonation in greater detail would be useful to investigate lung volume level effects, particularly with regard to the impact on vocal fold adduction and changes in the glottal source spectrum.

The present study examined healthy speakers in an artificial speech context under somewhat unusual respiratory conditions. Nevertheless, the data show that efforts aimed at modifying the activity of one speech subsystem--in this case respiration--can have varying degrees of impact on the phonatory and articulatory subsystems. These findings underline the need for researchers to be aware of potential carry-over effects between different parts of the speech production mechanism when the activity in one individual subsystem is manipulated. As more is learned about the overall coordination of speech production, it might be possible to develop more effective treatment approaches which take advantage of the natural interactions between the different subsystems (Dromey et al., 1995; Ramig et al., 1995).

References

Abbs, J.H., & Connor, N.P. (1991). Motorsensory mechanisms of speech motor timing and coordination. Journal of Phonetics, 19, 333-342.

Adams, S.G., Weismer, G., & Kent, R.D. (1993). Speaking rate and speech movement velocity profiles. Journal of Speech and Hearing Research, 36, 41-54

Aronson, A.E. (1990). Clinical voice disorders (3rd ed.). New York: Thieme Inc.

Barlow, S.M., Cole, K.J., & Abbs, J.H. (1983). A new head-mounted lip-jaw movement transduction system for the study of motor speech disorders. Journal of Speech and Hearing Research, 26, 283-288.

Cooper, M., & Cooper, M.H. (1977). Direct vocal rehabilitation. In M. Cooper & M.H. Cooper (Eds.), Approaches to vocal rehabilitation (pp. 57-72). Springfield, IL: Charles C. Thomas.

Dromey, C., Ramig, L.O., & Johnson, A.B. (1995). Phonatory and articulatory changes associated with increased vocal intensity in Parkinson disease: A case study. Journal of Speech and Hearing Research, 38, 751-764.

Dromey, C., Stathopoulos, E.T., & Sapienza, C.M. (1992). Glottal airflow and electroglottographic measures of vocal function at multiple intensities. Journal of Voice, 4, 44-54.

Flege, J.E. (1988). Effects of speaking rate on tongue position and velocity of movement in vowel production. Journal of the Acoustical Society of America, 84, 901-916.

Godfrey, S. (1974). Respiratory sensation and respiratory muscle activity. In B. Wyke (Ed.) Ventilatory and Phonatory Control Systems: An International Symposium (pp. 167-177). London: Oxford University Press.

Gracco, V. (1988). Timing factors in the coordination of speech movements. Journal of Neuroscience, 8, 4628-4646.

Gracco, V.L., & Abbs, J.H. (1986). Variant and invariant characteristics of speech movements. Experimental Brain Research, 65, 156-166.

Gracco, V.L., & Abbs, J.H. (1988). Central patterning of speech movements. Experimental Brain Research, 71, 515-526.

Gracco, V.L., & Abbs, J.H. (1989). Sensorimotor characteristics of speech motor sequences. Experimental Brain Research, 75, 586-598.

Greene, M.C.L. (1972). The voice and its disorders. Philadelphia: J.B. Lippincott Co.

Hanson, D.G., Gerratt, B.R., & Berke, G.S. (1990). Frequency, intensity and target matching effects on photoglottographic measures of open quotient and speed quotient. Journal of Speech and Hearing Research, 33, 45-50.

Hixon, T.J. (1973). Respiratory function in speech. In F. Minifie, T. Hixon, & F. Williams (Eds.), Normal aspects of speech, hearing and language (pp. 73-125). Englewood Cliffs, NJ: Prentice-Hall.

Hixon, T.J., Goldman, M.D., & Mead, J. (1973). Kinematics of the chest wall during speech production: Volume displacements of the rib cage, abdomen, and lung. Journal of Speech and Hearing Research, 16, 78-115.

Hixon, T.J., Klatt, D.H., & Mead, J. (1971). Influence of forced transglottal pressure on fundamental frequency. Journal of the Acoustical Society of America, 49, 105 (A).

Hixon, T.J., & Putnam, A.B. (1983). Voice abnormalities in relation to respiratory kinematics. Seminars in Speech and Language, 5, 217-231.

Hixon, T.J., Putnam, A.H., & Sharp, J.T. (1983). Speech production with flaccid paralysis of the rib cage, diaphragm, and abdomen. Journal of Speech and Hearing Disorders, 48, 315-327.

Hoit, J.D., Solomon, N.P., & Hixon, T.J. (1993). Effect of lung volume on voice onset time. Journal of Speech and Hearing Research, 36, 516-521.

Isshiki, N. (1964). Regulatory mechanism of voice intensity variations. Journal of Speech and Hearing Research, 7, 17-29.

Lieberman, P., Knudson, R., & Mead, J. (1969). Determination of the rate of change of fundamental frequency with respect to subglottal air pressure during sustained phonation. Journal of the Acoustical Society of America, 45, 1537-1543.

McFarland, D., & Smith, A. (1992). Effects of vocal task and respiratory phase on prephonatory chest wall movements. Journal of Speech and Hearing Research, 35, 971-982.

Netsell, R. (1969). Subglottal and intraoral air pressures during the intervocalic contrast of /t/ and /d/. Phonetica, 20, 68-73.

Newsom Davis, J.N. (1970). Supraspinal control. In E. J. M. Campbell, E. Agostoni, & J. Newsom Davis (Eds.), The Respiratory Muscles: Mechanics and Neural Control (pp. 234-270). Philadelphia, PA: W. B. Saunders Co.

Orlikoff, R.F., & Kahane, J.C. (1991). Influence of mean sound pressure level on jitter and shimmer measures. Journal of Voice, 5, 113-119.

Ostry, D.J., & Munhall, K.G. (1985). Control of rate and duration of speech movements. Journal of the Acoustical Society of America, 77, 640-648.

Putnam, A.H., & Hixon, T.J. (1983). Respiratory kinematics in speakers with motor neuron disease. In M. McNeil, J. Rosenbek, & A. Aronson (Eds.), The dysarthrias (pp. 37-67). San Diego, CA: College-Hill Press.

Ramig, L.O., Countryman, S., Thompson, L.L., & Horii, Y. (1995). A comparison of two forms of intensive speech treatment for Parkinson disease. Journal of Speech and Hearing Research, 38, 1232-1251.

Russell, N.K., & Stathopoulos, E.T. (1988). Lung volume changes in children and adults during speech production. Journal of Speech and Hearing Research, 31, 146-155.

Sant'Ambrogio, G., Mathew, O.P., Fisher, J.T., & Sant'Ambrogio, F.B. (1983). Laryngeal receptors responding to transmural pressure, airflow and local muscle activity. Respiratory Physiology, 54, 317-330.

Sapienza, C.M., & Stathopoulos, E.T. (1994). Respiratory and laryngeal measures of children and women with bilateral vocal fold nodules. Journal of Speech and Hearing Research, 37, 1229-1243.

Scherer, R.C. (1991). Physiology of phonation: A review of basic mechanics. In C.N. Ford & D.M. Bless (Eds.), Phonosurgery: Assessment and surgical management of voice disorders (pp. 77-93). New York: Raven Press Ltd.

Schulman, R. (1989). Articulatory dynamics of loud and normal speech. Journal of the Acoustical Society of America, 85, 295-312.

Shipp, T. (1973). Intraoral air pressure and lip occlusion in midvocalic stop consonant production. Journal of Phonetics, 1, 167-170.

Smith, A., & Denny, M. (1990). High-frequency oscillations as indicators of neural control mechanisms in human respiration, mastication and speech. Journal of Neurophysiology, 63, 745-758.

Solomon, N.P., & Hixon, T.J. (1993). Speech breathing in Parkinson's disease. Journal of Speech and Hearing Research, 36, 294-310.

Sperry, E.E., Hillman, R.E., & Perkell, J.S. (1994). The use of inductance plethysmography to assess respiratory function in a patient with vocal nodules. Journal of Medical Speech-Language Pathology, 2, 137-145.

Stathopoulos, E.T., Hoit, J.D., Hixon, T.J., Watson, P.J., & Solomon, N.P. (1991). Respiratory and laryngeal function during whispering. Journal of Speech and Hearing Research, 34, 761-767.

Stathopoulos, E.T., & Sapienza, C.M. (1993). Respiratory and laryngeal function of women and men during vocal intensity variation. Journal of Speech and Hearing Research, 36, 64-75.

Titze, I.R. (1989). On the relation between subglottal pressure and fundamental frequency in phonation. Journal of the Acoustical Society of America, 85, 901-906.

Titze, I.R. (1994). Principles of voice production. San Diego, CA: College-Hill Press.

Warren, D.W., Dalston, R.M., Morr, K.E., Hairfield, W.M., & Smith, L.R. (1989). The speech regulating system: Temporal and aerodynamic responses to velopharyngeal inadequacy. Journal of Speech and Hearing Research, 32, 566-575.

Widdicombe, J.G. (1974). Pulmonary reflex mechanisms in ventilatory regulation. In B. Wyke (Ed.) Ventilatory and Phonatory Control Systems: An International Symposium (pp. 131-143). London: Oxford University Press.

Williams, W.N., Brown, W.S., & Turner, G.E. (1987). Intraoral air pressure discrimination by normal-speaking subjects. Folia Phonatrica, 39, 196-203.

Wilson, D.K. (1987). Voice problems in children. Baltimore: Williams and Wilkins.

Winkworth, A.L., Davis, P.J., Adams, R.D., & Ellis, E. (1995). Breathing patterns during spontaneous speech. Journal of Speech and Hearing Research, 38, 124-144.

Winkworth, A.L., Davis, P.J., Ellis, E., & Adams, R.D. (1994). Variability and consistency in speech breathing during reading: lung volumes, speech intensity, and linguistic factors. Journal of Speech and Hearing Research, 37, 535-556.

NCVS Status and Progress Report - 10
November 1996, 69-77

Speech Characteristics Associated with Aging and Idiopathic Parkinson Disease in Men and Women

Cynthia M. Fox, M.A., CCC-SLP
National Center for Neurogenic Communication Disorders, The University of Arizona at Tucson
Lorraine Olson Ramig, Ph.D.
Department of Communication Disorders and Speech Science, The University of Colorado at Boulder
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts

Abstract

This study examined sound pressure level (SPL), self-perception of speech and voice, duration of maximum sustained vowel phonation, and sex differences that may exist among these variables in men and women with idiopathic Parkinson disease as compared to healthy individuals. Thirty subjects with Parkinson disease (PD; 15 men, 15 women) and 14 healthy comparison subjects (HC; 7 men, 7 women) participated in the study. Measures included SPL in a variety of speaking tasks, duration of maximum sustained vowel phonation, and self-ratings of perceptual characteristics pertaining to speech and voice. Variability was examined by having subjects repeat the data collection procedure on three different days. Results revealed that subjects with PD were 2.0 - 4.0 dB SPL lower than HC subjects. No significant differences for tasks, sex, and day-to-day variability were identified for SPL, or for duration of maximum sustained vowel phonation. Self-ratings of speech and voice characteristics revealed that men and women with PD rated themselves more severely impaired than HC men and women on variables such as loudness, hoarseness, and confidence in the voice. These results support the commonly reported perceptual characteristic of reduced vocal loudness in people with Parkinson disease with a related acoustical measure and provide an initial probe into self perception of speech and voice in this population.

Introduction

Parkinson disease is a degenerative condition resulting from a nigrostriatal dopamine deficiency (Hornykiewicz, 1966; Hornykiewicz & Kish, 1986). Approximately 75% of people with Parkinson disease have speech and voice characteristics that affect their communication abilities (Canter, 1965; Hartelius & Svensson, 1994; Logemann, Fisher, Boshes, and Blonsky, 1978; Ramig, Bonitati, Lemke, & Horii, 1994), including perceptual characteristics of reduced loudness, reduced pitch variability, imprecise articulation, and rate disturbances (Critchley, 1981; Darley, Aronson, & Brown, 1969a, 1969b). While these perceptual characteristics have been well documented, the description of speech and voice in this population remains incomplete. Examination of the existing speech and voice literature reveals differences between reported perceptual characteristics and related acoustical measures of people with Parkinson disease (Boshes, 1966; Canter, 1963; Logemann et al., 1978; Metter & Hanson, 1986). In addition, important descriptive variables, such as self-perception of speech and voice, and sex-related differences have not been considered. The purpose of this study is to address some of the incomplete aspects of the descriptive literature in order to provide a more comprehensive understanding of speech and voice characteristics in people with Parkinson disease.

Reduced vocal loudness is a commonly reported perceptual characteristic in people with Parkinson disease (Critchley, 1981; Logemann et al., 1978). SPL is a measure that closely relates to vocal loudness and has been used as its acoustic correlate (Boshes, 1966; Canter, 1963; Canter, 1965; Ludlow & Bassich, 1984; Metter & Hanson, 1984); however, studies of SPL have not reported significant differences between subjects with Parkinson disease and healthy comparison subjects at the normal loudness level (Boshes, 1966; Canter, 1963; Canter, 1965; Metter & Hanson, 1986). The lack of group differences for SPL is puzzling given the commonly reported characteristic of reduced loudness in people with Parkinson disease. Variability among study

designs, such as different sample sizes, population comparisons, or tasks used for speech samples, may be contributing to these discrepant results.

Perceptual studies of speech and voice characteristics of people with Parkinson disease have included large sample sizes (Darley et al., 1969; Logemann et al., 1978), but healthy comparison subjects were not used. Acoustical studies looking at SPL during speaking tasks have included comparison subjects, but smaller sample sizes were used (Boshes, 1966; Canter, 1963; Canter, 1965; Metter & Hanson, 1986). Lack of a comparison group in perceptual studies suggests that judges rating speech samples were aware of the subject's condition of Parkinson disease and may have been biased in their ratings. In contrast, SPL studies have included healthy comparison subjects, and no differences between subject groups were found.

Obtaining large sample sizes of a clinical and comparison population is difficult and expensive. The use of multiple data collection sessions on subjects has been suggested as an efficient and cost-effective alternative to increasing sample size (King, Ramig, Lemke, Horii, 1994). This repeated measures approach not only increases sampling, it also increases control over subject variability (Keppel, 1991). People with Parkinson disease have been reported to be variable in their speech and voice performance ability, especially when tasks are effort-dependent (Kent, Kent, & Rosenbek, 1987) and conducted in clinical testing situations (Weismer, 1984). In addition, King et al. (1994) documented a learning effect in people with Parkinson disease resulting in improved performance of speaking and voice tasks from the voice assessment experience alone. Examining day-to-day performance could help identify the nature of variability exhibited by people with Parkinson disease and the amount of learning effect associated with repeated task assessment.

Tasks used for speech samples may be another factor contributing to discrepancies between reports of reduced vocal loudness and SPL in people with Parkinson disease. Conversation or a combination of reading and conversation were used for speech samples in perceptual studies (Darley et al., 1969; Logemann et al., 1978), while reading and monosyllables were used for SPL studies (Boshes, 1966; Canter, 1963; Canter, 1965; Metter & Hanson, 1986). Examining a variety of tasks ranging from structured to spontaneous speaking would elicit different speech and voice abilities and provide information about how these different task demands affect SPL.

Little information exists on the self-perception of speech and voice characteristics in people with Parkinson disease. Sensory deficits have been reported in people with Parkinson disease in relation to bradykinesia, or slowness of movement, which is one of the four primary signs of Parkinson disease (Barbeau, 1986). This slowness of movement has been hypothesized to be the result of either a delay in the integration of sensory feedback or an inability to process sensory input which results in a decreased amplitude of motor output (Barbeau, 1986; Schneider, Diamond, & Markham, 1986; Schneider, Diamond, Markham, 1987). Barbeau (1986) contended that this deficit was evidenced in the loss of loudness, pitch, and intonation of the voice as well as other motor functions in people with Parkinson disease. Given that sensory feedback may be impaired in people with Parkinson disease, it is of great interest to assess self-perception in this population.

Sex differences in areas other than speech and voice have been reported in people with Parkinson disease. For example, a study of motor activity identified that men with Parkinson disease demonstrated a significantly lower amount of motor activity as compared to women with Parkinson disease and healthy subjects (Van Hilten, Hoogland, van der Velde, van Dijk, Kerkhof, and Roos, 1993). In a study by Pantelatos and Fornadi (1993), clinical features and medical treatment associated with age of onset of Parkinson disease were examined. Results indicated that women with young onset Parkinson disease had a significantly longer duration of the disease and that they were started on L-dopa treatment later than men.

The literature reveals that speech and voice data on people with Parkinson disease have given little attention to sex differences, and that existing speech and voice data are primarily from men (Canter, 1965; Forrest, Weismer, Turner, 1989; Kent et al., 1994; Ramig, et al., 1994; Metter & Hanson, 1986; Solomon & Hixon, 1993). A few studies of speech and voice characteristics of people with Parkinson disease have described observed sex differences. Kent et al. (1994) reported that laryngeal phonetic function for speech intelligibility appeared to be more disrupted in men as compared to women with Parkinson disease. In a study of spectrographic analysis of vowels by Hertrich and Ackermann (1995) it was reported that women with Parkinson disease demonstrated increased amounts of subharmonic energy in vowel production and more abrupt changes in fundamental frequency as compared to men with Parkinson disease.

Given that sex differences have been identified in various aspects of Parkinson disease and in some speech and voice characteristics, examination of sex differences is essential for understanding it's role in speech and voice data of people with Parkinson disease.

The present study was designed to examine SPL and self-rated perceptual speech and voice characteristics of men and women with Parkinson disease as compared to healthy men and women. Variables examined included SPL in various speaking tasks, duration of maximum sustained vowel phonation, and self-rated perceptual characteristics pertaining to speech and voice. Each subject repeated the data collection procedure on three different days within a four day period. Group, sex, and day-to-day differences were examined for all tasks.

Methods

Subjects

Forty-four subjects volunteered to participate in this study. Thirty subjects with idiopathic Parkinson disease (PD; 15 men, 15 women) and 14 healthy comparison subjects (HC; 7 men, 7 women) were included. Mean ages of the men and women with PD were 72.5 years (sd=8.7) and 66.7 years (sd=11.2), respectively. Mean ages of the men and women in the HC group were 71.7 years (sd=7.5) and 67.9 (sd=7.5) years, respectively. A one-way analysis of variance revealed no significant differences (p .05) in age among the four groups (F (3, 40) = 1.16, p = 0.335).

Subjects with PD were examined on additional variables, such as time post diagnosis and stage of Parkinson disease (Hoehn & Yahr, 1967). Subjects ranged in time post diagnosis from 1.5 to 20 years with a mean of 8.0 years (sd=4.9) for men, and 6 months to 19 years with a mean of 7.0 years (sd=6.3) for women. An independent groups t-test revealed no significant difference (p .05) between men and women with PD for the variable time post diagnosis (t (28) = -0.50, p = -0.62). Stage of disease ranged from 2.0 to 5.0 with a mean of 2.85 (sd=1.0) for men and 1.0 to 4.0 with a mean of 2.40 (sd=0.8) for women. Stage of disease was available for 10 men and 10 women with PD, thus, significance testing was based upon n=20. An independent groups t-test revealed no significant difference (p .05) between men and women with PD for the variable stage of disease (t (18) = -1.11, p = 0.28).

All subjects with PD were taking anti-Parkinson medications at the time of data collection. Subjects did not change medications during this period. Subjects were not always seen at the same time in their medication cycle due to the logistics of scheduling the sessions. Given that medication has been documented to have a limited affect on improving speech and voice in people with PD and significant differences in speech and voice abilities have not been reported at different times in subjects medication cycle (Hanson, Gerratt, & Ward, 1984; Larson, Ramig, & Scherer, 1988; 1994; Solomon & Hixon, 1993), this was judged not to be of great concern. The experimenters did record the time of each subjects last medication and next medication at the beginning of each session for reference in the event that large or unusual variability was observed in a subject's performance from session to session.

Healthy comparison subjects were free of any known condition that could affect their speech or voice, with the exception of one man in the comparison group who reported being an "occasional" smoker. To confirm that subjects were clear of any laryngeal pathology, videolaryngostroboscopic examinations of the larynx were conducted on all subjects by an otolaryngologist prior to the subject's participation in the study.

Procedures and Equipment

Subjects participated in three data collection sessions within a 4 day period. Sessions consisted of recording a variety of speaking and voice tasks, and completion of a self-rated perceptual form.

Audio recordings were made with subjects seated in a sound-treated booth. A head-mounted microphone (AKG C410) was fitted to each subject's head with mouth-to-microphone distance of approximately 5 cm remaining constant throughout the session. A sound level meter (SLM) (Bruel & Kjär 2236) was placed 30 cm in front of the subject's lips and maintained at that distance throughout the recording session. The microphone and SLM signals were recorded onto a digital audio tape (DAT) 8-channel recorder (Sony PC-208AUC). In addition, the experimenter hand recorded the peak SPL measures that were continuously displayed at 1 sec intervals from the digital output of the SLM during all speaking and voice tasks. The same experimenter collected all the hand written SPL data.

Speaking and Voice Tasks

SPL was recorded during four speaking and voice tasks. These ranged from structured tasks, such as maximum sustained vowel phonation and reading, to less structured tasks, such as picture description and monologue. Task requirements and instructions were as follows:

Six maximum duration sustained vowel phonations were elicited, four at the beginning of the recording session and two at the end. Subjects were instructed to "take a deep breath and say 'ah' for as long as you can." A clock with a second hand was provided for the subjects to watch and each subject was encouraged to monitor his or her performance. No instructions for loudness level were given for this task.

Subjects were asked to read aloud a phonetically balanced paragraph "The Rainbow Passage" (Fairbanks, 1960) at a normal pitch and loudness level. The reading passage was in large type and placed on a music stand in front of the subjects at a distance comfortable for them to read the words.

Samples of spontaneous speech were obtained by asking the subjects to, "Give me 30 seconds of monologue on a topic of your choice." If the subject could not generate a topic, the experimenter would provide a cue, such as "Tell me about what you are doing today" or "Tell me about a memorable vacation." No instructions were given for loudness level.

Subjects were asked to describe a standard picture, the "Cookie Theft" picture (Goodglass & Kaplan, 1983). The picture was placed on a music stand in front of subjects at a distance comfortable for them to see it. Subjects were instructed to describe the picture for 30 seconds. No instructions for loudness level were given.

Perceptual Self-Rating Task

Subjects were asked to complete a perceptual self-rating scale at each of the three recording sessions. A visual analog scale (Kempster, 1984; Schiffman, Reynolds, & Young, 1981) was used to obtain subject self-ratings on nine variables related to voice (loudness, shakiness, hoarseness, monotone), speech (slur, mumble), and spoken communication (understood by others, participate in conversation, and start conversation). A complete description of this scale has been provided previously (Ramig, 1992).

Data Analysis

SPL means were calculated using the continuously hand recorded peak SPL that was displayed at 1 sec intervals from the digital output of the SLM during all speaking and voice tasks. Comparison of mean SPL measures derived from hand recorded second-to-second peak SPL with mean SPL measures derived from a custom built software program analysis of SPL (Ramig, Countryman, Thompson, & Horii, 1995) has been previously reported to be comparable (Countryman & Ramig, 1993). Because the computer program incorporates the entire contour of SPL and the hand recorded method incorporates peak SPL, the latter method generates data approximately 1 to 2 decibels greater than the computer method of analysis. Given the large sampling of SPL in this study, use of the hand recorded peak SPL at 1 sec intervals was the preferred method for deriving SPL means. Since this method of analysis was used for both subject groups, any difference between subject groups would not be attributable to the analysis method used.

The SPL means for the maximum duration sustained phonation task were derived by first calculating the mean SPL of the six maximum phonations from each recording session. The mean SPL of these six phonations was then calculated to be the overall mean SPL of the maximum sustained phonation task for each of the three recording sessions. The SPL means for the reading passage, monologue, and picture description were derived by calculating the mean SPL for each speaking task for all three recording sessions.

Duration of maximum sustained vowel phonation was analyzed using a custom-built software program employing standard procedure (Ramig et al., 1995). The mean duration was analyzed for the six maximum sustained vowel phonations elicited at each session. These data were then used to calculate an overall mean duration of the maximum sustained vowel phonation task for the three recording sessions.

Given that this was an initial probe into self-perception of speech and voice characteristics and for ease of analysis, only session 3 of the perceptual data were examined. Session 3 data were chosen because it provided the most complete data set. Standard procedure for analysis of visual analog scales was used to examine perceptual data (Boeckstyns & Backer, 1989).

Intrasubject and intrameasurer reliability were calculated for SPL, duration of maximum sustained vowel phonation, and the self-rated perceptual scales. SPL and duration of maximum sustained vowel phonation data from sessions 1, 2, and 3 were correlated, and mean difference scores were calculated for intrasubject reliability. Self-rated perceptual data from sessions 2 and 3 were correlated and mean difference scores were calculated for intrasubject reliability. Intrameasurer reliability was determined by recalculating 25% of the SPL data, remeasuring 25% of the duration of maximum sustained vowel phonation data, and

Table 1.
Mean (and standard deviation) SPL (in dB SPL at 30 cm) for men and women with Parkinson disease (PD) and healthy comparison (HC) men and women across sessions and tasks.

	Sustained Phonation			Rainbow Passage			Monologue			Picture Description		
Group	Session 1	Session 2	Session 3	Session 1	Session 2	Session 3	Session 1	Session 2	Session 3	Session 1	Session 2	Session 3
PD Men n=15	69.11 (4.55)	69.53 (4.41)	70.19 (5.45)	71.71 (3.90)	72.64 (2.92)	72.71 (4.08)	70.24 (4.29)	69.84 (3.75)	70.10 (4.52)	69.80 (5.98)	70.53 (4.52)	71.35 (4.61)
PD Women n=15	67.60 (4.15)	68.54 (4.92)	68.78 (5.45)	69.96 (3.27)	70.17 (3.18)	70.35 (3.33)	68.01 (2.80)	68.75 (3.56)	68.24 (4.05)	68.86 (4.78)	68.39 (3.56)	68.38 (4.10)
HC Men n=7	73.30 (4.55)	74.59 (5.41)	75.06 (7.05)	73.64 (3.62)	73.45 (3.18)	74.20 (3.73)	72.39 (4.79)	71.56 (4.04)	72.45 (5.50)	72.22 (3.99)	72.46 (4.48)	73.15 (4.48)
HC Women n=7	72.33 (5.14)	71.18 (5.26)	71.73 (5.98)	73.30 (1.23)	73.74 (1.13)	73.43 (1.63)	71.42 (1.74)	71.44 (1.60)	72.31 (2.40)	71.00 (2.36)	71.66 (2.42)	71.94 (2.98)

dB SPL at 30 cm
* Complete data available for all subjects except one man with Parkinson disease who did not have session 3 data.

Table 2.
F and p values for significance testing on SPL data

Between Subject Effect	df	F Value	P Value
sex	1, 4	3.93	0.1185
group	1, 4	9.79	0.0352*
sexXgroup	1, 4	2.33	0.2020
Within Subject Effect			
task	1, 4	0.65	0.4656
taskXsex	1, 4	0.99	0.3769
taskXgroup	1, 4	1.69	0.2635
taskXsexXgroup	1, 4	2.77	0.1713
session	1, 4	1.24	0.3286
sessionXsex	1, 4	0.18	0.6937
sessionXgroup	1, 4	0.63	0.4710
sessionXsexXgroup	1, 4	0.08	0.7908
taskXsession	1, 4	1.33	0.3134
taskXsessionXsex	1, 4	0.15	0.7160
taskXsessionXgroup	1, 4	0.15	0.7160
taskXsessionXgroup	1, 4	0.59	0.4854

*Significance $p \leq .05$

remeasuring 25% of the session 3 visual analog scales. Correlation coefficients and mean difference scores were calculated for all intrameasurer reliability checks.

Results

Reliability

SPL intrasubject reliability resulted in correlation coefficients that ranged from 0.82 to 0.88, and mean difference scores that ranged from 0.28 dB SPL to 0.59 dB SPL. Duration of maximum sustained vowel phonation intrasubject reliability resulted in correlation coefficients that ranged from 0.91 to 0.93 and mean difference scores that ranged from 0.05 to 0.11 sec. The results of intrasubject reliability for the self-rated perceptual scale was a correlation coefficient of 0.86, with a mean difference score of 0.24%. These measures indicated good intrasubject reliability for SPL, duration of maximum sustained vowel phonation, and self-rated perceptual data. Recalculation of SPL data resulted in a correlation coefficient of 0.99 and mean difference score of 0.04 dB SPL. Intrameasurer reliability for duration of maximum sustained vowel phonation was a correlation coefficient of 0.99 and mean difference score of 0.15 sec. Intrameasurer reliability for perceptual data was a correlation coefficient of .99 and mean difference score of 0.07%.

SPL Data

Mean SPL (and standard deviation) for all subject groups, sessions, and tasks are given in Table 1. This table illustrates differences in SPL for groups, sex, and sessions. A repeated measures analysis of variance (ANOVA) with two within-subject factors (session and task) and two between-subject factors (group and sex) was conducted on the SPL data to determine significance of any SPL differences.

Table 3.
Overall mean differences (and standard deviation) of SPL for groups with sex and session data pooled.

Group	Sustained Phonation	Rainbow Passage	Monologue	Picture Description
PD (n=30)	68.96 (4.78)	71.25 (3.56)	69.18 (3.86)	69.55 (4.66)
HC (n=14)	73.03 (5.68)	73.62 (2.67)	71.93 (3.71)	72.07 (3.61)
Group Difference	4.07	2.37	2.75	2.52

dB SPL at 30 cm

Table 4.
Mean (and standard deviation) maximum duration sustained phonation times (in seconds) for men and women with Parkinson disease (PD) and healthy comparison (HC) men and women.

Maximum Duration Sustained Phonation

Group	Session 1	Session 2	Session 3
PD Men* n=13	17.98 (7.14)	18.06 (3.77)	18.46 (5.87)
PD Women n=15	15.02 (7.91)	14.67 (6.54)	15.74 (8.60)
HC Men n=7	17.27 (6.94)	17.38 (5.57)	16.07 (5.12)
HC Women n=7	17.62 (4.74)	17.82 (3.90)	17.94 (5.01)

* Complete data available for all subjects except two men with Parkinson disease who did not have session 3 data.

Data were entered into a statistical analysis computer program, (SAS, 1995), and a Type IV SS (Sum of Squares) was used for hypothesis testing as an estimated function to correct for the unbalanced design. Results are summarized in Table 2. A significant difference was identified for the main effect of group. No other significant main or interaction effects were found.

Given that a significant group difference for SPL was identified, the group differences for overall mean SPL across tasks with sex and session data pooled are summarized in Table 3. On average, the HC subjects produced speech that was 2.00-4.00 dB SPL greater than the PD subjects. Examination of group mean differences revealed that the greatest difference in SPL between the PD subjects and the HC subjects was with maximum sustained vowel phonation, followed by monologue, picture description, and the reading passage.

Duration of Maximum Sustained Vowel Phonation

Mean (and standard deviation) duration of maximum sustained vowel phonation for all subject groups across sessions are provided in Table 4. A repeated mea-

sures ANOVA with one repeated factor (session) and two between-subject factors (group and sex) was conducted for significance testing of duration time differences. No significant differences (p .05) were identified for the main effect of group ($F(1, 38) = .28$, $p = .60$), gender ($F(1, 38) = .12$, $p = .74$), and session ($F(1, 76) = .02$, $p = .98$), or for any of the related interactions. Duration of maximum sustained vowel phonation ranged from 5.74 to 34.85 sec for subjects with PD and from 8.47 to 26.23 sec for HC subjects.

Perceptual Data

A complete data set for the visual analog scale was not available because of some subjects' inability to complete the form or an error in completing the form, such as skipping an item. Data for 12 men and 13 women with PD and 6 men and 7 women in the HC group were used. Mean ratings for groups and sexes for each of the nine perceptual variables are provided in Table 5

Given the small sample sizes, significance testing was not conducted. To analyze group differences, 95% confidence intervals were constructed for each perceptual variable for all subject groups. Group differences were determined by identifying confidence intervals that did not overlap with each other. There were no differences between the men and women within the HC group or between the men and women within the PD group. Differences were identified when the PD men and women were compared with the HC men and women. Subjects with PD rated their speech and voice characteristics more severely impaired than the healthy subjects. These differences are summarized in Table 6. Variables for which both the men and women with PD rated themselves more severely impaired from the HC subjects included "shakiness" and "hoarseness" of the voice, "slurred speech", and "not being understood by others". Overall, the PD men differed from the HC men and women on more perceptual variables than the PD women. Additional variables for which the PD men rated themselves more severely impaired from the HC subjects included "loudness" and "monotone" for the voice, "mumbled speech", "participation in conversation", and "initiation of conversation".

Table 5.
Mean ratings (and standard deviation) for session 3 perceptual variables from the visual analog scale for men and women with Parkinson disease (PD) and healthy comparison (HC) men and women.

Variable	PD Men n=12	PD Women n=13	HC Men n=6	HC Women n=7
Loudness	54.08 (17.56)	61.77 (22.32)	85.00 (12.74)	84.14 (10.56)
Shaky	66.33 (20.17)	63.54 (22.92)	91.33 (10.35)	83.43 (8.22)
Hoarse	60.92 (18.15)	55.62 (21.36)	87.83 (10.94)	82.43 (11.66)
Monotone	62.17 (20.17)	67.54 (18.40)	88.50 (12.66)	83.43 (10.97)
Slur	69.25 (16.80)	67.69 (18.87)	90.33 (8.96)	86.00 (7.51)
Mumble	62.33 (16.22)	63.62 (20.81)	84.83 (9.37)	84.71 (8.56)
Understood by others	49.50 (14.91)	63.46 (20.31)	87.67 (7.20)	80.29 (9.76)
Participate in conv.	51.92 (14.38)	66.85 (21.12)	77.67 (12.79)	76.00 (14.82)
Start conversation	48.25 (12.62)	61.38 (19.64)	71.17 (16.73)	74.86 (16.03)

Perceptual characteristics are rated on a scale of 0-100% with 0% being the most severe and 100% the least severe.

Discussion

Results of this study identified a significant group difference for SPL between subjects with PD and HC subjects. Men and women with PD were found to be 2.0 - 4.0 dB SPL lower than HC men and women. This 2.0 - 4.0 dB SPL difference between subjects with PD and HC subjects could have a considerable impact on speech intelligibility given that an increase of 1 decibel at threshold can improve speech intelligibility approximately 10% for a listener (Scharf, 1978; Speaks, Parker, Harris & Kuhl, 1972). No significant differences for sex, session, task, or related interactions for SPL were found. In addition, no significant differences for duration of maximum sustained vowel phonation were identified for group, sex, sessions, or related interactions. Examination of self-rated perceptual characteristics revealed that subjects with PD rated themselves more severely impaired than HC on perceptual variables, such as hoarseness and being understood by others. Furthermore, men with PD were found to rate their speech and voice more severely impaired from HC subjects on a greater number of variables than women with PD.

The significant group difference for SPL between subjects with PD and HC subjects identified in this study is in contrast to previous studies that compared SPL between

Table 6.
Summary of differences between men and women with Parkinson disease (PD) and healthy comparison (HC) men and women on self-perceptual ratings of variables from the visual analog scale.

PD men & HC men	PD men & HC women	PD women & HC men	PD women & HC women
Loud	Loud	Shaky	Hoarse
Shaky	Mumble	Hoarse	
Hoarse	Understood by Others	Slur	
Monotone	Participate in Conversation	Understood by Others	
Slur	Start Conversation		
Mumble			
Understood by Others			
Participate in Conversation			

Variables listed are ones that the two groups differed on based on analysis of 95% confidence intervals.

subjects with PD and HC subjects (Boshes, 1966; Canter, 1963; Canter 1965; Metter & Hanson, 1986). While previous studies did not report significant group differences for SPL, they did report trends of lower SPL in subjects with PD as compared to HC subjects. However, these trends were not reliable enough to be statistically significant. The reason this study detected a significant group difference for SPL when others did not may be due to the extensive speech and voice sampling obtained through the use of repeated data collection sessions and a variety of speech and voice tasks.

Although variability has been reported in speech and voice performance in people with PD, this study found no significant variability in day to day performance across tasks within the PD group for SPL or duration of maximum sustained vowel phonation. Unlike the findings from King et al. (1994) where subjects with PD improved their speaking and voice performance from the voice recording experience alone, subjects with PD in this study did not demonstrate a significant learning effect for task performance. Thus, the nature of variability exhibited by subjects with PD in this study could be considered to fall within the range of normal clinical variability (Kent et al, 1987).

Examining a variety of speech and voice tasks was of interest in this study to determine the effects of task demand on SPL. While there were no significant differences across tasks for either subject group, examination of group means revealed some interesting trends. The reading task produced the largest mean SPL in both groups while monologue produced the lowest SPL in the HC subjects and next to lowest in the subjects with PD. Aronson (1985) suggested that people with PD may be able to produce a louder voice on demand, but not spontaneously. This may have been reflected in the SPL levels produced for reading and monologue tasks in this study. The reading task was more of a performance type task for which the subjects with PD may have tried to meet the demands of the performance and produced a louder voice even though they were instructed to read at a normal pitch and loudness. However, the spontaneous speech required of the monologue provided no performance demand, thus, subjects produced a lower SPL.

The ability of subjects with PD to produce a louder voice in a task such as reading was further evidenced by examining the group mean differences for SPL. Of all group mean differences for tasks, reading was the smallest difference suggesting that the subjects with PD were able to more closely approximate the SPL level of HC subjects in the reading tasks. In contrast, maximum sustained vowel phonation produced the largest mean SPL difference between the two groups, and was the task with the lowest SPL for subjects with PD. Maximum sustained vowel phonation, while being structured, may not have required the same performance demand on the subjects that the reading task did.

Overall, task demand did not statistically significantly affect SPL for the two subject groups; however, variability in group mean differences was observed. The greater difference in mean SPL between the PD and HC subjects in conversation as compared to reading indicates that a task such as conversation may be more sensitive to group differences than a reading task. Thus, it is important to include conversational samples in addition to reading samples when examining vocal loudness and SPL in people with PD. This allows for sampling a range of abilities, and will provide a more accurate description of overall speech and voice abilities.

Considerable variability was observed for the duration of maximum sustained vowel phonations in the subjects with PD and HC subjects. This variability was consistent with previous reports of maximum phonation times in people with PD (King et al, 1994; Metter & Hanson, 1986).

While sex differences have been identified in some aspects of speech and voice in people with PD (Kent et al, 1994; Hertrich & Ackermann, 1995), it did not play a significant role for SPL or duration maximum sustained vowel phonation in this study. Given this lack of sex difference, one could expect SPL to be similar in both men and women with PD, assuming that other characteristics of their disease were similar.

Based on the finding of significantly reduced SPL in subjects with PD as compared to HC subjects, the results of this study support a therapy program that would target increasing vocal loudness as measured by SPL. Recently, such a speech therapy program has been developed. This program, referred to as the Lee Silverman Voice Treatment (LSVT), has been documented to be efficacious in improving perceived vocal loudness, SPL, and overall communication abilities in people with PD (Countryman, Ramig, & Pawlas, 1994; Dromey, Ramig, & Johnson, 1995; Ramig, 1992; Ramig, 1995; Ramig et al., 1994; Ramig et al., 1995). Results of this study not only support the approach, but may be useful when considering appropriateness of a patient with PD for therapy. The mean SPL levels for different tasks and sexes in people with PD and healthy individuals could be useful as reference SPL levels when completing initial evaluations. In addition, documenting reduced SPL in a potential client based on these findings may assist in making a case for insurance reimbursement.

The perceptual information from this study serves as an initial indication of differences between self-rated perception of speech and voice in men and women with PD and HC men and women. Despite possible sensory deficits in people with PD that may affect their ability to put forth the appropriate degree of motor speech output (Barbeau, 1986; Schneider et al., 1986; Schneider et al., 1987), subjects with PD in this study did perceive some deterioration of their speech and voice performance abilities. This was indicated by the more severe ratings from subjects with PD than the HC subjects on variables from the visual analog scale. In addition, the variables that the subjects with PD rated more severely, such as loudness, hoarseness, and imprecise ar-

ticulation, were similar to reports of disordered perceptual characteristics from listener-rated studies (Darley et al. 1969; Logemann et al. 1978).

An interesting finding from examination of the self-rated perceptual variables was that men with PD rated themselves more severely impaired from the HC subjects on a greater number of variables than the women with PD. One explanation for the difference between men and women with PD on their self-ratings of perceptual speech and voice characteristics may be related to different life experiences. A majority of the men with PD in this study reported being career-oriented throughout their life time and had vocally dependent jobs, such as being a lawyer, professor, or salesman, for which they relied a on speaking abilities for success. In contrast, most of the women in the study did not have a career focused life. Thus, the men with PD may have been more attuned to their vocal abilities based on their life experience, and as a result, they were more critical of deterioration. This explanation would be interesting to follow as more women who have career oriented lives become part of the PD population. If this explanation holds true, then one would expect to see the self-rated perceptions of speech and voice characteristics in women with PD to become more severe over time.

Another possible explanation for the sex differences in self-rated perceptual variables in subjects with PD may be related to tolerance of symptoms. Pantelatos and Fornadi (1993) suggested that a sex-specific threshold for tolerance of symptoms may exist, which they used to explain why some women with young onset PD were started on pharmacological treatment later than men. Perhaps, this threshold for tolerance of symptoms in women with PD also impacts their tolerance of deteriorating speech and voice characteristics. If women with PD have a higher tolerance for deteriorating speech and voice abilities as compared to men, then they may be less critical and rate speech and voice symptoms less severe.

Conclusions

Several incomplete aspects of the descriptive literature of speech and voice characteristics of people with Parkinson disease have been addressed in this study. Results revealed that as a group subjects with PD were significantly lower in SPL than HC subjects, which supports the commonly reported perceptual characteristics of reduced loudness in people with PD with a related acoustical measure. Analysis of self-rated perceptual speech and voice data revealed that subjects with PD were aware of some speech and voice deterioration as indicated by more severe ratings from subjects with PD as compared to HC subjects. Finally, sex differences did not play a significant role for any measure of SPL; yet, it was a factor for self-rated perception of speech and voice where men with PD were more critical of speech and voice characteristics than women with PD.

Acknowledgments

This research was supported, in part, by National Multipurpose Research and Training Center Grants P60 DC-00976 and DC-01409, and Research Grant RO1 DC-01150 from the National Institute on Deafness and Other Communication Disorders. Appreciation and gratitude is extended to all subjects who volunteered their time and energy to participate in this study. The authors would like to thank the following persons for their contributions to this paper: from the University of Arizona - Tucson, Dr. Jeanette Hoit; from the Gould Voice Research Center of The Denver Center for The Performing Arts, Ms. Deborah Huhn and Mr. Geron Coale.

References

Aronson, A. (1985). Clinical Voice Disorders. New York: Thieme-Stratton.

Barbeau, A. (1986). Parkinson's disease: Clinical features and etiopathology. In P.J. Vinken, G.W. Bryn, & H.L. Klawans (Eds). Handbook of Clinical Neurology, Vol. 5(49): Extrapyramidal Disorders. (pp. 87-152). Elsevier Science Publishers B.V.

Boeckstyns, M.E. & Backer, M. (1989). Reliability and validity of the evaluation of pain in patients with total knee replacement. Pain, 38(1), 29-33.

Boshes, B. (1966). Voice changes in Parkinsonism. Journal of Neurosurgery, 21, 286-288.

Canter, G.J. (1963). Speech characteristics of patients with Parkinson's disease: I. Intensity, Pitch, and Duration. Journal of Speech and Hearing Disorders, 28(3), 221-229.

Canter, G. J. (1965). Speech characteristics of patients with Parkinson's disease: II. Physiological support for speech. Journal of Speech and Hearing Disorders, 31(1), 44-49.

Countryman, S. & Ramig, L.O. (1993). Effects of intensive voice therapy on voice deficits associated with bilateral thalamotomy in Parkinson disease: A case study. Journal of Medical Speech-Language Pathology, 1(4), 233-249.

Critchley, E.M.R. (1981). Speech disorders of Parkinsonism: a review. Journal of Neurology, Neurosurgery, and Psychiatry, 44, 751-758.

Countryman, S., Ramig, L.O., & Pawlas, A. (1994). Speech and voice deficits in Parkinsonian plus syndromes: Can they be treated? Journal of Medical Speech-Language Pathology, 2(3), 211-225.

Darley, F.L., Aronson, A., & Brown, J. (1969a). Differential diagnosis patterns of dysarthria. Journal of Speech and Hearing Research, 12, 246-249.

Darley, F.L., Aronson, A.E. & Brown, J.R., (1969b). Clusters of deviant speech dimensions in the dysarthrias. Journal of Speech and Hearing Research, 12, 462-496.

Dromey, C., Ramig, L.O., & Johnson, A. (1995). Phonatory and articulatory changes associated with increased vocal intensity in Parkinson disease: A case study. Journal of Speech and Hearing Research, 38, 751-764.

Fairbanks, G. (1960). Voice and articulation drill book. New York: Harper and Brothers.

Forrest, K., Weismer, G., & Turner, G.S. (1989). Kinematic, acoustic, and perceptual analyses of connected speech produced by Parkinsonian and normal geriatric adults. Journal of the Acoustical Society of America, 85(6), 2608-2622.

Goodglass, H. & Kaplan, E., 1983. Boston Diagnostic Aphasia Examination. 2nd ed. Philadelphia, PA: Lea & Febiger.

Hanson, D.G., Gerratt, B.R., & Ward, P.H. (1984). Cinegraphic observations of laryngeal function in Parkinson's disease. Laryngoscope, 94, 348-353.

Hartelius, L. & Svensson, P. (1994). Speech and swallowing symptoms associated with Parkinson's disease and multiple Sclerosis: A survey. Folia Phoniatr Logop, 46, 9-17.

Hertrich, I. & Ackermann, H. (1995). Gender-specific vocal dysfunctions in Parkinson's disease: Electroglottographic and acoustical analyses. Ann Otol Rhinol Laryngol, 104, 197-202,

Hoehn, M. & Yahr, M. (1967). Parkinsonism: Onset, progression and mortality. Neurology, 17, 427.

Hornykiewicz, O. (1966). Metabolism of brain dopamine in human Parkinsonism: Neurochemical and clinical aspects. In E. Costa, L. Cote, & M. Yahr (Eds.), Biochemistry and pharmacology of the basal ganglia. New York: Raven Press.

Hornykiewicz, O. & Kish, S. J. (1986). Biochemical pathophysiology of Parkinson's disease. In M. D. Yahr & K. J. Bergman (Eds.), Advances in Neurology, 45, 19-34.

Kempster, G. (1984). A multidimensional analysis of vocal quality in two dysphonic groups. Unpublished dissertation, Northwestern University, Evanston.

Kent, R.D., Kent, J.F., & Rosenbek, J.C. (1987). Maximum performance tests of speech production. Journal of Speech and Hearing Disorders, 52, 367-387.

Kent, R.D., Kim, H., Weismer, G., Kent, J., Rosenbek, J.C., Brooks, B.R., & Workinger, M. (1994). Journal of Medical Speech-Language Pathology, 2(3), 157-175.

Keppel, G. (1991). Design and Analysis: A Researcher's Handbook. New Jersey: Prentice-Hall, Inc.

King, J.B., Ramig, L.O., Lemke, J.H., Horii, Y. (1994). Parkinson's disease: Longitudinal changes in acoustic parameters of phonation. Journal of Medical Speech-Language Pathology, 2(1), 29-42.

Larson, K.L., Ramig, L.O., & Scherer, R. (1988). Acoustic analysis of voice: The on-off effect of medication in Parkinson's disease A paper presented at the Clinical Dysarthria Conference (San Diego).

Larson, K.L., Ramig, L.O., & Scherer, R. (1994). Acoustic and glottographic analysis during drug-related fluctuations in Parkinson disease. Journal of Medical Speech Language Pathology, 2(3), 227-239.

Logemann, J.A., Fisher, H.B., Boshes, B. & Blonsky, E.R. (1978). Frequency and coocurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. Journal of Speech and Hearing Disorders, 43, 47-57.

Ludlow, C. & Bassich, C. (1984). Relationships between perceptual ratings and acoustic measures of hypokinetic speech. In M. McNeil, J. Rosenbek, & A. Aronson (Eds.), The Dysarthrias: Physiology, Acoustics, Perception, Management (pp. 163-195). San Diego: College-Hill.

Metter, E. J. & Hanson, W. R. (1986). Clinical and acoustical variability in hypokinetic dysarthria. Journal of Communication Disorders, 19, 347-366.

Pantelatos, A. & Fornadi, F. (1993). Clinical features and medical treatment of Parkinson's disease in patient groups selected in accordance with age at onset. Advances in Neurology, 60, 690- 697.

Ramig, L.O. (1992). The role of phonation in speech intelligibility: A review and preliminary data from patients with Parkinson's disease. In R.D. Kent (Ed.), Intelligibility in speech disorders: Theory, measurement and management (pp. 119-156). Amsterdam: John Benjamin.

Ramig, L.O. (1995). Speech therapy for patients with Parkinson's disease. In W.C. Koller and G. Paulson (Eds.), Therapy of Parkinson's disease. (pp. 539-548) New York: Marcel Dekker.

Ramig, L., Bonitati, C., Lemke, J., & Horii, Y. (1994). Voice treatment f or patients with Parkinson disease: Development of an approach and preliminary efficacy data. Journal of Medical Speech-Language Pathology, 2(3), 191-209.

Ramig, L.O., Countryman, S., Thompson, L.L., & Horii, Y. (1995). Comparison of two forms of intensive speech treatment for Parkinson disease. Journal of Speech and Hearing Research, 38, 1232-1251.

SAS Software Release 6.11 (1995). Cary, NC: SAS Institute, Inc.

Scharf, B. (1978). Loudness. In E.C. Carterette & M.P. Friedman (Eds.), Handbook of Perception, Vol. 4: Hearing (pp.187-234). New York: Academic Press.

Schiffman, S., Reynolds, M.L., & Young, F.W. (1981). Introduction to multidimensional scaling: Theory, Methods and Applications. New York: Academic Press.

Schneider, J.S., Diamond, S.G., & Markham, C.H. (1986). Deficits in orofacial sensorimotor function in Parkinson's disease. Annals of Neurology, 19(3), 275-282.

Schneider, J.S., Diamond, S.G., & Markham, C.H. (1987). Parkinson's disease: Sensory and motor problems in arms and hands. Neurology, 37, 951-956.

Solomon, N.P. & Hixon, T.J. (1993). Speech breathing in Parkinson's disease. Journal of Speech and Hearing Research, 36, 294-310.

Speaks, C., Parker, B., Harris, C., & Kuhl, P. (1978). Intelligibility of connected discourse. Journal of Speech and Hearing Research, 15(3), 590-602.

Van Hilten, J.J., Hoogland, G., van der Velde, E. A., van Dijk, J. G., Kerkhof, G. A., & Roos, R. A. C. (1993). Quantitative assessment of Parkinsonian patients by continuous wrist activity monitoring. Clinical Neuropharmacology, 16(1), 36-45.

Weismer, G. (1984). Articulatory characteristics of Parkinsonian dysarthria. In M.R. McNeil, J.C. Rosenbek & A. Aronson (Eds.), The Dysarthrias: Physiology-Acoustic-Perception-Management, (pp. 101-130) San Diego: College Hill Press.

NCVS Status and Progress Report - 10
November 1996, 79-87

Perceptual Voice and Speech Characteristics in Patients With Idiopathic Parkinson Disease

Annette Arnone Pawlas, M.A., CCC-SLP
Lorraine Olson Ramig, Ph.D., CCC-SLP
Stefanie Countryman, M.A., CCC-SLP
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts

Abstract

This study was designed to describe the frequency of occurrence and type of perceptual voice and speech characteristics in 45 patients with idiopathic Parkinson disease. Three experienced listeners rated the presence or absence of 43 voice and speech characteristics from readings of the "Rainbow Passage." Disordered voice quality was identified in 91% and disordered articulation in 56% of the patients. Sex differences did not exist for frequency of occurrence or type of voice and speech characteristics. Disordered voice characteristics, such as a hoarse/harsh/rough voice quality, were prevalent early in the disease course while disordered speech characteristics, such as imprecise articulation, appeared more frequently later in the disease course. Both the frequency of occurrence and number of disordered voice and speech characteristics generally increased as duration of the disease and UPDRS motor section score increased. The early onset and frequent occurrence of disordered voice characteristics in patients with Parkinson disease supports the need for timely referrals for speech treatment. Since an effective speech treatment for patients with Parkinson disease now exists, speech treatment referrals made at the onset of disordered voice or speech characteristics will enable patients to preserve their functional oral communication for a longer period of time.

Parkinson disease effects over one million Americans.[1] At least 70% of these individuals have voice and speech disorders[2,3] which can negatively affect their employment and quality of life.[4,5] It is not uncommon for individuals with Parkinson disease to live ten to twenty years beyond their initial diagnosis[6] with every individual eventually developing voice and speech disorders.[7] While the etiology, neurophysiology, and physical pathologies of Parkinson disease have been extensively investigated,[8-10] comprehensive studies investigating voice and speech characteristics in this population are limited.[11] Consequently, voice and speech characteristics accompanying Parkinson disease may not be identified consistently and individuals may not receive timely referrals for speech assessment and effective behavioral speech treatment.[5,12]

Despite the high incidence of voice and speech disorders in patients with Parkinson disease (i.e., at least 70%), only 3% of these individuals currently receive behavioral speech treatment.[12,13] This is unfortunate because, while pharmacological interventions do not consistently alleviate voice and speech disorders,[14,15] an effective behavioral speech treatment program for patients with idiopathic Parkinson disease (IPD) does exist.[16-22] If the frequency of occurrence and type of voice and speech characteristics in IPD were clearly documented, physicians and other health care professionals could confidently identify these characteristics and make timely referrals for behavioral speech treatment.

Clarification of voice and speech characteristics in IPD could aid in the differential diagnosis of this disease from Parkinson Plus Syndromes (PPS). Frequently, individuals in the early stages of PPS, such as Progressive Supranuclear Palsy (PSP), exhibit neurological symptoms similar to individuals with IPD.[23-25] However, voice and speech characteristics may differ. For example, individuals with IPD typically exhibit reduced volume, and a breathy, hoarse voice quality,[3,26,27] which is distinct from the "strain-strangled" voice quality heard in individuals with PSP.[28] Furthermore, voice and speech disorders in PPS typically progress more quickly than those in IPD.[29] Therefore, documenting the frequency of occurrence and type of voice and speech characteristics in IPD could help distinguish

between individuals with IPD and PPS. As a result, an earlier and more accurate diagnosis could be made, leading to prompt medical and behavioral treatments.

The existing descriptive studies of voice and speech characteristics in Parkinson disease have methodological problems which limit their usefulness.[3,26,27,30,31] Small sample size, single sex design, biased and/or untrained listeners, and ratings made in uncontrolled clinical settings have lead to inconsistent results. For example, Darley et. al.[26] reported the pitch of male patients with Parkinson disease to be *lower* than non-disordered speakers, while Canter[30] described the pitch of male patients with Parkinson disease as *higher*. Moreover, current voice and speech studies have been completed while patients were "off" their medication[3] and have not reported completed diagnostic information (i.e. stage of disease, UPDRS score) for patients when "on" their medication.[32-35] Consequently, the frequency of occurrence and type of voice and speech characteristics associated with Parkinson disease remains unclear. The purpose of this study was to describe the frequency of occurrence and type of voice and speech characteristics in a group of patients with IPD.

Methods

Subjects

Forty-five individuals (33 male and 12 female) with IPD participated in this study. The diagnosis of IPD was determined by a neurologist specializing in movement disorders. Patient characteristics of age, stage of disease,[36] duration of the disease, and score on the motor section of the UPDRS are summarized in Table 1. All patients were receiving anti-parkinson medication except for three newly diagnosed patients. All data were collected while patients were "on" their medication.

Table 1.
Group Characteristics of 45 Patients With IPD

	Males (n=33)	Females (n=12)
Age		
Mean (Standard Deviation)	65.2 (8.90)	62.3 (14.00)
Minimum	49.0	32.0
Maximum	80.0	81.0
+ Stage of Parkinson Disease (1-5)		
Mean (Standard Deviation)	2.6 (.67)	2.3 (.92)
Minimum	1.0	1.0
Maximum	4.0	4.0
Duration of Parkinson Disease (In Years)		
Mean (Standard Deviation)	7.2 (5.70)	4.3 (4.20)
Minimum	1.0	1.0
Maximum	20.0	13.0
***UPDRS Motor Section Score**		
Mean (Standard Deviation)	27.9 (12.20)	20.4 (16.90)
Minimum	2.0	1.0
Maximum	47.0	48.0

+ Hoehn and Yahr, 1967
* Note. Higher scores on the UPDRS indicate greater disability. Scores can range from 0 - 108.

Data Collection of the Voice Samples

All patients were seated in an IAC sound-treated booth with a headset microphone (AKG 410) positioned 8 cm in front of their lips. The patients were asked to read aloud the phonetically balanced "Rainbow Passage."[37] After preamplification through an ATI-1000 amplifier, the microphone signal was recorded onto a Sony Digital PC-108M (DAT) eight-channel recorder. A Bruel and Kjaer Type 2230 sound level meter was placed in the booth 50 cm from the patient's mouth. The signal from the sound level meter also was recorded onto the DAT. All data were collected by the same researcher.

Development of the "Master" Audio Tapes

A computer random number generator determined the ordering of the 45 reading samples which were subsequently dubbed onto two "master" digital audio tapes (DAT).

Since intensity levels may confound perceptual judgments of voice quality and speech intelligibility,[38] intensity levels were normalized across all reading samples during the dubbing procedure. Means, standard deviation, and ranges of sound pressure level (SPL) data (i.e., intensity) before normalization are presented in Table 2.

Voice and Speech Characteristics

The voice and speech characteristics to be rated were chosen based upon previously developed perceptual frameworks.[26,39,40] In order to obtain detailed perceptual descriptions, those characteristics reported in other neurological disorders were also included in the listening procedure.[39] Definitions of the voice and speech characteristics rated are in Appendix A.

Table 2.
Sound pressure level (SPL) data in decibels (dB) at a microphone to mouth distance of 50 cm for the "Rainbow Passage" (Fairbanks, 1960) before normalization.

	Males (n=33)	Females (n=12)
Mean	66.36	65.14
Standard Deviation	.66	.73
Minimum	59.35	61.14
Maximum	75.98	69.57

Listeners

Three speech/language pathologists certified by the American Speech Language Hearing Association and having at least 5 years of clinical experience served as expert listeners. All were females with normal hearing and unfamiliar with the subjects.

Training

Prior to completing the listening procedure, the listeners participated in a three hour training session. The goal of the training was to review the: 1) listening procedure, 2) rating form, 3) computerized scanning form (i.e. answer sheet), 4) definitions and examples of the voice and speech characteristics to be rated, and 5) operation of the equipment. Finally, a practice listening session was conducted using two patient samples not included in the data pool.

Listening Procedure for Rating Voice and Speech Characteristics

Listeners individually rated the samples while seated in a IAC sound-treated booth. Listeners were informed that they were rating voices of patients with Parkinson disease. To reduce the potential for learning effects on the ratings, the order of the master tapes was randomized across listeners. To limit fatigue, listeners were instructed to rate a maximum of 2 hours per session and up to 6 hours per week.

The tapes were played on a Technics Digital Audio Tape Deck (SV-DA10) through a Technics Stereo Integrated Amplifier (SU-V303). On the master tapes, each patient was identified by a number, age and sex. (i.e., Patient #1, age 45, male). To insure that the master tapes were played at a constant intensity level, the listeners were instructed to adjust the intensity to a comfortable level at the beginning of each listening session and to keep this intensity level fixed throughout each session. The listeners were instructed to play each reading sample as often as necessary in order to accurately rate the voice and speech characteristics.

A "master" rating form was used by the listeners to rate the presence or absence of the voice and speech characteristics. The present/absent rating paradigm is consistent with previous descriptive voice and speech studies in Parkinson disease.[3] Severity was not rated because the goal of the study was to derive frequency of occurrence and type of voice and speech characteristics, not magnitude. Sample questions from the master rating form are in Appendix B.

Using the rating and computerized scanning forms (i.e. answer sheet), the listeners rated each patient's reading sample. If a voice or speech characteristic was *disordered*, the listener rated it as present (i.e., "true) and proceeded to answer more detailed questions related to that voice or speech characteristic. If the voice or speech characteristic was not *disordered*, it was rated as absent (i.e., "false"). A

Table 3.
Intrajudge percent agreement for each listener for the voice and speech characteristics.

Voice Characteristics	Listener A	Listener B	Listener C
Pitch	63	100	88
Monotone Pitch	63	75	88
Unsteady Pitch	88	100	88
Pitch Breaks	100	100	88
Nasal Resonance	75	100	88
Voice Quality	100	100	88
Stress Patterning	71	67	88
Prosody	86	100	88
Speech Characteristics	**Listener A**	**Listener B**	**Listener C**
Articulation	86	50	100
Rate	75	67	100
Fluency	100	86	88

Table 4.
Frequency of occurrence (%) for the disordered voice and speech characteristics rated from samples of the "Rainbow Passage" (Fairbanks, 1960) normalized for intensity. n=45

Disordered Voice and Speech Characteristics	Frequency of Occurrance (%)
Voice Quality	91
Articulation	56
Pitch	53
Rate	53
Stress	53
Fluency	44
Prosody	44
Nasal Resonance	13

Table 5.
Frequency of occurrence (%) for the specific *types* of disordered voice and speech characteristics rated from samples of the "Rainbow Passage" (Fairbanks, 1960) normalized for intensity. n=45

	Disordered Types of Voice and Speech Characteristics	Frequency of Occurrence (%)
1	Hoarse/Harsh/Rough	71
2	Imprecise Articulation	53
3	Monotone Pitch	49
4	Reduced Stress	49
5	Unnatural Prosody	40
6	Breathy	40
7	Glottal Fry	36
8	Mucus/Crackle	24
9	Rapid Rate	24
10	Vocal Tremor	20
11	Audible Prolongations	20
12	Omissions Of Phonemes	20
13	Phrase Repetitions	18
14	Pitch Too Low	16
15	Hypernasal	13
16	Short Rushes Of Speech	13
17	Initial Phoneme Repetitions	13
18	Whole Word Repetitions	13
19	Pitch Too High	11
20	Pressed Voice Quality	11
21	Slow Rate	9
22	Wet/Gurgle	7
23	Strain/Strangle	7
24	Part Word Repetitions	7
25	Unsteady Quality	4
26	Variable Rate	4
27	Unsteady Pitch	4
28	Pitch Breaks	2
29	Labored Articulation	2
30	Substitutions Of Phonemes	2
31	Inappropriate Silences	2
32	Palilalia	2
33	Excessive Stress	2
34	Bizarre Prosody	2
35	Hyponasal	0

true or false answer was recorded by darkening the appropriate circle on the computerized scanning form.

Data Analysis

To determine the frequency of occurrence for the voice and speech characteristics, each characteristics was considered present when two out of the three listeners agreed.

Reliability

Due to the binary nature of the responses (i.e. 1=true, 2=false), intrajudge reliability was calculated as

Table 6.
Frequency of occurrence (%) in males and females for disordered voice and speech characteristics rated from samples of the "Rainbow Passage" (Fairbanks, 1960) normalized for intensity.

Disordered Voice and Speech Characteristics	Frequency of Occurrence (%) Males (n=33)	Females (n=12)
Voice Quality	91	92
Articulation	58	50
Pitch	58	42
Rate	58	42
Stress	58	42
Fluency	42	50
Prosody	42	50
Nasal Resonance	15	8

percent agreement (PA) for 18% of the samples. Percent agreement for each listener on the voice and speech characteristics is listed in Table 3.

Results

All results reflect ratings made from speech samples normalized for intensity.

The disordered voice and speech characteristics and their frequency of occurrence are listed in Table 4. The most frequently occurring disorders were voice quality (91%), articulation (56%), pitch (53%), rate (53%), and stress (53%).

The frequency of occurrence for the specific *types* of disordered voice and speech characteristics are listed in Table 5. The most frequently occurring *types* of voice and speech characteristics were hoarse/harsh/rough voice quality (71%), imprecise articulation (53%), monotone pitch (49%), reduced stress (49%), unnatural prosody (40%), breathy voice quality (40%), glottal fry voice quality (38%), mucus crackle voice quality (24%), rapid rate (24%), and vocal tremor (20%), audible prolongations (20%) and omissions of phonemes (20%).

The frequency of occurrence for the disordered voice and speech characteristics in males and females are reported in Table 6. Results revealed that for both sexes, disorders of voice quality and articulation occurred the most frequently.

The frequency of occurrence for the disordered voice and speech characteristics according to duration of the

Table 7.
Frequency of occurrence (%) according to duration of Parkinson disease (in years) for the disordered voice and speech characteristics rated from samples of the "Rainbow Passage" (Fairbanks, 1960) normalized for intensity.

Disordered Voice and Speech Characteristics	Duration of the Disease 1-5 years (n=23)	10+ years (n=11)
Voice Quality	91	91
Articulation	48	82
Pitch	61	46
Rate	44	82
Stress	52	73
Fluency	30	82
Prosody	44	55
Nasal Resonance	13	9

Table 8.
Frequency of occurrence (%) according to the UPDRS motor section scores on the disordered voice and speech characteristics rated from samples of the "Rainbow Passage" (Fairbanks, 1960) normalized for intensity.

Disordered Voice and Speech Characteristics	UPDRS Motor Section Score 0-16 (n=11)	17-28 (n=11)	29-35 (n=10)	36-48 (n=10)
Voice Quality	82	91	100	100
Articulation	27	46	80	70
Pitch	46	36	70	70
Rate	27	54	70	80
Stress	27	46	60	80
Fluency	27	27	50	70
Prosody	27	27	60	60
Nasal Resonance	9	18	10	10

disease are listed in Table 7. Disordered voice and speech characteristics occurred early in the disease (i.e., <5 years) with the frequency of characteristics increasing, in most cases, as the duration of the disease increased. Individuals having the disease 5 years or less primarily exhibited disordered voice characteristics, specifically voice quality, pitch, and stress. Individuals having idiopathic Parkinson disease 10 years or more primarily exhibited a disordered voice quality in conjunction with disordered speech characteristics, specifically, articulation, rate, and fluency.

The frequency of occurrence for the disordered voice and speech characteristics according to patients' UPDRS motor section scores are listed in Table 8. Voice quality remained the primary disordered voice characteristic in all score ranges. Even those individuals with "low" UPDRS motor sections scores (<28) exhibited a variety of disordered voice and speech characteristics, including disordered voice quality. Generally, the frequency of voice and speech characteristics increased as UPDRS motor section scores increased.

Discussion

This study was designed to report the frequency of occurrence and type of disordered voice and speech characteristics in 45 patients with IPD. Three expert listeners rated the presence or absence of 43 voice and speech characteristics from each patient's reading sample.

With the reading samples normalized for intensity, voice quality was identified as the primary disordered voice and speech characteristic. Even with the groups separated by gender (Table 6), duration of the disease (Table 7), and UPDRS motor section scores (Table 8), voice quality remained the single most disordered characteristic. Articulation was generally the second most disordered voice and speech characteristic. These findings are consistent with previous studies.[3,26,27]

Further examination of the frequency of occurrence for the disordered *types* of voice and speech characteristics (Table 5) revealed eight of the first twelve types to be related to voice. These *types* included the disordered voice qualities of hoarse/harsh/rough, breathy, glottal fry, mucus/crackle, and vocal tremor, as well as monotone pitch, reduced stress, and unnatural prosody. Voice characteristics observed here, which have not previously been identified to occur frequently in patients with Parkinson disease included: unnatural prosody (40%) and glottal fry (36%) and mucus/crackle voice qualities (24%). Disordered voice characteristics have been frequently reported as the initial symptom in individuals diagnosed with Parkinson disease.[41] Therefore, the significance of disordered voice characteristics in the diagnosis of Parkinson disease should not be overlooked. These observations are consistent with reports of dysarthria as one of the earliest symptoms in patients with Parkinson Plus Syndromes (PPS), such as, Progressive Supranuclear Palsy (PSP)[23,24] and Shy-Drager Syndrome (SDS).[42-44]

When examining the frequency of occurrence for the disordered voice and speech characteristics in relation to duration of the Parkinson disease (Table 7), the voice characteristics of quality (91%), pitch (61%), and stress (52%) were the most prevalent characteristics heard early in the disease course (i.e. 1-5 years). This is in contrast to the disordered speech characteristics of articulation (82%), rate (82%), and fluency (82%) primarily heard later in the disease course (i.e. >10 years). These results support Logemann's[3] conclusion that laryngeal functioning in individuals with Parkinson disease may deteriorate first followed by articulatory functioning. These findings are consistent with observations of an overall amplitude scale down of output across the speech mechanism that is initially more apparent in phonatory output.[45]

Disordered voice characteristics in patients with Parkinson disease may be overlooked because voice symtoms such as hoarseness, breathiness, tremor and reduced loudness (i.e. intensity) are similar to the voice characteristics heard in normal aging individuals.[46] However, Fox and Ramig (unpublished observations in review), recently reported that patients with Parkinson disease exhibited a statistically significant difference in intensity and perceived their communication to be more disordered when compared to normal aging individuals. We hypothesize that voice changes in patients with Parkinson disease may not be as obvious to the physician as other motor symtoms such as limb tremor and rigidity. While the medical office environment (i.e. small quiet room, good lighting, one-to-one conversation) facilitates easy communication between the patient and the physician, it may actually mask the voice symptoms of patients with Parkinson disease. Furthermore, the physician may not have known the patient premorbidly and therefore does not have a pre-Parkinson disease voice comparison. As a result, a patient's report of voice changes in the early stages of Parkinson disease may go unnoticed and not be a primary management focus. This is unfortunate, since an effective speech treatment program supports the usefulness of early intervention to maintain functional oral communication.[16-22]

The high incidence of disordered voice early in the disease course suggests that referrals for behavioral speech treatment are necessary. Even in cases where voice changes in IPD are identified, physicians may be reluctant to make referrals for speech treatment because, historically, speech treatment for patients with Parkinson disease has been ineffective.[47-49] A recently developed program, the Lee Silverman Voice Treatment (LSVT), has been scientifically proven as an effective behavioral speech treatment for patients with idiopathic Parkinson disease.[16-22] Unlike other forms of speech treatment, which focused on articulation and rate, the LSVT focuses on voice. The program uses intensive, high effort voice treatment and sensory calibration to improve oral communication in patients with Parkinson disease. Ramig and colleagues have reported improvements in intensity, intonation, and intelligibility with patients (stages I-IV)[36] who are able to maintain treatment improvements from 6 months up to 24 months post-treatment without additional speech treatment.[16]

This study represents a first attempt at a more systematic and comprehensive description of the voice and speech characteristics in IPD. While listener reliability (i.e. percent agreement) continues to be a critical issue in perceptual ratings of voice and speech,[50,51] it is important to realize that for voice quality, the primary disordered voice and speech characteristic identified in this study, intrajudge percent agreement for each listener was between 88% (one listener) and 100% (two listeners).

If patients with Parkinson disease are referred for behavioral speech treatment early in the course of their disease, they can maintain functional oral communication longer. While physicians are initially concerned with managing pharmacological treatment for limb symptoms and rigidity,[5] pharmacological treatment does not consistently or significantly improve speech production.[14,15] The results of this study suggests patients with Parkinson disease have voice quality problems early in the disease course. Referring patients for behavioral speech treatment as soon as voice or speech changes occur will only serve to enhance the overall management and ultimately, the well-being of patients with Parkinson disease. Future research will address the frequency of occurrence and type of voice and speech characteristics in an age-matched control group.

Acknowledgment

This work was supported in part by: NIH grants #RO1DC01150, OE-NIDRR #H133G40108 and the Hearst Foundation. We thank our listeners: Susan Hensley, Kathe Perez, and Christina Taskoff and our patients who participated in this study. The assistance of Dr. Christopher Dromey is gratefully acknowledged.

AppendixA

Definitions of Voice and Speech Characteristics rated by three expert listeners on 45 patients with idiopathic Parkinson Disease[26,34,50,52-55]

Voice Characteristics

Quality

Hoarse/Harsh/Rough: A rough, coarse, husky quality of the voice.

Breathy: Audible escape of air resulting in a thin, weak phonation, related to afunctional inability to firmly adduct the vocal folds.

Glottal Fry: A crackling low-pitched phonation.

Mucus/Crackle: A wet sounding voice with a "crackle" in it.

Vocal Tremor: Rhythmic alterations in pitch or loudness.

Pressed: A voice that sounds like a mild squeezing of the voice through the glottis.

Wet/Gurgle: A liquid, gurgling sounding voice.

Strain/Strangled: A voice that sounds like an extremely effortful squeezing of the voice through the glottis.

Unsteady Quality: Non-regular variation in quality.

Pitch

Monotone: Voice characterized by little or no variation of pitch or loudness; pitch range is usually restricted to one of four semitones.

Too Low: The pitch of the voice is too low for the individual's age and sex.

Too High: The pitch of the voice is too high for the individual's age and sex.

Unsteady Pitch: Non-regular variations in pitch. The pitch of the voice is not consistently maintained at one pitch.

Pitch Breaks: A sudden abnormal shift of pitch during speech. The typical pitch break is one octave higher (ascending pitch break) or one octave lower (descending pitch break) than the normal voice.

Stress Patterning

The amount of force or strength of movement in the production of one syllable as compared with another; usually result in the syllable sounding longer and louder than other syllables in the same word.

Reduced: The proper stress on usually emphasized parts of speech is reduced.

Excess: There is excess stress on usually unstressed parts of speech.

Prosody

Refers to the melodic aspects of speech that signal linguistic and emotional features. It includes stress patterning, intonation, and rate-rhythm.

Unnatural: Speech that does not conform to the listener's standards of rate rhythm, intonation and stress patterning. It also does not conform to the syntactic structure of the utterance being produced.

Bizarre: A more severe form of unnatural speech/prosody.

Nasal Resonance

Hypernasality: An excessively undesirable amount of perceived nasal cavity resonance during phonation.

Hyponasality: Lack of nasal resonance resulting from a partial or complete obstruction in the nasal tract.

Speech Characteristics

Articulation

Imprecise: The production of consonants that have slurring, reduced sharpness and crispness and are distorted.

Omissions: During speech, phonemes are omitted from words.

Labored: A slow, effortful production of speech.

Substitutions: During speech, one phoneme is replaced by another.

Rate

Rapid: Speech that is produced at a rate greater than 160-170 wpm.

Slow: Speech that is produced at a rate less than 160-170 wpm.

Variable: Speech that is produced at a rate that fluctuates between slow, normal, and/or rapid.

Short Rushes: Rushes of speech with pauses in between.

Inappropriate Silences: During connected speech, there are silences between words that are not appropriate.

Fluency

Audible Prolongations: The lengthening of a speech sound or maintaining the posture of the lips, tongue, or other parts of the speech mechanism in an attempt to modify the stuttering pattern.

Repetitions: The repeating of a initial phoneme, syllable, word, or phrase before continuing with the rest of the speech.

Palilalia: A word, phrase, or sentence repeated many times, with increasing rapidity and with less distinctness so that the latter part may become almost inaudible.

Appendix B

Sample Questions From Master Rating Form

Instructions: Listen to each patient's "Rainbow Passage" as often as you like. Using the answer sheet (i.e., bubble form) answer the following questions by filling in the appropriate circle on your answer sheet.

Pitch

1. This person's pitch is DISORDERED based on his/her age and sex.

If you circled FALSE for #1, go to question #4
If you circled TRUE for #1, continue with the following questions:

2. The pitch is TOO HIGH for this person based on his/her age and sex.

3. The pitch is TOO LOW for this person based on his/her age and sex.

4. The pitch is MONOTONE.

5. The pitch is UNSTEADY.

6. PITCH BREAKS are heard.

Voice Quality

7. The voice quality is DISORDERED.

If you circled FALSE, go to the next section labeled NASAL RESONANCE
If you circled TRUE, continue with the following questions:

8. The voice quality is BREATHY.

9. The voice quality is PRESSED.

10. The voice quality is STRAINED/STRANGLED.

11. The voice quality is HOARSE/HARSH/ROUGH.

Nasal Resonance

References

1. Lieberman AN, Gopinathan G, Neophytides A, Goldstein M. Parkinson's disease handbook. New York: American Parkinson Disease Association, 1992.

2. Atarashi J, Uchida E. A clinical study of Parkinsonism. Recent Advances in Research in the Nervous System 1959;3:871-882.

3. Logemann JA, Fisher HB, Boshes B, Blonsky ER. Frequency and concurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. J Speech Hearing Dis 1978;42:47-57.

4. Oxtoby, M. Parkinson's disease patients and their social needs. London: Parkinson's Disease Society, 1982.

5. Mutch RJ, Strucwick A, Roy S, Downie AW. Parkinson's disease: Disability, review, and management. Brit Med J 1986;293:675-677.

6. Martilla RJ, Rinne VK. Changing epidemiology of Parkinson's disease: Predicted effects of levadopa treatment. Acata Neurol Scand 1979;59:80-87.

7. Selby G. Parkinson's disease. In: Vinken PJ, Bruyn GW, editors. Handbook of clinical neurology. Vol 6. Amsterdam: North Holland Publishing Company, 1969.

8. Jellinger, K. The pathology of parkinsonism. In: Marsden CD, Fahn S, editors. Movement disorders 2, London: Butterworths, 1990:124-165.

9. Koller WC, Paulson G, editors. Therapy of Parkinson's disease. New York: Marcel Dekker, 1995.

10. Paulus W, Jellinger K. The neuropathic basis of different clinical subgroups of Parkinson's disease. J Neuropath Exp Neurol 1991;50(6):747-755.

11. Critchley EMR. Speech disorders of Parkinsonism: A review. J Neurol Neurosurg Psychiat 1981;44:751-758.

12. Hartelius L, Svensson P. Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: A survey. Folia Phoniatr Logop 1994;46:9-17.

13. Vieregg P, Dethlefsen J. Physical therapy and speech therapy in Parkinson syndrome: A status assessment. Fortschr Neurol Psychiat 1992;60:369-374.

14. Larson K, Ramig LO, Scherer R. Acoustic and glottographic voice analysis during drug related fluctuation in Parkinson disease. J Med Speech-Language Path 1994;2:227-239.

15. Leanderson R, Meyerson BA, Persson A. Lip muscle function in parkinsonian dysarthria. Acta Otolaryngol 1972;74:354-357.

16. Ramig LO, Countryman S, O'Brien C, Hoehn M, Thompson L. Intensive speech treatment for patients with Parkinson disease: Short- and long-term comparison of two techniques. Neurology. In press.

17. Ramig LO, Countryman S, Thompson L, Horii Y. Comparison of two forms of intensive speech treatment for Parkinson disease. J Speech Hearing Res 1995;38:1232-1251.

18. Dromey C, Ramig L, Johnson A. Phonatory and articulatory changes associated with increased vocal intensity in Parkinson disease: A case study. J Speech Hearing Res 1995;38:751-764.

19. Countryman S, Ramig L, Pawlas A. Speech and voice deficits in Parkinson plus syndromes: Can they be treated? J Med Speech-Language Path 1994;2:211-225.

20. Countryman S, Ramig LO. Effects of intensive voice therapy on voice deficits associated with bilateral thalamotomy in Parkinson disease: A case study. J Med Speech-Language Path 1993;1:233-249.

21. Ramig LO, Bonitati C, Lemke J, Horii Y. Voice treatment for patients with Parkinosn disease: Development of an approach and preliminary efficacy data. J Med Speech-Language Path 1994;2:191-209.

22. Ramig LO, Pawlas AA, Countryman, S. The Lee Silverman Voice Treatment: A practical guide for treating the voice and speech disorders in Parkinson disease. Iowa City (IA): University of Iowa, National Center for Voice and Speech, 1995.

23. Maher ER, Lees AJ. The clinical features and natural history of the Steele-Richardson-Olszewski syndrome (progressive supranuclear palsy). Neurology 1986;36:1005-1008.

24. Golbe LI, Davis PH, Schoenberg BS, Duvoisin RC. Prevalence and natural history of progressive supranuclear palsy. Neurology 1988;38:1031-1034.

25. Jankovic, J. The relationship between Parkinson's disease and other movement disorders. In: Calne DB, editor. Handbook of experimental pharmacology. Berlin: Springer-Verlag, 1989:227-270.

26. Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. J Speech Hearing Res 1969;12:246-269.

27. Darley FL, Aronson AE, Brown JR. Clusters of deviant speech dimensions in the dysarthrias. J Speech Hearing Res 1969;12:462-496.

28. Kluin KJ, Foster NL, Berent S, Gilman S. Perceptual analysis of speech disorders in progressive supranuclear palsy. Neurology 1993;43:563-566.

29. Wenning GK, Shlomo YB, Magalhães M, Daniel SE, Quinn NP. Clinical features and natural history of multiple system atrophy: An analysis of 100 cases. Brain 1994;117:835-845.

30. Canter GJ. Speech characteristics of patients with Parkinson's disease: Intensity, pitch and duration. J Speech 1961;28:221-229.

31. Canter GJ. Speech characteristics of patients with Parkinson's disease II: Physiological support for speech. J Speech Hearing Dis 1965;30:44-49.

32. Connor NP, Ludlow CL, Schulz GM. Stop consonant production in isolated and repeated syllables in Parkinson's disease. Neuropsychologia 1989;27:829-838.

33. Forrest K, Weismer G, Turner GS. Kinematic, acoustic, and perceptual analyses of connected speech produced by Parkinsonian and normal geriatric adults. J Acoust Soc Amer 1989;85:2608-2622.

34. Ludlow CL, Bassich CJ. Relationships between perceptual ratings and acoustic measures of hypokinetic speech. In: McNeil M, Rosenbek J, Aronson A, editors. The dysarthrias: Physiology, acoustics, perception, management. San Diego: College Hill, 1984:163-196.

35. Metter EJ, Hanson WR. Clinical and acoustic variability in hypokinetic dysarthria. J Communication Dis 1986;19:347-366.

36. Hoehn M, Yahr M. Parkinsonism: Onset, progression and mortality. Neurology 1967;19:427-442.

37. Fairbanks G. Voice and articulation drill book. New York: Harper, 1960.

38. Rostolland D. Acoustic features and shouted voice. Acoustica 1982;50:118-125.

39. Darley FL, Aronson AE, Brown JR. Motor speech disorders. Philadelphia: WB Saunders; 1975.

40. Gerratt HBR, Till JA, Rosenbek JC, Wertz RT, Boysen AE. Use and perceived value of perceptual and instrumental measures in dysarthria management. In: Moore CA, Yorkston KM, Teuhelman DR, editors. Dysarthria and apraxia of speech. Baltimore: Paul H. Brookes, 1991:77-93.

41. Aronson AE. Clinical voice disorders. New York:, Thieme, 1990.

42. Bassich CJ, Ludlow, CL, Polinsky RJ. Speech symptoms associated with early signs of Shy-Drager syndrome. J Neurol Neurosurg Psychiat 1984;47:995-1001.

43. Bawa R, Ramadan HH, Wetmore SJ. Bilateral vocal cord paralysis with Shy-Drager syndrome. Otolaryngol 1993;109:911-914.

44. Kew J, Gross M, Chapman P. Shy-Drager syndrome presenting as isolated paralysis of vocal cord abductors. Brit Med J 1990;300:1441.

45. Muller F, Stelmach GE. Scaling problems in Parkinson's disease. In: Requin J, Stelmach GE, editors. Tutorials in motor neuroscience. Netherlands: Kluwer Academic Publishers, 1991:161-174.

46. Linville SE. The sound of senescence. J Voice 1996;10:190-200.

47. Allan CM. Treatment of non-fluent speech resulting from neurological disease: Treatment of dysarthria. Brit J Disord Comm 1970;5:3-5.

48. Green MCL. The voice and its disorders. London: Pitman Medical, 1980.

49. Sarno MT. Speech impairment in Parkinson's disease. Arch Phys Med Rehabil 1968;49:269-275.

50. Bassich CJ, Ludlow CL. The use of perceptual methods by new clinicians for assessing voice quality. J Speech Hearing Dis 1986;51:125-133.

51. Kreiman J, Gerratt B, Kempster G, Erman A, Berke GS. Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. J Speech Hearing Res 1993;36:21-40.

52. Prater RJ, Swift RW. Manual of voice therapy. Boston: Little, Brown and Company, 1984.

53. Nicolosi L, Harryman E, Kreshech J. Terminology of communication disorders speech language hearing. Baltimore: Williams & Wilkins, 1978.

54. Yorkston KM, Beukelman DR, Bell KR. Clinical management of dysarthric speakers. Boston: Little, Brown & Company, 1988.

55. Critchley M. On Palilalia. J Neurol Psychopath 1927;8:23-31.

NCVS Status and Progress Report - 10
November 1996. 89-94

Modelling Biphonation — The Role of the Vocal Tract

Patrick Mergell, M.S.
Ear, Nose and Throat Clinics, The University Erlangen/Nurnberg, Erlangen, Germany
Hanspeter Herzel, Ph.D.
The Institute for Theoretical Biology, Humboldt University, Berlin, Germany

Abstract

Instabilities of the human voice source appear in normal voices under certain conditions (newborn cries, vocal fry, creaky voice) and are symptomatic of voice pathologies. Vocal instabilities are intimately related to bifurcations of the underlying nonlinear dynamical system. We analyse in this paper bifurcations in 2-mass models of the vocal folds and study, in particular, how the incorporation of the vocal tract effects bifurcation diagrams. A comparison of a simplified model [Steinecke & Herzel 1995] with an extended version including vocal tract resonances reveals that essential features of the bifurcation diagrams (as e.g. frequency locking of both folds and toroidal oscillations) are found in both model versions. However, vocal instabilities appear in the extended model at lower subglottal pressures and even for weak asymmetries.

Introduction

Under normal conditions the voice source can be regarded approximately as a periodic excitation — a limit cycle. However, under certain conditions various gross voice instabilities are observed. Examples are found in newborn cries [Sirvio & Michelsson 1976, Mende et al. 1990], non-cry vocalisations of infants [Robb & Saxman 1988], Russian lament [Mazo et al 1995], and also in normal conversational speech [Dolansky & Tjernlund 1968, Kohler 1996]. In particular, vocal fold lesions, paralysis, and other pathological conditions may induce subharmonic vocalisation, biphonation (two independent pitches), and deterministic chaos [Herzel & Wendler 1991, Herzel et al. 1994].

The theory of nonlinear dynamics provides the appropriate framework for these instabilities [Herzel 1993; Titze et al. 1993]. Stationary signals can be related to attractors (steady state, limit cycle, torus, chaotic attractor) and qualitative changes due to parameter variations can be classified as bifurcations (Hopf bifurcation, period-doubling, ...). Recommendable introductions to this concept are Berge et al. 1983, Glass and Mackey 1988, or Kaplan and Glass 1994.

Attractors and bifurcations have been analysed also in aerodynamical-biomechanical models of the voice source [Herzel et al. 1991, Berry et al. 1994, Steinecke & Herzel 1995]. In this paper we focus on modelling biphonation using asymmetric two-mass models of the vocal folds.

The simultaneous appearance of two pitches (biphonation) has been reported in newborn cries [Sirvio & Michelsson 1976], non-cry vocalisations of infants [Robb & Saxman 1988], in a child's voice [Herzel & Reuter 1996b] (see Fig. 1), in excised larynx experiments [Berry et al. 1996], and in intense high-pitched vocalizations of young woman [Herzel & Reuter 1996a, Tigges et al. 1996].

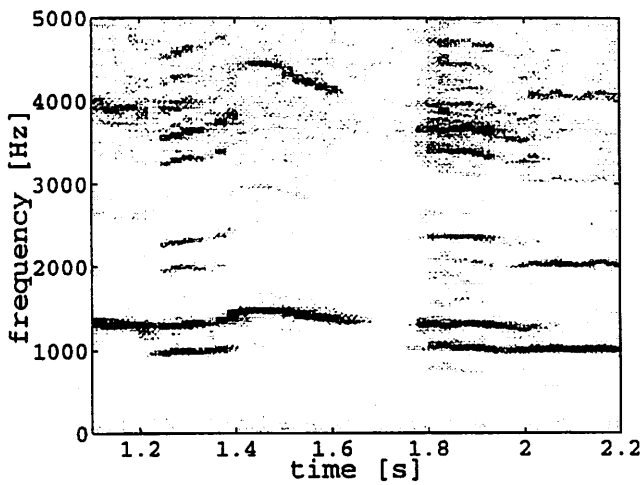

Figure 1. Biphonation in a child's voice. A nine year old boy with a healthy voice was able to phonate above 1000 Hz in a whistle register (see [Herzel & Reuter 1996b] for details). In the transition region between whistle register and and falsetto biphonic episodes were found (1.2-1.4 s and 1.75-2.05 s). The spectrogram is based on short-term spectra from segments of 1024 points (about 50 ms) using Hanning windows and a shift of 512 data points. A boost of 10 dB per decade is applied to enhance higher harmonics.

In one case of a normal healthy female voice a thorough analysis of biphonation with high speed glottography was possible. It turned out that there was a glottal gap during biphonic phonation and pronounced left-right asymmetry of the folds. Spectral analysis of the amplitudes of both folds revealed that they were vibrating with different frequencies $f_{left} \approx 820$ Hz and $f_{right} \approx 680$ Hz. Moreover, there was a strong modulation of the signal with the beat frequency $f_{left} - f_{right}$.

These observations gave us essential clues for modelling biphonation with asymmetric two-mass models. We will show in the following that weak asymmetry of the folds, a glottal gap, and a pitch near vocal tract resonances leads to biphonation comparable to the observations.

The Simplified 2-Mass Model

Details of the model approach can be found elsewhere [Ishizaka & Flanagan 1972, Steinecke & Herzel 1995]. The elongations of the lower mass x_1 and the upper mass x_2 are governed by the usual mechanical equations of coupled oscillators:

$$m_i\ddot{x}_i + r_i\dot{x}_i + k_i x_i + \Theta(-a_i)c_i\left(\frac{a_i}{2l}\right) + k_c(x_i - x_j) = F_i(x_1, x_2). \quad (1)$$

The Θ-function (the unit step function) is related to an additional restoring force during closure of the glottis ($a_i = a_{i0} + 2lx_i < 0$; a_{i0}: rest area; l: length of the glottis). Nonlinearities of the elastic forces have been neglected.

The driving forces F_i can be derived as follows: We assume constant pressure below the glottis (termed subglottal pressure P_s) and above the glottis (vocal tract input pressure P_T=0). Moreover we assume, that at the point of minimum area a_{min} a jet is formed which induces an immediate pressure decay to zero. Consequently, the driving force of the upper mass F_2 is identically zero for all glottal configurations.

$F_1 = l_{d1}P_1$ (d_1: thickness of the lower mass) is the force exerted by the pressure P_1 on the lower part of the glottis. The corresponding pressure P_1 can be obtained from the Bernoulli law:

$$P_s = \frac{\rho}{2}\left(\frac{U}{a_{min}}\right)^2 = P_1 + \frac{\rho}{2}\left(\frac{U}{a_1}\right)^2. \quad (2)$$

Here ρ and U denote the air density and the glottal volume flow, respectively. Using

$$U = \sqrt{\frac{2P_s}{\rho}}a_{min}\,\Theta(a_{min}) \quad (3)$$

one obtains

$$P_1 = P_s\left[1 - \Theta(a_{min})\left(\frac{a_{min}}{a_1}\right)^2\right]\Theta(a_1). \quad (4)$$

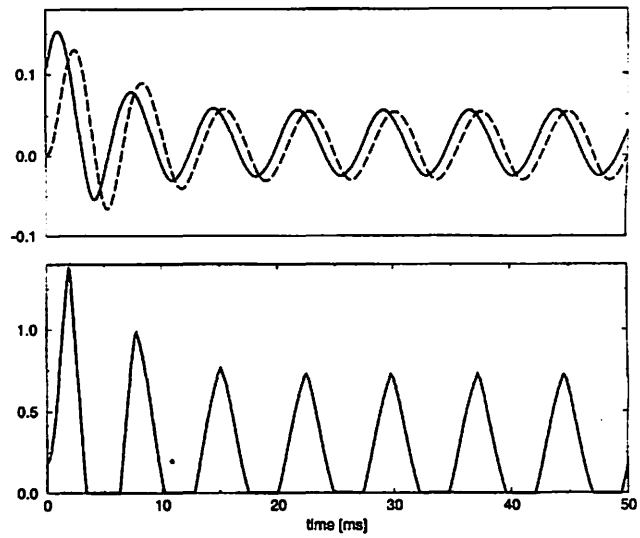

Figure 2. Simulation of normal phonation with the simplified model. Upper graph: Elongations of the lower mass (solid line) and the upper mass (dashed). Lower graph: Glottal volume flow.

Newtons equations (1) and the algebraic pressure equation (4) constitute the simplified model version. It has been shown [Steinecke & Herzel 1995] that despite the drastic simplifications the waveforms, phonation threshold, and parameter dependences appear realistic. Along the lines of Ishizaka and Flanagan 1972 we choose the following parameters to model a normal voice.

$$\begin{array}{ll} m_1=0.125 & c_1=3k_1 \\ m_2=0.025 & c_2=3k_2 \\ r_1=0.02 & r_2=0.02 \\ d_1=0.25 & a_{01}=0.05 \\ d_2=0.05 & a_{02}=0.05 \\ k_1=0.08 & k_2=0.008 \\ k_c=0.025 & l=1.4 \\ P_s=0.008 & \rho=0.00113 \end{array} \quad (5)$$

All units are given in cm, g, ms and their corresponding combinations.

Figure 2 displays the amplitudes of the two masses and the corresponding glottal volume flow. One can see the characteristic phase shift between upper and lower mass. Since the glottal flow is proportional to the minimal area (compare Eq. (3)) no skewness of the air pulses is observed.

Bifurcations in the symmetric model version have been studied in [Herzel & Knudsen 1995]. As an attempt to model high-pitched phonation with incomplete closure we rescaled stiffness coefficients and masses by a factor of seven, increased the glottal rest areas, and decreased the damping coefficients. Moreover, moderate asymmetry of the lower mass pair is modelled using an asymmetry factor Q (see [Ishizaka & Isshiki 1976, Smith et al. 1993, Steinecke & Herzel 1995] for a detailed discussion of laryngeal

asymmetries). The modified parameter values are as follows (the subscripts r and l refer to the right and left fold, respectively):

$$\begin{aligned} k_{1r}&=Q\ k_{1l}=Q\ 0.56\\ k_{2r}&=k_{2l}=0.056\\ m_{1r}&=\tfrac{m_{1l}}{Q}=\tfrac{0.018}{Q}\\ m_{2r}&=m_{2l}=0.0037\\ r_{1r}&=r_{1l}=0.005\\ r_{2r}&=r_{2l}=0.005\\ a_{01r}&=a_{01l}=0.08\\ a_{02r}&=a_{02l}=0.08 \end{aligned} \qquad (6)$$

Since the eigenfrequencies of oscillators are proportional to $\sqrt{\frac{k}{m}}$ the above rescaling from (5) to (6) increases the pitch almost by a factor of seven. Due to the asymmetry factor Q the eigenfrequencies are detuned. Bifurcations due to a varying parameter Q have been discussed elsewhere [Tigges et al. 1996]. Below we fix Q at 0.74 or 0.8 and vary the subglottal pressure.

A Simple Model of Source-Tract Coupling

So far, we have assumed vanishing vocal tract input pressure P_T. A generalisation of the Bernoulli law (2) reads:

$$P_s = P_1 + \frac{\rho}{2}\left(\frac{U}{a_1}\right)^2 = P_T + \frac{\rho}{2}\left(\frac{U}{a_{min}}\right)^2 \quad . \qquad (7)$$

In a one tube approximation the vocal tract is represented by a lumped inertance I_T and a reflection coefficient at the tube outlet for modelling the mouth. We apply the wave reflection model to describe the supraglottal pressure in a coherent way

$$P_T = I_T\frac{dU}{dt} = P_T^+ + P_T^- \quad , \qquad (8)$$

where $P_T^{\pm}$ are the wave functions for the upward (+) and downward (-) propagating pressure wave. The inertance I_T is inversely proportional to the cross section area of the tube A_T. Integration of this equation with the initial condition $P_T^-(0)=0$ leads to the following expressions (a detailed derivation will be published elsewhere):

$$\begin{aligned} P_T^-(t) &= \frac{r_M\rho c}{A_T}U(t-T_T)\Theta(t-T_T) \quad ,\\ P_T^+(t) &= \frac{\rho c}{A_T}[U(t)+r_M U(t-T_T)\Theta(t-T_T)] \quad , \end{aligned} \qquad (9)$$

where r_M denotes the mouth reflection coefficient. The reflected wave P_T^- becomes nonzero after a tract cycle $T_T=\frac{2L}{c}$ where L and c denote the vocal tract length and sound velocity, respectively. For $t > T_T$ one obtains

$$U(t) = \left[-c\left(\frac{a_{min}(t)}{A_T}\right) + \sqrt{c^2\left(\frac{a_{min}(t)}{A_T}\right)^2 + \frac{2P_s}{\rho} - \frac{4r_M c}{A_T}U(t-T_T)}\right] a_{min}(t) \qquad (10)$$

We state here, that the area A_T controls the coupling strength between source and resonator. It is clearly visible in Eq.(10) that a constricted tract above the glottis is closely related to a strong interaction between source and resonator. This agrees with the results of the studies by Titze and Story (1996) predicting a lowered phonation threshold and a strong interaction between glottal oscillations and the vocal tract resonances for the case of a constricted epilarynx. As A_T is increased the coupling strength decreases and we find in the limit

$$\begin{aligned} \lim_{A_T\to\infty} P_T &= 0 \quad ,\\ \lim_{A_T\to\infty} U &= \sqrt{\frac{2P_s}{\rho}}a_m \quad , \end{aligned} \qquad (11)$$

which corresponds to the pressure and flow equations in the simplified 2-mass model. Furthermore, the memory effect of the tract merges to the surface in form of the Eq.(10). The self-consistent flow equation relates the function with itself one tract cycle shifted in the past. Here should be mentioned that all further steps towards resonator complexity such as a two tube approximation of the tract or the coupling of the glottal system to a one tube trachea would lead to a complicated system of pressure and flow equations. Therefore, a mathematical analysis and a transparent discussion of the interaction between source and resonator would be much more difficult.

In contrast to the more complex glottal flow formula derived by Titze (1984) using the transmission line model for a coupled sytem of subglottal as well as supraglottal resonances, we assumed a constant lung pressure P_s and we could determine the pressure function P_T analytically in a one tube approximation.

Bifurcations Without Vocal Tract

The asymmetric simplified model has been analysed extensively in [Steinecke & Herzel 1995] using two-dimensional bifurcation diagrams, phase portraits, and Poincare sections. Instabilities for default parameters (5) have been found for Q<0.6 and subglottal pressure above P_s=0.013. Typically, at the borderline of normal phonation abrupt jumps to subharmonic regimes are observed.

Figure 3 shows a bifurcation diagram for the parameter configuration (6). Phonation onset is found only at about 0.014. For subglottal pressures between 0.014 and 0.025 we observe periodic vibrations of the vocal folds (a limit cycle) despite the asymmetry ($Q = 0.74$). In the range

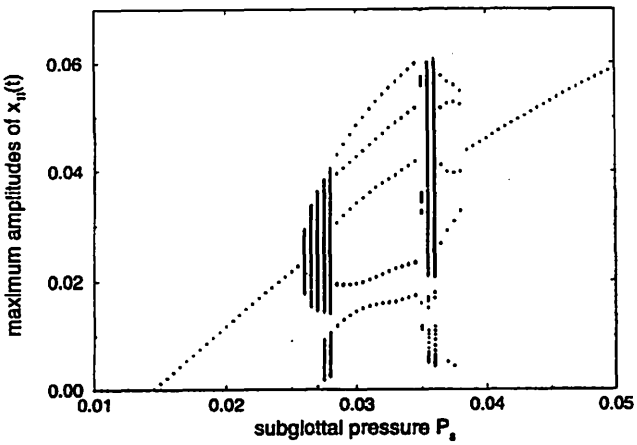

Figure 3. Bifurcations in the asymmetric 2-mass model for increasing subglottal pressure. At each parameter value an initial transient of 2 s was discarded. Then during another second the maxima of $x_{1l}(t)$ were plotted. In this way a torus (two independent frequencies) leads to a continuum of points and periodic vibrations appear as a discrete number of points.

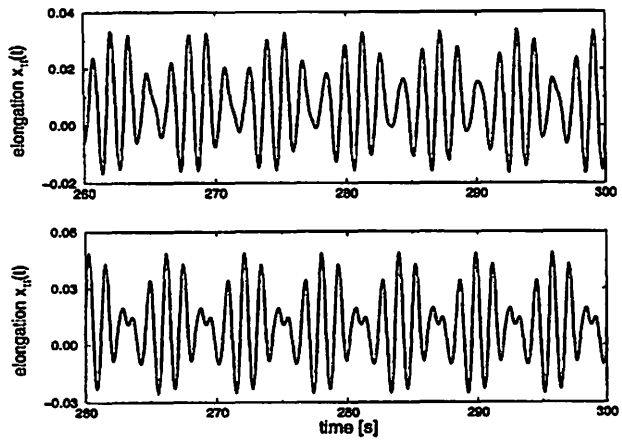

Figure 4. Elongations of the lower left mass at P_s=0.0265 (toroidal oscillations, upper graph) and P_s=0.03 (5:4 frequency locking, lower graph).

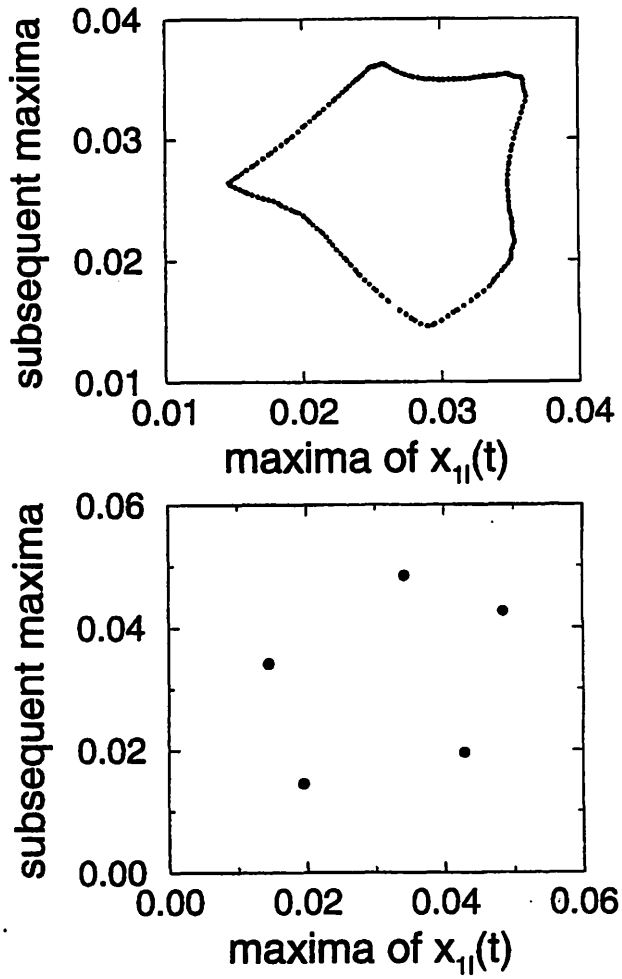

Figure 5. Plot of consecutive maxima for a torus (upper graph) and 5:4 frequency locking (lower graph).

from 0.025 to 0.039 various instabilities occur. There are toroidal oscillations around 0.027 and 0.037 and complex periodic patterns between 0.028 and 0.035. This region, where five maxima occur, can be characterised as a 5:4 frequency locking region, i.e we find during one long period five maxima of $x_{1l}(t)$ and four maxima of $x_{1r}(t)$. Representative time series are presented in Fig. 4. The corresponding power-spectra show beside the two peaks at f_{left} and f_{right} spectral components at multiples of the beating frequency f_{left}-f_{right}. Such a harmonic series of the beating frequency is also characteristic for experimental observations of biphonation [Herzel & Reuter 1996a, Tigges et al. 1996, Herzel & Reuter 1996b] (compare Fig. 1).

Another powerful technique for the analysis of complex time-series are next-amplitude plots that are topologically equivalent to Poincare sections [Berge et al. 1986]. The torus in Fig. 4 leads to a closed curve in this representation (Fig. 5, upper graph) whereas frequency locking is characterised by a set of discrete points (Fig. 5, lower graph).

The Role of the Vocal Tract

In the 2-mass models the phase shift of upper and lower masses (see Fig. 2) is the dominant mechanism to get self-sustained oscillations of the vocal folds. It models the wave-like motion of the folds that is clearly visible in stroboscopy. However, the interaction with sub- and supraglottal resonances provides another mechanism to compensate energy losses [Titze 1994]. In fact, vibrations of the 1-mass model of Landgraf and Flanagan [Flanagan & Landgraf 1968] are possible due to that mechanism.

It turns out that for our parameter configuration (6) the vocal tract supports vocal fold vibrations. Even for standard damping coefficients r_i=0.02 we find already sustained oscillations at $P_s > 0.002$ (see Fig. 6). Fig. 6 reveals a

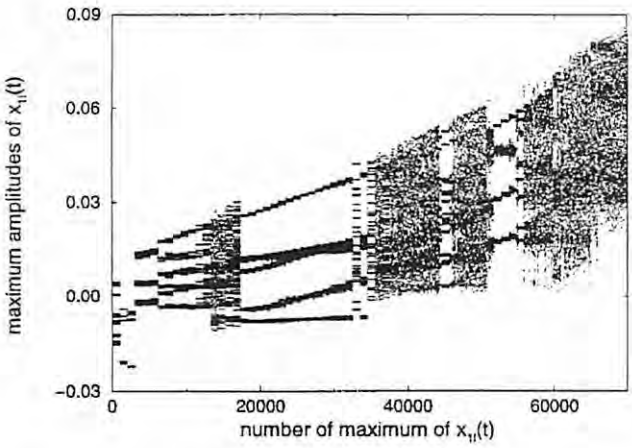

Figure 6. Bifurcation diagram with vocal tract and weak asymmetry (Q=0.8). As tract parameters we used A_T=0.6, r_M=-0.8, and T_T=1. The subglottal pressure has been increased in steps of 0.0004 from zero (left) to 0.028 (right). At each parameter value a transient of 2 s was discarded and the maxima during the following second are plotted against the index. Such a version of a bifurcation diagram is comparable to an amplitude contour.

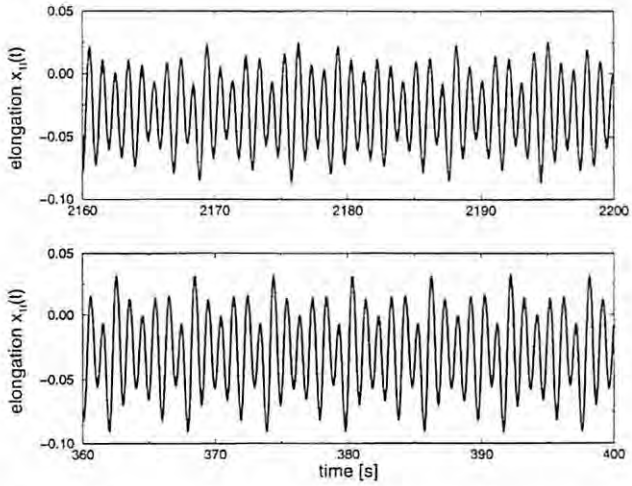

Figure 7. Time series of the left lower mass for P_s=0.006 (upper graph) and P_s=0.01 (lower graph).

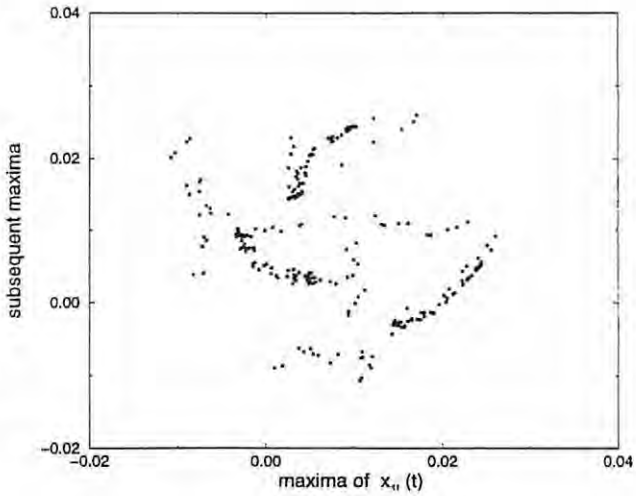

Figure 8. Next-maximum plot for the nonperiodic oscillations shown in the upper graph of Fig. 7.

rather rich bifurcation structure. There are alternating regions of frequency locking, tori, and deterministic chaos. A nonperiodic time-series and a signal with multiple maxima per period are shown in Fig. 7. The maximum plot in Fig. 8 shows a complicated pattern that indicates chaotic dynamics. We have chosen a rather small cross sectional area of the tube to get a strong source-tract interaction. In this way we get instabilities even for low subglottal pressures and weak asymmetries. (In fact, we found biphonation even for Q = 0.95.) A more detailed analysis of the bifurcations of the extended model will be discussed elsewhere.

Discussion

The core of our paper was a comparison of a simplified 2-mass model with an extended version with a 1-tube approximation of the vocal tract. Vocal instabilities have been found in the simplified model for moderate asymmetry and high subglottal pressures. Simulations of the more realistic model with vocal tract have shown comparable instabilities. This can be regarded as a justification of analysing strongly reduced models as a first approach.

However, several aspects of voice production are reflected more realistically by the extended model. For examples, skewing and ripples of the glottal flow pulses well known from inverse filtering can be reproduced. Moreover, the phonation onset threshold was significantly reduced due the vocal tract resonator.

Details of the bifurcation diagrams of both models are quite different. This is not surprising since nonlinear dynamical systems depend sensitively on small parameter changes and, hence, the source-tract interaction strongly effects bifurcation structures.

There are, however, also common features of both bifurcation diagrams such as alternations of toroidal oscillations and frequency locking. In particular, entrainment zones of left and right vocal fold were found in both model versions. The 6:5 frequency locking visible in Figs. 6 and 7 resembles closely observations of the biphonic voice analysed in [Tigges et al. 1996].

Acknowledgments

Support was provided by the Deutsche Forschungsgemeinschaft. We thank I. R. Titze, D. Berry, and B. Story at the National Center for Voice and Speech in Iowa City for their warm hospitality and many fruitful discussions.

References

P. Berge, Y. Pomeau, and C. Vidal (1986), Order within Chaos (Wiley, New York).

D. A. Berry, H. Herzel, I. R. Titze, and K. Krischer (1994), " Interpretation of Biomechanical Simulations of Normal and Chaotic Vocal Fold Oscillations with Empirical Eigenfunctions", J. Acoust. Soc. Am., Vol. 95, pp. 3595-3604.

D. A. Berry, H. Herzel, and I. R. Titze (1996), "Bifurcations in excised larynx experiments", J. Voice, Vol. 10, pp.129-138.

L. Dolansky & P. Tjernlund (1968) "On Certain Irregularities of Voiced—Speech Waveforms", IEEE Trans., Vol. AU—16, pp. 51-56.

J. L. Flanagan & L. Landgraf (1968) "Self-oscillating source for vocal tract synthesizers", IEEE Trans., Vol. AU-16, pp. 57-64.

L. Glass & M. Mackey (1988), From Clocks to Chaos (Princeton University Press).

H. Herzel & J. Wendler (1991), "Evidence of Chaos in Phonatory Samples", Proc.EUROSPEECH, Genova, 1991 (ESCA), pp. 263-266.

H. Herzel, I. Steinecke, W. Mende, and K. Wermke (1991), "Chaos and bifurcations during voiced speech", in Complexity, Chaos and Biological Evolution, ed. by E. Mosekilde and L. Mosekilde (Plenum Press, New York), pp. 41-50.

H. Herzel (1993), "Bifurcations and chaos in voice signals", Appl. Mech. Rev., Vol. 46, pp. 399-413.

H. Herzel, D. A. Berry, I. R. Titze and M. Saleh (1994) "Analysis of Vocal Disorders with Methods from Nonlinear Dynamics", J. Speech Hearing Res., Vol. 37, pp. 1008-1019.

H. Herzel & C. Knudsen (1995), "Bifurcations in a vocal fold model", Nonlinear Dynamics, Vol. 7, pp. 53-64.

H. Herzel & R. Reuter (1996), "Biphonation in Voice Signals", in Nonlinear, Chaotic, and Advanced Signal Processing Methods For Engineers and Scientists, ed. by R. A. Katz, T. W. Frison, J. B. Kadtke, and A. R. Bulsara (American Institute of Physics, Woodbury).

H. Herzel & R. Reuter (1996), "Whistle Register and Biphonation in a Child's Voice", Folia phoniatrica, submitted.

K. Ishizaka and J. L. Flanagan (1972), "Synthesis of voiced sounds from a two-mass model of the vocal cords", Bell Syst. Techn. J., Vol. 51, pp. 1233-1268.

K. Ishizaka and N. Isshiki (1976), "Computer simulation of pathological vocal-cord vibrations", J. Acoust. Soc. Am., Vol. 60, pp. 1194-1198.

D. Kaplan & L. Glass (1995), Understanding Nonlinear Dynamics (Springer, Berlin).

K. J. Kohler (1996), "Articulatory Reduction in German Spontaneous Speech", Proc. 4th Speech Prod. Seminar, Autrans, 20-24 May 1996, ed. by P. Perrier (ESCA), pp. 1-4.

M. Mazo M, D. Erickson, and T. Harvey (1995), " Emotion and expression: temporal data on voice quality in Russian lament", Vocal Fold Physiology, ed. by O. Fujimura & M. Hirano (Singular Publ Group., San Diego), pp. 173-178.

W. Mende, H. Herzel, and K. Wermke (1990), "Bifurcations and Chaos in Newborn Cries", Phys. Lett. A, Vol. 145, pp. 418-424.

J. B. Robb & J. Saxman (1988), "Acoustic observations in young children's vocalizations", J. Acoust. Soc. Am., Vol. 83, pp. 1876-1882.

P. Sirvio & K. Michelsson (1976), "Sound-spectrographic cry analysis of normal and abnormal newborn infants", Folia phoniat., Vol. 28, pp. 161-173.

M. E. Smith, G. S. Berke, B. R. Gerrat, and J. Kreimann (1992), "Laryngeal paralyses: theoretical considerations and effects on laryngeal vibration", J. Speech Hear. Res., Vol. 35, pp. 545-554.

I. Steinecke & H. Herzel (1995), "Bifurcations in an asymmetric vocal fold model", J. Acoust. Soc. Am., Vol. 97, pp. 1874-1884.

M. Tigges, P. Mergell, H. Herzel, T. Wittenberg, and U. Eysholdt (1996) "Observation and modelling of glottal biphonation", Acustica, in press.,

I. R. Titze (1984), "Parametrization of the glottal area, glottal flow, and vocal fold contact area", J. Acoust. Soc. Am., Vol. 75, pp. 570-580.

I. R. Titze, R. Baken, and H. Herzel (1993), "Evidence of Chaos in Vocal Fold Vibration", in Vocal Fold Physiology: Frontiers in Basic Science, ed. by I. R. Titze (Singular Publishing Group, San Diego), pp. 143-188.

I. R. Titze (1994), Principles of Voice Production, (Prentice Hall, Englewood Cliffs).

I. R. Titze & B. H. Story (1996) "Acoustic interactions of the voice source with the lower vocal tract", J. Acoust. Soc. Am., submitted.

NCVS Status and Progress Report - 10
November 1996, 95-101

A Simplified Model for Simulation and Transformation of Speech

Brad H. Story, Ph.D.
Department of Speech Pathology and Audiology, The University of Iowa
Ingo R. Titze, Ph.D.
Department of Speech Pathology and Audiology, The University of Iowa
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts
Darrell Wong, Ph.D.
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts

Abstract

This paper explores a model that reduces speech production to the specification of four time varying parameters; F1 and F2, voice fundamental frequency (Fo), and a relative amplitude of the voice. The trajectory of the first two formants, F1 and F2, is treated as a series of coordinate pairs that are mapped from the F1F2 plane into a two-dimensional plane of "coefficients". These coefficients are multipliers of two empirically-based orthogonal basis vectors which, when added to a neutral vowel area function, will produce a new area function with the desired locations of F1 and F2. Thus, area functions and voice parameters extracted at appropriate time intervals can be fed into a speech simulation model to recreate the original speech. A transformation of the speech can also be imposed by manipulating the area function and voice characteristics prior to the recreation of speech by simulation. The model has initially been developed for vowel-like speech utterances but the effect of consonants on the F1F2 trajectory is also briefly addressed.

Introduction

Any computational scheme for the synthesis of speech requires that the speech production process be reduced to some simplified set of variable parameters that specify, at least, the pitch and amplitude of the voice and the vocal tract resonance characteristics. The voice source can be represented as simply as an idealized glottal flow pulse or as complicated as a finite element simulation of the vocal fold vibration. Assuming a "source-filter" theory of speech production [1], the output of the voice source can be considered to be "injected" into the vocal tract where it is shaped by the tract resonances (or formants). The method in which those resonances are imposed can differ greatly between various types of synthesizers.

A formant synthesizer attempts to replicate the vocal tract resonances with a cascaded (or sometimes parallel) set of resonant digital filters [2], the characteristics of which are specified by desired formant and bandwidth values. Articulatory synthesis, which is more of a simulation of the speech process, is often governed by a midsagittally-based articulatory model in which positions of articulators (tongue, velum, lips, etc.) are manipulated to generate a givenutterance. The synthesis is realized by transforming the articulator positions into an area function (i.e. the cross-sectional area of the vocal tract as a function of the distance from the glottis) via an empirically-derived midsagittal-to-area transformation and then using the area function to create some form of a digital filter with the appropriate resonance pattern. Examples of such midsagittally-based models can be found in Lindblom and Sundberg [3], Mermelstein [4], Coker [5], and Browman and Goldstein [6].

Other highly compact articulatory models are those of Stevens and House [7] and Fant [1], both of which represented the vocal tract with only three parameters; the place and cross-sectional area of the main vocal tract constriction and a ratio of lip protrusion to lip open area. With these parameters, the entire area function from just above the glottis to the lips can be constructed by empirically-based rules.

It is the purpose of this paper to explore a highly simplified model that can be used to synthesize (or preferably, *simulate*) speech. The model is based on the decomposition of a set of vocal tract area functions obtained from magnetic resonance imaging for ten vowels [8],[9] into a set of empirical orthogonal basis vectors (or "modes")and a mean area function [8],[9]. Four time-varying parameters are required to drive the speech simulation: the first two formants, F1 and F2, and the amplitude (U_o) and fundamental frequency (F_o) of the voice. Unlike a formant synthesizer, F1 and F2 are treated as a coordinate pair that can be mapped into pair of "articulatory coefficients". These coefficients are multipliers of the two most significant empirically-derived orthogonal basis vectors (or modes) which, when added to a neutral vowel area function, will produce an area function with acoustic resonances at the desired locations of F1 and F2. Since the decomposition of the area functions is performed in the "area domain", no transformation from midsagittal dimension to cross-sectional area is required. Thus, the specification of F1 and F2 is used to create the appropriate vocal tract area function while values of the voice fundamental frequency (F_o) and amplitude (U_o) are used to produce the voice source. The four parameters are fed into a speech simulator based on the wave-reflection approach [11],[8] in which losses due to yielding walls, viscosity, heat conduction, and radiation are taken into account. The voice parameters drive a glottal flow pulse model [12].

The specific aims of the paper are to demonstrate how the mapping between the F1-F2 coordinate values and the empirical modal coefficients can be used to generate a sequence of area functions of accuracy sufficient to mimic human speech in the case of some simple utterances.

Summary of the Model

Parameterization of the Vocal Tract Shape

The decomposition of the vocal tract area functions for ten vowels from Story et al. [9] allows each area function to be compressed from a 44 section cylindrical tube representation into a few coefficients that are vowel-specific multipliers of the empirically-determined orthogonal basis vectors. Inclusion of just the two most significant orthogonal basis vectors in a subsequent reconstruction of the area functions explained 88% of the total variance in the set, thus each vowel can be reasonably represented with two coefficients [8],[10]. Figure 1 shows the two basis vectors and the mean area function onto which perturbations can be superimposed. Any of the area functions in the original vowel set can be reconstructed with the following equation,

$$A(x) = A_0(x) + \sum_{i=1}^{2} c_i \varphi_i(x) \qquad (1)$$

where $A_o(x)$ is the mean area function, c_i are the vowel specific coefficients, and $\varphi_i(x)$ are the orthogonal basis

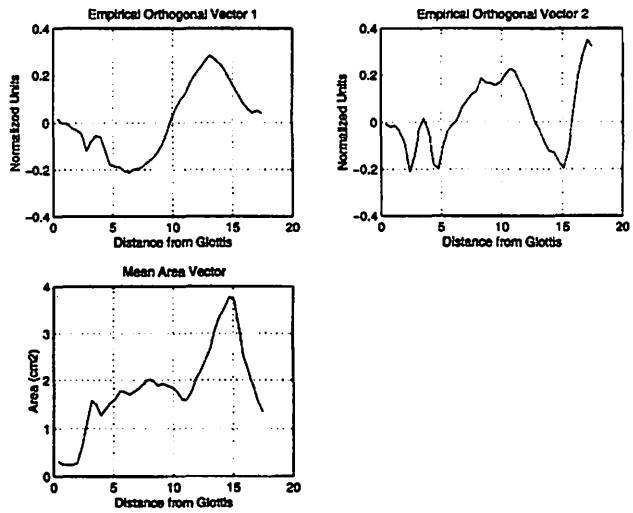

Figure 1. Two empirical orthogonal basis vectors and mean area function from a ten vowel set [10].

vectors. A 50x50 grid of coefficient pairs was used to generate 2500 new area functions. The frequency response of each area function was computed with a frequency domain transmission line method [13] and the first two formants (F1 and F2) were determined by peak-picking and parabolic interpolation [14]. Figure 2 shows the grid of coefficients and the corresponding deformed grid of F1F2 pairs. Each line connecting formant pairs in this grid represents a constant value of either the first or second modal coefficient; i.e. each line is an "iso-coefficient" line. The coefficient and formant pairs corresponding to each of the original ten vowels are shown with solid dots.

Within a large range of the F1F2 plane, a given formant pair can be uniquely mapped back into the coefficient grid. This property suggests that if F1 and F2 can be adequately determined from a speech signal, they can be mapped into a pair of coefficients to generate a physiologically realistic area function. Thus F1 and F2 can be transformed into the articulatory parameters, C1 and C2. This feature will be exploited in a later section.

Voice Source Model

As mentioned previously, the voice source used in the speech simulator is of the glottal flow pulse type [15],[16],[12]. With such a model, vocal fold vibration is not simulated but only the resultant glottal flow signal is generated. The potential naturalness of a self-oscillating vocal fold model is eliminated but a glottal flow model allows precise control over the fundamental frequency (F_o), amplitude (U_o), and waveshape which is not easy to achieve with the self-oscillating type of model. The particular glottal flow pulse model used in this study was given in Titze et al. [12] and requires as inputs F_o, U_o, frequency modulation, amplitude modulation, open quotient, skewing quotient,

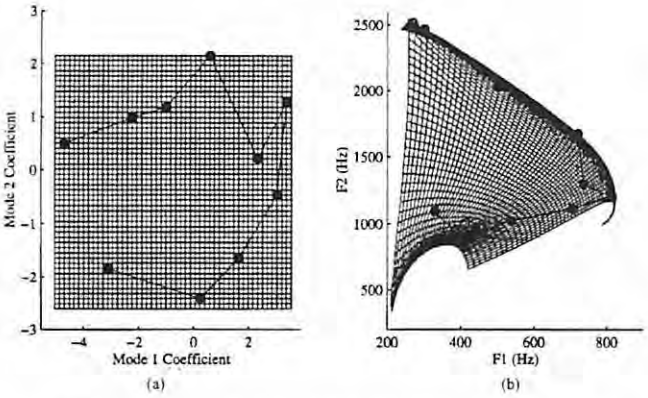

Figure 2. Mapping from articulatory to acoustic parameters where the black dots (in both articulatory and acoustic domains) represent the locations of the original ten vowels used to generate the orthogonal basis vectors: a) grid of empirical basis function coefficient pairs, and b) deformed grid of corresponding F1-F2 pairs.

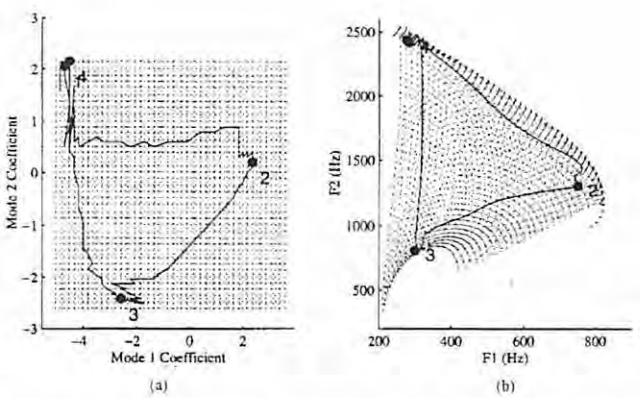

Figure 3. Mapping of F1-F2 pairs extracted from the utterance /iɑui/ back to corresponding "articulatory coefficients, a) coefficient trajectory corresponding to the F1F2 trajectory shown in (b), b) F1-F2 trajectory of the utterance /iɑui/. The solid lines represent the mapping of the utterance while the two grids from Figure 2 are shown in the background. The numerical symbols represent specific analysis frames.

and dc flow amplitude. However, to reduce the number of control variables for the voice, all of the parameters have been set to be constant except for F_0 and U_0.

Vocal Tract Model

The type of vocal tract model used for simulation of speech sounds is a wave-reflection analog (or equivalently a wave digital filter). The model requires that the three-dimensional vocal tract shape be discretized into a finite number of equal length cylindrical sections. Reflection coefficients based on the relative areas of adjoining sections are calculated at each cylinder junction and waves are propagated through the system by using the reflection coefficients to compute the incident and reflected components of the pressure or flow waves at each junction at each step in time. This has the effect of transporting a wavefront from section to section. It is an attractive method for acoustic modeling of the vocal tract because computations are performed serially in time-synchrony with the acoustic wave propagation. Computations are efficient because the equations describing the wave propagation become a digital filter structure in their final form. Additionally, energy losses due to the yielding properties of the vocal tract walls, fluid viscosity, and radiation from the mouth have been incorporated into the model.

The simulation is sampled at a frequency of 44.1 kHz and each finite section of the area function represents a tube length of 0.396 cm. The vocal tract length has been standardized to be 17.5 cm, thus each area function is composed of 44 individual cross-sectional areas. The simulation of speech is performed by injecting the parameterized glottal flow waveform (Figure 3) into the glottal end of the vocal tract.

Interpolation of Parameters

A given set of vocal tract and voice parameters are fed into the speech simulator with a "time stamp" attached to them. The time stamp specifies the point in time that the model parameters should attain a desired value; each set of parameters is essentially a vocal and articulatory target, that target should be achieved at the point in time specified by the time stamp. This time code for the parameter target can specify new parameter values at fixed time intervals or highly non-uniform intervals. In either case, the parameter values are interpolated between time stamps to provide model inputs for every time sample (44100 per second); e.g. if two consecutive parameter sets are specified at 0.4 and 0.5 seconds an interpolation of parameter values is performed to fill in the 441 time sample between the two targets. The voice parameters are linearly interpolated while the articulatory coefficients are subjected to a second order filter in which undershoot or overshoot of a target could occur. For each interpolated set of coefficients an area function is reconstructed using equation (1).

Extraction of Vocal Tract and Voice Parameters From the Acoustic Signal

The model described in the previous section can derive its inputs from either a "script" that dictates desired formant values and voice parameters as a function of time, or they can be extracted from a recorded speech waveform, in which case the desire would be to recreate (or simulate) the natural speech. The "scripted" method of imposing the input parameters is similar to the rule-based approach used for driving a formant synthesizer in a text-to-speech system [2]. The second method, which extracts input parameters from recorded speech, lends itself to a speech-to-speech system. Such a system could be used for compression/

transmission of speech signals or for a transformation of the original speech into a new or different speech/voice quality. The emphasis in this study will be on the latter application, speech-to-speech simulation, and hence this section will be devoted to extracting the model parameters from the speech waveform.

Vocal Tract Parameters: Formant to Coefficient Mapping

To extract the first two formants, a 55 pole, linear predictive coding (LPC) technique was used with window length of 20 msec to generate a pseudo-formant spectrum at intervals of 5 msec. The frequency locations of the first two formants can then be determined by peak-picking and parabolic interpolation [14]. Thus, an F1-F2 pair is generated every 5 msec, creating a trajectory in the F1-F2 plane. The time interval could easily be reduced or increased depending upon the desired time resolution of the formant pairs.

Once the F1-F2 trajectory has been obtained, each formant pair is matched, in a least squares sense, to the closest F1-F2 pair in the grid shown in Figure 2b. These formant pairs can then be mapped to their corresponding coefficient pair in Figure 2a and when required, an area function can be constructed with equation (1). This technique works only if the F1-F2 trajectory stays within the bounds of the formant grid. If it does not, the closest match between the measured F1-F2 pair and the grid pair may actually be quite far apart. It has been shown that highly or overarticulated speech tends to push an F1-F2 trajectory beyond the bounds of the formant grid [10]. This is, however, expected since the original area functions derived from MRI were suspected to be slightly centralized due to subject fatigue during the long image acquisition [9]. Additionally, the decomposition of the area functions into empirical orthogonal modes required that all vowels be normalized to one length. Thus any vocal tract length changes such as lip-rounding/spreading or larynx raising/lowering are not represented by the empirical orthogonal modes. Overarticulated speech would almost certainly use these articulatory maneuvers to excessive degrees. Nonetheless, for conversational type speech the mapping from F1F2 pairs back to modal coefficient pairs seems to be useful.

Voice Parameters: Extraction of Fundamental Frequency and Amplitude

The time-varying fundamental frequency (F_0)of the voice is determined by first downsampling the speech recording from 44.1 kHz down to 1 kHz (with the appropriate low pass filtering prior to downsampling). A low pass filter with a cutoff frequency of 200 Hz is then applied to the downsampled signal to eliminate frequency components above the F_0; this cutoff frequency will need to be adjusted for female and child voices. The F_0 contour over time is then determined as the unwrapped phase of the Hilbert transform of the filtered signal. A final operation is to low pass filter the resulting F_0 contour with a cutoff frequency of 15 Hz. This cutoff retains slowly changing voice characteristics such as vibrato (frequency modulation) which typically occurs at about 5 Hz. The F_0 contour can be sampled at intervals of 5 msec to be compatible with the time resolution of the vocal tract coefficients discussed above.

An estimate of the amplitude of the voice source is obtained by a frame-based RMS operation on the recorded speech signal. The signal is first low-pass filtered below 3500 Hz to eliminate possible contributions to the amplitude by turbulent sources (i.e. fricatives). The RMS value of the filtered signal is computed every 1 msec over a window size of 25 msec, providing an amplitude signal effectively sampled at 1000 Hz.

Testing the Model with Natural Speech

In this section, several simple utterances recorded from the same subject who was imaged for the original MRI acquired area functions (see [9]), were selected to be analyzed with the methods outlined in previous sections. The following utterances were recorded; /iɑui/, /uɑ/, and /udɑ/. The last two utterances were chosen to demonstrate the effect of a consonant element. At this point, the formant-to-coefficient mapping has been developed specifically for vowel-like utterances, however, the formant transitions during the onset and offset of consonants should also be captured in the F1-F2 trajectory.

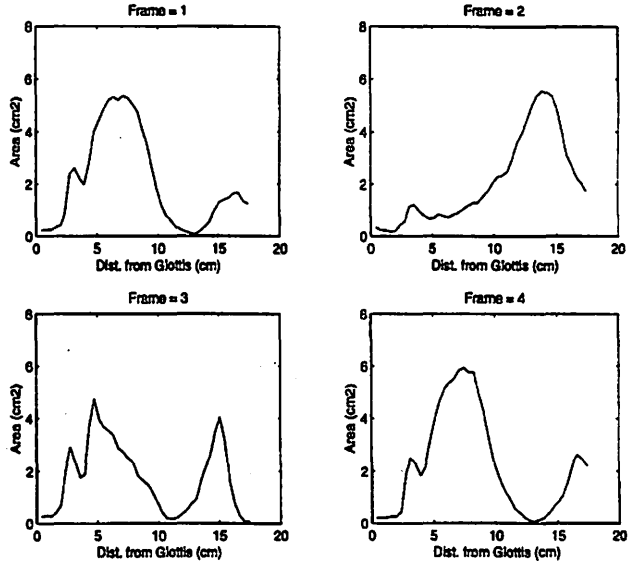

Figure 4. A sampling of four area functions constructed with equation (1) after mapping the F1-F2 trajectory of the utterance /iɑui/ to empirical basis function coefficients. The numbers at the top of each area function correspond to the numerical symbols indicated in Figure 3.

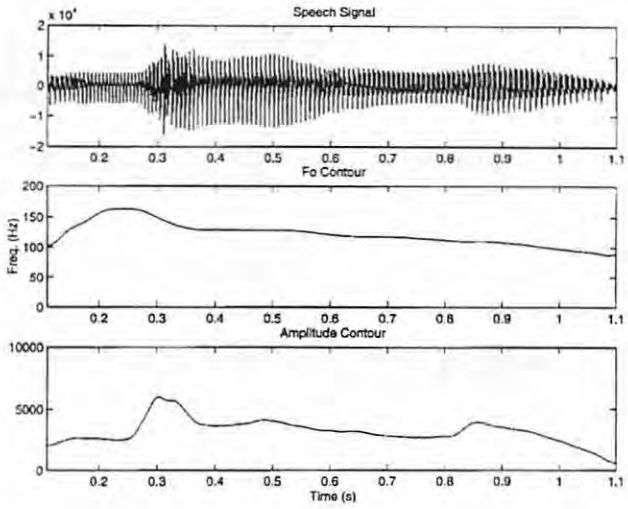

Figure 5. Original acoustic signal for /iɑui/ and the extracted fundamental frequency and RMS amplitude contours.

Each utterance was recorded direct to disk at a sampling frequency of 44.1 kHz via a Digidesign Audiomedia board installed in a MacIntosh Quadra 950. The recordings were made in a sound treated room using an AKG C410 microphone. The audio files were later read into MATLAB where the extraction of F1 and F2, fundamental frequency, and amplitude were performed.

Utterance: /iɑui/

Figure 3b shows the F1-F2 trajectory of /iɑui/ superimposed on the F1-F2 grid mesh (now shown with dotted lines so that the F1-F2 trajectory can be better seen). The "1" and "4" symbols represent the beginning and end of the utterance, respectively, and the other two numerical symbols represent selected analysis frames that are discussed later. The F1-F2 trajectory lies comfortably within the formant mesh except for the beginning and end of /i/ which occupies the region at the upper left corner of the formant grid. In this region the mapping between formant and coefficient pairs is not one to one, resulting in some ambiguity for determining the appropriate coefficients with which to reconstruct area functions. The rest of the F1-F2 trajectory can easily be mapped into corresponding articulatory coefficients. Figure 3a shows the coefficient trajectory corresponding to the F1-F2 trajectory in Figure 3b. Again the "1" represents the beginning of the utterance, "4" the end, and the numerical symbols and adjacent dots are the coefficient pairs that correspond the frame numbers in Figure 3b. The ambiguity of the mapping between formants and coefficients for the initial and final /i/ vowels is reflected in the oscillation of the coefficient trajectory near the beginning and end points shown in Figure 3a. The coefficient trajectory has a jagged characteristic due to forcing each F1-F2 pair extracted from the speech utterance to a discrete

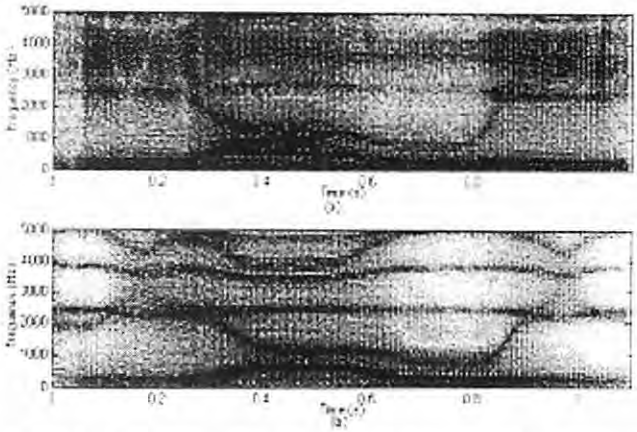

Figure 6. Spectrograms of /iɑui/: a) original speech, and b) simulated version.

point in the F1-F2 mesh. A smoother mapping could be realized by interpolating within mesh cells, but for this preliminary study a simple minimum distance criterion was assumed to be adequate. The general shape of the coefficient trajectory is similar to that of the F1-F2 trajectory but rotated by approximately +45 degrees.

Figure 4 shows a series of four area functions reconstructed with equation (1) that have been sampled from the F1-F2 trajectory of /iɑui/. The numerical symbols shown in Figure 3b are the sampled points and their corresponding number is indicated at the top of each area function plot. The first area function is the initial "i" -like vowel which then evolves into the "ɑ" -like shape shown in the second frame. The third frame shows the tract shape becoming an "u" and transitioning back to the final "i" - like vowel in the fourth frame.

Figure 5 shows the speech signal along with the contours of fundamental frequency and amplitude for the /iɑui/ utterance. There is raising followed by a rapid lowering of F_0 during the during the initial /i/ and then F_0 gradually falls throughout the remainder of the utterance.

The wide band spectrograms in Figure 6 demonstrate the time-varying spectral content of the original sentence (Fig. 7a) and a simulated version obtained by using the four parameters extracted from the speech signal. The spectrograms are quite similar in terms of the shape and frequency location traversed by the formant trajectories of F1 and F2 except that it appears that the simulated version does not reach the vowel formant targets as rapidly as the natural speech. The third formant also seems to be well matched between the two even though no control was exercised over F3. It is also apparent that the formant bandwidths generated by simulation are slightly narrower than natural speech, possibly contributing to a more "mechanical" type of sound quality.

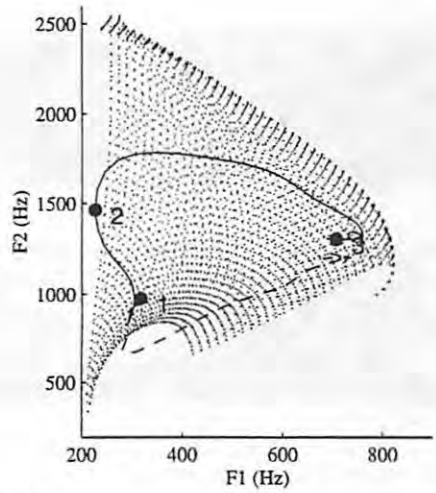

Figure 7. F1-F2 trajectories for /uɑ/ (dashed) and /udɑ/ (solid). The numerical symbols and corresponding dots signify specific frames used for later analysis.

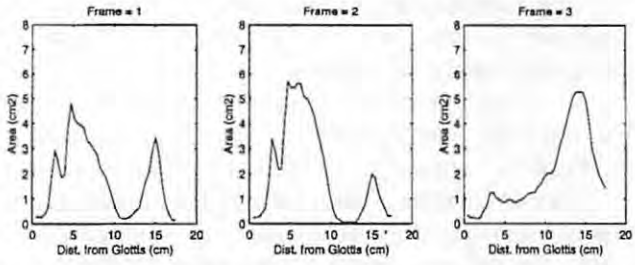

Figure 8. A sampling of three area functions constructed with equation (a) after mapping the F1-F2 trajectory of utterance /udɑ/ to empirical basis function coefficients. The numbers at the top of each area function correspond to the numerical symbols indicated in Figure 7.

Utterances: /uɑ/, and /udɑ/

F1-F2 trajectories are shown for the /uɑ/ and /udɑ/ utterances in Figure 7 (the F1-F2 grid is again plotted in the background). The /uɑ/ trajectory (dashed line) indicates a direct movement from the /u/ configuration to the /ɑ/ while for /udɑ/ the trajectory is diverted upward in the direction of F2 and downward in F1 before making the transition to /ɑ/. Since a portion of the trajectory escapes the F1-F2 grid, the mapped coefficient pairs will correspond to the closest formant pairs on the left vertical edge of the formant grid. Reconstructed area functions for the marked points on the /udɑ/ trajectory are given in Figure 8. The first and third area functions are the /u/ and /ɑ/ vowels, respectively. The second area function shows an approximation to the consonant /d/. It shows the formation of a broad constriction that is centered at about 5 cm behind the lips. The constriction location is farther behind the lips than a typical /d/ and may actually be closer to simulating a /g/.

Spectrograms of the natural recorded and simulated versions of /udɑ/ are shown in Figures 9a and 9b,

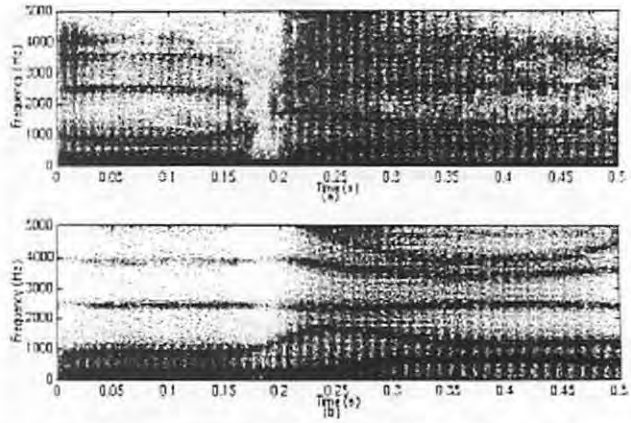

Figure 9. Spectrograms of /udɑ/; a) original speech, and b) simulated version.

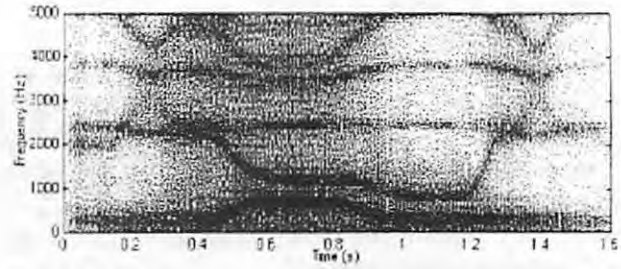

Figure 10. Spectrograms of a simulated version of /iɑui/ in which the fundamental frequency has been uniformly lowered by 20 percent and the rate of articulation has been decreased by 40 percent (i.e. slowed speech).

respectively. Again F1 and F2 are quite similar between the two versions except that the simulated version does not provide the break in the formants during the /d/.

Transformation of Speech and Voice Qualities

The model proposed in this paper, lends itself to performing voice transformation in addition to recreating the original speech utterance. Once the four parameters have been determined, they can be manipulated prior to using them in the final simulation. For example, the fundamental frequency could be uniformly (or nonuniformly) increased or decreased across the utterance. Once the area functions have been derived, they could be lengthened or shortened and various regions of the vocal tract could expanded or compressed by various multipliers. Additionally, rate of articulation could be altered by changing the time code that is "stamped"on the parameters when they are extracted.

The spectrogram in Figure 10, show the case of uniformly slowing down the rate of speech by 40 percent (each time stamp was multiplied by 1.4) and scaling the F_0 contour by a factor of 0.8 (F_0 decreased by 20 percent).

Conclusion

At this point, the model has many serious limitations. The results obtained were based on ten vowel area functions obtained by magnetic resonance imaging of one adult male speaker (native of the midwestern United States) making the results speaker dependent. All of the vowels were normalized to one standard length which destroyed any information about vocal tract length changes such as lip rounding/spreading or larynx raising/lowering. Additionally, since only ten vowels were subjected to the analysis, the effect of consonantal area functions on the results and subsequent conclusions are unclear. Another limitation is that using only the first two formants will nearly guarantee failure for utterances in which F3 is important (e.g. for utterances containing /r/ or /l/). The voice parameters are also very limited in that by using only the fundamental frequency and a relative voice amplitude, the ability to recreate an appropriate voice quality is quite limited. However, it is useful to explore all of the possibilities of this simple model before moving on to more sophisticated versions.

Even with the limitations cited, the mapping of F1-F2 pairs into "articulatory coefficient" pairs provides a compact system by which to specify physiologically realistic area functions that can be used to simulate dynamic vowel-like utterances. The quality of the simulation can be enhanced by directly mapping from F1F2 pairs extracted from natural speech to physiologically realistic area functions and using those area functions to simulate the original speech. Future work will include understanding the effect of consonants on the formant-to-coefficient mapping process as well as an attempt to provide an articulatory control parameter for F3.

Acknowledgements

This work was jointly supported by Grants P60 DC000976 and R01 DC02532 from the National Institute on Deafness and other Communication Disorders. The authors would also like to thank Julie Lemke for her assistance in preparing this manuscript.

References

[1] Fant, G., The Acoustic Theory of Speech Production, Mouton, The Hague, 1960.

[2] Klatt, D. H., "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am., 67(3), 971-995, 1980.

[3] Lindblom, B., and Sundberg, J., "Acoustical consequences of lip, tongue, jaw, and larynx movement," JASA, 4(2), 1166-1179, 1971.

[4] Mermelstein, P., "Articulatory model for the study of speech production," JASA, 53(4), 1070-1082, 1973.

[5] Browman, C., and Goldstein, L., "Gestural specification using dynamically-defined articulatory structures," Haskins Lab. Stat. Rep. on Speech Res., SR-103/104, 95-110, 1990.

[6] Coker, C. H., "A model of articulatory dynamics and control," Proc. IEEE, 64(4), 452-460, 1976.

[7] Stevens, K. N., and House, A. S., "Development of a quantitative description of vowel articulation," JASA, 27(3), 484-493, 1955.

[8] Story, B. H., Speech Simulation with an Enhanced Wave-Reflection Model of the Vocal Tract, Ph. D. Dissertation, University of Iowa, 1995.

[9] Story, B. H., Titze, I. R., and Hoffman, E. A., "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am., 100(1), 537-554, 1996.

[10] Story, B. H., and Titze, I. R., "Parameterization of vocal tract area functions by empirical orthogonal modes," J. Acoust. Soc. Am., (in review), 1996.

[11] Liljencrants, J., "Speech Synthesis with a Reflection-Type Line Analog," DS Dissertation, Dept. of Speech Comm. and Music Acous., Royal Inst. of Tech., Stockholm, Sweden, 1985.

[12] Titze, I. R., Mapes, S., and Story, B., "Acoustics of the tenor high voice," J. Acoust. Soc. Am., 95(2), 1133-1142, 1994.

[13] Sondhi, M. M., and Schroeter, J., "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. ASSP, ASSP-35(7), 1987.

[14] Titze, I. R., Horii, Y., and Scherer, R. C., "Some technical considerations in voice perturbation measurements," JSHR, 30, 252-260, 1987.

[15] Rosenberg, A. E., "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am., 49(2), 583-590, 1971.

[16] Fant, G., Liljencrants, J., Lin, Q-g, "A four-parameter model of glottal flow," Paper presented at the French-Swedish Symposium, Grenoble, April 1985.

NCVS Status and Progress Report - 10
November 1996. 103-110

Voice Transformation With Physiologic Scaling Principles

Ingo Titze, Ph.D.
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts
Department of Speech Pathology and Audiology, The University of Iowa
Darrell Wong, Ph.D.
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts
Brad Story, Ph.D.
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts
Department of Speech Pathology and Audiology, The University of Iowa
Russell Long, Ph.D.
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts

Abstract

This study begins to explore the importance of the physiological domain in voice transformation. A general approach is outlined for transforming the voice quality of sentence-level speech while maintaining the same phonetic content. Transformations will eventually include gender, age, emotional state, disordered state, or impersonation, but only a specific voice quality, twang, is described in detail here. The basic question is: relative to pure signal processing, can voices be transformed more effectively if biomechanical, acoustic, and anatomical scaling principles are applied? Two approaches are used to answer the question, a Linear Predictive Coding signal approach and an articulatory biomechanical simulation approach.

Introduction

Voice transformation is defined as the purposeful change of perceived age, gender, identity, or personality of an individual on the basis of his or her voice. It can be done behaviorally, surgically, or electronically, but only electronic voice transformation is of interest here. Applications include variations of instructional synthetic speech (to make it more interesting), simulation of multiple telephone voices, and correction of the effects of illness, fatigue, or pathology in any electronically assisted vocal communication. An earlier introduction to the topic was given by Childer's et al. (1989) in this journal.

The task of voice transformation consists of three phases: 1) analysis of the speech into a set of parameters that allows manipulation of voice quality, 2) transformation of the parameters into a set that describes a different voice quality, hopefully based on anatomical and physiological variations, and 3) a reconstruction of the speech utilizing the new parameters with the original articulatory (phonetic) content. The first phase of the process requires that the speech be separated into its articulatory component (that which largely determines *what* is being said) and its voicing component (that which largely determines *how* and by *whom* it is being said).

But this is not a simple source-filter separation. Much about voice quality is determined by lower vocal tract filtering (the pharynx, the epilarynx tube, the piriform sinuses, and the velopharyngeal port). Qualities such as *twang*, *ring*, and *sob* (Colton and Estill, 1978) are based on more than vocal fold adjustments. For these qualities, basic vocal registers (distinct sound qualities based on modes of vibration of vocal fold tissues) are augmented by adjustments of the lower vocal tract structures.

The present study summarizes two schemes for analyzing speech and compares them for the purpose of voice transformation (Figure 1). Method 1 uses a traditional frame-by-frame linear predictive coding (LPC) strategy performed on the microphone signal. The LPC coefficients are mapped to a pseudo-area function sequence, or into a

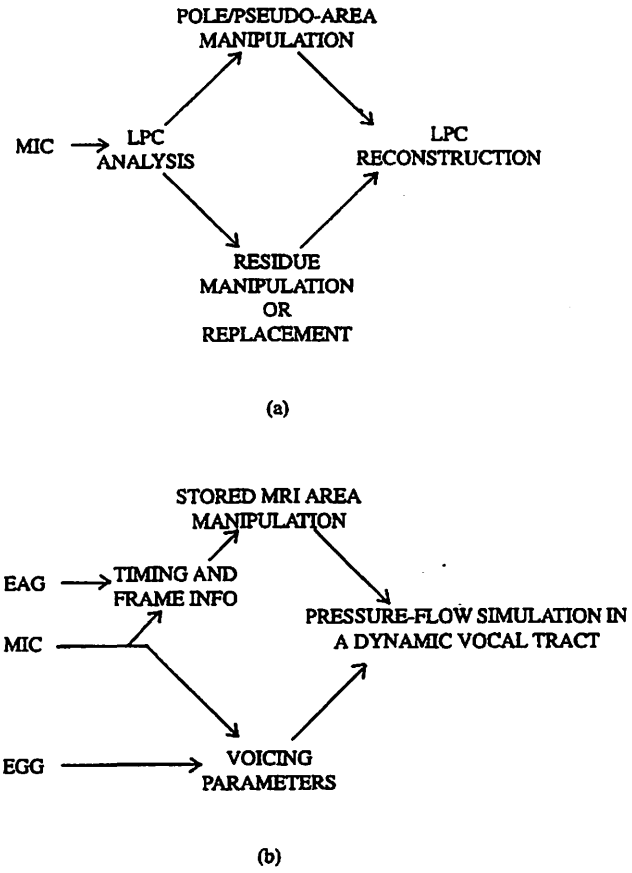

Figure 1. Simplified flow charts of two transformation strategies. (a) the LPC approach and (b) the simulation approach.

pole-zero representation. For the voicing characteristics, an estimate of the glottal flow wave is derived from the LPC residual using two different glottal pulse models--one based on a glottal area and glottal flow parameterization (Titze, Story, Mapes 1994) and the other based on statistical data fitting of the flow derivative in each cycle (Childers and Hu, 1994; Milenkovic, 1993). For the area/flow model, the voicing parameterization includes an identification of the voicing (on/off) state, an estimation of pitch period markers from the glottal flow residual (F_o), and a measurement of the intensity of phonation in each voiced cycle (the maximum flow declination rate measured from the LPC residual). For the statistical model, only the cycle markers and a set of polynomial function coefficients are retained. A synthetic flow wave is then constructed cycle-by-cycle, with the systematic addition or deletion of cycles performed as necessary for the modification of pitch. In the case of the voiceless portions of the utterance (or any interval which lacks sufficient periodicity in excitation), the full LPC residual is retained. The composite (voiced/voiceless) excitation signal is then introduced frame-by-frame into the time-varying LPC filter which yields the transformed speech.

The second method of voice transformation (bottom of Figure 1) uses measured vocal tract areas rather than the pseudo-area or pole-zero representation derived from LPC; henceforth this will be called the simulation method. The characteristics of voicing for the input speech are analyzed frame-by-frame from the LPC residual as in method 1, but a combination of the microphone signal and EGG (for more accurate cycle marking) is used. The area/flow model is again used as a voice source. Three-dimensional MRI images of the vocal tract, measured specifically for the speaker (Story, Titze and Hoffman, 1996), are used to construct a list of the target phonemes for the utterance. These 3-D images are converted to area functions and then time-aligned (manually) with the microphone and EGG signals obtained from the actual test sentence. A third transducer, the electroarticulograph EAG, is also used to facilitate this manual alignment (Karlsson and Nord, 1970). After time-alignment of the phoneme target area functions, an interpolation between targets is generated as a way of estimating the dynamic movement of articulators through the allophonic and coarticulated sounds.

Excitation Models

The glottal area and flow model was originally described by Titze (1983) and in more detail by Titze, Mapes and Story (1994). It defines a glottal flow pulse y and the glottal area a as follows:

$$u_n = av_o(\pm(1 + (\gamma a v_o)^2 + 2\gamma u_{n-1})^{1/2} - \gamma a v_o) \quad (1)$$

where

$$\gamma = 0.080(Q_s - 1) \quad (cgs\ units) \quad . \quad (2)$$

$$v_o = (2P_L/(k_t\rho))^{1/2} \quad , \quad (3)$$

$$a = a(\theta) = \max[0, \sin^{1+Q_o}\theta] \quad , \quad (4)$$

and

$$\theta = \pi n/(Q_o T_o) \qquad 0 < \theta < \pi, \quad 0 < n < T_o \quad (5)$$

Parameters in this model are the fundamental period T_o, the open quotient Q_o, the skewing quotient Q_s, and the lung pressure P_L. The air density ρ is a constant (0.00114 g/cm^3)m and a transglottal pressure coefficient k_t can either be chosen as a constant (1.1) or varied with glottal shape during the cycle (Scherer and Guo, 1991).

Although the well-described LF model (Fant, Liljencrants and Lin, 1985) could have been adopted for this purpose, the parameters used in this model are more in tune

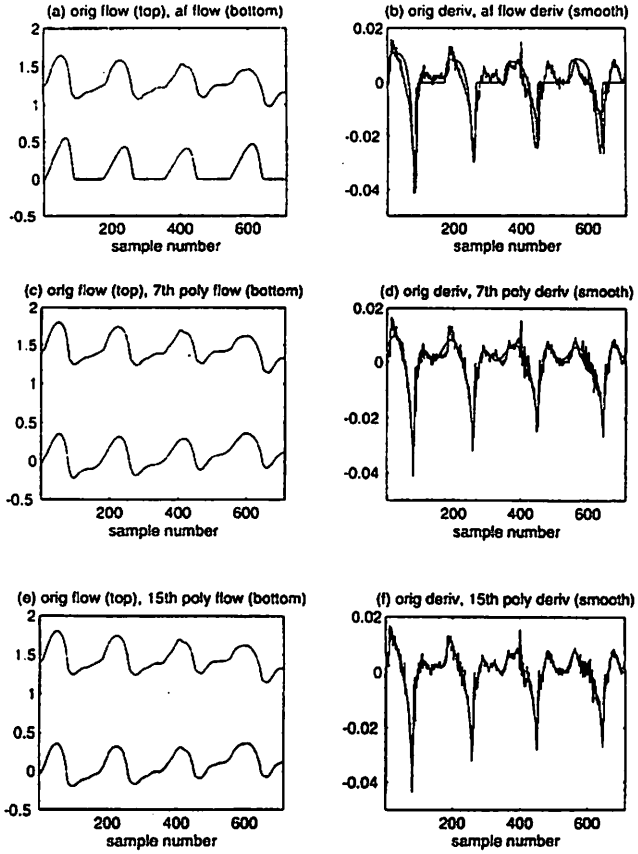

Figure 2. (a) Original LPC residual pseudoflow (top) and a/f model approximation (bottom) from a steady state portion of a sentence, (b) overlaid derivatives of flows from Figure 2a, (c) some vowel with 7th order polynomial model at the bottom, (d) flow derivatives of waveforms at left, (e) same vowel with 15th order polynomial model at the bottom, (f) flow derivatives of waveforms at left.

with the large data sets available on adult males, females and children (e.g. Holmberg, Hillman and Perkell, 1988; Perkell, Hillman and Holmberg, 1994; Stathopoulos and Sapienza, 1993 a,b). In these experimental studies on human subjects, lung pressure, open quotient, skewing and fundamental frequency were explicitly reported. Hence, our desire is to incorporate the measured parameters and scale them appropriately.

In the analysis phase, T_o, Q_o, and Q_s are estimated from the LPC residual (pseudoflow) and applied as control parameters for transformation. This skewing quotient emulates the effect of the vocal tract inertance; i.e., the acoustic load. Note that when $Q_s = 1$, $\gamma = 0$ and $u_n = av_o$, the glottal area multiplied by the "no-load" particle velocity v_o. For values of Q_s greater than 1.0, the flow skews to the right. The model can also be extended to include vocal fold length and maximum glottis diameter as variables (Wong et al. 1996), allowing age and gender differences in the structure of the larynx to be captured. Additionally, the glottal noise from flow turbulence may be included by using pseudorandom noise with an appropriate spectrum whenever orifice and flow conditions require it. Figure 2a shows the LPC residual flow pulse and the model approximation, and Figure 2b shows the overlaid flow derivatives (model and data).

The second excitation model parameterizes the glottal flow derivative rather than the glottal area, and is identical to a model proposed by Childers and Hu (1993). The flow derivative, represented by the LPC residual, is modeled by a high order polynomial for each cycle.

$$p(x) = p_1 x^n + p_2 x^{n-1} + \ldots p_n x + p_{n+1} \qquad (6)$$

An optimization procedure is used to fit the polynomial of degree n to the data in a least squares sense. The procedure requires the cycle to begin and end at the maximum negative spike locations in each cycle. The polynomial equation attempts to fit all the major low frequency humps in each cycle, thus requiring a polynomial coefficient set for every cycle. Polynomial orders from 7 to 15 have been used for this study, with a 15th order being more accurate, but a 7th order polynomial more efficient. Turbulence can be added as an additional contribution after the polynomial has been fitted (Wong et al., 1996). Milenkovic (1993) also proposed a model of this nature. These models grew out of the one proposed by Imaizumi, Kiritani and Sato (1991), who modeled the derivative with three piece-wise sections using a parabola, a constant, and a cubic, rather than using a single polynomial. The single polynomial provides a better fit to the data since it does not assume a closed phase in the flow (the constant section in Imaizumi's model). Figures 2c and 2d show the residual flow and flow derivative for a 7th order polynomial fit and Figures 2e and 2f show the same thing for a 15th order polynomial.

The polynomial model fits the flow derivative, since this signal is assumed to be the excitation signal for the vocal tract. It does not, however, use physiological measurements of the shape and structure of the vocal fold in its definition, and thus does not provide any insight as to the influences that parameters such as lung pressure and vocal fold length have on the system.

Vocal Tract Models

For the LPC method, all vocal tract characteristics are contained in the predictor coefficients. The procedure for transformation and scaling is discussed below.

For the simulation approach, a wave-reflection analog is used. The model requires that the three-dimensional vocal tract shape be discretized into a finite number of equal length cylindrical sections. Reflection coefficients based on the relative areas of adjoining sections are calculated at each cylinder junction and waves are propagated through the system by using the reflection coefficients to compute the incident and reflected components of the

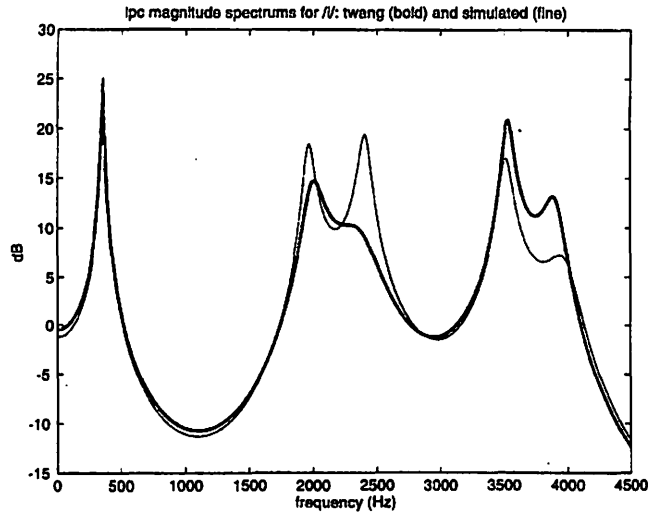

Figure 3. LPC spectra for a normal and twang speech vowel /i/.

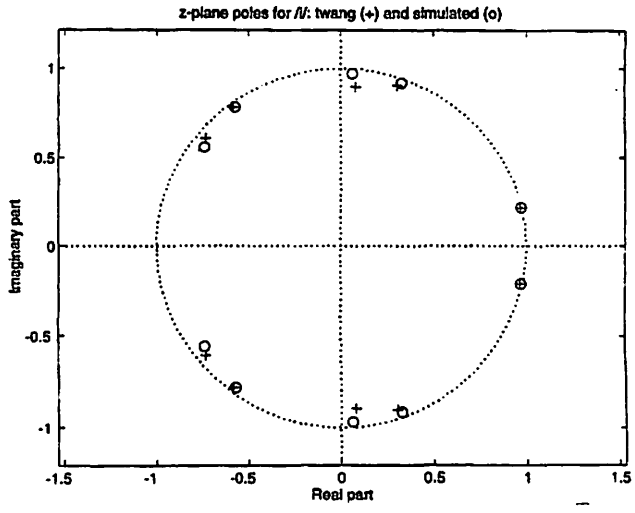

Figure 4. Pole representation in the complex plane for natural twang and simulated twang.

pressure or flow waves at each junction at each step in time. Additionally, energy losses due to the yielding properties of the vocal tract walls, fluid viscosity, and radiation from the mouth have been incorporated into the model.

The simulation is sampled at a frequency of 44.1 kHz and each finite section of the area function represents a tube length of 0.396 cm. The vocal tract length has been standardized to be 17.5 cm for an adult male, thus each area function is composed of 44 individual cross-sectional areas. The simulation of speech is performed by injecting the parameterized glottal flow waveform into the glottal end of the vocal tract.

A sequence of vocal tract area functions and voice parameters are fed into the speech simulator as an instruction set which dictates the time course of vocal and articulatory events. Each "instruction" specifies the point in time that the area function and voice model parameters should attain a desired value. This time coded instruction set is allowed to specify new parameter values at fixed time intervals or highly non-uniform intervals. Both area functions and voice parameters are interpolated in time, between consecutive target values to provide inputs to the speech simulator at every time sample (44100 per second). The voice parameters are linearly interpolated while the area functions are subjected to a second order filter for which understood or overshoot of a target could occur.

Transforming the Vocal Tract

Vocal tract scaling can be applied to gender and age, as well as to certain voice quality modifications, e.g. *twang*, *sob* and *ring*. This scaling may involve length changes as well as localized area changes.

Linear and non-linear length scaling using MRI extracted areas (Yang and Kasuya, 1995), has shown that differences in F_1 and F_2 resulting from each type of scaling often fall below the 5 percent perceptual difference linear. Thus, for experiments on gender and age changes, application of linear length changes are recommended.

For the example described here, the voice quality *twang* is of interest. This does require nonlinear manipulations of localized areas of the vocal tract. The study began with steady state vowels. An adult male subject, RS, a voice scientist and amateur singer, produced several vowels at a comfortable pitch. For each vowel, he was able to imitate normal speech and twang. These were recorded on DAT at a sampling frequency of 44.1 kHz, digitized into a Power Macintosh, and read into the Matlab software environment.

LPC Method

The LPC method is capable of generating a pseudo area function for each analysis frame, but its physiologic relevance is questionable. Thus, because the sensitivity of local area changes of pseudo-area vectors is not fully understood (vis-a-vis the equivalent changes in MRI data), we attempted the transformation in the pole-zero domain instead.

The speech was downsampled to 10 kHz, preemphasized by differentiation, and a 10th-order linear prediction algorithm was applied to model the vocal tract filter function. The linear prediction coefficients were then averaged over the duration of the vowel. Figure 3 shows the LPC spectra for the normal and twang cases. By using a low order model, the angle of the poles in the resulting LPC polynomial closely match the formants in the speech. The residual error was retained for synthesis of the glottal source signal from one of the excitation models.

The formant frequency ratios of a given imitated quality to normal speech were computed for all formants below 4500 Hz. A 4th-order polynomial estimate of these

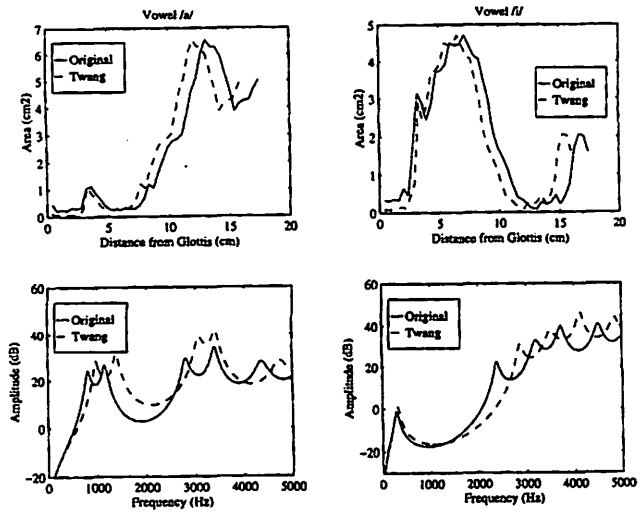

Figure 5. Vocal tract area functions (top) and frequency response functions (bottom) for the original measured male vowels (solid line) and the transformation to a twang-like quality (dashed lines).

ratios, as a function of the normal formant frequencies, was then derived. This enabled us to alter the first four or five formants of the voiced speech, and allowed the transformation process to be independent of sampling frequency.

To accomplish a transformation from a normal vowel to a vowel of another quality, we first found the complex conjugate poles of the normal vowel corresponding to frequencies between 100 and 4500 Hz. These were rotated by evaluating their angles in the 4th-order polynomial. Figure 4 shows the close proximity of the poles for natural twang and those obtained by the algorithm. The modified poles were recombined with any unaltered poles and used to generate a new LPC polynomial. Finally, the derivative flow source signal was filtered by the new vocal tract function to create the output speech.

Simulation Method

The simulation method allows direct manipulation of the vocal tract area functions that correspond to various speech sounds. For example, to achieve the twang quality, the adult male area functions reported in Story et al (1996) were first uniformly shortened by 8 percent and then the cross-sectional areas of the epilaryngeal section (approximately from the glottis to 2.5 cm above the glottis) were reduced by 75 percent. Figure 5 shows area function scaling for /a/ and /i/ vowels (top) and frequency response functions (bottom). The values of F_1 and F_2 are shifted upward in both cases. Also, the third and fourth formants have clustered together to give a spectral prominence around 3000-4000 Hz.

The area function manipulations and equivalently, the pole rotations for the LPC method are effectively "spatial" transformations as they alter the shape or size of an articulatory structure. The simulation method allows for a simpler temporal transformation of a speech signal, however. The rate of articulation can be altered by changing the time code that drives the specification of area function targets and voice parameters.

Sentence-Level Processing

Sentences spoken by a 31 year old adult male (BS), were recorded. The adult male had a mid-Western native-born American accent, was 5'6" in height, a non-smoker, with no history of voice problems.

The sentences recorded were:

1) "Goldilocks ran as fast as she could through the woods."
2) "We were away a year ago."
3) "The blue spot is on the key again."

In the examples given here, only sentence (3) was used for analysis and reconstruction. The recordings, made in an anechoic chamber, were initially recorded to digital audio tape (DAT) at a sampling rate of 44.1 kHz (Panasonic SV3700). The files were then digitized at 20 kHz using CSPEECH and downloaded to the MATLAB environment where other processing took place. Files were analyzed using pitch asynchronous LPC analysis.

Once the excitation and articulation components were captured, the residual (excitation component) was parameterized. Cycle marks corresponding to the maximum negative spike in each cycle in the pseudoflow derivative signal were extracted as a representation of the pitch period. The markings were stored with accuracy to the nearest sample. Since the residual is an approximation to the flow derivative, the amplitude of the negative spikes was stored as a representation of the maximum flow declination. The marking algorithm was based on autocorrelation and minimum mean squared error calculations. In some cases, the EGG was used to augment the marking algorithm. When the area/flow model is applied to the voiced excitation, the open quotient and skew quotient are measured from an arbitrary section of the flow residual, and these values are then held constant throughout the sentence.

LPC Method

The LPC/pole rotation method for voice transformation to the twang quality for the /i/ vowel was extended to sentence-level speech. The two sentences were analyzed using pitch asynchronous frames with 50% overlap. The LPC polynomial of each frame was then factored to obtain the poles, adjusted by the rotation polynomial, and reconstructed to produce the modified vocal tract filter function. No attempt was made to treat vowel and consonant segments of the sentences separately at the filter.

Simulation and Transformation

The techniques for extracting and transforming the F_o and amplitude contours from a speech signal specified for

the LPC method can also be used to derive the voicing parameters for the simulation method. However, the specification of appropriate area function targets and their corresponding time code are at this point manually obtained from the speech signal. For fricative consonants, a noise source based on a Gaussian random number generator is located just downstream of the vocal tract constriction for a given consonant. A trial and error process is used to fine tune the time code for all of the speech parameters.

Figure 6 shows three spectrograms of the sentence "the blue spot is on the key again"; white represents large amplitude (hence formants) while black is silence. The top spectrogram represents the natural recorded speech and the middle spectrogram is the simulation of the original speech. The bottom shows the attempt at transforming the speech to a twang-like quality by shortening the vocal tract and reducing the epilaryngeal cross-sectional areas.

In the region of F_1 and F_2, the simulation matches the original speech reasonably well. However, the upper formants are much more apparent in the simulation than the original, implying that the energy loses in the vocal tract model require some adjustment. The high amplitudes of the upper formants could also be due to a shallow spectral drop off of the glottal flow signal; thus, adjustment of the glottal flow parameters may be necessary. In the transformation to the twang quality, the formants are visibly shifted upward; this was also shown in Figure 5 for the /a/ vowel. Both simulations show appropriate timing of the plosive and fricative consonants, but formant transitions into and out off a consonant are sometimes exaggerated relative to the original speech.

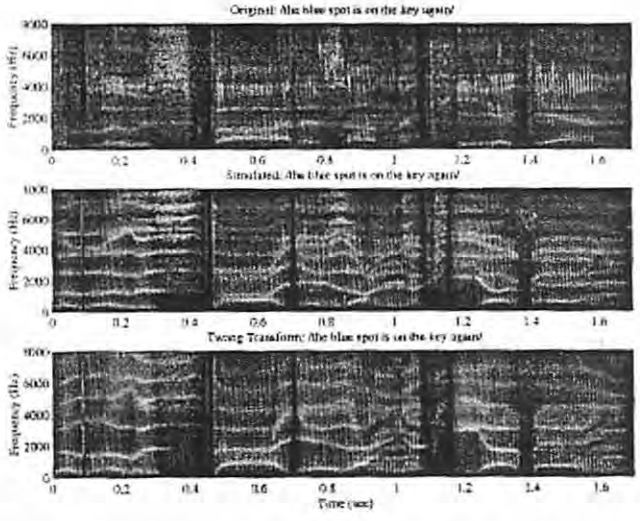

Figure 6. Three spectrograms of the sentence "the blue spot is on the key again" spoken by subject BS. The top spectrum represents the original speech, the middle is a simulation attempt to replicate the original and the bottom shows the transformation to a twang-like quality.

Discussion

Informal listening tests suggested to us that we had successfully produced natural sounding excitation models. Both the area/flow and the polynomial models produced acceptably natural speech when the articulation was not modified. The polynomial model captured the nuances of the excitation cycle better than the area/flow model, since it attempted to fit the dynamics of the cycle, whereas the area/flow model enforced a closure period. Using pitch asynchronous LPC as a decomposition technique may have made the area/flow model even less suitable, since this method of LPC is most susceptible to errors due to bandwidth estimation and 'leakage' between the excitation and articulation estimates. Perhaps a pitch synchronous or closed phase LPC method would produce a pseudoflow pulse that is more like the flow pulse expected by the area/flow model. While the area/flow model did not capture the dynamics of the excitation as well as the other model, further experiments not reported here, using age and gender-related measurement data from the literature indicated that the area/flow model successfully takes into account age and gender-related lung pressure, vocal fold length and vibrational amplitude measurements as variables.

It should be noted that the estimation of the open quotient in the area/flow model is quite difficult. Matching the magnitudes of the spectral harmonics of the flow pulses is the best criteria, although it requires an iterative process that combines an estimation of a threshold in the time domain with error minimization in the spectral domain. In this study, this optimization process was not automated.

Keeping the open quotient constant does not decrease the naturalness of the speech, in agreement with studies by Ananthapadmanabha (1995) and Childers (1989). This concept is very useful for very efficient encoding, as a single polynomial coefficient set can be transmitted to represent a sentence or even a conversation, along with an accurate intensity contour.

In informal listening tests, the LPC method was able to produce acceptable voice qualities in the sentence level speech. Although there was a slight loss of naturalness, the speaker BS was clearly recognizable, and the target qualities were easily identified. More natural and convincing results could probably be achieved by extending the analysis algorithm to include vowels other than /i/.

Informal listening of the simulated speech indicated that the speech was highly intelligible but was not a true "copy" of the original speech. The simulated speech seems to take on a character of its own. Thus, even when a replication of the original speech is attempted, a transformation actually results. The transformation to a twang-like voice quality seemed to produce a reasonably convincing result but much work remains to fine tune the vocal tract energy losses and timing/interpolation schemes for traversing the vocal and articulatory targets.

Conclusions

The physiological domain transformation is presently in the initial stages of development and implementation. It provides a means for coordinating a combination of different vocal effects at various levels of production, and it appears to be a natural choice for simulating a variety of qualities exhibited in pathology and vocal performance.

The area/flow model has been successfully used to describe changes due to age and gender. Nominal parameter values produce ac flow and flow declination values that are within the expected ranges of reported data, and successfully characterize the differences due to gender and age. This is reported in a study by Wong et al. (1996).

At present, the matching of MRI image targets to the acoustic signal sequence is performed manually. In the future, an (automated) neural network mapping from acoustic signal frames to vocal tract area functions will be used. A multi-layer perceptron (Rahim, 1994) will be used to provide a mapping to continuous area function estimates for voiced intervals. For unvoiced sounds, a learning vector quantizer (Kangas, Torkkola and Kokkonen, 1992) may be used to classify consonant speech to normalized consonant-area estimates.

Acknowledgment

This work is supported by research grant number 5 R01 DC0232-01 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

References

T.V. Ananthapadmanabha (1995), "Acoustic factors deterining perceived voice quality," in *Vocal Fold Physiology: Voice Quality Control*, ed. by O. Fujimura and M. Hirano (Singular Publishing Group, San Diego, California, USA), pp. 113-126.

T.V. Ananthapadmanabha and J. Estill (1989), "Analysis and synthesis of six voice qualities," *J. Acoust. Soc. Amer.*, Vol. 86, p. S36.

D.G. Childers & H.T. Hu (1994), "Speech synthesis by glottal linear prediction," *J. Acoust. Soc. Amer.*, Vol. 96, pp. 2026-2036.

D.G. Childers, K. Wu, D.M. Hicks & B. Yegnanarayana (1989), "Voice conversion," *Speech Communication*, Vol. 8, pp. 147-158.

R. Colton & J. Estill (1978), "Mechanisms of voice quality variation: Voice modes", in *Transcripts of the Seventh Symposium Care of the Professional Voice*, ed. by V.L. Lawrence, (New York, NY: The Voice Foundation), pp. 71-79.

G. Fant, J. Liljencrants & Q. Lin (1985), "A four parameter model of glottal flow," *STL-QPSR 4/1985, Speech Transmission Laboratory*, (Royal Institute of Technology KTH, Stockholm, Sweden), pp. 1-13.

E.B. Holmberg, R.E. Hillman & J.S. Perkell (1988), "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voice". *J. Acoust. Soc. Amer.*, Vol. 84, pp. 511-529.

S. Imaizumi, S. Kiritani, & S. Saito (1991), "Perceptual evaluation of a glottal source model for voice quality control," in *Vocal Fold Physiology: Acoustic, Perceptual and Physiological Aspects of Voice Mechanisms*, ed. by J. Gauffin & B. Hammarberg (San Diego, CA: Singular Publishing Group), pp. 233-242.

J. Kangas, K. Torkkola, & M. Kokkonen (1992, March), "Using SOMS as feature extractors for speech recognition". *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 341-344),

I. Karlsson & L. Nord (1970), "A new method of recording occlusion applied to the study of Swedish stops", *Speech Transmission Laboratory-Quarterly Progress and Status Report 2/3*, 8-18.

P.H. Milenkovic (1993), "Voice source model for continuous control of pitch period,"*J. Acoust. Soc. Amer.*, Vol. 93, pp. 1087-1096.

J.S. Perkell, R.E. Hillman & E.B. Holmberg (1994), "Group differences in measures of voice production and revised values of maximum airflow declination rate," *J. Acoust. Soc. Amer.*, Vol. 96, pp. 695-698.

M.G. Rahim (1994), *Artificial Neural Networks for Speech Analysis/ Synthesis* (London, Chapman and Hall).

E.T. Stathopoulos & C.M. Sapienza (1993a), "Respiratory and laryngeal function of women and men during vocal intensity variation," *J. Speech Hear. Res.*, Vol. 36, pp. 64-75.

E.T. Stathopoulos & C.M. Sapienza (1993b), "Respiratory and laryngeal measures of children during vocal intensity variation," *J. Acoust. Soc. Amer.*, Vol. 94, pp. 2531-2543.

B.H. Story, I.R. Titze, & E. Hoffman (1996), Vocal tract area functions form magnetic resonance imaging. *J. Acoust. Soc. Amer.*, Vol. 100, pp. 537-554.

I.R. Titze (1983), "Synthesis of sung vowels using a time-domain approach", in *Transcripts of the Eleventh Symposium: Care of the Professional Voice*, ed. by Van L. Lawrence, (New York: The Voice Foundation), pp. 90-98.

I.R. Titze & B.H. Story (1996), "Acoustic interactions of the voice source with the lower vocal tract," submitted to the *J. Acoust. Soc. Amer.*, July, 1996.

I.R. Titze, S. Mapes & B.H. Story (1994), "Acoustics of the tenor high voice," *J. Acoust. Soc. Amer.*, Vol. 95, pp. 1133-1142.

D. Wong, R.C. Lange, R.K. Long, B.H. Story, & I.R. Titze (1996), "Age and gender related speech transformations using linear predictive coding", submitted to *J. Acoust. Soc. Amer.*

E. Yanagisawa, J. Estill, S. Kmucha & S. Leder (1989), "The contribution of aryepiglottic constriction to 'ringing' voice quality - A videolaryngoscopic study with acoustic analysis," *J. Voice*, Vol 3, pp. 342-350.

C-S Yang & H. Kasuya (1995), "Uniform and non-uniform normalization of vocal tracts measured by MRI across male, female and child subjects," *IECE Trans. Inf. And Systems*, Vol. E78-D, June, pp. 732-737.

NCVS Status and Progress Report - 10
November 1996, 111-126

Age and Gender Related Speech Transformations Using Linear Predictive Coding

Darrell Wong, Ph.D.
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts
Robert C. Lange, Ph.D.
Texas Instruments, Dallas, Texas
Russel K. Long, Ph.D.
Wilbur James Gould Voice Research Center, The Denver Center for the Performing Arts
Brad H. Story, Ph.D.
Ingo R. Titze, Ph.D.
Department of Speech Pathology and Audiology, The University of Iowa

Abstract

An effective transformation of sentence level speech from aged to young, or male to female, requires the parameterization and modification of various components in the speech. In this paper, experiments in voice-to-voice transformation which modify speech from a male adult to that of a non-specific male child, and from a male adult to a non-specific female adult, are described. A Linear Prediction based approach is used, permitting an initial parameterization into excitation and articulation components. Both the excitation signal and the articulatory components are then either modified or replaced with target-related equivalents. This study compares two models of the glottal pulse for excitation replacement. The first model includes the vocal fold length, maximum tissue displacement, and lung pressure as variables in the calculation of the flow pulse, thus permitting adjustment for age and gender-related growth differences. The second model uses a polynomial fitting procedure applied to the LPC residual. Articulation data, represented as LPC coefficients, can be transformed by projection to the pole-zero domain, followed by modifications to reflect vocal tract length differences. Modifications to pitch and intensity contours are also discussed. An attempt has been made in this study to relate the proposed modifications in articulation and excitation to published physiological measurements.

Introduction

An interesting area within speech research is speech transformation and mimicry. The generation of speech that perceptually evokes different accents, voice qualities, ages, and genders, derived from only one normative database of speech, would be useful in the telephony and entertainment fields. A thorough understanding of the age and gender related parameters that influence voice quality would also be useful from a clinical perspective. Comprehensive simulation models of speech might then be used as predictive tools in the treatment of voice and speech disorders.

An examination of the anatomy of the larynx and articulatory organs for males and females of different ages suggests that transforming speech may not be a simple linear rescaling of pitch and tract size. A recent Magnetic Resonance Imaging (MRI) study of an adult male, an adult female, and a ten year old boy, by Yang and Kasuya (1994) has indicated that within the vocal tract, the proportions associated with the epiglottis, pharynx and mouth regions vary with age and gender. Many studies of the larynx have also noted significant size, tissue morphology and framework disproportionalities between males and females and between young and old (Hirano 1983).

A question that must be answered prior to constructing a nonlinear scaling model is whether these structural disproportionalities produce measurable differences in acoustic or physiologic variables. Stathopoulos and colleagues have conducted a number of studies making mea-

surements of acoustic and respiratory function, comparing men, women and children (Stathopoulos and Weismer, 1986; Stathopoulos and Sapienza, 1993a,b, 1996; Tang and Stathopoulos, 1995). Perkell et al. (1994) have conducted similar studies comparing adult males and females. The results indicate that parameters such as F_0, flow pulse height, open quotient, and maximum flow declination rate exhibit age or gender differences. Titze (1989) studied the acoustic differences arising from the laryngeal size differences between male and female larynges. His results indicated that F_0 is scaled according to the membranous length of the vocal folds, while mean airflow and amplitude of vibration are scaled according to overall larynx size.

Regarding the vocal tract, Yang and Kasuya (1994, 1995) found that the relative length relationships between the epiglottis, pharynx, and oral regions were different for males, females, and children. They linearly and nonlinearly scaled the child and female vocal tracts to the male vocal tract length. The linear scaling preserved the child's (or female's) regional length relationships, while the nonlinear scaling produced the male's. The scaling techniques produced formant shifts for F_1 and F_2 that were less than 5% apart i.e. below the perceptual difference limen levels defined by Flanagan (1972). Yang and Kasuya (1995) suggested that scaling for growth differences (other than length) is a secondary contributor to the explanation of formant differences due to age and gender. Other studies on male-female differences (e.g. Hogberg 1995), however, have found that the growth adjustments do contribute, but can account for only a portion of the formant differences. It thus appears that the necessity of vocal tract scaling (beyond simple length scaling) is somewhat controversial.

It is apparent that some of these effects need to be modeled if high quality speech transformations are to be successful. One approach is to use a biomechanical model of the larynx coupled to a biomechanical model of the vocal tract. These models, which would solve all the partial differential equations for air and tissue motion, would contain the boundary dimensions for any type of scaling. Unfortunately, these models are presently incomplete. The leap to sentence level speech is difficult, since input data describing the time varying vocal tract and larynx structures are difficult to obtain empirically. Additionally, the computational burden is extremely high for this type of model.

A simpler approach, in which the source function is controlled by external parameters, is used for the current study. Two models are compared and contrasted. The first is a glottal area and flow model, originally introduced by Titze (1982) and further described by Titze, Mapes and Story (1994). This model prescribes a glottal area by equation. The pitch period and the open quotient (the proportion of time that the glottis is open) are measured from the LPC residual pseudoflow, and used to construct the shape of the glottal area pulse. Rothenberg's low frequency model of the glottis and vocal tract system (Rothenberg 1981) is then used to determine the glottal flow. The skewing of the glottal flow pulse is also estimated from the pseudoflow and applied as a control parameter for the flow. This skewing quotient emulates the effect of the vocal tract inertance. The model has been extended in this paper to include vocal fold length and maximum glotttis diameter as variables, allowing age and gender differences in the structure of the larynx to be captured. Additionally, the glottal noise from flow turbulence is emulated using pseudorandom noise with an appropriate spectrum whenever orifice and flow conditions require it.

The second excitation model parameterizes the glottal flow derivative rather than the glottal area, and is identical to a model proposed by Childers and Hu (1993). The flow derivative, represented by the LPC residual, is modeled by a high order polynomial for each cycle. Turbulence can be added as an additional contribution after the polynomial has been fitted. Milenkovic (1993) also proposed a model of this nature. These models are also similar to the one proposed by Imaizumi, Kiritani and Sato (1991), except that Imaizumi modeled the derivative with three piece-wise sections, using a parabola, a constant, and a cubic, rather than a single polynomial. The single polynomial provides a better fit to the data since it does not assume a closed phase in the flow (the constant section in Imaizumi's model).

Another parametric model of the glottal flow was developed by Fant, Liljencrants and Lin (1985). Known as the LF model, four parameters were used to define various timing and slope observations in the inverse-filtered flow derivative. Fujisaki and Ljungqvist (1986) developed a more complex seven parameter version of the LF model. The LF model will not be reviewed in this study, since it has been well documented in the literature (for a recent review, see (Fant, 1995)).

Both the LF-type and the polynomial-type models were developed to describe the flow derivative, since this derivative signal is assumed to be the excitation signal for the vocal tract. They do not, however, make optimal use of physiological measurements of the shape and structure of the vocal fold in their definition, and thus do not provide enough insight into the influences that parameters such as lung pressure and vocal fold length have on the system.

In this study, Linear Predictive Coding (LPC) has been used to obtain articulatory information. This information is in the form of filter coefficients, held constant over 25 millisecond frames, which are excited by a the LPC residual extracted during analysis. Pole rotations are used for overall length rescaling. The age dependent epiglottal, pharyngeal, and mouth length relationships described by Yang and Kasuya were not used to warp the vocal tracts, following their recommendations. An experiment using Yang's length relationship data, applied to new MRI data was conducted to determine the generalizability of Yang's conclusions.

Public Health Department or the State Education Department. P.L. 99-457 mandates services from birth through age five, with P.L. 94-142 mandating services for children from age five through twenty-one. Other governmental agencies (i.e. Head Start) also will provide screening free of charge to qualified families and refer for further diagnostic testing.

Summary

Pediatric patients with voice or speech problems should usually receive a team assessment where communication between the pediatrician or primary care physician, the otolaryngologist and speech pathologist occurs. Although speech or voice problems may prompt an otolaryngologic evaluation, the voice or speech problem may simply be the manifestation or symptom of a larger or more complex disease process. Whether that is the case of hypernasal speech eventually leaving to the diagnosis of velocardiofacial syndrome or bilateral vocal fold paralysis eventually leading to the diagnosis of hydrocephalus, it is apparent that a speech or voice disorder may eventually require multi-disciplinary evaluation.

The outlook for children with speech and voice difficulties is better then ever. Recent equipment advances such as the flexible laryngoscopy, videostroboscopy and nasometry for detection, evaluation and management of speech problems has created a better environment then ever existed for care of these problems. Much research is being performed in the area of pediatric voice and speech problems. Both the National Institute of Deafness and Communicative Disorders, as well as the National Institute of Dental

Table 5. Physician's Checklist for Referral

	The Child with Normal Disfluencies Age of Onset: 1½ - 7 years	**The Child with Mild Stuttering** Age of Onset: 1½ - 7 years	**The Child with Severe Stuttering** Age of Onset: 1½ - 7 years
Speech behavior you may see or hear:	Occasional (not more than once in every 10 sentences), brief (typical ½ second or shorter) repetions of sounds, syllables or short words, e.g. li-li-like this.	Frequent (3% or more of speech), long (½-1 second repetions of sounds, syllables, or short words, e.g. li-li-li-like this. Occasional prolongations of sounds.	Very frequent (10% or more of speech), and often very long (1 second or longer) repetions of sounds, syllables or short words. Frequent sound prolongations and blockages.
Other behavior you may see or hear:	Occasional pauses, hesitations in speech or fillers such as "uh", "er", or "um", changing words or thoughts.	Repetitions and prolongations begin to be associated with eyelid closing and blinking, looking to the side, and some physical tension in and around the lips.	Similar to mild stutters only more frequent and noticeable; some rise in pitch of voice during stuttering. Extra sounds or words used as "starters".
When problem is most noticeable:	Tends to come and go when child is: tired, excited, talking about complex/new topics, asking or answering questions or talking to unresponsive listeners.	Tends to come and go in similar situations, but is more often present than absent.	Tends to be present in most speaking situations, far more consistent and non-fluctuating.
Child reaction:	None apparent	Some show little concern, some will be frustrated and embarassed.	Most are embarrassed and some sare also fearful of speaking.
Parent reaction:	None to a great deal	Most concerned, but concern may be minimal.	All have some degree of concern.
Referral decision:	Refer only if parents moderately to overly concerned.	Refer if continues for 6 to 8 weeks or if parental concern justifies it.	Refer as soon as possible.

Reprint courtesy of Stuttering Foundation of America
(800) 992-9392

A previous study by Childers, Wu, Hicks and Yegnanayarana (1989) on gender transformation used a similar LPC-based methodology to the one described here. They used a variety of excitation models, such as the LF model, the electroglottograph derivative, and one-impulse and three-impulse models to excite the LPC filters. Although the transformation of articulation parameters uses similar methods, the major difference in our study is the use and comparison of more sophisticated excitation models. It was hoped that this would lead to higher voice quality in both mimicry and transformation.

This paper begins with a description of the LPC analysis procedure. The two excitation models are then described. The glottal area/flow model is extended to include age and gender dependent parameters and a model of aspiration. Anatomical data from the literature are used to develop models of the child and female glottal pulse, which are then compared to flow measurements reported elsewhere. Following this, experiments in voice mimicry and transformation are described. We initially mimic sentence level speech produced by an adult male. A transformation from the male's speech to that of a child is then attempted, followed by a transformation from the adult male to an adult female. Recordings from a 10 year old boy and a 33 year old female are used as aids in this process. In the following sections, the methodology and the experiments are presented.

Methods

LPC Analysis

Sentences spoken by a 31 year old adult male BS, a 10 year old boy, KM, and a 33 year old female, KB were recorded. The adult male had a mid-Western native-born American accent, was 5'5" in height, a non-smoker, with no history of voice problems. The boy had a native-born Western Canadian accent, was 4'10" in height, with no history of voice problems. The adult female had a mid-Western native-born American accent, was 5'6" in height, a non-smoker, with no history of voice-problems.

The sentences recorded were

(1) "Goldilocks ran as fast as she could through the woods."
(2) "We were away a year ago."
(3) "The blue spot is on the key again."
(4) Steady state /a/, /i/, and /u/ phonations.

Only sentence number (1) was used for analysis and reconstruction in this study.

The recordings, made in a recording booth, were initially recorded to Digital Audio Tape at a sampling rate of 44.1 kHz (Panasonic SV3700). The microphone used was an AKG C410 held 8 cm from the lips (45 degrees off-normal axis). The files were then digitized at 20 kHz using CSPEECH and downloaded to the MATLAB environment on an Apple PowerMac 8500, where all off the processing took place. Files were analyzed using pitch asynchronous LPC analysis. The LPC coefficient vector for each frame was stored without quantization. The speech signal was pre-emphasized prior to analysis so that the residual error signal resembled a glottal flow derivative signal at the output of the analysis. The parameters for the LPC analysis appear in Table I.

Once the excitation and articulation components were captured, the residual (excitation component) was parameterized. Cycle markers corresponding to the maximum negative spike in each cycle in the pseudoflow derivative signal were extracted as a representation of the pitch period. The markings were stored with accuracy to the nearest sample. Since the residual is an approximation to the flow derivative, the amplitude of the negative spike was stored as a representation of the maximum flow declination rate. The marking algorithm was based on autocorrelation and minimum mean squared error calculations.

Figure 1a shows the microphone signal for the word "Goldilocks" from the sentence "Goldilocks ran as fast as she could through the woods". The residual obtained from

Table I.
Parameters for LPC analysis

Parameter	Value
Fs (sampling rate)	20 kHz
frame size	25 ms, fixed size
window	Hanning
window overlap	50 %
number of LPC coefficients	23
pre-emphasis	by numerical differentiation
rumble filter	5 Hz high pass

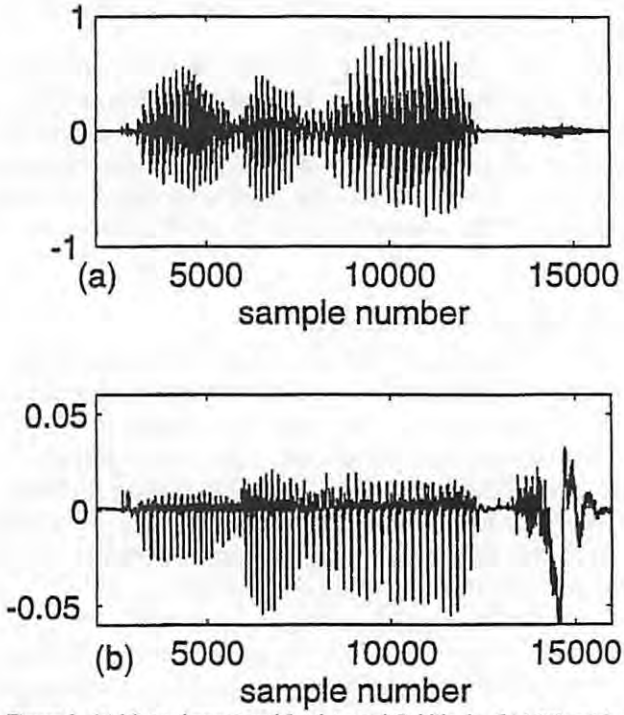

Figure 1. (a) Microphone signal for the word Goldilocks. (b) LPC residual extracted from the word Goldilocks.

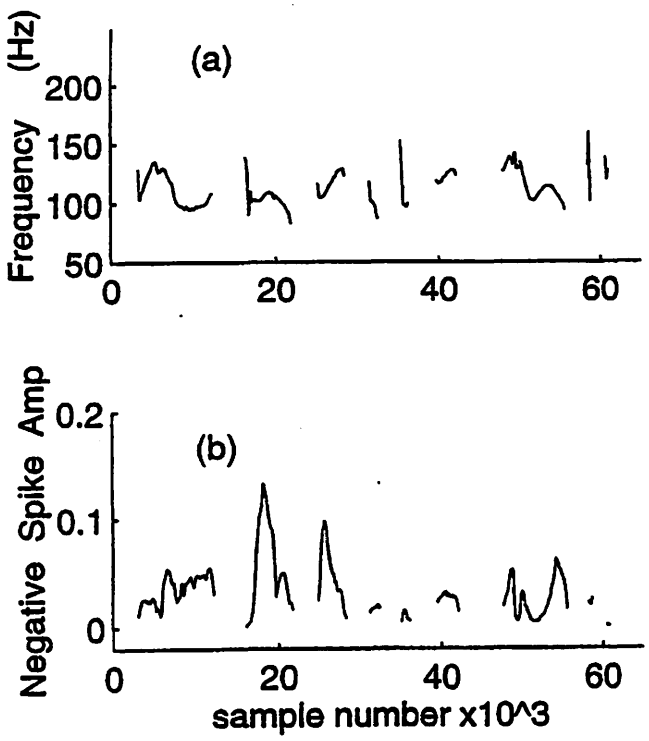

Figure 2. (a) F_0 contour for the sentence "Goldilocks ran as fast as she could through the woods", spoken by subject BS. (b) the flow derivative spike contour (equivalent to mfdr) extracted from the LPC residual for the same sentence and subject.

the LPC analysis appears in Figure 1b. The microphone signal illustrates the voiced nature of most of the speech, with frication from point 13000 onwards. Figure 1b shows how the amplitude of the negative spike train approximately follow the envelope of the speech in Figure 1a, suggesting a good correlation with intensity.

In Figure 2, the F_0 and intensity (maximum flow declination rate) contours are plotted for the sentence. The F_0 is extracted as the inverse of each pitch period. The contours display the voiced portions only. The unvoiced portions were eventually reconstructed from unmodified residual segments. Full speech reconstruction used an overlap-add procedure.

The Glottal Area/Flow Model

After decomposition, the reconstruction process was carried out by substituting the voiced excitation sections with flow models derived from the parameterization described above. In this section the glottal area/flow model is described. Data presented in Titze (1989), Titze et al. (1994), Perkell et al. (1994), and Stathopoulos et al. (1996), are used to confirm the validity of the model. In section C the polynomial excitation model is discussed.

A thorough description of the area/flow model appears in the paper by Titze, Mapes and Story (1994). Here we summarize the equations and then add or extend where necessary for the purposes of this study. The reader is referred to their paper for further details.

A model for glottal flow may be constructed using the analogy of an electric circuit with a time-varying resistance (Rothenberg, 1981). The equation summing the voltage drops is

$$P_L = R_g u + I u' \quad (1)$$

$$= 1/2\, k_t \rho |u| u / a^2 + I u' \quad (2)$$

where P_L is lung pressure, R_g is the nonlinear kinetic resistance representing the time-varying glottal impedance, and I is the lumped inertance of the vocal tract air column i.e. the vocal tract impedance. The flow and its time derivative are represented by u and u'. In equation (2), k_t is the translaryngeal pressure coefficient, ρ is the density of air, and a is the minimum glottal area. The area a and flow u are functions of time, k_t is a function of a, and I is constant over the F_0 cycle.

A discrete solution for u in (2) may be obtained by replacing u with u_n and u' with the backward difference formulation $(u_n - u_{n-1})/\Delta$. The parameter n is the n^{th} sampling period, and Δ represents the sampling interval. This leads to a recursive solution for u:

$$u_n = a v_0 (\pm (1 + \delta^2 + (2\delta / a v_0) u_{n-1})^{1/2} - \delta) \quad (3)$$

where

$$v_0 = (2P_L/(k_t \rho))^{1/2} \quad (4)$$

$$\delta = (I/(2P_L\Delta)) a v_0 = \gamma a v_0 \quad (5)$$

The quantity v_0 is the no-load (I=0) particle velocity, and δ is the inertial load factor that delays the buildup of flow when the glottis opens. The index n represents the time from the beginning to the end of a pitch period.

The glottal area function a is approximated by

$$a(\theta) = max(0, sin^{\beta}(\theta)) \quad (6)$$

where

$$\theta = \pi n / (Q_0 T_0) \qquad 0 < \theta < \pi,\ 0 < n < T_0 \quad (7)$$

The pitch period is T_0, the open quotient Q_0 is the fraction of the cycle when the glottis is open, and the exponent β is a shaping parameter that measures the softness of onset and offset of the pulse. A rule defining the covariance of Q_0 and β is

$$\beta = Q_0 + 1 \quad (8)$$

A skewing quotient Q_s defines the ratio of rising to falling portions of the glottal pulse i.e. $Q_s = T_p/T_n$, where T_p is the period of rising flow and T_n is the period of falling flow. The skew has been empirically related to γ by the equation

$$\gamma = 0.080(Q_s - 1) \quad (9)$$

As stated by Titze et al. (1994), Q_0 governs the spectral slope and Q_s affects the depth of the spectral valleys. The scale factor 0.080 differs from the value of 80,000 published by Titze because of the cgs units employed here. Using (5), δ can be related to γ instead of I, resulting in a modified Equation (3):

$$u_n = av_0\left(\pm \left(1 + (\gamma a v_0)^2 + 2\gamma u_{n-1} \right)^{1/2} - \gamma a v_0 \right) \quad (10)$$

By choosing values of Q_0 and Q_s, the parameters β and γ are first computed, followed by $a = a(\theta)$. The no-load velocity v_0 is calculated from a nominal lung pressure P_L and an empirical estimate of k_t, and then the flow u_n is obtained from (10). In Titze's 1994 study, k_t was assumed constant at 1.1 as a first approximation, while P_L was nominally chosen to represent a male's lung pressure. In the following section, we describe how these parameters and a may be modified according to age and gender.

Extensions to the Model

The following paragraphs describe the additional equations, and the age and gender-specific data, required to apply the model to female and child target transformations.

Orifice dependent translaryngeal coefficient: The k_t parameter varies as the glottal opening area a changes through the pitch period. Scherer and Lange (1996) empirically developed an equation for calculating k_t using the data in Table I of Scherer and Guo (1991), (reproduced in Table II below). In Table II, the k_t values at translaryngeal pressures of 5, 10, and 15 cm H_2O are listed for seven different orifice diameters(d). Each element in the table represents the mean value of k_t for the angles -40, -20, -10, -5, 0, 5, 10, 20, and 40 degrees.

Table II.
Orifice diameter d versus k_t From Scherer and Guo (1991).

Diameter d (cm)	k_t at 5 cm H_2O	k_t at 10 cm H_2O	k_t at 15 cm H_2O
0.32	1.03	-	-
0.16	1.01	1.02	-
0.08	1.05	0.984	0.985
0.04	1.05	0.961	0.948
0.02	1.30	1.20	1.15
0.01	2.89	1.98	1.71
0.005	11.7	6.54	4.49

The data in the table were reformulated as a new coefficient

$$T_k = 1/(k_t - 1/d) \quad (11)$$

T_k was then least-squares fitted producing the following third order polynomial:

$$T_k = -2.112d^3 - 0.8477d^2 - 1.004d - 0.0000993 \quad (12)$$

T_k is a function of d, the time varying diameter of the glottis. The table data were used assuming that pressure was not a variable. The translaryngeal coefficient at any time in the glottal cycle was then extracted by the inverse transformation

$$k_t = 1/T_k + 1/d \quad (13)$$

The value of k_t can be made gender and age dependent via d, as will be seen in the following discussion.

Vocal fold length: A normalized area function a was specified by equation (6). This function must be scaled to produce the correctly sized glottal orifice as a function of time. The glottal area may be parameterized as a function of the effective glottal diameter d (the lateral gap between the open vocal folds), and the membranous vocal fold length during motion, L,

$$a = dL \quad (14)$$

assuming for simplicity a rectangular glottis. L is assumed to be constant during a cycle.

An equation for L derived by Titze (1989) from data produced by Hollien and Moore (1960) shows L to be proportional to the square root of the fundamental frequency F_0 (externally provided from the pitch extraction analysis) and a power of the vocal fold rest length L_0

$$L = 0.038(L_0)^{1.6}(F_0)^{1.} \quad (15)$$

where L_0 is set to 1.6 cm for adult males, 1.0 cm for adult females and 0.9 cm for the ten year old child in this study. These values were obtained from Titze (1989), in which data from Hirano (1983) were used to describe the membranous length of the vocal fold as a function of age and gender.

Vocal fold maximum amplitude: The other factor that determines the maximum value of a is the glottal orifice diameter d. This diameter is related to the maximum displacement amplitude A of tissue vibration. According to Titze (1989), a typical male/female ratio of A is 1.2, based on flow data measured by Holmberg (1988). Assuming that a typical vibration amplitude for a male vocal fold is 0.07 cm and that a typical prephonatory displacement (from the mid-line) is about 0.02 cm (Wong 1991), the mid-line to maximum displacement is then 0.09 cm, close to the 1mm sug-

gested by Titze (1991). The maximum glottal diameter d_{max} (for normal phonation) is therefore 0.18 cm. Using the previously defined A ratio, a female would therefore have a glottal diameter of 0.15 cm. No data for a child's d_{max} are available, so we provide an estimate. Since children have smaller vocal fold lengths, they are likely to have smaller tissue amplitudes. Our initial estimate is 80% of the female value of A. Applying the ratio 0.8/1.2 to a nominal 0.18 cm yields a glottal diameter d_{max} of 0.12 cm and an amplitude A of 0.04 for the child. The normalized area a from (6) may then be scaled to a maximum Ld_{max}. As the output ac flow for children has been measured by Stathopoulos and Sapienza (1996), the output flow from the model can be compared to this data, and the appropriateness of the A value for the child may be inferred.

Lung pressure: The lung pressure P_L influences the flow calculation in (10) via the no-load particle velocity v_0 calculated in Equation (4). Stathopoulos and Sapienza (1996) have measured tracheal pressure for male and female adults and children at a number of ages. For a medium loudness level, mean tracheal pressures were 5.23 cm H_2O for adult males, 4.07 cm H_2O for adult females, and 8.66 cm H_2O for 10 year old boys. According to Stathopoulos et al. these values were lower than previously reported (e.g., Perkell's male and female (adult) data were 5.9 and 5.5 cm H_2O, respectively). For this study, the averages between Perkell and Stathopoulos' data for adults were used, 5.57 and 4.78 respectively, while retaining the child's lung pressure value of 8.66.

Aspiration noise during voicing: Glottal aspiration is assumed to be triggered by the magnitude of the Reynolds number of the flow u. The leakage or dc flow in the glottis contributes to this. The Reynolds number is then

$$Re = (u + u_{dc})\rho / (L\mu) \qquad (16)$$

where μ is the air viscosity. The maximum range of dc flow magnitude has been experimentally determined (during this study) as 1600 cm^3/s. A value of 0% has no noticeable hoarseness, while 100% sounds very hoarse. In the model, u_{dc} is entered as a control parameter like Q_0 and Q_s, specified as a percentage of the maximum. The dc flow helps in creating breathiness in the voice by lifting the pulse so that the Reynolds number threshold is exceeded for a longer portion of each cycle.

Turbulent noise flow is generated in two steps. First, a pseudo-random number sequence is generated with appropriate spectral characteristics. The sequence is then scaled in proportion to the Reynolds number of the non-aspirated flow (equation (16)) according to the following equation, which proportions and triggers the flow noise:

$$s_{noise} = (Re^2 - 1800^2)/s_{noisemax} \qquad Re > 1800 \qquad (17)$$
$$s_{noise} = 0 \qquad Re < 1800$$

Equation (17) generates a sequence of scale factor values for each point in the cycle. The sequence is shaped in time to be maximum at the instant of peak flow. Figure 3a displays the scaling parameter s_{noise} for one cycle. The pseudorandom number sequence is obtained from the LPC residual to maintain the residual's spectral characteristics. The residual is first numerically integrated so that it resembles a flow wave. The flow above a 90% threshold of the maximum in each cycle is then fitted with a fourth order polynomial. For every voiced cycle, the error between the fitted polynomial and the residual above the threshold is then extracted, zero-meaned, Hanning-windowed, and zero-padded to 1024 points. An FFT is calculated for each cycle, and the mean FFT across all cycles is calculated. This mean FFT spectrum represents the aspiration noise with appropriate spectral characteristics.

The magnitude of the mean FFT is assumed to be associated with a linear phase spectrum. To create a noise source with the same spectral magnitude characteristics as the mean FFT, the phase values are reordered using a uniformly distributed sampling procedure to produce a pseudorandom phase. The inverse FFT, using the new phase with the mean FFT magnitude, is calculated to give a new pseudo-random time sequence. The zero padding performed in the analysis thus ensures an adequate length for the sequence. All this is done prior to speech reconstruction.

During reconstruction, for each cycle, a pitch period of the pseudo-random sequence is randomly extracted. Figure 3b shows the pseudo-random noise sequence. The data are then scaled by multiplication with the sequence generated from Eq. (17) to produce a noise sequence

$$u_{noise}(n) = s_{noise}(n)\, r(n) \qquad (18)$$

where $r(n)$ is the pseudo-random sequence element. It is also scaled to counteract the effect of zero padding. This is shown

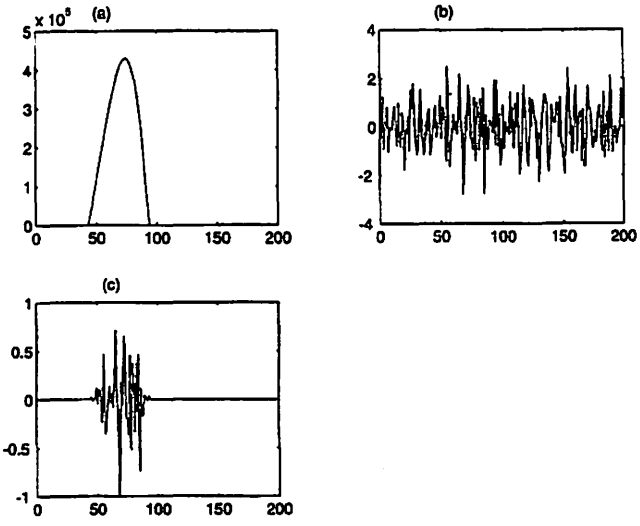

Figure 3. (a) Reynolds number scaling factor signal, extracted from the area/flow model assuming an adult male pulse. (b) turbulent noise sequence. (c) noise scaled by Reynolds number scaling factor signal.

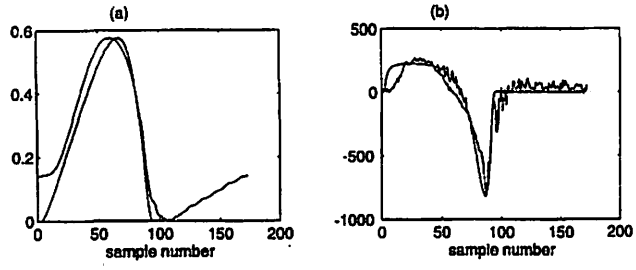

Figure 4. (a) flow u *plus turbulent noise, after scaling in preparation for reconstruction. The signal beginning closest to zero is the area/flow model approximation to a flow pulse generated form the residual integral. (b) the derivatives of the flow signals. The smooth signal is the area/flow model approximation.*

in Figure 3c. The aspiration u_{noise} is then added to the u_n calculated from (10), after u_n has been appropriately scaled to match the LPC residual magnitude.

Open and skew quotients: The open and skew quotients are parameters of the model. For transforming speech, it is useful to know if there are gender or age differences in these quotients that should be applied to the model. Stathopoulos (1996) measured Q_0 of 0.55, 0.70 and 0.66, respectively, for adult males, females and ten year old boys. Thus, given a specific male with a Q_0 of 0.50, the ratio 0.50(0.66/0.55) can be used to produce a child's voice.

The skewing quotient Q_s is generally changed by alterations in speaking style (articulatory changes affect the vocal tract load) or effort level. In our model, a nominal male value of 1.9 is first used, and the maximum flow declination rate (mfdr) described earlier modifies the glottal pulse and alters the final skew.

Scaling the model to suit the LPC residual: Figures 4a and 4b plot the flow and its derivative after scaling according to the mfdr, followed by the addition of the aspiration. Scaling is done according to our mfdr ratio (described in the following section) for any gender or age transformations required. Note that the units are arbitrary, as the LPC analysis produces signals that have been scaled due to data conditioning requirements.

Comparison Data

The data used to model the desired age and gender changes appear in Table III. The Q_s value is nominally set to 1.9. In Figures 5a and 5b, the glottal diameter and glottal areas are plotted for an assumed F_0 of 100 Hz, and an open quotient as specified in the table. The solid, bold and dashed lines represent adult male, adult female and child responses respectively.

Figure 5c shows the k_t values over a cycle. The values are approximately constant for diameters greater than 0.05 cm, but can be very large for smaller diameters. The average large diameter k_t values were 1.033, 1.026, and 1.016 for adult male, adult female and child. These values are slightly smaller than the 1.1 value used in the literature (e.g.

Table III.
Data Used for Age and Gender Transformation

	male	female	child
Rest length L_0 (cm)	1.6	1.0	0.9
Amplitude d_{max} (cm)	0.18	0.15	0.12
Lung press P_L (cm H_2O)	5.57	4.78	8.43
Open Quotient Q_o	0.55	0.70	0.66

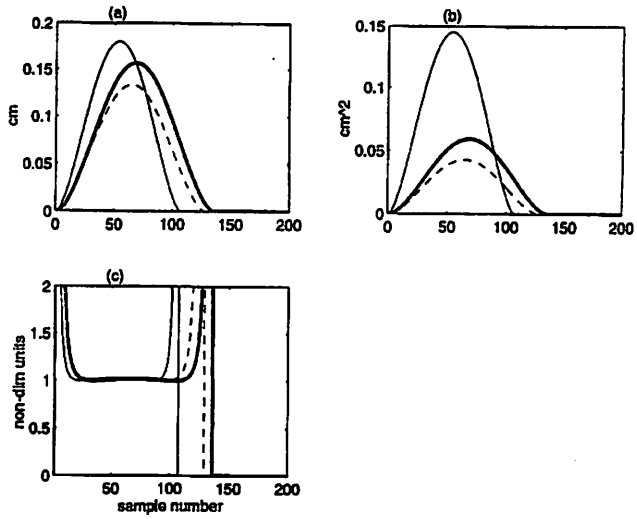

Figure 5. Area/flow model plots, with the following legend: adult male (solid), adult female (bold) and 10 year old boy (dash). (a) Glottal diameter d, *(b) Glottal area* a, *(c) Translaryngeal coefficient* k_t *(zoomed in).*

Titze, 1973), which has been derived from the difference in entry and exit coefficients first used by Van den Berg et al. (1957). The large values of k_t reduce v_0 near closure as shown in Figure 6a, in which the value of v_0 approaches zero more rapidly than the value of a in Figure 5b. Although k_t is large for only brief moments in time, these are important events near closure. It is possible that k_t could control the skew, and hence the presence of certain spectral harmonics. This could affect voice quality (Titze 1994). Note that the child's no-load particle velocity is larger than the female's or male's, due to the larger lung pressure. Figures 6b and 6c show the no-load flow and the loaded flow (the skewing is due to the inertive vocal tract).

Perkell et al (1994) recorded male and female ac flow values (sinusoidal or *alternating current* analogy). They measured a male/female (m/f) ac flow ratio of 330/160 (numbers reported here in cm^3/s, but in liters/s in their study), i.e. a normalized ratio of 2.06. Stathopoulos and Sapienza (1996) measured a male/female/child (m/f/c) ac flow ratio of 310/160/130, i.e. 1.94/1.0/0.81 (normalized to the female value). These data appear in Table IV. In Figure 6c, the maximum flows from the model are 345, 157, and 146, which gives a 2.16/1/0.93 ratio for m/f/c. In absolute terms, the numbers are well within a standard deviation of the mean

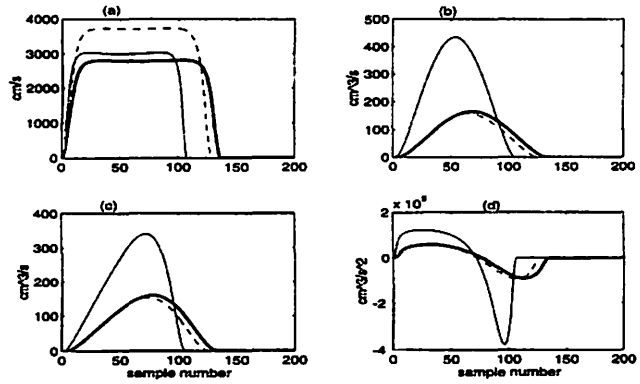

Figure 6. Area/flow model plots, with the following legend: adult male (solid), adult female (bold) and 10 year old boy (dash). (a) No-load particle velocity v_0, (b) No-load flow u_0, (c) Loaded flow u, (d) derivative of loaded flow u'.

values measured by Stathopoulos et al. and Perkell et al. As normalized m/f/c ratios, the model data also appear reasonable.

Figure 6d shows the flow derivative for the three cases. The mfdr values are the peak negative values in the plots. These values are 360,000, 87,000, and 89,000 cm^3/s^2, assuming a 100 Hz F_0 (and discarding the negative sign). Converted to l/s^2, these values are 360, 87 and 89, respectively. Perkell's mfdr data are in the m/f ratio of 337/184, i.e. a normalized ratio of 1.83. The measurements were made at an F_0 of 112 Hz for the male, and 204 Hz for the female. If the model data are adjusted for F_0, the data are then 360 (112/100) = 403 for the male and 177 for the female. This is a m/f ratio of 2.27, larger than Perkell's measured ratio. The F0 adjusted ratios are also larger than Stathopoulos' data (see Table IV). Note, however, that the model data are still within a standard deviation of Stathopoulos' and Perkell's data. In addition, the model produced the increased mfdr of the child relative to the female exhibited by Stathopoulos' data. It thus seems likely that the estimated tissue amplitude chosen for the child is reasonable.

Table IV.

AC Flow and mfdr results comparison. Numbers in brackets are standard deviation data. Numbers after the @ are F_0 of phonation at which data was gathered or simulated. The last two rows represent the third to last row after pitch adjustment. The data taken from Perkell are the 1993 measurements at the normal loudness level. The data taken from Stathopoulos refer to 1996 reference.

Parameter	male	female	child	m/f/c ratio	data source
ac flow (cm^3/s)	310 (90)	160 (70)	130 (40)	1.94/1/0.81	Stathopoulos
	330 (70)	160 (50)	-	2.06/1	Perkell
	345	157	146	2.16/1/0.93	model
mfdr (l/s^2) @ F_0	285@105 (95)	194@206 (71)	210@230 (80)	1.47/1/1.08	Stathopoulos
	337@112 (127)	184@204 (63)	-	1.83/1	Perkell
	360@100	87@100	89@100		model
	378@105	179@206	205@230	2.11/1/1.15	model (F_0 adj)
	403@112	177@204	-	2.27/1	model (F_0 adj)

The Polynomial Fitted Flow Derivative Model

Milenkovic (1993) developed a model of the flow derivative based on fitting a polynomial equation to the data for each cycle. This model required a combination of four basis functions, each function of order four. Childers and Hu (1994) simplified this model to a single 7th order polynomial. The attraction of both of these models is that the fitting procedure can be automated so that no human intervention is necessary (e.g. no estimation of Q_0 and Q_x). The procedure requires the cycle to begin and end at the maximum negative spike locations in each cycle. The polynomial equation attempts to fit all of the major low frequency humps in the data. This is useful because LPC often produces a pseudoflow (or pseudoflow derivative) signal which contains no obvious closed phase. Models such as the LF, and the area/flow model described above, expect a closed phase. As a result, these more stylized models do not always match the shape of the flow derivative signal as accurately as the polynomial model. A drawback to the polynomial representation, however, is an inability to interpret the coefficients in relation to events in the cycle or measurements from sensor signals (Milenkovic, 1993), (Childers, 1994).

The Polynomial Equation

The Matlab procedure polyfit was used to find the coefficients of the polynomial $p(x)$ of degree n that fits the data $p[x(i)]$ to $y(i)$ in a least squares sense, where $x(i)$ is the time sequence and $y(i)$ is the LPC pseudoflow derivative data. Thus

$$p(x) = p_1 x^n + p_2 x^{n-1} + \ldots p_n x + p_{n+1} \qquad (19)$$

In this study, polynomials of order (degree) ranging from 7 to 15 were used to capture the low frequency nuances in each cycle.

Reconstruction

The polynomial coefficients and the cycle period (in samples) are stored until reconstruction. A sequence of evenly sampled time values, equal in length to the period (the $x(i)$ above), are then entered into the polynomial function to reconstruct the pseudoflow derivative cycle. Note that the fundamental period is specified as an integral number of sample points. Examples are shown in Figures 7(d) and 7(f). The smooth plots are from the polynomial fit.

If instead of mimicry, a transformation from an adult to a child or female is desired, several changes must be made. The polynomial for each cycle fits only that cycle. Consider a pitch contour modification that increases or decreases the number of cycles. Since there is a one to one correspondence between polynomials and cycles, polynomials must be created or eliminated. The simplest way would be to use one polynomial (p) and repeat it for every cycle in the sentence. This ignores the variations in shape that the poly-

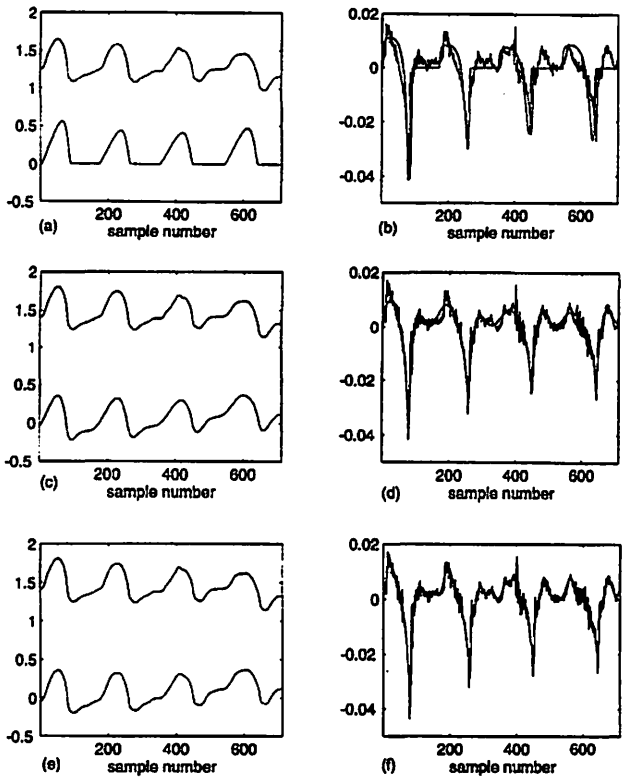

Figure 7. (a) Original LPC residual pseudoflow (top) and a/f model approximation (bottom) from steady state portion in Goldilocks sentence, (b) derivatives of flows from Figure 7a: original (noisy plot) and a/f model (smooth plot). (c) as for Figure 7a, but bottom curve is 7th order polynomial model. (d) as for Figure 7b, but for 7th order polynomial model. (e) as for Figure 7c, but bottom curve is 15th order polynomial model. (f) as for Figure 7e, but for 15th order polynomial model.

nomials can capture. The changing period length can be dealt with by resampling p with a higher or lower density relative to the period associated with p. Changes in intensity can be captured by scaling the reconstructed cycle so that it has a negative minimum equal to the value defined in the spike derivative contour. In addition, the m/f/c mfdr scale factors discussed above can be applied to the spike derivative if gender or age changes are required.

Choosing a single polynomial that is used for every cycle would prevent this model from capturing the subtle qualities inherent in a persons speech. An alternative is to fit each of the polynomial coefficients in a similar manner to the pitch contour (described below), modeling each coefficient value over time using another polynomial. The number of cycles required for the sentence or phrase could then be changed by resampling the coefficient contour at a different density. The computational effort to do this is high, however. Ananthapadmanabha (1995) has pointed out that an assumption of constant Q_0 and Q_s for a phrase or sentence does not harm the naturalness of the voice significantly. The polynomial fitting method, as applied to the polynomial coefficient contours, is therefore not discussed further in this paper.

Transforming Pitch and Intensity Contours

To transform the voice, the F_0 and flow-derivative spike (mfdr) contours in Figure 2 must be modified to represent the increased pitch produced by females and children, while maintaining the basic prosodic contours of the adult. In the system described here, the contours can be companded (compressed or expanded) to shift the F_0 or the intensity up or down by arbitrary amounts. The dynamics of the contours can also be exaggerated or flattened. This is done while keeping the timing for each word the same.

The F_0 contour for the whole sentence is made up of a series of voicing islands. Initially, the contour for the whole sentence is zero-meaned and the desired dynamic scaling parameter is applied. This multiplicative factor exaggerates (factor >1) or flattens (factor <1) the shape of the contour. Following this, the contour is shifted to its original mean F_0, plus an additional shift up or down.

Since the pitch markers represent F_0 cycle periods, a shift up or down in F_0 implies that more or fewer cycles are needed to span the same word-time intervals. To add or subtract cycles, the markers needed to be considered as discrete values representing a function. The pitch markers are examined to determine where the voicing 'islands' exist (i.e. where the marking algorithm has determined that voicing has been turned on or off). Each of these islands is fitted to a polynomial which can then be resampled to change the number of cycles. Each island can vary significantly in its dynamics, so the order of the polynomial is allowed to vary with the number of points in the island to be fitted. The order is chosen as N/4, where N is the number of original cycle markers to be fitted. The order is arbitrarily limited to a maximum of 15.

Once the polynomial for each island is obtained, it is resampled so that the cumulative time from the new markers does not exceed the time interval for the original island, yet is within half an F_0 cycle. All of the new markers are adjusted to the nearest sample, and the time interval of the island is readjusted according to the previous half cycle criterion. Once this is done, each of the markers (starting from the beginning of the island) has one sample added to its period length until the original island time interval has been achieved.

Intensity changes are produced in a similar manner. The flow derivative spike values are adjusted to reflect the desired dynamics. Although scaling and shifting of the intensity contour does not change the time interval of an island, any F_0 contour changes must be accompanied by intensity modifications. The F_0 manipulations determine the number of new cycles in each island. The polynomial fitting procedure used to modify intensity uses the newly determined number of cycles as a guide to resampling the intensity polynomial. It is necessary to ensure the intensity values are greater than zero (assuming an intensity value is positive), so that sound is not lost.

Transforming the Vocal Tract

Since previous studies on nonlinear vocal tract warping have produced contradictory conclusions, it was deemed useful to determine whether Yang's results could be generalized. MRI extracted areas, previously measured from subject BS (Story, 1996), were used as a comparison set. Yang's procedure for nonlinear scaling was carried out using the length relationships measured from their MRI data. (The epiglottal, pharyngeal, and oral regions of the male were compressed or dilated to match the female and child equivalent regions, using data from the first three columns of Table I from Yang and Kasuya (1995)). The F_1-F_2 formant shifts, shown in Table V, were generally within the 5% bound for the m/f scaling, except the F_2 for /u/. For the male-to-child scaling, the /u/ and /o/ F_2 and /e/ F_1 differences exceeded the bound. Note that F_3 and F_4 exceeded this bound more frequently. It was decided that this experiment confirmed the generality of Yang's observations, using his criteria for difference limen. As a consequence, only linear modifications of the vocal tract length were used.

Linear modification of the vocal tract can be achieved in several ways. One method is to spline fit the pseudoarea function obtained from the LPC coefficients, and then resample the area function at a higher or lower density (for longer or shorter tracts). If speech reconstruction is carried out at the original resampling rate, the filter will represent a tract in which each area element has the same length, but there will be fewer or more area elements (and LPC coefficients). This technique has a benefit in that it permits nonlinear modification through uneven resampling of the pseudoarea function. A drawback, however, is that the pseudoareas do not resemble the MRI areas for the same sound, due to bandwidth estimation problems in the LPC process. The pseudoareas may therefore have a different sensitivity to area manipulation than those derived from the MRI data.

An easier method is to project the LPC coefficients to the pole-zero domain, and then rotate the angle of each pole inversely proportional to the length change. This method has been used by Childers et al. (1989), although they used a function which rotated less for higher frequency poles. In this study, all poles were rotated by the same factor. One problem with this technique occurs when poles rotate past π (the Nyquist frequency). In listening tests, it was found that poles appearing near the Nyquist frequency combined with their conjugate poles very close by. This caused the pole skirts to add, resulting in boosted signals near the Nyquist frequency. When the magnitudes were close to the unit circle, this boost was heard as a chirp. The effect was prevented by reducing the magnitude of all poles in which the new frequency would be greater than 5 kHz to 80% of the original value. Poles rotating past π were placed at the origin.

Table V.
Formant frequencies (Hz) and % differences (in brackets) from the male reference for five vowels. The male/female and male/child formants are for the male tract modified to approach the female or child's proportional length relationships of epiglottal, pharyngeal, and oral cavities.

	Vowel	F1	F2	F3	F4
male	/a/	809	1156	2813	3402
male/female		817 (0.93)	1172 (1.35)	2647 (-5.86)	3675 (8.03)
male/child		817 (0.93)	1142 (-1.24)	2609 (-7.23)	3924 (15.35)
male	/i/	287	2375	3196	3720
male/female		284 (-0.76)	2484 (4.6)	2989 (-6.48)	3859 (3.74)
male/child		278 (-2.9)	2412 (1.6)	2984 (-6.6)	3942 (5.9)
male	/u/	332	1193	2558	3875
male/female		332 (0)	1110 (-6.9)	2685 (4.95)	4150 (7.11)
male/child		331 (-0.34)	1066 (-10.7)	2734 (6.86)	4363 (12.60)
male	/e/	604	1843	2344	3017
male/female		581 (-3.75)	1845 (0.12)	2463 (5.1)	3294 (9.2)
male/child		561 (-7.1)	1763 (4.34)	2629 (12.2)	3433 (13.8)
male	/o/	416	865	2487	3747
male/female		412 (-1.1)	848 (-1.9)	2398 (-3.6)	4017 (7.2)
male/child		396 (-4.9)	801 (7.4)	2722 (9.4)	3994 (6.6)

Experiments

Male Mimicry

The Goldilocks sentence was simulated using both the area/flow (a/f) and the polynomial models. The F_0 and intensity contours were not altered at all, and the filter data representing articulation was not modified. The a/f model assumed Q_s = 1.9 and Q_o = 0.53. The open quotient was determined by examining a set of pseudoflow cycles obtained from the residual integral. It was experimentally determined that this value of Q_o produced a voice quality closest to the original speech. This was done by manually optimizing the spectral harmonic amplitudes of the flow pulse. The nominal Q_s was changed for each cycle in reconstruction when the mfdr spike was scaled to equal the intensity contour. Figure 7a shows the pseudoflow (top) and the model approximation, and Figure 7b shows the pseudoflow derivative (residual) and the a/f approximation.

The a/f model produced speech which was very natural, but qualitatively different from the original. It appeared to have a boosted F_1. This is illustrated in Figure 8a, in which the spectra for the a/f model (gray line) and the original residual (black line) are presented. The spectra have been adjusted to have equal magnitude at F_0. It can be seen that the first two harmonics are well matched, while the next seven harmonics (the F_1 band) are too large. There was also a lack of high frequencies from 3 to 5 kHz (not shown).

The polynomial model makes no assumptions about the open quotient, fitting the data between the cycle markers. Figures 7c, 7d, and 8b show the fit with a 7th order model (the same order used by Childers and Hu). After reconstruction and playback, the sentence using this polynomial model was perceptually closer to the original than the a/f model, although lacking some midrange frequencies (500 to 1500

Hz). This produced slightly muffled speech. A polynomial of order 15 produced speech much closer to the original. This is shown in Figures 7e and 7f, where the dynamics in the signals are matched more closely to the original, and in Figure 8c, where the first 5 harmonics are perfectly matched,

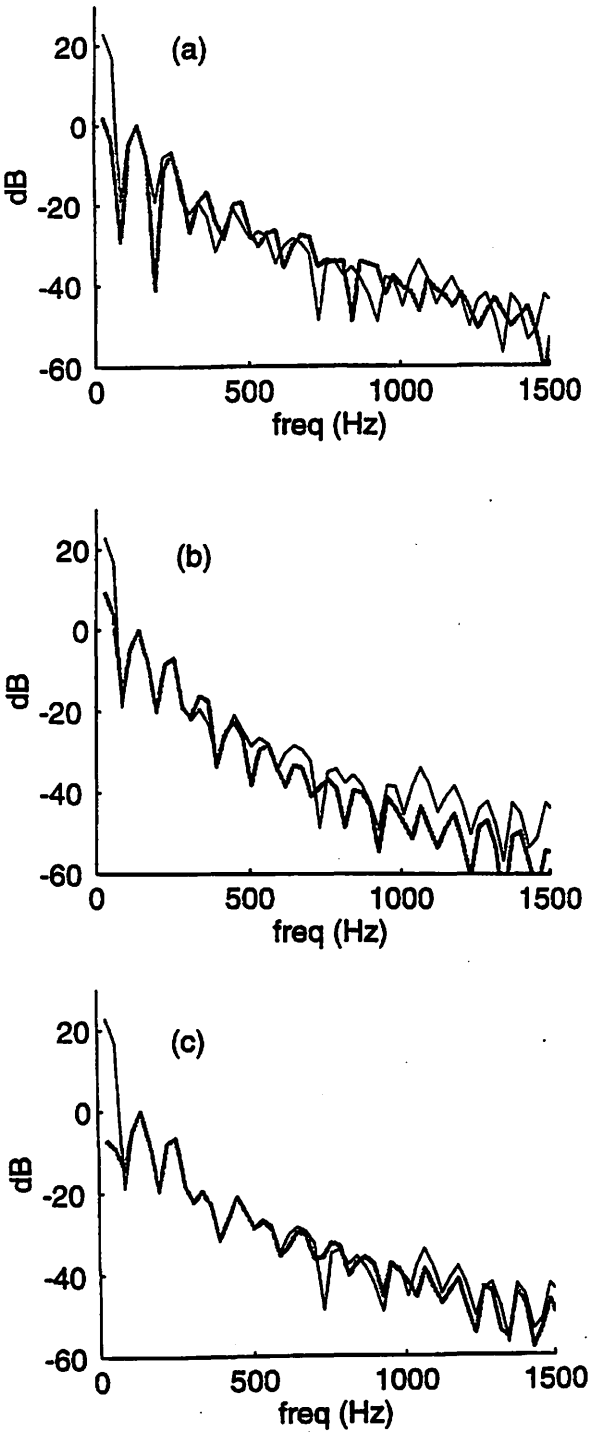

Figure 8. (a) spectrums of flow signals: original pseudoflow (solid), a/f approximation (gray). (b) comparing original pseudoflow (solid) and 7th order polynomial (gray). (c) comparing original pseudoflow (solid) and 15th order polynomial (gray).

and the harmonics from 500 to 1500 Hz also match better than the other models. Another technique for improving the quality was based on Imaizumi's observation that it is critically important to match the magnitude of the pseudoflow derivative markers. The 7th order polynomial fit frequently underestimated the magnitude of the markers. The 15th order polynomial produced a better fit for the marker magnitudes as well as the internal dynamics within the cycle. A compromise technique that can be used in reconstruction is to substitute the marker value for the first sample in the reconstructed flow derivative cycle. This guarantees a correct flow derivative minimum, at minimal computational cost. The resulting voice quality is significantly closer to the 15th order polynomial speech than the 7th order polynomial speech. Spectrally, the substitution technique more closely matched the 15th order spectra than the 7th order spectra.

It should be noted that there were some problems with the turbulence noise. In all three cases presented, the spectrum from 3 kHz to 10 kHz was below the true magnitude. Adding aspiration produced a flow pulse spectrum that matched the residual flow well between 3 and 10 kHz. Perceptually, however, even though the noise was applied during a small time window in the cycle (see Figure 4) with similar magnitude to the 'error' from the residual signal, the effect on the speech was more pronounced than in the original speech, sounding faintly hoarse, as if phlegm was present. It sounded natural, but was slightly more prominent than expected.

Male to Child Transformation

The transformation of speech from male adult to male child required modifications of all the speech components. The strategy employed was to analyze KM's excitation residual using LPC, and then adjust the a/f and polynomial models to simulate KM's flow cycle. These flow cycles replaced the adult male excitation. BS's vocal tract was then linearly shortened, and the pitch and intensity contours were modified to emulate a child's speech pattern. Since the intention was to use the adult prosodic contours and vocal tract articulation (with tract length adjustments), there was no target child speech for comparison.

Figure 9a shows the original F_0 contour produced by the male (lower) and the modified contour. The mean F_0 for the original contour was 110 Hz. Analysis of KM's speech indicated a mean pitch of 200 Hz. The upper contour was then lifted by 90 Hz. The dynamics of the contour were amplified by a factor of 3 to simulate a child-like 'fairy tale reading' voice, since BS originally produced the sentence with a somewhat monotonic delivery. The intensity contour (lower contour) in Figure 9b was amplified by a factor of 1.5 (without zero-meaning the data first).

Figures 10a and 10b show the flow derivative and the flow pulse spectra for the child and for the a/f model. The match is good for the first six harmonics in the spectrum. An

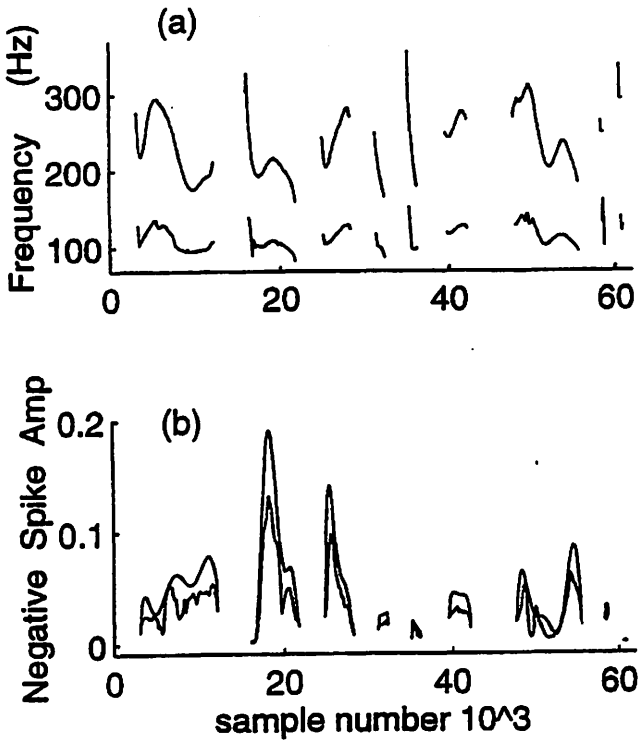

Figure 9. (a) Companded F_0 contour for the Goldilocks sentence for subject BS (male to child transformation). The modified F_0 (upper contour) has been shifted up by 90 Hz, and a gain of 3.5. (b) The mfdr spike contour has a gain of 1.5 applied.

open quotient of 0.74 and skew quotient of 2.4 were measured from KM's pseudoflow. The signal was taken from a short steady state section of the sentence. Figures 10c and 10d are equivalent plots using the 15th order polynomial model. The spectral match is better than for the a/f model, but both models fit well.

The vocal tract was linearly modified in length using pole rotations. From Yang's data, the male-child/male-adult vocal tract length ratios for each of the vowels /a/, /i/, /u/, /e/, and /o/ were calculated. This produced a mean of 0.782. This scale factor was applied to all LPC filter frames. Figure 11 shows the formant spectra for an LPC frame. The rotated poles produced the gray spectral plot. The average formant shift was 27.8%, as expected.

After speech reconstruction, each of the models produced perceptually very similar speech, as expected from Figure 11. The contour changes produced an animated speech style quite suited to the material. Informal judgments of the transformed speech are discussed in section D.

Male to Female Transformation

Transforming from the adult male to an adult female followed the same procedure as for the male-to-child process. A mean F_0 of 203 Hz was measured from KB's speech. The pitch contour was thus shifted up by 93 Hz, with a dynamic scale factor of 3.5 to give an animated style. A Q_0

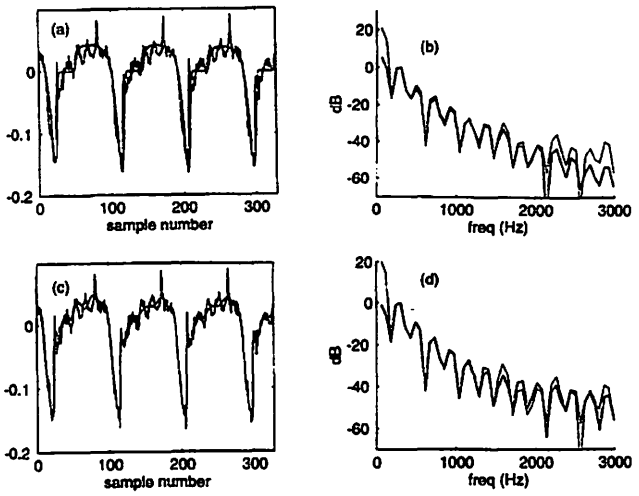

Figure 10. (a) LPC residual (pseduoflow derivative) (noisy plot) and a/f model approximation (smooth plot) to match the child subject KM for steady state voicing in Goldilocks sentence. (b) Flow spectra for Figure 10a - LPC residual (solid) and a/f model (gray). (c) LPC residual (noisy) and 15th order polynomial model (smooth). (d) Flow spectra for Figure 10c - LPC residual (solid) and polynomial model (gray).

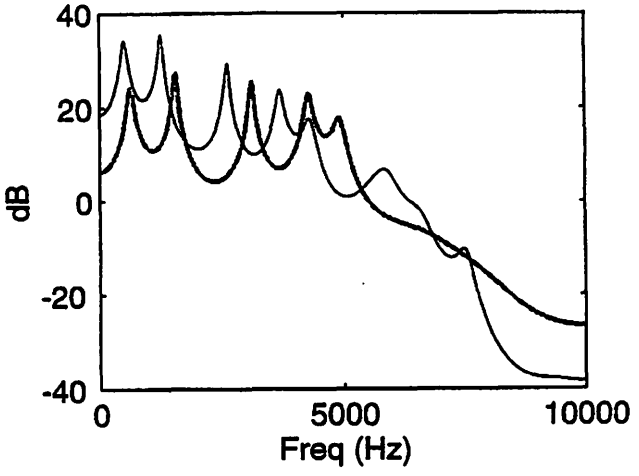

Figure 11. LPC spectral frame for original male phonation (solid), and after pole rotation (gray) for adult to child transformation.

and Q_s of 0.58 and 1.5 were estimated from a steady state portion of KB's pseudoflow. The flow derivative and flow spectra plots in Figures 12a and 12b show a good fit for the first seven harmonics. Figures 12c and 12d show the polynomial fit, which is a better fit, especially for the harmonics between 2 and 3 kHz.

The vocal tract was rotated by a female/male factor of 0.83, estimated using the male and female vocal tract length data from Hogberg (1995). Yang's male/female data was used initially, but the 0.79 ratio produced a child-like voice, since this value is close to the 0.782 used for the male-to-child transformation. The ratio measured from Hogberg's data produced a more adult-like voice.

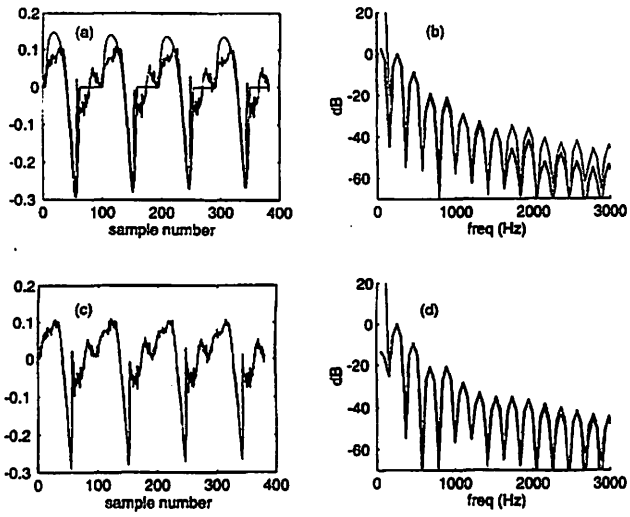

Figure 12. (a) LPC residual (pseduoflow derivative) (noisy plot) and a/f model approximation (smooth) to match the adult female subject KB for steady state voicing in Goldilocks sentence. (b) Flow spectra for Figure 13a - LPC residual (solid) and a/f model (gray). (c) LPC residual (noisy) and 15th order polynomial model (smooth). (d) Flow spectra for Figure 13c - LPC residual (solid) and polynomial model (gray).

Perceptual Testing of Resulting Transformations

An informal perceptual test was conducted with four listeners. A number of sentences using the Goldilocks passage were created using speech from BS, KB and KM. In some of the sentences, the excitation was replaced by one of the models, and the articulation was not modified. In other sentences the contours and vocal tract were also altered in an attempt to transform the speech, as described in sections B and C. The sentences were played in random order.

Mimicry

All listeners thought that the adult male (BS) speech, after the excitation was replaced with either the a/f model or the polynomial model, was highly natural. When this was done to the adult female's speech (KB) the opinion was that the sentence produced from the polynomial model was highly natural, but the a/f model produced speech that was not as good, although acceptable. The speech obtained from KM's tract and the two excitation models was judged most natural using the polynomial model, but again, the a/f model was still acceptable.

Pulse Substitution

Sentences using KB and KM's polynomial model pulses were inserted into BS's sentence, while retaining BS's contours and vocal tract length. This was an attempt to find out if a female or child's voice quality affected judgments of gender or naturalness. All listeners were able to identify the speaker as male and adult for both sentences. The sentences were all judged as highly natural.

Constant Open Quotient

A sentence was created from BS using a single polynomial pulse extracted from BS's excitation. The pitch period was varied by resampling the polynomial. This method kept the open quotient constant throughout the sentence. All listeners considered this sentence to be highly natural.

Male to Child Transformation

The a/f and polynomial parameter estimates extracted from KM were applied to the sentence produced by BS. The vocal tract was shortened and the pitch contour modified as described in experiment B above. Three of the four listeners identified the modified speech as coming from a child. The speech was considered to be only somewhat natural, however. Neither of the two models produced speech that was noticeably more natural than the other.

Male to Female Transformation

The procedure for male-to-child transformation was repeated using KB's model estimates, as described in experiment C. The listeners were confused with respect to gender and age. The speech was considered to be only somewhat natural, and less convincing than the male-to-child transformation.

Discussion

The informal listening test results, although preliminary, suggest that we have successfully produced natural sounding excitation models, while articulatory changes were less successful. Both the area/flow and the polynomial source models produced acceptably natural speech when the articulation was not modified. The polynomial model captured the nuances of the excitation cycle better than the area/flow model, since it attempted to fit the dynamics of the cycle, whereas the area/flow model enforced a closure period. Using pitch asynchronous LPC as a decomposition technique may have made the area/flow model even less suitable, since this method of LPC is most susceptible to errors due to bandwidth estimation and 'leakage' between the excitation and articulation estimates. Perhaps a pitch synchronous or closed phase LPC method would produce a pseudoflow pulse that is more like the flow pulse expected by the area/flow model.

While the area/flow model did not capture the dynamics of the excitation as well as the polynomial model, age and gender-related measurement data from the literature indicated that the area/flow model successfully takes into account age and gender-related lung pressure, vocal fold length and vibrational amplitude measurements as variables. The ac flow and mfdr results, which are gross measures of pulse height and shape, are all within a standard deviation of the mean measured values measured by Stathopoulos et al

and Perkell et al. This suggests that the area/flow model can be used to model changes due to age and gender.

Keeping the open quotient constant does not decrease the naturalness of the speech, in agreement with studies by Ananthapadmanabha and Childers (1989). This concept is very useful for very efficient encoding, as a single polynomial coefficient set can be transmitted to represent a sentence or even a conversation, along with an accurate intensity/mfdr contour.

Modifying the vocal quality by substituting excitation models from a female or child's voice does not affect estimations of naturalness, age, or gender, although it does change the quality. The articulation and the pitch contour appear to be more significant determinants of age and gender.

It should be noted that the estimation of the open quotient in the area/flow model is quite difficult. Matching the magnitudes of the spectral harmonics of the flow pulses is the best criteria, although it requires an iterative process that combines an estimation of a threshold in the time domain with error minimization in the spectral domain. In this study, this optimization process was not automated.

The turbulent noise model produces perceptually realistic noise, but the perceptual effect is slightly magnified. Although the magnitude range of the added noise is similar to the original signal, and the window of application is very small, the turbulence is easily perceived.

For the polynomial model, a seventh order model has problems matching harmonics 3 to 8, although this can be improved by substituting the first point with the mfdr intensity value specified in the intensity contour. The best accuracy has been achieved with a 15^{th} order polynomial.

The age transformation from the adult male to an arbitrary male child has met with limited success. Most of the listeners thought it was speech from a child, but the speech was judged to be less than convincing in its characterization of a child. The gender change was even less successful. The substitution of female or child pulses into the male's speech did not degrade the naturalness, but the modifications to the tract and F_0 contour created unconvincing speech. Perhaps the animated style applied to the F_0 contour was too suggestive of a child, although in the authors' opinion, the tract modifications appear to be a greater contributor to the unconvincing characterization. It is possible that Yang's conclusions regarding nonlinear tract changes may have to be reexamined, especially as Table V indicates that F_3 and F_4 often exceeded the 5% threshold of just noticeable differences. It is known that F_3 and F_4 contribute to the perception of the twang and sob speech qualities, and that these qualities can be generated by manipulating the epiglottal end of the vocal tract (Titze and Story, 1996).

Conclusions

This study has demonstrated that the LPC methodology, combined with either of two models of excitation, can produce sentence-level speech of high quality and naturalness. The polynomial fitting method is the best match to the LP filter articulation because it can be automated, and because it makes no assumptions about the closed phase of the vocal fold pulse. It therefore matches the residual well and ignores any errors in the decomposition process. The area/flow model is not as successful because it requires an estimate of the open quotient which is not easily and optimally measured. In fact, the LPC decomposition may not produce a closed phase at all. Note, however, that in articulatory synthesis, where the vocal tract shape is estimated using techniques such as neural nets, the area/flow model may be a better input model. Methods of LPC which can determine the glottal pulse with more accuracy may also find that the area/flow model would more closely fit the residual.

The area/flow model has successfully been extended to describe changes due to age and gender. The nominal parameter values used here produce ac flow and maximum flow declination rate values that are within the expected ranges of reported data, and successfully characterize the differences due to gender and age. Our model of turbulent noise produces perceptually natural results, although further work is needed on the nature and structure of the high frequency content in the residual signal. Perhaps placement of the aspiration is as important as spectrum and magnitude. Childers and Lee (1991) placed noise just after the glottis closed for the simulation of breathy voices.

The age and gender transformations have produced only limited success, with the male-adult-to-boy transform faring best. The success of the excitation models points to the vocal tract modifications as the source of the problem. We used Childers (1991) method of pole rotation, in which higher frequency poles were rotated less, without any significant improvement. The formant data calculated in Table V indicated that the nonlinear tract changes may change F_3 and F_4 more than expected, which we believe can alter the quality of the perceived vowels. Further simulations using more MRI data, and perceptual listening tests using the manipulated tracts, need to be conducted. The epiglottal, pharyngeal, and oral markers taken from Yang's data need to be identified on our data, so that there is no mismatch. Even if nonlinear changes are warranted, the modifications may have to vary from vowel to vowel. For sentence level speech, this would require identification of the vowel prior to manipulation. In addition, this technique would not be well suited to the LPC method since the pseudoareas estimated from the LPC filter coefficients may have different sensitivities with respect to area element manipulation as compared to MRI areas. Equivalent manipulations in the pole-zero domain are probably more useful in the context of LPC

synthesis, if the effects of the nonlinear growth relationships can be easily related to pole rotations.

Acknowledgement

This work was supported by Grant No. R01 DC0232-01 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

References

Ananthapadmanabha, T.V. (1995). "Acoustic factors determining perceived voice quality," *Vocal Fold Physiology: Voice Quality Control*, edited by O. Fujimura and M. Hirano, (Singular Publishing Group, San Diego, California, USA), pp. 113-126.

Berg, J.W. van den, Zantema, J.T., and Doornenbal, P. Jr., (1957). "On the air resistance and the Bernoulli effect of the human larynx," Journal of the Acoustical Society of America, vol. 29, pp. 626-631.

Childers, D.G., Wu, K., Hicks, D.M., and Yegnanarayana, B. (1989). "Voice conversion," Speech Communication 8, pp. 147-158.

Childers, D.G., and Lee, C.K. (1991). "Some acoustical, perceptual, and physiological aspects of vocal quality," in *Vocal Fold Physiology*, edited by J. Gauffin and B. Hammarberg, (Singular Publishing Group, San Diego, California, USA), pp. 233-242.

Childers, D.G., and Hu, H.T. (1994). "Speech synthesis by glottal linear prediction," Journal of the Acoustical Society of America, vol. 96 (4), pp. 2026-2036.

Fant, G., Liljencrants, J., and Lin, Q. (1985). "A four parameter model of glottal flow," STL-QPSR 4/1985, Speech Transmission Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden, pp. 1-13.

Fant, G. (1995). "The LF-model revisited. Transformations and frequency domain analysis," STL-QPSR 2-3/1995, Speech Transmission Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden, pp. 119-156.

Flanagan, J.L. (1972). *Speech Analysis, Synthesis and Perception.* (2nd ed. Springer-Verlag, New York).

Fujisaki, H., and Ljungqvist, M. (1986). "Proposal and evaluation of models for the glottal source waveform," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1986, Tokyo, Japan.

Hirano, M. (1983). "The structure of the vocal folds," in *Vocal Fold Physiology*, edited by K. Stevens and M. Hirano (University of Tokyo, Tokyo, Japan), pp. 33-43.

Hogberg, J. (1995). "From sagittal distance to area function and male to female scaling of the vocal tract," STL-QPSR 4/1995, Speech Transmission Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden, pp. 11-53.

Hollien, H., and Moore, P. (1960). "Measurements of the vocal folds during changes in pitch," J. Speech Hear. Res. 3 92), pp. 157-165.

Holmberg, E., Hillman, R., and Perkell, J. (1988). "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voice," Journal of the Acoustical Society of America, vol. 84, pp. 511-529.

Imaizumi, S., Kiritani, S., and Saito, S. (1991). "Perceptual evaluation of a glottal source model for voice quality control," in *Vocal Fold Physiology*, edited by J. Gauffin and B. Hammarberg, (Singular Publishing Group, San Diego, California, USA), pp. 225-232.

Milenkovic, P.H. (1993). "Voice source model for continuous control of pitch period," Journal of the Acoustical Society of America, vol. 93, pp. 1087-1096.

Perkell, J.S., Hillman, R.E., and Holmberg, E.B. (1994). "Group differences in measures of voice production and revised values of maximum airflow declination rate," Journal of the Acoustical Society of America, vol. 96 (2), pp. 695-698.

Rothenberg, M. (1981). "Some relations between glottal airflow and vocal fold contact area," *Proceedings of the Conference on the Assessment of Vocal Pathology*, ASHA Rep. (Rockville, MD, no. 11), pp. 88-96.

Scherer, R.C., and Guo, C.G. (1991). "Generalized translaryngeal pressure coefficient for a wide range of laryngeal configurations", in *Vocal Fold Physiology*, edited by J. Gauffin and B. Hammarberg, (Singular Publishing Group, San Diego, California, USA), pp. 83-90.

Scherer, R.C., and Lange, R.C. (1996). Personal communication.

Stathopoulos, E.T., and Sapienza, C.M. (1993a). "Respiratory and laryngeal function of women and men during vocal intensity variation," Journal of Speech and Hearing Research, vol. 36, pp. 64-75.

Stathopoulos, E.T., and Sapienza, C.M. (1993b). "Respiratory and laryngeal measures of children during vocal intensity variation," Journal of the Acoustical Society of America, vol. 94, pp. 2531-2543.

Stathopoulos, E.T., and Sapienza, C.M. (1996). "Cross-sectional study of children's speech," submitted to the Journal of the Acoustical Society of America.

Stathopoulos, E.T., and Weismer, G. (1986). "Oral airflow and air pressure during speech production. A comparative study of children, youths, and adults," Folia Phoniatrica, 37, pp. 152-159.

Story, B.H., and Titze, I.R., and Hoffman, E.A. (1996). "Vocal tract area functions from magnetic resonance imaging," Journal of the Acoustical Society of America, vol. 100 (1), pp. 537-554.

Tang, J., and Stathopoulos, E.T. (1995). "Vocal efficiency as a function of vocal intensity: A study of children, women, and men," Journal of the Acoustical Society of America, vol. 97, pp. 1885-1892.

Titze, I.R., (1983). "Synthesis of sung vowels using a time-domain approach," in *Transcripts of the 11th Symposium: Care of the Professional Voice*, edited by V.Lawrence, (The Voice Foundation, New York), pp.90-98.

Titze, I.R. (1989). "Physiologic and acoustic differences between male and female voices," Journal of the Acoustical Society of America, vol. 85 (4), pp. 1699-1707.

Titze, I.R. (1991). "Mechanisms underlying the control of fundamental frequency", in *Vocal Fold Physiology*, edited by J. Gauffin and B. Hammarberg, (Singular Publishing Group, San Diego, California, USA), pp. 129-138.

Titze, I.R., Mapes, S., and Story, B.H. (1994). "Acoustics of the tenor high voice," Journal of the Acoustical Society of America, vol. 95 (2), pp. 1133-1142.

Titze, I.R., and Story, B.H. (1996). "Acoustic interactions of the voice source with the lower vocal tract," submitted to the Journal of the Acoustical Society of America, July, 1996.

Wong, D., Ito, M.R., Cox, N.B., and Titze, I.R. (1991). "Observation of perturbations in a lumped-element model of the vocal folds with application to some pathological cases," Journal of the Acoustical Society of America, vol. 89 (1), pp. 383-394.

Yang, C-S., and Kasuya, H. (1994). "Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects," *Proc. ICSLP94*, pp. 623-626.

Yang, C-S., and Kasuya, H. (1995). "Uniform and non-uniform normalization of vocal tracts measured by MRI across male, female and child subjects," IECE Trans. Inf. and Systems, vol. E78-D, no. 6, June, pp. 732-737.

NCVS Status and Progress Report - 10
November 1996, 127-132

Populations in the U.S. Workforce Who Rely on Voice as a Primary Tool of Trade

Ingo R. Titze, Ph.D.
Julie Lemke
Doug Montequin, B.A.
Department of Speech Pathology and Audiology, The University of Iowa

Abstract

The *United States Bureau of Labor Statistics* and other sources were consulted about the percentages of the working population who we identified as professional voice users. The largest percentage may be in sales and sales-related occupations (13%), but the exact breakdown of those who approach their clients vocally rather than by mail is still uncertain. The second largest population is teachers, who contribute 4.2% percent of the U.S. workforce (1994 statistic). Teachers have been identified as having the greatest incidence of voice disorders.[1,2] Population data are also given for professional voice users who could present a significant hazard to public safety if their vocal communication skills were severely impaired.

Introduction

When one thinks of professional voice users, singers and actors immediately come to mind. However, there are many other professionals for whom voice is a primary tool of trade. Telemarketers, teachers, receptionists, emergency vehicle dispatchers and broadcasters would find it virtually impossible to interact with their clients, students, or audiences without a well-functioning and enduring voice. Thus, vocal problems can affect careers and reduce profit for a company. In addition, vocal problems can jeopardize the safety of the general public if there is miscommunication of key facts and directives, particularly in law enforcement and the Armed Forces.

The primary purpose of this study is to identify the numbers of professional voice users in various work settings. These population statistics will establish a priority list for potential voice clients. We define professional voice users as: 1) those who depend on a consistent, special, or appealing voice quality as a primary tool of trade, and 2) those who, if afflicted with dysphonia or aphonia, would generally be discouraged in their jobs and seek alternative employment.

Specifically, then, this study addressed the questions: (1) which professionals are typically treated in voice clinics and what causes them to seek therapy; (2) what percentage of the U.S. workforce can be called professional voice users; and (3) what percentage of the U.S. workforce requires a healthy voice for public safety reasons?

Voice Professionals Seen in Clinics

The factors believed to contribute most to abnormal voicing are summarized by Johnson[3] in Table 1 (following page). It is logical to assume that these factors (primarily misuse and overuse) are present, in part or in total, among the voice professionals to be discussed in this paper. Personal history forms in clinics everywhere confirm these underlying factors.

Herrington-Hall and colleagues[4] found that of 73 different occupations represented in a sample of 1,262 clients, the ten most frequently seen groups were retired persons (former occupation unknown), homemakers, factory workers, unemployed persons, executive managers, teachers, students, secretaries, singers and nurses. All ten of these groups had disorders related to vocal abuse, except for retired persons.

In data obtained from 1,484 new clients seen in eight major departments of phoniatrics in Sweden[2], teaching was the most common occupation, followed by office work. Large representations were also found among persons occupied in social work, law and clerical work. Phonasthenia (weak voice) was found to be the most prevalent diagnosis, followed by vocal fold edema and recurrent nerve palsy.

Table 1.
Factors Contributing to Abnormal Voice*

Misuse	Loud talking, yelling, screaming
	Hard glottal attacks
	Singing or speaking outside acceptable physiological range
	Speaking in a noisy environment
	Excessive coughing and throat clearing
	Grunting (as in exercising and lifting)
	Excessive talking
	Loud, hard, abusive laughing
	Producing voice when laryngeal tissues are inflamed
Exposure	Alcohol consumption
	Medication
	Caffeine
	Recreational drugs
	Smoke
	Reflux of stomach contents
Psychogenic Causes	Musculoskeletal tension

* (After Johnson, 1994)

Between 1991-1993, a quality of life questionnaire was distributed to 174 voice clinic patients seeking treatment from the Department of Otolaryngology, University of Iowa Hospitals and Clinics, and the Division of Otolaryngology - Head and Neck Surgery, University of Utah. Of those completing the questionnaire, 16.4% were teachers, 4.4% each were entertainers (including actors and singers), sales agents, office managers and secretaries, or waiters; 3% each were bill collectors, buyers, ministers, or craftspersons/machine operators. Most of the patients seen were employed in occupations requiring the use of voice on an ongoing basis in their daily work activities.[5] The study is continuing, and there are now about 460 patients.

The occupations identified in the above studies appear similar to a second (informal) two-month survey done on 121 patients seen at the University of Iowa Hospitals and Clinics, Otolaryngology Speech and Hearing Clinic (unpublished data, 1996). Of the 70% who reported their occupations, 12% were teachers, 11% were retired (former occupation unknown), 9% were unemployed; 7% each were salespeople, singers, students and farmers; and 5% were homemakers.

The largest data set we found is being accumulated at the University of Wisconsin (unpublished data, 1996). The last patient count was 1,593. Teachers constitute about 20% of the clinic load, followed by singers (11%), salespeople (10%), clerks (9%), administrators and managers (7%) and factory workers (6%). Since there is considerable information about employment, and since this data set was easily assessable to us, we used this set as the primary source in constructing our own data tables (shown later).

Voice Professionals as Percentage of Workforce

Unless otherwise noted, data on the size of occupational groups were obtained from government documents, specifically the *Statistical Abstracts of the United States*[6] and from the *United States Bureau of Labor Statistics, Employment and Earnings*.[7,8] For the remaining occupational groups, estimates were obtained from national organizations to which individuals in these occupations belong. In all instances, effort was made to obtain 1994 figures ($\pm$ 1 year) to maintain consistency. Numbers were rounded to the nearest thousandth and percentages calculated on the basis of a total workforce of 123,060,000 in the United States in 1994.

Results are shown in Table 2. The first column shows populations in thousands, the second column the percent of the labor force, and the third column (whenever available) the percent of the Wisconsin clinic load. We will briefly mention some of the main groups and subgroups.

Factory Workers

Factory workers are tabulated as a kind of "control group". We have no *a priori* reason to believe that factory workers abuse or overuse their voices, except perhaps a small percentage who shout in noisy environments and another small percentage who are exposed (recall Table 1). Factory workers do show up in voice clinics (about 6% of the clinic load), but considerably less than would be expected as the basis of their 15% representation of the workforce. The majority of factory workers would not be considered professional voice users under the definition given in the Introduction.

Salespeople

The largest sector of professions identified as possibly being professional voice users are those in sales and sales related occupations, numbering 15,956,000. This group constitutes about 13% of the total workforce. First impressions of customers are critical for all those who are in the sales professions. These customer impressions are based not only on an outgoing and motivated personality, but also an effective voice. Those who are in the telephone sales must rely solely on their voice personality, without the benefits of body language or written communication skills.

As a percentage of the clinic load, salespeople in general are not remarkable. Their clinic load percentage is about equal to their workforce percentage. There are some interesting subgroups, however. Telephone marketers, of which there were 955,000 in 1994[9], constitute only 0.78% of the U.S. workforce, but make up 2.3% of the clinic load. We consider this a remarkable disproportion. The data do not

Table 2.
Occupations and Their Representation in Voice Clinics

Occupation (16 yrs and older)	Total Number (in thousands)	% of U.S. working population	% of clinic load
Factory Workers	17,876	14.53	5.6
Salespeople	15,956	12.97	10.3
Telephone Marketers	955	.78	2.3
Door-to-Door Salespeople	335	.27	
Ticket Reservation/Travel Agents	260	.21	0.4
Auctioneers	12	.01	
Stock Traders	7	.006	
Other	14,387	11.69	
Clerical Workers	13,004	10.57	8.6
Teachers	5,168	4.20	19.6
Special Education	308	.25	
Prekindergarten/kindergarten	496	.40	
Elementary	1634	1.33	
Secondary	1197	.97	
Higher Education	826	.67	
Other	695	.56	
Receptionists	931	.76	
Lawyers/Judges	861	.70	
Judges	40	.03	
Lawyers	821	.67	
Clergy	371	.30	
Psychologists	280	.23	
Counselors	237	.19	1.6
Telephone Operators	165	.13	
Interviewers/Recruiters	158	.13	
Public Relations Specialists	142	.12	
Speech-Language Pathologists	92	.07	
Actors/Directors	86	.07	
Broadcasters	77	.06	
Singers	23	.02	11.5
Classical	3	.002	
Other	20	.02	
TOTAL	55,427	45.05	57.2

make a distinction between those individuals who conduct marketing surveys and those who actually sell products, but that is immaterial for this study.

Door-to-door salespeople, totaling 335,000 in 1994, are also a small subset. This group constitutes 0.27% of the U.S. workforce and is actually declining in size. Clinic load data were not available to us.

Ticket and travel agents, totaling 260,000, spend large portions of their day speaking with clients and airline personnel, both face-to-face and over the telephone. They constitute 0.21% of the U.S. workforce, but had twice the representation in our voice clinic sample.

One would guess that auctioneers would be prime candidates for vocal fatigue and, ultimately, voice disorders. Their occupation demands the ability to speak (or chant) quickly for hours at a time with little intermittent rest time. In 1994, auctioneers numbered 12,000 (0.01% of the workforce). Clinic populations are not known.

Stock and commodity traders are in a similar must-speak situation. Many of them give oral instructions over long distances on a large, noisy floor. This profession comprises nearly 7,000 people according to the Securities Traders Association and the Corporate Transfer Agents Association[10]. Data are unknown as to the breakdown of those individuals actually buying and selling on the floor versus those working in more quiet venues over the phone.

All total, then, the subgroups of heavy voice users in sales identified here constitute about 1.3% of the working population. We suspect that the percentage of the clinic load will be about 3-4%, given that telephone marketers alone make up over 2% of a clinic load sampled here.

Clerical Workers

Clerical workers constitute another type of "control group" in the sense that they are not considered professional voice users. There may be a few secretaries and receptionists among them who speak much during the day, but they do not affect the group at large. Note that their percentage of the clinic load (9%) is similar to their percentage of the workforce (11%), indicating that the group as a whole is not remarkable.

Teachers

Teachers represent 4.2% of the U.S. workforce. This includes all levels of teaching: special education (308,000), preschool (496,000), elementary (1,634,000), secondary (1,197,000), and higher education (826,000). Interestingly, teachers constitute about 20% of the voice clinic load, a five-fold disproportion. This is quite remarkable.

A recent questionnaire distributed to 242 elementary and secondary education teachers in Utah and Nevada revealed that teachers were more concerned than a control group over options for future career if suffering from voice problems. A related study, conducted jointly by the University of Iowa and the University of Utah, confirmed that teachers represent the highest occupational group seen in voice clinics. Twenty percent of the teachers surveyed reported missing between one day and one week of work per year because of their vocal conditions. The most frequently cited vocal symptoms were hoarseness, vocal breathiness, weakness, tiredness, effortfulness, and a low-speaking voice.[1] It is possible that teachers represent the highest occupational group seen in voice clinics because they are aware of health issues and where treatment may be sought. They may not necessarily be in a profession of greatest risk, but they are the best educated in options for help.

Many studies continue to be conducted to help teachers become less susceptible to vocal fatigue. Reverberation times, sound levels and Rapid Speech Transmission Index values have been measured in occupied and unoccupied classrooms in order to plan the acoustics in new

schools and renovation of older schools.[11,12] Gotass and colleagues[13] compared teachers who experienced problems with fatigue with those who did not. He discovered that teachers who fatigue "tend to spend more time on activities that appear to be vocally demanding."

Receptionists and Public Relations Specialists

Receptionists and public relation specialists are the up-front faces and voices of their organizations. They are responsible for greeting people and serving as the interfaces to the medial and general public. According to the 1994 statistics, receptionists represent 0.8% of the workforce and public relations specialists represent 0.12% of the workforce. We do not know at this point what fraction of a clinic load they represent, but given their appreciable voice user under a high visibility profile, we would guess 3-4%.

Lawyers/Judges

According to the 1994 statistics, there were 821,000 lawyers and 40,000 judges certified in the United States. Less than half of the lawyers (about 350,000) are American Bar Association (ABA) members. This association, when polled by us, claimed that 60,000 ABA members are involved regularly in courtroom litigation. Including the non-ABA members, we estimate that 120,000 lawyers speak regularly in court. This is about 0.1% of the workforce. Many lawyers belong to specialty organizations. The ABA Section of Criminal Justice has a membership of 9,790 attorneys. Lawyers with 12 or more years of experience may belong to the International Academy of Trial Lawyers, which has a membership of 554.[10] These two sub-groups will be contacted in the future for further information on voice use.

Clergy, Psychologists, Counselors, and Speech-Language Pathologists

One-on-one oral communication is an important part of the workload for the clergy, psychologists, counselors and speech-language pathologists. In addition to one-on-one counseling, many clergypersons deliver regular sermons and some of them chant or sing. Speech-language pathologists often serve as role models for vocal production. In combination, these professionals constitute 0.8% of the workforce. In the case of counselors, we found that while they alone constituted only 0.2% of the workforce, they made up 1.6% of a clinic population. This is, again, an interesting disproportion, and suggests that voice misuse or overuse may be a problem in this group.

Telephone Operators

Telephone operators (165,000 or 0.13% of the workforce) are similar to telephone marketers in their daily voice use. Due to automated telephone service being implemented in recent years, the role of many telephone operators has become more diverse. These individuals may not only assist in placing long distance calls for current customers, but have the added responsibility of acting as telephone solicitors for prospective customers (U.S. Link, personal communication, March 14, 1996). Their appearance in the clinic (0.4%), is disproportionate in the same ratio (about 3:1) as telephone marketers.

Interviewers/Recruiters

According to the 1994 statistics, interviewers and recruiters comprised approximately 158,000 (0.13%) of the workforce. These individuals are similar to receptionists and public relations specialists in that they often serve as a first contact for their organizations. In another sense they are similar to counselors and psychologists in that they spend much time in one-on-one interaction. Specific clinic load data are not available at present.

Actors, Directors, Broadcasters, and Singers

In 1994, there were 86,000 actors and directors (approximately 0.07% of the U.S. workforce). A study in Prague, conducted by Novak and colleagues[14] on well-trained actors, suggests that there is considerable vocal fatigue in this group of voice professionals. In the researchers' opinion, the fatigue is caused by either high vocal and physical effort, or by emotional stress. Exact clinic population data are unknown for actors.

Statistics obtained from the Radio and TV Correspondent's Association and the American Federation of Television and Radio Artists indicate that there are approximately 77,000 radio and TV broadcasters in the United States.[10] Although this is a small percentage of the workforce (.06%), broadcasters are an important sector of voice professionals. The category we researched included not only those who report news, weather, traffic, and sports, but also talk-show hosts, of which there are an ever-increasing number.

There are approximately 23,000 professional singers in the United States[15], of which an estimated 3,500 are classical singers (American Guild of Musical Artists, personal communication, May 28, 1996). Further breakdown into country, rock, gospel, jazz, or blues is difficult because many professional singers fit into multiple categories. The concern singers have over their voices is evidenced by the fact that 11.5% of our sampled clinic load were singers, a huge disproportion in relation to the fraction of the workforce (0.02%).

Voice and Public Safety

We have identified a few occupations for which poor oral communication could pose a public safety risk (Table 3). According to our findings, approximately 3% of the U.S. workforce are in this category, but it is not at all clear

Table 3.
Occupations in Which Voice Use is Necessary for Public Safety

Occupation (16 yrs and older)	Total Number (in thousands)	% of U.S. working population
Military	1,691	1.37
Officers	*256*	*.21*
Enlisted	*1435*	*1.17*
Police Officers	833	.67
City	*368*	*.30*
Suburban	*220*	*.18*
County	*193*	*.16*
State	*52*	*.04*
Construction Supervisors	704	.57
Dispatchers	226	.18
Police and Fire	*27*	*.02*
Others	*199*	*.16*
Firefighters/Fire Prevention	210	.17
Air Pilots/Navigators	104	.08
Air Traffic Controllers	17	.01
TOTAL	3,785	3.05

which individuals give critical vocal commands and how devastating miscommunication due to a poor voice can be in this context.

For example, pilots and navigators in aircraft (104,000) and air traffic controllers (17,000; Federal Aviation Association, personal communication, April 3, 1996), give many vocal commands, but perhaps there is enough redundancy in these commands that vocal quality or loudness may not be key considerations. Is there evidence of frequent command repeats for some pilots or some traffic controllers? We have no answers at this point.

Firefighters (210,000) and police officers (833,000)[16] must possess the ability to project commands over the roar of sirens and other loud noises. Military personnel are in a similar situation, particularly officers (256,000)[6] who give critical commands over radio links, or enlisted personnel who give vocal commands in trenches or over short distance electronic links.

Construction supervisors (704,000) confront similar problems. These individuals must have the ability to instruct over noisy equipment, often using megaphones, blow-horns, or walkie-talkies to assist them. We do not know the extent to which errors occur.

Dispatchers (226,000) play an important role in public safety. Their ability to correctly transmit instructions over the radio to emergency personnel is crucial for timely arrival and execution in crisis situations. In particular, police and fire dispatchers (27,000 or 0.02% of the workforce) play the most critical roles. Polling the city of Washington, DC, we found one dispatcher (police and fire) per every 3,540 inhabitants (Washington DC Communication Division, personal communication, July 30, 1996). In Minneapolis, Minnesota there is one per 6,133 (Minneapolis Emergency Communications Department, personal communication, July 30, 1996) and in Cedar Rapids, Iowa, one per 7,333 (Cedar Rapids Police Department, personal communication, July 30, 1996).

Conclusions and Future Studies

We have identified occupations for which voice is a primary tool of trade. In these occupations, one encounters long speaking times, high intensity speech, emotional speaking and singing, and speaking in noisy environments. Limited data collected from clinics indicate a need to target certain sub-populations and make them more knowledgeable of proper vocal use. We hope that this study will stimulate clinicians to modify their patient history forms to be more specific about occupation and voice use. It would then seem advantageous to promote good vocal health through company-sponsored workshops and professional organization newsletters.

At this time, there is not enough evidence in the literature to show specific data on phonation times and intensities in heavy voice-user sub-populations. Further studies with voice accumulators that calculate phonation time and intensity in a typical day, and studies that count the frequency of miscommunication in critical situations, would appear to be on top of the priority list.

Acknowledgments

This study was supported by a grant from the National Institutes of Health, Grant No. P60-DC00976. The authors are grateful to Debra Walters-Smith for initial statistic collection and to Drs. Michael Karnell and Diane Bless for tabulating clients seen at the University of Iowa Hospitals and Clinics and the University of Wisconsin-Madison, respectively.

References

1. Smith, E., Gray, S., Dove, H., Kirchner, L., & Heras, H. (in press). Frequency and effects of voice problems in teachers. *Journal of Voice.*

2. Fritzell, B. (1995). Occupation and voice problems. *Proceedings from XXIII World Congress of IALP* [Abstract], p. 70.

3. Johnson, A. (1994). *Vocal Arts Medicine: The Care and Prevention of Professional Voice Disorders.* (In: Benninger, M.S., Jacobson, B.H., & Johnson, A.F., Eds.) New York: Thieme Medical Publishers, Inc. (page 155).

4. Herrington-Hall, B., Lee, L., Stemple, J., Niemi, K., & McHone, M. (1988). Description of laryngeal pathologies by age, sex, and occupation in a treatment-seeking sample. *Journal of Speech and Hearing Disorders,* 53, 57-64.

5. Smith, E., Verdolini, K., Gray, S., Nichols, S., Lemke, J., Barkmeier, J., Dove, H., & Hoffman, H. (in press). Effect of voice disorders on quality of life. *Journal of Medical Speech-Language Pathology.*

6. *United States Bureau of the Census Statistical Abstracts of the United States*. (1995) (115th Ed.). Washington DC: U.S. Government Printing Office.

7. *United States Bureau of Labor Statistics: Employment and Earnings*. (1994, January). Washington, DC: U.S. Department of Labor.

8. *United States Bureau of Labor Statistics: Employment and Earnings*. (1995, January). Washington, DC: U.S. Department of Labor.

9. Direct Marketing Association (1995, October). *Economic Impact: U.S. Direct Marketing Today*. [Commissioned research study by Whanton Economic Forecasting Associates]. New York, NY.

10. Schwartz, C.A. & Turner, L.A. (Eds.). (1995). *Encyclopedia of Associations* (29th ed., Vol. 1, Pt. 3). Detroit, MI: Gale Research Incorporated.

11. Pekkarinen, E., & Viljanen, V. (1991). Acoustic conditions for speech communication in classrooms. *Scandinavian Audiology*, 20, 257-263.

12. Sala E., & Viljanen, V. (1995). Improvement of acoustic conditions for speech communication in classrooms. *Applied Acoustics*, 45, 81-91.

13. Gotass, C., & Starr, C.D. (1993). Vocal fatigue among teachers. *Folia Phoniatrica*, 45, 120-129.

14. Novak, A., Dlouha, O., Capkova, B., & Vohradnik, C. (1991). Voice fatigue after theater performance in actors. *Folia Phoniatrica*, 43, 74-78.

15. *Musical International Directory of Performing Artists*. (1996). Hightstown, NJ: K-III Directory Corporation.

16. *Uniform Crime Reports for the United States* (1994). Washington DC: U.S. Government Printing Office, pgs. 290-295.

Part II

Tutorial reports and updates

NCVS Status and Progress Report - 10
November 1996, 133-149

Voice Disorders in Children

Steven D. Gray, M.D.
Division of Otolaryngology/Head and Neck Surgery, University of Utah Medical Center and
Primary Children's Hospital
Marshall E. Smith, M.D.
Department of Otolaryngology/Head and Neck Surgery, University of Colorado Health Sciences Center and The Children's Hospital
Wilbur James Gould Voice Research Center, The Denver Center For The Performing Arts

Introduction

The voice is a primary means of expression and oral communication, and has lifelong importance to social well being. The cry of the infant eventually becomes the voice of the teacher, lawyer, singer, or receptionist; it provides a means of livelihood for many. The voice is intimately related to most individuals' sense of self-identity. It is also an indicator of health, emotion, age, and gender. "Voice" may be defined in a broad (synonymous with speech) or narrow sense[60]. In a narrow sense "voice" refers to vocalization, production of the sound created by vocal fold vibration. Phonation is the physical and physiological process of vocal fold vibration.

The components of the entire speech production system (respiration, phonation, articulation, resonance) are of relevance to voice disorders; however in this report, only the anatomy and physiology of phonation and resonance as they pertain to the growing child will be presented. The discussion then turns to "voice" clinically; the disorders of phonation (or "dysphonias") seen in the pediatric age group. Common causes of voice problems are reviewed, and their evaluation and management.

Developmental Anatomy and Physiology of Phonation and Resonance

An appreciation of the growth and development of the larynx and vocal tract is helpful to gain perspective on voice changes throughout infancy and childhood. In the newborn, the larynx is positioned high in the neck with the cricoid at C3 to C4[23]. This arrangement facilitates simultaneous respiration and swallowing during feeding[5]. The larynx gradually descends to the level C6 to C7 by fifteen years. The effect of these changes on the voice are not related to phonation but to the vocal tract resonances[5,51]. The frequency of vocal tract formants drops as the vocal tract enlarges. Changes with the vocal tract during puberty are different for male and female; males increase the size of the pharynx relative to the oral cavity more than females[51].

The structure of the larynx also changes. Extensive measurements of the anatomic dimensions of the larynx in infancy and childhood found that laryngeal growth relates to age as overall body growth, that is a sigmoidal curve with acceleration between birth and three years, then deceleration, then rapid growth phase during puberty, especially in males. The thyroid ala in infancy are positioned in a curving semicircle of about 130 degrees. This narrows to 120 degrees in the prepubertal female and 110 degrees in the male[38]. Kahane also documented the changes in external laryngeal anatomy resulting from puberty[38,39]. Significant regional growth localized to the anterior aspect of the thyroid cartilage was measured in laryngeal specimens of pubertal males, i.e. formation of "Adam's apple". This results both in the increase in length of the anterior vocal folds, and change in the angle of the thyroid ala to 90 degrees. Other external laryngeal measurements showed less dramatic differences between pubertal males and females.

The studies of Hirano described the development of the phonatory larynx[34]. Up to 10 years, the length of the vocal fold does not vary much between males and females. At ten years of age the length of membranous portion of the vocal fold, about 6 to 8 mm, increases in females to 8.5 to 12 mm by age 20, but in males grows to 14.5 to 18 mm, more than double in length. The cartilaginous (posterior) portion of the vocal fold, formed by the arytenoid body and vocal process, is at birth already about half the adult length. In children, a larger portion of the glottis (space between the vocal folds) comprises the posterior glottis. This has been termed by Hirano et al as the "respiratory" glottis[33]. Indeed,

respiratory and protective functions of the larynx play a larger role than phonation in infants and children. The membranous portion of the vocal folds is more susceptible to edema than adults, yet because the membranous folds (the anterior or "phonatory glottis") comprise a smaller percentage of the entire glottal area these obstructive effects are minimized, serving as a relative protection.

The acoustic output of the phonatory larynx is produced by the vibration of the anterior membranous edges of the vocal folds which periodically interrupt the airstream from the lungs. This fundamental aspect of phonation is creation and maintenance of mucosal traveling waves and their entrainment with the airflow[4]. Vibration, or oscillation, is self-sustained by 1) the elastic recoil of vocal fold mucosa and, 2) by alternating pressures within the glottis that separate and bring together the folds[59,14,60]. This traveling mucosal wave, observed with high speed cinematography or stroboscopy, is created by the interaction of the airflow with the vocal fold mucosal cover. The inferior-to-superior movement of the traveling wave is influenced by the pressure of the airstream (lung pressure), the thickness of the vocal folds, the approximation of the vocal folds, and the elasticity of vocal fold tissue.

The most notable feature of the pediatric voice is pitch change. The pitch drops throughout infancy, childhood, and adolescence, for both males and females. The frequency of the infant's cry is about 500 Hz, and this drops by about one-half by age eight to ten. The male adolescent voice goes through a transition, usually between 13 and 14 years of age, where the pitch drops about 50 Hz. This is due to the anterior growth of the thyroid cartilage in response to testosterone, that causes an increase in vocal fold length. An additional change in laryngeal structure is an increase in bulk of the thyroarytenoid muscle. This causes increase in vertical thickness of the vocal fold as well as bulging of its medial contour[60]. With this change glottal closure occurs over a larger portion of the glottal cycle, and amplitude of vocal fold vibration increases, resulting in a richer quality to the voice. The pitch of both male and female voices continues to gradually drop through the rest of adolescence.

The cry of an infant or scream of a distressed child attest to the fact that children can create loud voices. Further consideration reveals that levels of acoustic power (averaging 70 decibels for conversational speech) comparable to adults are generated with a much smaller phonatory and respiratory mechanism. Several physiological principles underlie this; including the dependence of vocal intensity on frequency and lung pressure, and differences in the pediatric respiratory system. Titze explained that vocal intensity increases about 8 to 9 decibels per octave increase in fundamental frequency[60]. However, other issues play into the ability to drive shorter vocal folds at a faster rate, namely lung pressure. In recent work, Stathopoulos and Sapienza studied vocal intensity variations during phonation in 4 year old and 8 year old children and adults[58]. They found that for comparable soft, comfortable, and loud phonation tasks the children generated lung pressures 50 to 100 percent greater than the adults. In association with this, rib cage excursion for 4 year olds was equal to and for 8 year olds was nearly twice that of adults. Because lung volume excursion in children is about half that of adults, they explained that children are required to move their rib cages more to achieve the same lung volume displacement, and have greater lung volume excursion relative to vital capacity during phonation. These findings lead to the conclusion that children work harder than adults to use their voice. This increased expended respiratory effort declines to adult like patterns by 10 years of age[35].

Besides the larynx, other sites in the vocal tract where airflow is modified to create speech sounds include the velopharynx, oral cavity, and lips. The functions of articulation and resonance are intimately associated with these structures. Though a comprehensive discussion of these aspects of speech production is beyond the scope of this report, it is helpful to review the structure and function of the velopharynx, abnormalities of which may be associated with voice disorders. The velopharynx is a musculomembranous valve that, in addition to roles in deglutition and respiration, acts as a sphincter to control airflow between oral and nasal passages during speech. Its muscles are attached to the hard palate and portions of the sphenoid and temporal bones. Several paired muscular slings; palatopharyngeus, palatoglossus, levator veli palatini, superior constrictor, and musculus uvulae work in concert to elevate, depress, constrict, and relax the soft palate and pharyngeal walls that modify airflow between the oral and nasal cavities. Resonance, the amplification and filtering of sound waves produced by vocal fold vibration, is affected by the size and shape of these and other resonator cavities in the vocal tract. The opening and closure of the velopharyngeal valve is related to the consonant and vowel sounds produced. In the English language, the three nasal consonants /m/, /n/, /ng/ require opening of the VP valve, but in production of all other consonants and vowels it closes.

Evaluation of Voice Problems

Voice problems can be described as abnormalities in quality, pitch, loudness, and resonance. The most common voice quality problem is hoarseness. Causes of various voice disorders are discussed below. The approach to evaluation of a child with a voice problem differs from the adult in several respects[51]. Voice disorders in children often co-occur with speech, language, and developmental delays. Children have limited ability to cooperate, so that any procedures or assessment techniques must be performed in the least-invasive and nonthreatening manner. Parents, family members, and other care-givers are also included in

the acquisition of historical information and involved with therapy.

The assessment of voice problems requires a comprehensive medical and behavioral evaluation. Of necessity, this brings together specialists from a variety of disciplines; otolaryngology-head and neck surgery, speech-language pathology, pediatrics, psychology, and social work. Other medical specialties that may be involved include neurology, gastroenterology, genetics, and endocrinology.

The history of a voice problem often involves individuals other than the child; parents, school teachers, etc. In detailing the history of a voice problem we find useful the questionnaire reported by Maddern et al[47]. Complete examination is performed by the otolaryngologist. A hearing test is obtained. In cases of adolescent male dysphonias, physical maturity is assessed[1].

Laryngeal and velopharyngeal examination are crucial aspects of the voice evaluation. The flexible fiberoptic laryngoscope has become the tool of choice for the evaluation of voice problems, in adults and children[43]. Even very young children can tolerate this examination well with proper preparation of the child and parent, use of topical nasal anesthesia, and an engaging, nonthreatening atmosphere. These methods have been well described[19,11,45]. The use of video equipment has advantages of: engaging the child's attention, providing visual feedback for parent, and facilitating communication between physician, speech pathologist, and others. Several points may be emphasized in the course of this examination. After the fiberscope is passed beyond the turbinates, the velopharynx is examined for velopharyngeal insufficiency. The adenoid pad and tonsils are seen as the fiberscope passes beyond the nose into the pharynx. In the hypopharynx, the laryngeal and pharyngeal mucosa are inspected for erythema and mucosal thickening, especially the arytenoid mucosa and posterior glottis. Diffuse edema of the laryngeal mucosa, "cobblestoning" of the posterior pharynx, and edematous nasal mucosa may be a sign of allergy. Laryngeal mobility is inspected, including coughing and sniffing maneuvers. The larynx is observed during connected speech and repetition of phrases, such as counting to ten. These tasks are helpful for observing supraglottal hyperfunction, seen both in functional dysphonias and as compensatory behavior in organic lesions[43]. The vocal folds themselves are examined for irregularities, swelling, or lesions. Such abnormalities of the vocal fold when seen unilaterally should be suspected for congenital cysts[48,7]. During sustained vowel /I/ phonation the glottal closure is seen. Incomplete posterior glottal closure (posterior "chink") is common in children as it is in many adult females[28,32]. During fiberscopic visualization, a determination is made regarding whether stroboscopic lighting is helpful to elucidate the problem, eg. suspicion of abnormal mucosal wave vibration (due to scar, sulcus) in the absence of nodules, unilateral cord lesion, etc. Laryngostroboscopy has limitations in young children, who may not be able to phonate an adequate length of time (at least 5 seconds) to obtain stroboscopic pictures. Older children generally can tolerate rigid oral telescopic examination of the larynx very well.

The voice assessment conducted by the speech-language pathologist is ideally conducted during the same visit as the otolaryngologist. The child's voice is rated perceptually during a variety of speech tasks and sustained vowels. The voice is recorded for acoustic measures. Aerodynamic and glottographic measurements may also contribute information to the problem. These are easily conducted with current instrumentation. They have certain advantages: documentation of the problem, corroboration of findings with the laryngeal imaging examination, documentation of treatment efficacy, use as biofeedback during therapy.

There are occasions when direct visualization of the larynx under anesthesia is required for diagnosis of pediatric dysphonia. Indications for this include: 1) inability to examine the larynx fiberscopically, 2) when detailed visualization is required to determine the presence of glottic web, laryngeal cleft, arytenoid fixation, or vocal fold lesion such as papilloma, sulcus vocalis, cyst, etc. 3) perform laryngeal electromyography for paralysis assessment[42].

Causes of Hoarseness

Laryngeal Papillomas

Recurrent respiratory papillomatosis (RRP) is the most common benign laryngeal neoplasm in children (Fig.1). These papillomas can cause hoarseness, stridor, and respiratory distress, which may necessitate surgical intervention to maintain an adequate airway. Impairment of the voice may occur as a result of RRP or the treatment. While the disease is present the voice may be affected because of interference of the papilloma with vocal fold closure. However, after the disease is in remission, the voice may be affected as a complication of the surgical treatment previously used.

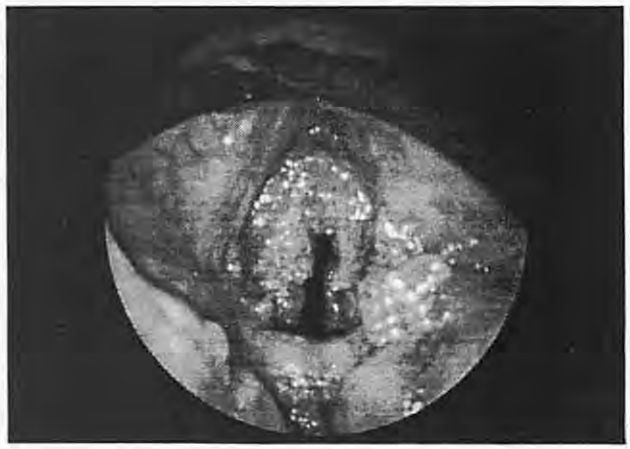

Figure 1. Laryngeal Papilloma

Crockett et al described glottic complications as a result of surgical intervention for RRP[17]. Glottic scarring, anterior and posterior glottic webbing occurred in patients who had frequent and multiple procedures.

One of the difficulties in treatment of RRP is that aggressive removal of the papillomas to provide the best airway may result in eventual injury to the vocal folds. Since severe papilloma disease obscures normal anatomic laryngeal landmarks and structures, it can be difficult to limit the surgical excision to the epithelial layers. This can be particularly true in the anterior glottis, the posterior glottis, and the membranous vocal folds. Wetmore and colleagues reported that the incidence of soft-tissue complications increased in patients who required six or more laser operations[63]. They concluded that deep vaporizations of papilloma resulted in higher incidence of glottic complications. Papilloma removal in the 1980's focused on the concept of total or near-total eradication of papilloma with each surgery[17,63]. In the late 1980's we conducted a study of voice characteristics in eight patients who had experienced extensive, repeated surgery for removal of papilloma and were then in remission for at least two years[29]. Results demonstrated reduced frequency and intensity range, and laryngeal stroboscopic findings consistent with stiff vocal folds, due to scar. Subjectively, patients did not feel their voice performed normally. Such findings raised concerns about whether the voice difficulty experienced by these patients arose from the disease or from the treatment.

An awareness of these voice disorders has led some surgeons to a more conservative approach for papilloma removal[3]. This approach is directed at two points of surgical technique. The first is use of more precise instrumentation for papilloma to minimize damage to the underlying subepithelial tissues, eg. lamina propria. Advances have been made in reducing the laser spot size, the use of a micromanipulator, and adjusting laser parameters to minimize tissue destruction[52]. The second consideration involves sparing a small area of papilloma unilaterally at the anterior and posterior glottis if that area is involved. Small cup forceps may be used to debulk papilloma from these areas, avoiding laser injury[3]. Using this approach, Ossoff et al presented 22 patients (14 children and 8 adults) where the delayed soft-tissue complications were significantly reduced[52]. Only three children had problems related to posterior glottic web or vocal fold scar. From this study, they concluded that these problems in patients undergoing papilloma surgery are related more to the surgical technique used and not to the number of procedures performed. This report underscores the need to consider the effect of these endoscopic laryngeal procedures on the voice, though the focus of the treatment involves airway and papilloma management.

Congenital Glottic Web

Congenital glottic webs are rare in children, and are nearly always located in the anterior glottis (Fig. 2). In the largest series reported by Cohen, 51 patients with congenital anterior glottic webs were seen over a 32 year period[12]. Voice problem at birth was the most commonly reported symptom. Cohen categorized glottic webs based on observation and estimation of web extent in the glottic lumen[12]. Those involving less than 35% often required no treatment. Airway symptoms increased with increasing web size. Only four of thirty-two patients with 50% or less of the glottis involved had a tracheotomy, but all sixteen patients with greater than 50% involvement had a tracheotomy. Treatment of the web involved a combination of endoscopic treatment with dilation, division or laser. Twelve patients underwent laryngofissure and placement of NcNaught or Silastic keel. Dedo reported success in managing a newborn with anterior glottic web using a Teflon® keel placed endoscopically[21]. In this patient, however, problems with development of posterior glottic stenosis were encountered from keel irritation in the posterior commissure and arytenoids. Cohen recommended waiting until 3 years of age when the larynx is of sufficient size to avoid this complication of keel placement[12]. He also emphasized that normal voice quality was rarely attained in his series. Most patients had hoarse or husky voice, weak voice, or "double voice". The comments of Zalzal et al on voice problems following pediatric laryngotracheal reconstruction also apply to these patients; the voice problems from the underlying abnormality may be compounded by surgery to correct the disease[66].

Gastroesophageal Reflux Disease

Gastroesophageal reflux disease (GERD) is common in infants and children, as well as adults. It has been implicated as a causal or etiological co-factor in many cases

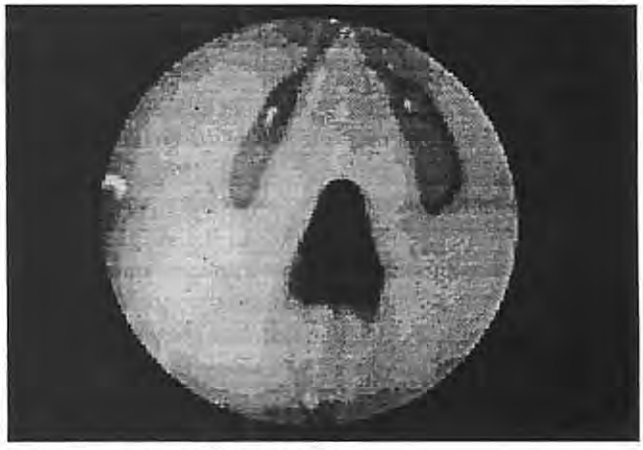

Figure 2. Anterior web of vocal folds

of adult voice problems[44]. GERD is associated with a variety of pediatric problems such as failure to thrive, asthma and SIDS. The incidence of GERD in pediatric voice disorders is unknown. A case of hoarseness associated with clinical laryngeal findings of reflux-related chronic laryngitis in a young child has been reported and several such cases in children and adolescents have been seen by the authors (personal observations)[55]. Diagnosis is inferred from laryngeal examination findings of erythema of arytenoid and posterior glottic mucosa and pooling of thickened secretions in the pyriform and/or post-cricoid region. Two channel esophageal pH probe may also be performed to confirm the diagnosis although it is expensive, uncomfortable, and does not entirely rule out GERD with a negative study in cases of intermittent reflux[44]. The treatment of hoarseness from chronic reflux laryngitis involves behavioral anti-reflux management and H_2 blockers. The association of GERD with other vocal pathologies, such as vocal nodules, is unknown at this time.

Vocal Nodules

Vocal nodules refer to a swelling, usually bilateral, present in the mid-membranous portion of the vocal fold. Histopathological studies show basement membrane zone disorganization and an abnormal extracellular matrix of the lamina propria. These vocal fold nodules impair the normal vibratory pattern of the vocal folds, thus producing what we acoustically hear as hoarseness. The size of these benign vocal fold swellings may fluctuate and be aggravated by vocal overuse or inflammatory conditions such as laryngitis (Fig. 3).

The incidence of vocal nodules among children is not clear. However, studies have estimated that vocal nodules are responsible for 38-78% of children with chronic hoarseness. This would make vocal nodules the most common laryngeal lesion. Vocal nodules are more common in boys than girls, with a ratio of about 2 or 3 to 1.

Figure 3. Vocal fold nodules

It is generally felt that nodules are the result of harsh mechanical contact between the two vibrating edges of the vocal fold. The forceful vibratory contact produces both a shearing and plunging injury. The extent of the injury is related both to the loudness or intensity of the voice produced, as well as the duration of speaking. Consequently, most children who develop vocal nodules have personality characteristics which would make them loud or incessant talkers. Toohill studied 77 children with vocal nodules and found that a high percentage of these children do have aggressive tendencies[61]. The parents of many of these children describe their children as screamers with aggressive hyperactive tendencies. The concept that pediatric vocal nodules may be entirely an abusive phenomenon may not be entirely correct. More recent studies about vocal nodules suggest that the biological composition of vocal folds to withstand stress has some genetic and consequently individual determinants. This suggests that some children may simply be able to scream and talk more without any injury to the vocal folds, while others develop vocal nodules more easily.

The presence of vocal nodules is not considered serious laryngeal disease. The natural history of pediatric vocal nodules, as opposed to vocal nodules in adults, is that of eventual resolution. Resolution of pediatric vocal nodules usually occurs by the time the child has gone through puberty. The vocal fold lengthens during this period of time and associated with a lower pitch in voice probably reduces the stress on the vocal fold and a subsequent resolution of the nodule occurs. Regardless of whether or not this theory is correct, it is clinically apparent that most pediatric vocal fold nodules do resolve without surgical treatment.

Since vocal nodules do not constitute a serious or emergent illness, it is important to differentiate those conditions which can be confused with vocal nodules but may constitute a more serious illness. Recurrent respiratory papillomatosis may also present initially as hoarseness. Other rare laryngeal conditions such as laryngeal malignancy, vocal polyps or laryngeal webs may also present initially as hoarseness. These conditions usually cause some respiratory difficulty, a symptom which is not present in vocal fold nodules. Should respiratory distress, auditory breathing or stridor be associated with hoarseness then a prompt referral for diagnosis should be obtained. Progressive hoarseness is also a worrisome sign, since vocal nodules do not usually cause progression and would indicate a more serious disease. Diagnosis of nodules is usually performed by obtaining a look at the vocal folds. This is obtained most commonly by using a flexible fiberoptic laryngoscope passed through the nose. This is an office based procedure which most children tolerate surprisingly well.

Treatment

Treatment of pediatric vocal nodules is performed with behavioral voice therapy. In the United States most children have access to voice therapy through their local school system. In most cases a note to the local school system indicating the diagnosis of vocal nodules will allow the speech pathologist within that school system to provide behavior voice therapy at no expense to the child. Although there is strong data suggesting that voice therapy is efficacious in the treatment of adult vocal nodules, there is less data to document the efficacy of the behavioral treatment of vocal nodules in children. Since most children really do not care about how their voice sounds, the motivation for following certain voice and speech modifications is lacking. Particularly when children are playing on the playground or in team sports or other social activities. Due to this inability to follow behavioral voice techniques some authors have advocated that no behavioral voice therapy is really necessary since pediatric nodules eventually spontaneously resolve. However, studies looking at behavioral treatment programs for vocal nodules have reported success. Common behavioral therapy strategies for vocal nodules are outlined in approaches by Maddern, Wilson and Andrews[47,64,1].

Studies are not available which have indicated efficacy for surgical treatment of pediatric vocal nodules, nor have studies indicated better outcome than voice therapy. Nevertheless, there may be some indications in which surgery is warranted. Bouchayer and Cornut have indicated that children have a higher incidence of congenital cysts, polyps and sulci, seen only on microlaryngoscopy in children previously diagnosed with vocal nodules[7]. When a unilateral vocal fold lesion is found it may be reasonable to recommend excision since typically vocal nodules come in pairs. In cases where it is clear that the child has reduced their vocal abuse and is following recommended speech behavior but the nodules are not resolving, then surgery may also be considered. It was suggested by Bouchayer and Cornut that these conditions are usually not met until the child has at least reached the ages of 9-11[7].

Even though nodules are thought to be essentially related to vocal abusive behavior, other conditions may lead to a predisposition for vocal nodules. Particularly inflammatory conditions of the larynx may predispose the vocal folds towards injury. An acute episode of inflammation caused by laryngitis could be a precipitating factor for the development of nodules. However, chronic inflammatory conditions are more likely to be associated with vocal fold nodules. The most common of these chronic inflammatory conditions is the coexistence of gastroesophageal reflux which spills into the larynx, thus causing chronic inflammation of the vocal folds. In some instances the hoarseness may be simply due to the presence of reflux, without the presence of vocal nodules. In other cases both vocal nodules and reflux may exist. Treatment of the reflux may lead to complete resolution of the vocal fold nodules and hoarseness, or a combination treatment including control of the reflux and behavioral voice therapy may be necessary. Another condition associated with nodules is the presence of velopharyngeal insufficiency. This condition occurs most often in children with cleft palates.

Occasionally children may have other functional voice problems in which the vocal folds look normal but hoarseness exists. These functional voice problems are best treated with voice therapy.

Vocal Fold Paralysis

Vocal fold paralysis refers to the inability of the vocal fold to move in an adductory (closure) or abductory (open) pattern. In the pediatric population, this lack of vocal fold motion is usually neurogenic in origin. Vocal fold closure and opening is due to the intrinsic muscles of the larynx acting upon the arytenoid and cricoid. The cricoarytenoid joint allows the arytenoid to swing medially in an adductory pattern or laterally and posteriorly in an abductory pattern. Adduction of the vocal folds occurs during voicing, swallowing, or airway protection events such as coughing. Abduction of the vocal folds occurs during breathing or sniffing. Although there are some unusual instances in which cricoarytenoid joint fixation is responsible for impairment of vocal fold motion, the large majority of pediatric vocal fold motion problems are due to neurological conditions in which either the nerve has been disrupted or impaired, or the central nervous system is impaired [36,6]. Therefore, the rest of this chapter will really refer to problems with vocal fold motion as laryngeal or vocal fold paralysis. In discussing vocal fold paralysis, one should remember that the anatomy of the innervation of the larynx is different on the left than it is on the right. Although neurological signals to the vocal fold musculature are carried in the tenth cranial nerve and later in the recurrent laryngeal nerve, these two nerves follow different courses. On the right side the recurrent laryngeal nerve passes the larynx, hooks around the subclavian artery to double back toward the larynx. On the left side the recurrent laryngeal nerve comes off of the vagus in the thorax, at the level of the aortic arch and then heads back up towards the larynx. On both sides the recurrent laryngeal nerve ascends in the groove between the trachea and esophagus until it enters the larynx just posterior to the cricothyroid joint. The route of the left recurrent laryngeal nerve predisposes it towards many injuries of thoracic origin.

Symptoms

Symptoms of vocal fold paralysis partially depend on whether it is unilateral or bilateral. Unilateral laryngeal paralysis refers to one vocal fold not moving while the other is normal in motion. These patients usually present with a

soft, breathy, weak or absent cry or voice. Aspiration and feeding difficulties may or may not be present, depending on how incompetent the larynx is in closure during swallowing[67]. Aspiration pneumonia may occur in severe cases, although most infants with unilateral vocal fold paralysis do not have severe aspiration problems. Auditory breathing may be present, stridor which may be inspiratory or biphasic in nature, is not usually present[2].

Bilateral vocal fold paralysis refers to the absence of motion of both vocal folds. More recent reports show that detailed examination of the larynx in many pediatric patients who were initially diagnosed with bilateral vocal fold paralysis may in fact have some adductory motion, but abduction is impaired or absent[65,30]. Patients with bilateral vocal fold paralysis or paresis present with stridor and airway difficulties. Stridor is prominently inspiratory or biphasic. It may be associated with other laryngeal problems such as laryngomalacia. Aspiration and feeding difficulties may exist and a breathy and weak cry may exist, although these symptoms often are not as prominent as they are in children with unilateral vocal cord paralysis. The subset of children that have abductory paralysis are symptomatic from airway distress, but usually have fairly normal feeding capabilities and normal voice.

Interestingly, many infants with bilateral vocal fold paralysis are able to get by surprisingly well during the first few months of life. However, as their activity increases at around 6-12 months of age, their demand for rapid oxygen exchange due to increased activity makes them more symptomatic. Many infants who do not need a tracheostomy during the first few months of life may eventually require one during the later months of the first year of life.

It is important to determine that the symptoms are in fact related to the vocal fold paralysis. Particularly in the newborn, inspiratory or biphasic stridor can be caused by many conditions of the larynx [16]. Subglottic stenosis may also cause biphasic stridor and of course laryngomalacia is the most common cause of inspiratory stridor. Other less common conditions may cause inspiratory or biphasic stridor and need to be diagnosed. The presence of stridor, airway distress, recurrent aspiration, feeding difficulties or weak voice should prompt an examination of the vocal folds. This is done quite easily using a flexible laryngoscope or in some cases a direct laryngoscopy. A flexible laryngoscopy or bronchoscopy will allow assessment of the upper airway or lower airway during breathing so the vocal fold motion can be assessed and possible presence of other associated laryngeal or airway conditions can be assessed.

Etiology of Paralysis

Laryngeal paralysis can be congenital or acquired and unilateral or bilateral. The following table lists conditions associated with laryngeal paralysis (Table 1; see following page).

Most acquired left sided unilateral vocal fold paralysis is the result of traumatic injury or compression of the left laryngeal nerve as it passes through the chest. In the neonate this is most often due to cardiovascular surgery and the most likely operation is ligation of a patent ductus arteriosis. Metastatic chest disease is also a frequent cause of left sided vocal fold paralysis.

Bilateral vocal fold paralysis in the neonate is most often related to central neurologic disorders. The most common of those being the presence of hydrocephalus or Arnold-Chiari malformations [6]. The finding of bilateral vocal fold paralysis should prompt an evaluation of the central nervous system for the presence of hydrocephalus, Arnold-Chiari Malformation, cerebral agenesis, or if the child has had a ventricular shunt placed then a possibility of shunt dysfunction must be considered[20]. Occasionally bilateral vocal fold paralysis is the first indication of shunt dysfunction. Fortunately, congenital bilateral vocal fold paralysis due to central nervous system causes have a moderate recovery rate, either spontaneously or after treatment of the underlying condition [25].

Bilateral vocal fold paralysis from syndromes, trauma or other thoracic diseases are less likely to recover spontaneously. A table of genetic syndromes associated with vocal fold paralysis is given[25] (Table 2).

Management

Management of vocal fold paralysis depends on the severity of the symptoms and the underlying cause for the paralysis. Many cases of unilateral vocal fold paralysis do not need treatment as the larynx compensates for the problem or the symptoms are not severe enough to warrant intervention. As some cases of bilateral or unilateral paralysis may spontaneously recover, in the pediatric population it is recommended that surgical treatment be delayed until it is certain that spontaneous recovery will not occur. This period of delay may be from nine months to two years. A child with significant airway distress needs emergent management. On the other hand, a child with a slightly weak voice, but is otherwise asymptomatic, can afford to wait years before management.

The optimal method of diagnosis is with the use of a flexible laryngoscope. There are occasional instances where direct laryngoscopy is useful to determine whether cricoarytenoid joint fixation is present. At present, the use of electromyography for the diagnosis and management of vocal fold paralysis is not clinically helpful. As further studies are done evaluating this method of evaluation and the information it provides, its role will be better established.

Maintenance of the airway and protection of the airway during feeding is of prime importance. The ability to breath normally and to eat are critical, especially in the first few months of life. Failure to thrive can be associated with an impaired airway. If significant airway difficulty is present

Table 1. Conditions Associated with Laryngeal Paralysis

Congenital	Acquired
Central nervous system	*Trauma*
Cerebral agenesis	Birth injury
Hydrocephalus	Post-surgical correction of cardiovascular or esophageal anomalies
Encephalocele	
Meningomyelocele	
Meningocele	*Infections*
Arnold-Chiari malformation	Whooping cough encephalitis
Nucleus ambiguous dysgenesis	Polyneuritis
Associated multiple congenital anomalies	Polioencephalitis
Mental retardation	Diphtheria
Down Syndrome	Rabies
Other cranial nerve palsies	Syphilis
	Tetanus
Peripheral nervous system	Botulism
Congenital defect in peripheral nerve fiber at neuromuscular junction, as in myasthenia gravis	Tuberculosis
Platybasia	Guillain-Barre
	Supranuclear and nuclear lesions
Cardiovascular anomalies	Kernicterus
Cardiomegaly	Multiple sclerosis
Interventricular septal defect	
Tetralogy of Fallot	
Abnormal great vessels	
Vascular ring	
Dilated aorta	
Double aortic arch	
Patent ductus arteriosus	
Transposition of the great vessels	
Associated with other congenital anomalies	
Tumors or cysts of mediastinum (bronchogenic cyst)	
Malformation of the tracheobronchial tree	
Esophageal malformation	
Cyst	
Duplication	
Atresia	
Tracheoesophageal fistula	
Diaphragmatic hernia	
Erb Palsy	
Cleft palate	
Laryngeal anomalies	
Laryngeal cleft	
Subglottic stenosis	
Laryngomalacia	

Reprint Courtesy of W.B. Saunders Company,
Philadelphia, PA; 1996.

then a tracheostomy is still the gold standard for treatment of bilateral vocal fold paralysis. Cavanaugh and Holinger reported a fifty percent occurrence of tracheostomy in children with bilateral vocal fold paralysis[10,36]. It is uncommon to have to do a tracheostomy in unilateral vocal fold paralysis. The tracheostomy may be removed when the vocal fold paralysis resolves, the condition is treated, or if the condition is permanent laryngeal surgery is done to improve the airway.

For permanent bilateral vocal fold paralysis there are many procedures designed to improve the laryngeal airway or bypass it. As mentioned earlier, tracheostomy is a gold standard in which the vocal folds are left intact and the larynx is simply bypassed for breathing. Although a tracheostomy is acceptable in the pre-school and occasionally elementary school child, tracheostomy becomes an increasingly socially difficult problem in the older child and in the teenager. Therefore, if it is apparent that bilateral vocal fold paralysis is permanent, other solutions besides a tracheostomy are sought for as the child enters the school environment. These other options include unilateral arytenoidectomy, unilateral or bilateral cordotomy, arytenoid separation surgery, and arytenoid lateralization surgery. All of these surgeries aim at providing a larger laryngeal airway, by either removing or positioning arytenoid or vocal fold tissue laterally. The tradeoff of all of these surgeries is a more breathy voice. An arytenoidectomy refers to the removal of the arytenoid, thus providing a larger glottic airway. This has been quite successful in allowing children to be decannulated from their tracheostomy [8]. A cordotomy refers to the sectioning or releasing of the membranous vocal fold from the vocal process[40]. This procedure has been shown to be quite effective in adults, although its use in children has been limited. One of the advantages to this procedure is that it can be done in gradation so that this procedure can be employed in combination with other procedures when just a little bit

Table 2. Genetic Syndromes Associated with Voice Disorders

Syndrome	Etiology	Phenotypic Characteristics	Laryngeal & Vocal Phenotype	References
Cri-du-chat Syndrome	Chromosomal 5p-	Micrognathia, hypotonia, mental retardation, midline oral clefts	Unilateral adductor paralysis; asymmetrical epiglottis, displaced aryepiglottic folds	13
Gerhardt Syndrome	Autosomal dominant	CNS involvement, tracheostomy required	Adductor paralysis; hoarseness, stridor	27, 49, 50
Private familial syndrome	X-linked recessive or autosomal	Marked physical and mental retardation	Abductor paralysis	53, 62
Private familial syndrome	X-linked recessive		Abductor paralysis	18
Private familial syndrome	Chromosome 6 linkage to HLA, glyoxalase I (GLO)	None reported	Bilateral adductor paralysis, progressive hoarseness	46
Private familial syndrome	Unknown genesis	Cricopharyngeal achalasia, cyanosis	Abductor paralysis, stridor, high-pitch low-volume cry	31

Reprint Courtesy of John S. Rubin, M.D., Robert T. Sataloff, M.D., D.M.A., Gwen S. Korovin, M.D., and Wilbur J. Gould, M.D., from the book Diagnosis and Treatment of Voice Disorders, Igaku-Shoin Medical Publishers, New York, NY; 1995.

more airway width is needed. Lateralization arytenoidectomy procedures refer to repositioning the arytenoid or vocal process in a permanently lateralized position, thus creating a larger airway[58]. Arytenoid separation surgery has been fairly recently described and is employed for those patients who have some residual adductory motion but impaired or absent abduction [30]. This is occasionally seen in patients with central nervous system origins for their laryngeal paralysis. The surgery does not interfere with arytenoid motion and hence adductory motion may be preserved while the arytenoids are positioned in a further lateral resting position. The theoretical advantage of this operation is that if adduction is preserved then voice may remain normal. Further experience with this operation will determine its role in treatment of this disorder.

Since unilateral vocal fold paralysis often recovers, conservative management in treating this problem is recommended. Unfortunately, in the pediatric population most unilateral vocal fold paralysis is traumatically or iatrogenically caused or is a result of chest malignancy[25,20]. These are generally cases that do not spontaneously recover. There is no clear consensus as to the best way to treat this condition for improvement of their voice. Historically this condition has been treated with Teflon® (polytetrafluoroethylene) injection. Over the last ten years the complications and drawbacks for this material for injection have been detailed [22,41]. This is primarily due to the inflammatory reaction which Teflon® can cause in the vocal folds. Most pediatric otolaryngologists no longer employ Teflon® as a primary means of treatment for this disorder. Unfortunately, there are no other good materials for injection, although other materials have been tried. Fat injection into the vocal fold has been described and is oftentimes employed. The advantage of fat injection to augment the vocal fold in unilateral vocal fold paralysis is biocompatibility. The disadvantage is that frequently the fat reabsorbs or is phagocytized and the resulting voice gains are not permanent. Another material used has been gelfoam, although gelfoam lasts only for a short while (3-6 weeks).

Laryngeal thyroplastic operations are commonly employed in the adult population, but may have somewhat limited use in the pediatric population[19,56]. These procedures refer to medialization of the vocal fold from an external approach. In this procedure the vocal fold is medialize by pushing in a block of thyroid cartilage next to the vocal fold. This procedure has been successful in eliminating the aspiration problems of unilateral vocal fold paralysis in the pediatric population [37]. Another surgery, arytenoid adduction, is a procedure used very successfully in the adult population for correction of unilateral vocal fold paralysis. This procedure has been successfully employed in the teenage population, but it is not recommended for patients who do not have a mature pharynx since the operation may fix the distance between the arytenoid and thyroid cartilage[56].

Velopharyngeal Inadequacy or Hypernasality

Velopharyngeal inadequacy (VPI) is a generic term used to designate any type of abnormal velopharyngeal function which results in inadequate separation of the oral and nasal cavities during speech production. During normal speech the nasal chamber should be separated from the oral chamber. This separation occurs from elevation of the soft palate towards the posterior pharyngeal wall and adenoid area and also from medial movement of the lateral pharyngeal walls at the same level. This results in a sphincteric closure or separation of the oral from the nasal space. This closure also occurs during swallowing and during certain activities such as blowing and sucking. During speech there are only three consonants in which air is sent purposefully through the nose. These are "m", "n" and "ng". Otherwise all of the English language is produced with air traveling from the larynx through the mouth.

When there is an abnormal amount of air traveling through the nose during speech the speech is characterized by a particular resonance pattern which is referred to as hypernasal speech. Many cleft palate individuals have this characteristic and hence the speech is often referred to as cleft palate speech. Hyponasality or denasality is characterized by the lack of a normal amount of air resonating through the nose. This characteristic of speech is often seen in people with huge, obstructing adenoids.

Hypernasality and the associated VPI may be due to a number of conditions including cleft palate, submucous cleft palate, congenital palatal incompetence, and many neurologic diseases which effect the ability of the palate to precisely and timely close and open the nasopharynx during articulated speech sentences. A submucous cleft palate may be hard to diagnose, but can be detected by the presence of a bifid uvula, a palpable notch at the posterior edge in the midline of the hard palate, and an associated bluish colored area or furrow down the middle of the hard palate which corresponds to the lack of muscle across the middle of the soft palate and is referred to as the zona pellucida. Congenital palatal incompetence is a term used to designate children in which the palate is often structurally normal but simply does not work adequately. Children with congenital palatal incompetence may have short palates or palates of adequate length but do not have adequate motion of their palate in their attempts to separate their nose from their mouth during speech.

In cases where congenital palatal incompetence is expected or a submucous cleft may exist, the diagnosis of syndromes which may cause palatal incompetence or palatal clefting should be entertained. One such syndrome that bears mentioning is that of velocardiofacial syndrome. Velocardiofacial syndrome is an under recognized condition in which velopharyngeal insufficiency is associated with submucous cleft palate, poor pharyngeal motion referred to

as pharyngeal hypotonia, cardiac abnormalities, learning disabilities and small stature[57].

Patients with congenital palatal incompetence or submucous cleft are at risk for VPI following any surgical intervention which affects the nasopharynx. The most common operation performed in this area is an adenoidectomy. An adenoidectomy may precipitate velopharyngeal dysfunction and may unmask or lead to hypernasal speech. To an untrained listener it is sometimes difficult to determine whether the speech characteristics reflect hypernasality or hyponasality. Performing an adenoidectomy in patients with hyponasal speech will result in improved speech. Performing an adenoidectomy in someone with hypernasal speech will lead to worsening of the condition. Therefore, a correct diagnosis of the problem is imperative. This can sometimes be facilitated by consultation with a speech pathologist trained in diagnosing and treating resonance disorders like velopharyngeal insufficiency.

Frequently the degree of hypernasality or hyponasality is an issue and may be a factor in treatment options planned. The degree of hyponasality or hypernasality is assessed by instrumentation known as nasometry[26]. The nasometer is a microcomputer based instrument designed for assessment of treatment for individuals with nasality prob-

Table 3. Referral Considerations for Velopharyngeal Inadequacy

Speech treatment	Trial Period of Speech Treatment	Medical/Surgical Management
Etiologies based on mislearning (i.e. phoneme specific nasal emissions) Normal subjective ratings of nasality Nasometric scores within 2 standard deviations of the normative Nasoendoscopy reveals adequate velopharyngeal functioning	Subjective ratings of borderline to mild hypernasality Nasoendoscopy shows inconsistent velopharyngeal closure Nasometric stimulability testing shows an improvement in nasalance scores by controlling the demands of speech using strategies such as: Slowing speech rate Reducing linguistic complexity of speech sample Using speech stimuli with low and mid vowels with non-pressure consonants Utilizing biofeedback Limited expressive speech Limited sound repertoire Inability to complete adequate assessment of the velopharyngeal port	Subjective ratings of moderate to severe hypernasality Subjective ratings of significant hyponasality Nasoendoscopy shows a consistent lack of velopharyngeal closure across a varied speech sample Nasometric stimulability testing shows a lack of improvement in abnormal nasalance scores Correct oral placement for sounds accompanied by consistent hypernasality and/or nasal emissions

lems. The instrument measures the sound energy emitting from both the nasal and oral cavities. It is well tolerated by children. Normative values have been collected for children and each child's score can be compared against these normative values. Scores outside of two standard deviations are indications of an abnormality of the velopharyngeal mechanism.

Treatment, further diagnosis and evaluation of velopharyngeal insufficiency is performed with a flexible nasopharyngoscope. In this examination the flexible nasopharyngoscope is used to examine the sphincteric action of the velopharyngeal area during speech. Based on this examination, one can determine if velopharyngeal insufficiency exists, the type of closure pattern present, the severity of the anatomic abnormality, and possible treatment options.

Treatment

Treatment of velopharyngeal insufficiency is usually based on the extent and severity of the speech problem (see Table 3). There are three possible management options: speech therapy, surgical management or prosthetic management. Speech therapy is utilized when it is felt that the patient may be instructed in obtaining better velopharyngeal closure. These patients are usually those in which during the nasopharyngoscopic examination complete closure can be seen inconsistently or consistently. Generally the nasometer scores are not as severely outside the standard deviation of normals. A trial period of speech therapy should be given in those cases of borderline or mild hypernasality. If the child does not have the normal anatomy which can functionally close the velopharyngeal port then the child will become frustrated and discouraged with speech therapy. In these patients a surgical solution should be sought. The time course for speech therapy is usually 3-6 months for an evaluation and treatment. If the child is continuing to improve after 6 months of therapy then further speech therapy should be utilized. If a child does not improve with speech therapy reaches a plateau where no further improvement is noted then a repeat evaluation with nasoendoscopy is warranted.

Surgical treatment is reserved for those with anatomic abnormalities of the velopharyngeal mechanism or those patients with hypernasality ratings of moderate to severe, which usually corresponds with nasometer scores outside of two standard deviations of the normative scores. Nasopharyngoscopy shows consistent lack of velopharyngeal closure across the varied speech sample. Surgical therapy for VPI depends upon the defect found in the velopharyngeal area. Repair of the cleft palate musculature may be needed in those patients with submucous cleft palate. Pharyngeal flap surgery refers to transposing tissue from the posterior pharyngeal wall and attaching it to the soft palate so that it obturates the middle part of the velopharyngeal port or opening (Fig. 4). This was the standard way to treat velopharyngeal problems for the past 20 years. More recently, another approach known as the sphincter pharyngoplasty has become popular in treating velopharyngeal disorders. In the sphincter pharyngoplasty, tissue from the posterior tonsillar pillars is transposed and placed transversely across the posterior pharyngeal wall about at the level of the adenoid pad. This creates an augmented sphincter which partially obturates the velopharyngeal port. Both of these procedures have been shown to be successful in treating velopharyngeal insufficiency. The pattern of closure of the velopharynx partially dictates which of these procedures is chosen and this is

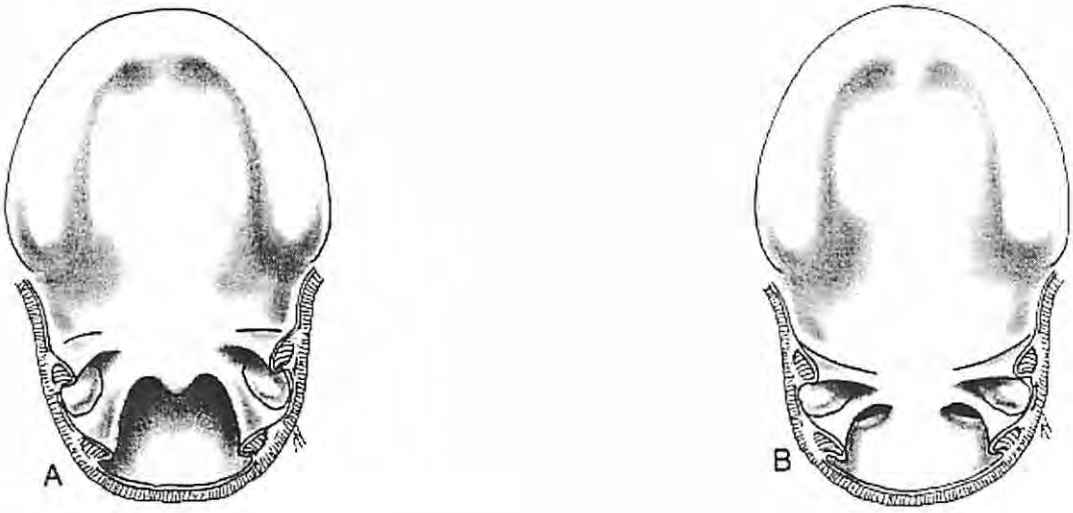

Figure 4. Schematic representation of a central pharyngeal flap. Note that the flap of tissue connects the posterior pharyngeal wall to the soft palate and obturates the middle of the pharynx.

determined during flexible nasoendoscopy. Occasionally augmentation of the posterior pharyngeal wall in the area of the adenoid pad is useful in correcting the velopharyngeal insufficiency. Tissue from the posterior pharyngeal wall can be used for this purpose, although other materials such as Teflon® or Proplast® have been used as well.

Stuttering

Stuttering has historically been one of the most confounding of communicative disorders to parents, teachers, physicians, and the individuals themselves who stutter. There are a number of different schools of thought as to the cause, appropriate onset of therapy, and type of therapy to be

Table 4. Suggestions for Parents of Children who Stutter

1. **Speak with your child in an unhurried way, pausing frequently.** Wait a few seconds, after your child finishes speaking, before you begin speaking.

2. **Reduce the number of questions you ask your child.** Children speak more freely and if they are expressing their own ideas rather than answering an adults questions. Instead of asking questions, simply comment on what your child has said, thereby letting him know you heard him.

3. **Use your facial expressions and other body language to convey to your child,** when they stutter, that you are listening to the content of the message and not to how they are talking.

4. **Set aside a few minutes at a regular time each day when you can give your undivided attention to your child.** During this time, let the child choose what he would like to do. Let him direct you in activities and decide himself whether to talk or not. When you talk during this special time, use slow, calm, and relaxed speech, with plenty of pauses. This quiet, calm time can be a confidence-builder for younger children, serving to let them know that a parent enjoys their company.

5. **Help all members of the family learn to take turns talking and listening.** Children, especially those who stutter, find it much easier to talk when there are few interruptions and they have the listeners attention.

6. **Observe the way you interact with your child.** Try to increase those times that give your child the message that you are listening to her and she has plenty of time to talk. Try to decrease criticisms, rapid speech patterns, interruptions, and questions.

7. **Above all, convey that you accept your child as he is.** Your own slower, more relaxed speech and the things you do to help build his confidence as a speaker are likely to increase his fluency and diminish his stuttering. The most powerful force, however, will be your support of him whether he stutters or not.

Reprint courtesy of Stuttering Foundation of America
(800) 992-9392

implemented. The most common ages during which we see children stutter are between 2½ to 4 years of age. Males generally stutter more often than females. The onset can be sudden and/or severe. The peak of stuttering is usually reached during the first two to three months of the onset of the disorder. The onset, as well, may be associated with an emotional event. Most often there is an integration of factors which may include physiologic, situational and linguistic features. It is important to understand that parents do not cause stuttering, rather, it is the environment which can and often does affect stuttering.

There is a difference between normal nonfluencies and actual stuttering. The characteristics, frequency, and occurrence is very different between these speech patterns. More typical types of normal nonfluencies include the following characteristics:

*5-10% of total speech samples are dysfluent
*Interjections (um, well, etc.)
*Hesitations
*Whole word repetitions of three units or less (ex. "mom, mom, mom")
*Whole phrase repetitions (ex. "I want, I want a ball")
*Highly episodic
*Improving

True stuttering usually includes the following characteristics:

*More than 10% of speech sample is dysfluent
*Part word repetitions of three units or more ("I wantwa-wa-wa-water")
*Sound prolongations
*Blocks
*Whole word repetitions of three units or more
*Struggle
*Avoidance behaviors for specific words or situations
*Occurrence is less episodic
*Not improving or worsening

Table 4 is a resource provided by the Stuttering Foundation of America which you may use as a checklist for referring your patients to a speech-language pathologist. Typically children who are exhibiting any awareness of the difference in their speech pattern will benefit from, at least, a speech and language evaluation with counseling provided to both the child (if appropriate) and the parents. In addition, Table 4 provides suggestions which you may give to parents of children who stutter.

Other Communication Disorders

Other areas included in the development of communication skills in children include articulation, phonology, language, and hearing. Articulation and phonological disorders are terms typically used interchangeably, although they are actually quite different. Articulation is the coordinated process of the oral structure and musculature to produce speech sounds. Articulation disorders may have either a sensorimotor or structural basis (dysarthria, cleft palate, syndrome related). Phonology examines the study of the sound structure of language and the sounds as they are used systematically. Phonological disorders are the results of a delayed suppression of processes and/or use of atypical (non-developmental) processes or patterns. Typically, a child is speaking in generally intelligible (understandable) two to four word phrases by age three. Speech sound development will continue through age six, however, most of the building blocks are acquired well before that age. As a rule of thumb, an unfamiliar listener is able to understand 50% of a child's speech at 24 months, 75% at 36 months, and nearly 100% of speech at 48 months.

Language is a socially acquired, primarily aural, inductively acquired symbol system used to communicate between and within individuals. It is a system of rules, a body of knowledge and a dynamic process. Babies begin learning the social rules of language in their first months of life. Possible etiologies of communication disorders may be related to prematurity, brain injury, social/emotional factors, temporary (ie. ear infections) or permanent hearing loss, poor prenatal care, congenital/genetic disorders, and psychiatric disorders. There are also many children who exhibit communication impairments of no known etiology. Table 5 is a checklist entitled "Physician's Checklist for Referral", a reprint courtesy of Stuttering Foundation of America. You may use this to help in the decision making process towards making a referral to a speech-language pathologist. Early intervention is of the utmost importance. A variety of resources have cited research findings which indicate that early intervention has a positive effect on future outcomes relating to school performance and to the number of dollars saved in later special education and therapy services [15,24,9].

You or the families with whom you work, may contact a speech-language pathologist in your area through a variety of resources. For children of all ages, the American-Speech-Language Hearing Association can provide you with a list of certified professionals in your geographic area. They may be contacted at:

American Speech-Language Hearing Association
10801 Rockville Pike
Rockville, MD 20852-3279
(301) 897-5700

In addition, the yellow pages of your telephone book will list specific professionals and agencies who provide services to children with speech and language impairments. Public Laws 94-142 and 99-457 require agencies to provide screening and referrals services for children at risk and those with communication disorders. The administering agency varies from state to state, however, is often either the

Research, have funded and currently fund many projects in these areas. Many pediatric hospitals now have voice or speech disorder clinics in which multiple disciplines are brought together to evaluate children with these problems. The child will obviously benefit best when speech and voice problems can be managed in an interdisciplinary setting when necessary and by professionals who have experience and training in these specialized pediatric problems. Given the local, professional and national resources that are expended towards recognition and treatment toward speech disorders in children it is truly a tragedy when those resources cannot be brought to assist children with voice and speech problems. Although recognition of voice and speech problems usually occurs by the parent or concerned family members, it may rest upon the pediatrician or other primary care giver to recognize these problems.

Acknowledgments

This work was supported by grant P60 DC00976, the National Center for Voice and Speech from the National Institute on Deafness and other Communicative Disorders (NIDCD). The authors express their appreciation to Helene Schneider for her assistance and contributions, and to Jennifer Lehnherr for her assistance in preparation of this chapter.

References

1. Andrews ML: *Voice Therapy for Children.* San Diego, CA, Singular Publishing Group, 1991.1
2. Benjamin B: Congenital disorders of the larynx. In Cummings CW [ed]:*Otolaryngology-Head and Neck Surgery*, ed 2. St Louis, Mosby Year Book, 1993, p 1831.
3. Benjamin B, Parsons DS: Recurrent respiratory papillomatosis: a ten-year study. *J Laryngol Otol* 102:1022-1028, 1988.4.
4. Berke GS, Gerratt BR: Laryngeal biomechanics: an overview of mucosal wave mechanics.*J Voice* 7:123-128, 1993.
5. Bloom LA, Rood SR. Voice disorders in children: structure and evaluation. *Pediatr Clin North Am* 28:957-963, 1981.
6. Bluestone CD, Delerme AN, Samuelson GH. Airway obstruction due to vocal cord paralysis in infants with hydrocephalus and meningomyelocele. *Ann Otol* 81:778, 1972.
7. Bouchayer M, Cornut G: Microsurgical treatment of benign vocal fold lesions: indications, technique, results. *Folia Phoniatr* 44:155-184, 1992.
8. Bower CM, Choi SS, Cotton RT: Arytenoidectomy in children. *Ann Otol* 103:271-278, 1994.
9. Capute A: Using language to track development. *Patient Care* 11:60, 1987.
10. Cavanaugh F. Vocal palsies in children. *J Laryngol Otol* 69:399, 1955.
11. Chait DH, Lotz WK: Successful pediatric examinations using nasoendoscopy. *Laryngoscope* 101:1016-1018, 1991.
12. Cohen SR. Congenital glottic webs in children: a retrospective review of 51 patients. *Ann Otol Rhinol Laryngol* 94 (Supplement 121):1-16, 1985.
13. Colovers J, Lucas M, Comley JA, et al: Neurological abnormalities in the "cri du chat" syndrome. *J Neurol Neurosurg Psychiatry* 35:711-719, 1972.
14. Cooper DS: The laryngeal mucosa in voice production. *Ear, Nose, and Throat J* 67:332-352, 1988.
15. Coplan J: Parental estimate of child's developmental level in a high-risk population. *Amer J of Disorders in Childhood* 136:101, 1982.
16. Cotton RT, Reilly JS: Stridor and airway obstruction. In *Pediatric Otolaryngology*, ed.3 Philadelphia, W.B. Saunders Company, 1996, chapter 79, p 1275.
17. Crockett DM, McCabe BF, Shive CJ: Complications of laser surgery for recurrent respiratory papillomatosis. *Ann Otol Rhinol Laryngol* 96:639-644, 1987.
18. Cunningham MJ, Eavey RD, Shannon DC: Familial vocal cord dysfunction. *Pediatrics* 76:750-753, 1985.
19. D'Antonio LL, Chait DH, Lotz WK, et al: Pediatric videonasoendoscopy for speech and voice disorders. *Otolaryngol Head Neck Surg* 94:578-583, 1986.
20. Dedo DD, Dedo HH: Neurogenic diseases of the larynx. In *Pediatric Otolaryngology*, ed.3 Philadelphia, W.B. Saunders Company, 1996, chapter 85, p 1352.
21. Dedo HH: Endoscopic Teflon® keel for anterior glottic web. *Ann Otol Rhinol Laryngol* 88:467-473, 1979.
22. Dedo HH: Injection and removal of Teflon for unilateral vocal cord paralysis. *Ann Otol Rhinol Laryngol* 101 (1): 81-86, 1992.
23. Dejonckere PH: Pathogenesis of voice disorders in childhood. *Acta Otorhinolaryngologica (Belgica)*, 38:307-14, 1984.
24. Eilers B, Nirmals S, Wilson, M, et al: Classroom performance and social factors of children with birth weights of 1250 grams or less. *Pediatrics* 77:203, 1986.
25. Emery PJ, Feron B: Vocal cord palsy in pediatric practice. A review of 71 cases. *Int J Pediatr Otorhinolaryngol* 8:147-154, 1984.
26. Fletcher SG: "Nasalance" vs listener judgments of nasality. *Cleft Palate J* 13:31, 1976.
27. Gacek RR: Hereditary abductor vocal cord paralysis. *Ann Otol* 85:90-93, 1976.
28. Glaze LE, Bless DM, Susser RD: Acoustic analysis of vowel and loudness differences in children's voice. *J Voice* 4:37-44, 1990.
29. Gray SD, Barkmeier J, Shive C, et al: Vocal function in papilloma patients. Presented at the American Society of Pediatric Otolaryngology, Waikoloa, HI, May 5-6, 1991.
30. Gray SD, Kelly SM, Dove H: Arytenoid separation for impaired pediatric vocal fold mobility. *Ann Otol Rhinol Laryngol* 103:510-515, 1994.
31. Grundfast KM, Milmore G: Congenital hereditary bilateral abductor vocal cord paralysis. *Ann Otol Rhinol Laryngol* 91:564-566, 1982.
32. Hirano M, Bless DM: *Videostroboscopic Examination of the Larynx.* Singular Publishing Group, Inc, San Diego, CA, 1993.
33. Hirano M, Kurita S, Kiyokawa K, et al: Posterior glottis: morphological study in excised human larynges. *Ann Otol Rhinol Laryngol* 95:576-581, 1986.
34. Hirano M, Kurita S, Nakashima T: Growth, development, and aging of the human vocal folds, in Bless DM, Abbs JH (eds.): *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, pp 22-43. San Diego, College-Hill Press, 1983.
35. Hoit JD, Hixon TJ, Watson PJ, et al: Speech breathing in children and adolescents. *J Speech Hear Res* 33:51-69, 1990.
36. Holinger LD, Holinger PC, Holinger PH. Etiology of bilateral abductor vocal cord paralysis-a review of 389 cases. *Ann Otol Rhinol Laryngol* 85:428, 1976.
37. Isaacson G: Extraluminal arytenoid reconstruction: Laryngeal framework surgery applied to a pediatric problem. *Ann Otol Rhinol Laryngol* 98:135-140, 1989.
38. Kahane JC: A morphological study of the human prepubertal and pubertal larynx. *Am J Anat* 151:11-20, 1978.
39. Kahane JC: Growth of the human prepubertal and pubertal larynx. *J Speech Hear Res* 25:446-455, 1982.
40. Kashima HK: Bilateral vocal fold motion impairment: pathophysiology and management by transverse cordotomy. *Ann Otol* 100:717-721, 1994.
41. Kasperbauer JL, Slavit DH, Maragos NE: Teflon® granulomas and overinjection of Teflon: a therapeutic challenge for the otorhinolaryngologist. *Ann Otol Rhinol Laryngol* 102(10): 748-751, 1993.
42. Koch BM, Milmoe G, Grundfast KM: Vocal cord paralysis in children studied by monopolar electromyography. *Pediatr Neurol* 3:288, 1987.
43 Koufman JA: Approach to the patient with a voice disorder. *Otolaryngol*

Clin North Am 24:989-998, 1991.
44. Koufman JA. Gastroesophageal reflux and voice disorders. In Rubin JS, Sataloff RT, Korovin GS, Gould WJ (eds). *Diagnosis and Treatment of Voice Disorders*, Igaku-Shoin, New York, 1995, pp 161-175.
45. Lotz WK, D'Antonio LL, Chait DH, et al: Successful nasoendoscopic and aerodynamic examinations of children with speech/voice disorders. *Int J Pediatr Otorhinolaryngol* 26:165-172, 1993.
46. Mace M, Williamson E, Morgan D: Autosomal dominantly inherited adductor laryngeal paralysis-A new syndrome with a suggestion to linkage to HLA. *Clin Genet* 14:265-270, 1978.
47. Maddern BR, Campbell TF, Stool S. Pediatric voice disorders. *Otolaryngol Clin North Am*, 24:1125-1140, 1991.
48. Monday LA, Cornut G, Bouchayer M, et al: Epidermoid cysts of the vocal cords. *Ann Otol Rhinol Laryngol* 92:124-127, 1983.
49. Morelli G, Mesolella C, Cavaliere ML, et al: Autosomal dominant inheritance of Gerhardt's syndrome in three generations of a family (letter). *J Neurol Sci* 47:325, 1980.
50. Morelli G, Mesolella C, Costa F, et al: Familial laryngeal abductor paralysis with presumed autosomal dominant inheritance. *Ann Otol Rhinol Laryngol* 91:323-324, 1982.
51. Morrison M, Rammage L, Nichol H, et al: Pediatric voice disorders: special considerations. In *The Management of Voice Disorders*. San Diego, CA, Singular Publishing Group, 1994, pg 120-140.
52. Ossoff RH, Werkhaven JA, Dere H: Soft-tissue complications of laser surgery for recurrent respiratory papillomatosis. *Laryngoscope* 101:1162-1166, 1991.
53. Plott D: Congenital laryngeal-abductor paralysis due to nucleus ambiguous dysgenesis in three brothers. *N England J Med* 271:593-597, 1964.
54. Priest RE, Ulvestad HS, Van de Water F, et al. Arytenoidectomy in children. *Ann Otol Rhinol Laryngol* 69:869, 1966.
55. Putnam PE, Orenstein SR: Hoarseness in a child with gastroesophageal reflux. *Acta Pediatr* 81:635-6, 1992.
56. Smith M, Gray S: Laryngeal Framework Surgery in Children. *Adv in Oto Head and Neck Surg* 8:91-106, 1994.
57. Sprintzen RJ, Goldberg RB, Lewin ML, et al. A new syndrome involving cleft palate, cardiac anomalies, typical facies, and learning disabilities: velocardiofacial syndrome. *Cleft Palate J* 15:56, 1978.
58. Stathopoulos ET, Sapienza C: Respiratory and laryngeal measures of children during vocal intensity variation. *J Acoust Soc Am* 94:2531-2543, 1993.
59. Titze IR: Comments on the myoelastic-aerodynamic theory of phonation. *J Speech Hear Res* 23: 495-510, 1980.
60. Titze IR: *Principles of Voice Production.* , Englewood Cliffs, NJ, Prentice-Hall, 1994.
61. Toohill RJ: The psychosomatic aspects of children with vocal nodules. *Arch Otolaryngol* 101:591-595, 1975.
62. Watters GV, Fitch N: Familial laryngeal abductor paralysis and psychomotor retardation. *Clin Genet* 4:429-433, 1973.
63. Wetmore SJ, Key JM, Suen JY: Complications of laser surgery for laryngeal papillomatosis. *Laryngoscope* 95:798-801, 1985.
64. Wilson DK: *Voice Problems in Children*, 3rd edition. Baltimore, MD, Williams and Wilkins, 1987.
65. Woodman D, Pennington CL. Bilateral abductor paralysis. *Ann Otol Rhinol Laryngol* 85:437, 1976.
66. Zalzal GH, Loomis SR, Derkay CS, et al: Vocal quality of decannulated children following laryngeal reconstruction. *Laryngoscope* 101:425-429, 1991.
67. Zitsch RP, Reilly JS. Vocal cord paralysis associated with cystic fibrosis. *Ann Otol Rhinol Laryngol* 96:680, 1987.

NCVS Status and Progress Report - 10
November 1996, 151-153

The Singing Voice

Ingo R. Titze, Ph.D.
Department of Speech Pathology and Audiology, The University of Iowa

Traditional Voice and Speech Analysis

Traditional voice analysis is based on extraction of temporal and spectral features from a microphone signal. This "observation at a distance" via an airborne signal has its limitations, especially if sound production at the larynx and sound transmission through the airways are to be studied separately. The microphone signal is an unfortunate mixture of the combined properties of the source of sound and the propagation of sound through the airways (known as the filter). Thus, speech scientists have been looking for additional information to augment acoustic recordings of human voices. The source-filter theory of voice production could then be studied in greater detail. In particular, the use of fiberoptic viewing of the vocal folds has provided important information about the source, while Magnetic Resonance Imaging (MRI) and Electron Beam Computed Tomography (EBCT) has produced three-dimensional shapes of the vocal tract airways (the filter). Both of these techniques still have limitations with temporal and spectral resolution, but results are promising.

Fiberoptic Imaging of Vocal Folds

Fiberoptic imaging of the larynx, and more specifically fiberoptic imaging of the vocal folds that vibrate to produce the sound, falls into two categories: (1) stroboscopic imaging at normal video frame rates (30 frames per second) and (2) high speed video imaging (1000-10,000 frames per second). Since the vocal folds typically vibrate at 100-1000 Hz in speech and singing (male and female ranges included), it is clear that a 30 Hz frame rate will at best capture one view of the folds every 3 to 4 cycles, and at worst one view every 30 to 40 cycles (Figure 1). This is acceptable if the vibration pattern is periodic. A stroboscopic flash can then be synchronized with the frequency of the vocal folds (as detected on the skin of the neck), and the phase of the flash can be moved around to observe specific points within the vibration cycle. The frequency of the strobe flash can also be detuned slightly from the frequency of the vocal

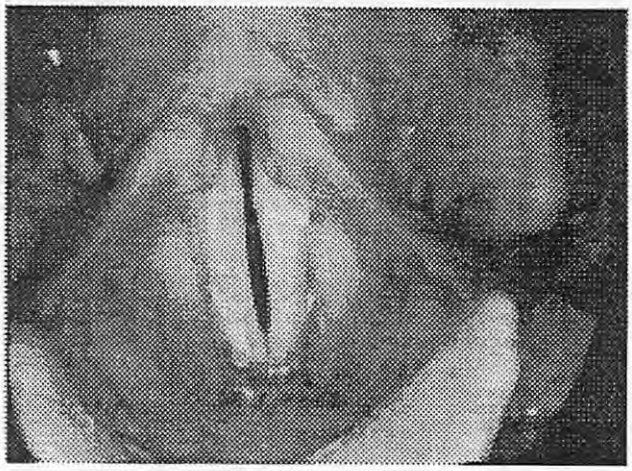

Figure 1. Fiberoptic image of the larynx (top view) during phonation. The dark vertical slit in the middle is the glottis, an airspace between the vocal folds. The vocal folds are the bright bands on either side of the glottis, which vibrate at frequencies from about 100-1000 Hz.

folds to observe the vibration in slow motion. This type of observation has given clinicians useful information about amplitude of vibration, degree of contact (collision) between the folds, and the modes of vibration. But when the modal pattern is complex and variable from cycle to cycle, the stroboscopic technique fails and high speed imaging must be invoked. With this technique, 10-100 images can be obtained within each cycle of vibration, depending on the frequency of vibration and the frame rate of the high speed video system. This temporal resolution is often adequate, however, to capture transient responses (such as in a cough or a sudden pitch jump) and complex vibratory modes (such as in a hoarse voice or a high whistle register of a singer).

Imaging of the Vocal Tract

MRI and EBCT images of the vocal tract are basically static, affording no temporal resolution of dynamically changing airway structures in speech or singing. But

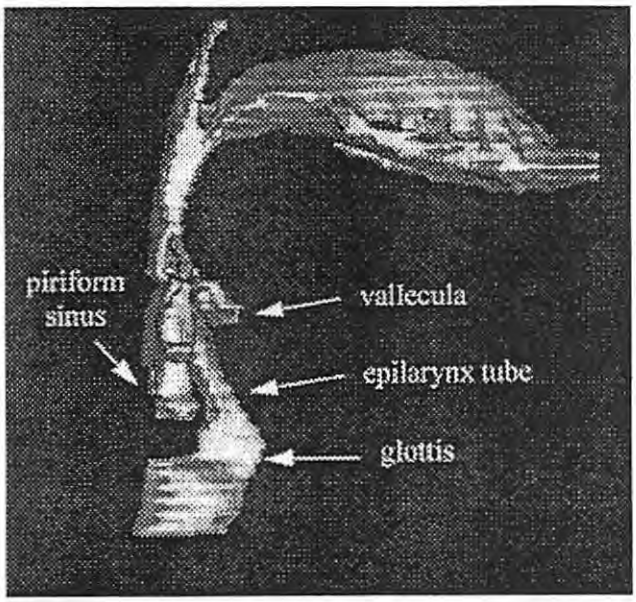

Figure 2. Sagittal (side) view of the human airway as obtained from electron beam computed tomography (EBCT). The vowel shape is an /ɑ/ as in father. The larynx is at the bottom and the mouth is on the top right.

the spatial resolution is so attractive that even these static shapes help to uncouple the propagation characteristics from the source characteristics. A subject typically spends many hours in the MRI machine to "map out" his or her articulatory space by mimicking the vowels and consonants in speech and holding each steady for several minutes (Figure 2). A complete volumetric scan, with millimeter accuracy, is obtained for each shape, including the oral cavity, the nasal cavity with the sinuses, the pharynx, the larynx, and the trachea. In and around the larynx are small pockets of air that make interesting resonators for sound (see arrows in Figure 2).

Special Features of the Singing Voice

Recent research has shown that trained singers are able to optimize their source-filter interactions to obtain easier and more efficient voice production. First, everyone has noticed that singers open their mouths wide, particularly at high notes. This has the acoustic benefit of producing a better impedance match from the airway to free space, much like the flare of a trumpet. Impedance (the ratio of acoustic pressure to airflow) is high at the larynx and low at the mouth, requiring a megaphone-like transformation. A wide mouth also has the benefit of getting the jaw out of the way of the larynx. Often a tense jaw restricts the larynx in finding a position that can be held over a wide range of pitches.

This brings up the second point about trained singers. The frequent practice of vocalises (little nonsense songs) on a variety of chosen vowels and consonants and over wide pitch ranges builds a certain source-articulator independence that untrained vocalists don't have. Voice quality, pitch and loudness can then be changed or held constant (at will) when the words of a song impose their own vowel and consonant structure. Speech articulation thus becomes a mere modulation of an instrument-like carrier sound that is little affected by this modulation.

Another benefit of training is the use of vocal tract resonances to enhance the acoustic output power of selected partials (overtones). Males and females use somewhat different strategies. In males, there is generally a paucity of acoustic energy in the 3000 Hz region because the fundamental frequency is low (100-500 Hz). At 100 Hz, for example, the 30th partial is at 3000 Hz, whereas at 500 Hz the 6th partial is at 3000 Hz. These partials, as produced by the larynx, have considerably less energy than the fundamental, thus tilting the overall frequency spectrum toward the low end. In terms of a woofer-tweeter analogy, the woofer carries excessive amounts of power. By creating a small resonance tube above the vocal folds (in the region of the ventricular folds) the singer can boost the high frequency energy selectively in the 3000 Hz region. This is perceived as the operatic "ring" in the voice, one of the most beloved qualities in tenors such as Enrico Caruso or Placido Domingo.

Female singers have less of a need to boost the high frequency portion of their acoustic spectrum because their fundamental frequencies are generally twice as high as males in song. Nevertheless, they sometimes use vocal tract resonance to boost selective portions of the spectrum. In particular, when the soprano uses a high "lifting" or "floating" quality, there is a desire to make the sound pure, almost free of harmonic structure. The natural resonance of the entire length of the vocal tract (from larynx to the lips) is then used to enhance the fundamental rather than an overtone. The result is a flute-like sound rather than a brass-like sound.

What Makes A Premier Singing Voice?

The singing voice characteristics mentioned so far, plus others such as the use and control of vibrato, are achievable by most people with voice training. But what is the make-up of a premier voice, one that comes along only once in a decade and can be recognized by everyone as special? The answer to this question is still speculative, but some hints are beginning to emerge. Naturally great voices tend to have a non-encumbrance of skeletal structures in and around the larynx. This involves, at a minimum, the shoulders, the neck, and the jaw. As a wide range of pitches and loudnesses are accessed, soft and hard tissues don't interfere with each other. There is enough room for expansion, contraction, and linear displacement of the larynx, ribcage, abdomen, diaphragm, and the airways, as necessary. This is not to say that they move a lot - on the contrary, a firm equilibrium position (posture) is desirable, but the critical

movements must be unencumbered. This applies particularly to the use of opposing muscles (agonist-antagonist pairs). They must not fight each other, but rather be able to turn on and off gradually (like a dimmer switch) to move structures and change tensions precisely and differentially. Jerky on-off movements are seldom seen in a premier singer. Rather, there is a death-like calmness on the surface, underneath which huge muscular efforts are expended.

Within the larynx, there are likely to be some morphological differences between ordinary and premier singers, although direct verification by inspection of the organs of deceased singers has not been possible. Scientists have relied on simulation, therefore, to test the "optimal" structures for sound production. Symmetry between the left and right vocal folds seems to play an important role. In principle, the two vocal folds have their own characteristic modes of vibration (like drums, bells, or strings). These modes depend on the viscoelastic properties of the vocal fold tissues and the boundaries that surround the tissues (the cartilages). If either the boundary structures or the internal tissue properties of the vocal folds are asymmetric, different modes (with different natural frequencies) can be excited. These modes can fight each other. A common airflow between the vocal folds does help to entrain the modes, but there is a limit to this entrainment. If large ranges of pitch and loudness are to be achieved, a highly symmetric pair of vocal folds has a much better chance of avoiding chaotic oscillation.

Computer simulation and physical construction of self-oscillating models of the vocal folds have also shown that a large benefit is obtained by having a thick, pliable mucosa as a covering of the vocal folds. This mucosa propagates a surface wave while the vocal folds are vibrating. In fact, it is the surface wave that facilitates the energy transfer from the airstream between the vocal folds to the tissue itself, thereby producing self-oscillation. Highly gifted singers probably have the genetic construct of a thick and pliable vocal fold mucosa, although direct histological verification is yet pending.

Underneath the loose, pliable mucosa must be a tough ligament that can support large tensions, much like a piano or violin string. For high pitches, this ligament absorbs most of the tension in the vocal folds. The amount and the type of collagen and elastin fibers that make up this vocal ligament may again be genetically determined. Thus, some people may be "born" with better material properties than others, much like certain woods or metals are more desirable for musical instrument design.

Bibliography

Sundberg, J. (1987). *Science of the Singing Voice.* DeKalb IL: Northern Illinois University Press.

Titze, I.R. (1994). *Principles of Voice Production.* Needham Heights, MA: Allyn & Bacon.

Titze, I., Mapes, S., & Story, B. (1994). Acoustics of the tenor high voice. *Journal of the Acoustical Society of America, 94*(2), 1133-1142.

Story, B., Titze, I., & Hoffman, E. (1996). Vocal tract area functions form magnetic resonance imaging. *Journal of the Acoustical Society of America, 100*(1), 537-554.

NCVS Status and Progress Report - 10
November 1996, 155-157

Continuing Education Update

Julie Ostrem, Continuing Education Coordinator
Department of Speech Pathology and Audiology, The University of Iowa

The escalating popularity of computer software, satellite transmissions and the Internet have forever changed the way we communicate in America. These technological strides allow communicators to exchange information more quickly, cheaply, and ecologically, and often better tailored to individual needs. Obviously, these advances have been a boon to the NCVS and other organizations providing continuing education. While NCVS's primary goal in continuing education remains the same - to distribute cutting-edge research findings in formats most usable to the clinician - the means to accomplish this mission continue to evolve.

CD-ROM Development

A 1994 survey revealed that most master's students in speech pathology programs receive limited exposure to normal voice production in their curricula (VanMersbergen, Ostrem, Titze, 1994, in review). Because of this knowledge gap, continuing education programs focusing on normal voice production can play an important role in improving the quality of care for the voice client. Jeff Fields has been hired at Iowa to develop a CD-ROM that explains the mechanisms of voice production. Jeff's background includes operatic performance and computer science, and thus, he is ideally suited to bridge the scientific principles of the voice mechanisms to practical applications.

The singer's formant was the first topic Jeff developed. Through text and other enhancements, the section explains why an operatic voice can be heard over the sound of a larger (and seemingly louder) orchestra. A synopsis of Dr. Johan Sundberg's research on opera singers and his models help explain the phenomenon. Also, MRI scans of a subject's vocal tract - the work of Dr. Brad Story - are included. These scans help the user better understand how the airspace of the vocal tract changes for two vowel sounds. Finally, users are linked to an animation that demonstrates with a simple tube model how the singer's formant works. The user hears audio changes as the open end of the acoustic tube (the vocal tract) is constricted and enlarged.

A second topic , the source filter theory of vowels, is also now fully developed. Four "rules" for modifying vowels are described:

- All formant frequencies decrease uniformly as the length of the vocal tract increases.
- All formant frequencies decrease uniformly with lip rounding and increase with lip spreading. An audio sample demonstrates how this sounds.
- A mouth constriction lowers the first formant and raises the second formant.
- A pharyngeal constriction raises the first formant and lowers the second formant.

Once the user understands these concepts, s/he is linked to an animated construction of an F_1-F_2 vowel chart.

In all, one dozen topics will be developed in a similar manner: new research will be integrated with explanations of basic scientific principles. Animations, movies, sounds and graphics will help the learner break down complex ideas into manageable learning modules. Future topics include vocal fold oscillation, voice classification, vocal fatigue, control of fundamental frequency and voice fluctuations. The development phase will be complete by January 1999. At that point, test CD-ROM's will be mailed to instructors of speech pathology and other voice professionals for evaluation and feedback. Dr. Titze's Principles of Voice Production course also will use the software as a learning lab. Once the feedback has been integrated and design phase complete, the CD-ROM will be made available for sale January 1, 2000.

Teaching via the Satellite

In a cooperative venture with the National Center for Neurogenic Communication Disorders at The University of Arizona, Dr. Ingo Titze presented Telerounds #32 October 16, 1996. Its title was "Control Mechanisms of the Larynx Under Normal and Paralytic Conditions." Dr. Titze

presented a live talk, accompanied by video clips and other graphics to more than 339 speech-language pathologists, students and other professionals. In his talk, Dr. Titze described the mechanisms used by individuals to vary their voices: pitch, loudness, tightness, register and resonance. These variables were presented in detail for the normal voice and a voice affected by unilateral vocal fold paralysis.

The Information Highway

The NCVS staff at Iowa has created a homepage for the Center (http://www.shc.uiowa.edu/ncvs_home.html). This site is linked to the other three NCVS sites: The Denver Center for the Performing Arts, University of Wisconsin-Madison and University of Utah. The NCVS sites are, in turn, linked to other sites providing voice and speech information. Thus, a user may gather information from a variety of institutions on a specific topic related to voice or speech. For example, the user may begin at the NCVS homepage, select the hypertext for The University of Wisconsin, choose the Department of Otolaryngology/Head and Neck Surgery, click on the "links" button, select the hypertext notation "Laryngeal Cancer", and arrive at the website of the National Cancer Institute. This inter-connectivity gives the user one-stop shopping for information about voice and speech.

At the primary NCVS site, the browser can learn about NCVS personnel, vocal health and current research, and be serenaded by Pavarobotti at the "fun stuff" page. Many of the written publications created as NCVS Information Dissemination and Continuing Education projects appear in electronic formats at the NCVS website. New topics are continuously added, often by students. Marisa Davis, a speech pathology student at Iowa, is creating web pages describing vocal qualities as a one-credit research course. Joe Tojek, a doctoral student at Wisconsin, is collaborating with Dr. Diane Bless to create three-dimensional animations of the voice and speech systems. In the near future, volumes of the status and progress reports will be made available in electronic formats, allowing subscribers to download only the manuscripts relevant to their research interests.

Since May 1995, 4,044 individuals have accessed the NCVS homepage.

Other types of electronic teaching and learning are interactive, however. Dr. Michael Karnell manages a listserver for professionals interested in voice and voice disorders. It is cosponsored by the American Speech-Language-Hearing Association Special Interest Division 3 and The University of Iowa Department of Otolaryngology-Head and Neck Surgery. The service promotes discussion among health care professionals, scientists and professional voice users about voice and voice disorders. There are currently 331 subscribers worldwide.

Other investigators have discovered that they can contribute to electronic dialogues about voice and speech disorders on various listservers, such as those for Parkinson disease, stuttering, and other communication difficulties. Often these requests come from family members of individuals with these disorders. NCVS researchers can direct them to current journal articles that these individuals may otherwise not locate. Other requests to NCVS investigators are initiated by practitioners looking for scientific bases for challenges they encounter in the clinic.

Conferences, Workshop, Written Materials

NCVS investigators are not abandoning the more traditional means of teaching, however. Conferences, workshops and written publications still play an important role in communicating research findings to practitioners.

In January 1996, Dr. Harry Hoffman organized a Clinical Laryngology Update at The University of Iowa Hospital and Clinics. Approximately 75 otolaryngologists, speech pathologists, scientists and students from Iowa, Illinois and Wisconsin attended the day-long program. Dr. Peak Woo, an otolaryngologist from Boston, was the featured speaker. Dr. Woo presented sessions on the use of miniplates in laryngeal framework surgery and endoscopic microlaryngeal surgery. Other faculty members were from Iowa's Department of Otolaryngology and the National Center for Voice and Speech.

In July 1996, Drs. Diane Bless and Charles Ford organized the fourth biennial Phonosurgery Symposium at The University of Wisconsin-Madison. Approximately 125 otolaryngologists, voice scientists and speech-language pathologists heard presentations on diagnosis and assessment, phonosurgical treatment and evolving research. Ten workshops accompanied the formal lectures, with topics ranging from laser surgery to laryngeal EMG to computer technology in the clinic. Program sponsors were the Division of Otolaryngology-Head and Neck Surgery, Department of Surgery, and Continuing Medical Education, University of Wisconsin Medical School; University of Wisconsin-Extension; the NCVS; and Meriter Hospital-Park. The next Phonosurgery Symposium will be held in July 1998.

Check out the NCVS site on the
World Wide Web!
http://www.shc.uiowa.edu/ncvs_home.html

The Lee Silverman Voice Treatment has been proven as a highly effective method for the treatment of patients with Parkinson disease. Dr. Lorraine Ramig, one of the cofounders of the LSVT, is actively teaching the method to speech-language pathologists in the United States and abroad. In the past year, Dr. Ramig and her colleagues at the Denver Center for the Performing Arts have presented 11 such workshops. Because Dr. Ramig could not possibly meet all requests for these one- to two-day workshops, the method has been published in a 126-page guidebook, which is sold through the Iowa offices of the NCVS. In its first year of publication, more than 500 copies of the LSVT guidebook were sold. Currently, Dr. Ramig is preparing a second edition which will be distributed by Singular Publishing Group, Inc.

Without exception, every NCVS investigator has served at least once as a presenter or coordinator for a continuing education activity in the past year. Some of the venues are large, professional meetings such as the American Association of Otolaryngologists - Head and Neck Surgeons. Other opportunities involve smaller audiences, such as stuttering workshops conducted for speech-language pathologists in small Midwestern communities. Other endeavors include book editing, writing of book chapters and columns. In all, more than 17,000 practitioners have been directly reached by NCVS continuing education efforts since August 1995.

NCVS Status and Progress Report - 10
November 1996, 159-160

Information Dissemination Update

Cynthia Kintigh, Dissemination Coordinator
Wilbur James Gould Voice Research Center, The Denver for the Performing Arts

Specific Aims

The goals of the dissemination project are to distribute information to the general public about care of the voice as well as prevention, detection, and treatment of voice and speech disorders. Methods for reaching the public include the use of media (electronic and print), educational presentations (workshops, seminars and lectures), interactive exhibits and shared information with other professional organizations dedicated to voice care, speech and training.

Public Service Announcements

The television Public Service Announcement produced under the first grant period was sent to an additional eleven secondary markets in March, 1996. (See NCVS Status and Progress Report #8, July 1995). Again we chose some cities with a high number of performers (Austin, Lousiville) or proximity to a voice research center (Baltimore). Orlando was chosen because of the high number of vacationers in the area. The phone number for the Denver site of the NCVS appears on the screen at the end of the spot. Calls from outside the Denver area have been referred to an appropriate agency in the viewer's area.

We have produced a follow-up PSA, "The Voice Doctor," and will distribute it this winter. We hope to continue our success with this clever PSA which emphasises the point that your voice is your "vocal fingerprint" through the use Jerry Lewis, Jimmy Stewart, Clint Eastwood, and Elvis Preslsy -- all via an impressionist. A voice doctor gives tips for vocal care, transforming himself into these famous personalities. As the second PSA is conceived as a "next step" to the first, it will be sent to the cities in the first PSA distribution as a follow-up.

Second Distribution of Public Service Announcement

MARKET	NUMBER OF TELEVISION HOUSEHOLDS
Sacramento, CA	1,100,000
Cleveland, OH	1,142,000
*Tampa/St. Petersburg, FL	1,384,000
*Pittsburgh, PA	1,141,000
Phoenix, AZ	1,017,000
Orlando, FL	998,000
Baltimore, MD	980,000
Portland, OR	933,000
Indianapolis, IN	925,000
Kansas City, MO	780,000
*Louisville, KY	533,000
Austin, TX	417,000
Total potential exposure:	11,350,000

Source: February 1996 A. E. Nielson report or *1995 Medium Market Guide, as indicated.

Note: A "television household" is measurement used by the A. E. Nielson company in measuring ratings and is the standard by which television markets are ranked. The measurement represents the number of households within a city, not the number of occupants or televisions.

Newsletters, Journals and Public Relations Efforts

Dr. Brad Story was a guest in two installments of a syndicated national radio program *Pulse of the Planet* in November, 1996. This story also was the featured topic of the *Pulse of the Planet* website in November.

The April, 1996 "Breakthroughs" section of *Discover* magazine ran a photograph and a "mini-article" on Dr. Brad Story's research in extracting vocal tract shapes using MRI and voice simulation.

An article on Dr. Story's MRI research appeared in the June 24, 1996, issue of *Advance for Radiologic Science Professionals* : "MR Studies Show How Sound Is Produced."

A feature article entitled "Recreating the Human Voice," appeared in the January 19, 1996 issue of *The Chronicle of Higher Education.*

A feature article entitled "A Human Touch-Computer Voices Get Emotional," appeared in the February 18, 1996, issue of the *Des Moines Sunday Register.*

The cover story of the February 29, 1996 issue of *Advance*

for Speech Language Pathologists highlighted the Lee Silverman Voice Treatment and the work underway in Denver using this treatment.

Dr. Ingo Titze's study of the performance voice was the cover story of the September 23, 1995 issue of *New Scientist* magazine, "What's In A Voice."

An article on the Lee Silverman Voice Treatment was distributed nationally to 200 newspapers via the Maturity News Service; reaching 2-3 million readers. The article was also made available to "Senior Net" a service of the internet service provider America On Line. To date we have notification from Maturity News Service that the article has appeared in the following newspapers:

Newspaper and City
Canton, Ohio *Repository*
Woonsocket, RI *Call*
Naples, FL
Sun City, AZ
Medford, OR

Pavarobotti

A new version of Pavarobotti has been completed and presented in performance in San Diego in February, 1996. The robot now has shoulder and elbow joints which can be programmed. The present performance uses an animated face and sings. A Linear Predictive Coding (LPC) system has been used to analyze, synthesize, and modify sentance level speech. We hope to use it with Pavarobotti to demonstrate speech modification techniques. The modified speech is ready, but an animated face needs to be constructed to accompany the voice. Work continues on the robot as we expand the speech and singing capabilities and explore exhibit and performance venues.

We are currently programming the robot to sing the national anthem and a new duet with Dr. Titze.

We have received addtional outside funding for a touring middle school program for the robot. The program is being developed by the principal pianist for the Colorado Symphony and will feature the pianist, one or two additional singers, and Pavarobotti. As Pavarobotti uses synthesized speech technology to speak and sing, his voice can be manipulated from male to female, can show the effects of bad habits on the voice such as shouting and smoking, and can even demonstrate what is called "source sound" - the sound of the vocal folds vibrating without the resonance created by the head. There is also the possibility that one of the engineers who developed the robot would be available to talk about the construction and programming of Pavarobotti.

Web Site

A World Wide Web site on the internet has been established. The content includes background on research projects undertaken as part of the NCVS, information on educational materials and workshops, and research training opportunities. The URL of the site is http://web1.dcpa.org.

Workshops

The public voice workshops continue to be one of the NCVS's most visible outreach efforts. Four public voice workshops were presented. This year a new workshop was designed as a more intensive follow up for voice users who have previously attended one of our voice workshops. This advanced workshop was presented once. Again this year, as an additional community outreach effort, two of the general workshops were presented at suburban arts centers, The Arvada Center for the Arts and Humanities and at the University of Colorado - Boulder, Imig School of Music. One workshop was presented through the Denver Center for the Performing Arts Latino Task Force and Public Radio station KUVO-FM.

This year, Lee Silverman Voice Treatment workshops have been presented in Boston, Dayton, Milwaulkee, Phoenix, Tuscon, Boston, Seattle, Marshfield, WI, Lorraine, OH, and Sydney and Melbourne Austrialia. A Lee Silverman Voice Treatment workshop was presented at the Denver Center for the Performing Arts May 17-18.

Training Update

Patricia Zebrowski, Training Coordinator
Department of Speech Pathology and Audiology, The University of Iowa

Predoctoral Trainees

Elisa Mordue, M.A.

Elisa earned a bachelor's degree in Communication Disorders in 1990 and a master's degree in Speech Pathology and Audiology in 1992 from Mankato State University and St. Cloud State University, respectively. She traveled between four skilled nursing facilities during her clinical supervision fellowship year (CFY). Following her CFY, she was Director of Speech Pathology at Immanuel-St. Joseph's Hospital for nearly one year. In 1994, she came to the University of Iowa to pursue her doctoral degree. She completed her comprehensive examinations in the fall of 1996. Currently, Elisa is completing her pre-dissertation and developing her thesis research in the area of adult neurogenic disorders under the direction of Dr. Donald Robin.

John A. Nelson, M.Aud.

John earned his bachelor's degree from the University of Minnesota-Duluth and master's degree from the University of South Carolina. Following two years of clinical and research training at the Veteran's Affairs Medical Center in Augusta, Georgia, he began doctoral training at the University of Iowa in the fall of 1993. A primary focus of John's training has been in the area of electrical engineering with an emphasis in digital signal processing techniques. He successfully completed his comprehensive examinations during the summer of 1995. His dissertation research investigates the effect of current digital processing algorithms on perceptions.

Helen Sharp, M.S.

Helen enrolled as a doctoral student in speech pathology in the fall of 1996. She earned her master's degree in speech from the University of Pittsburgh in 1992. Helen worked at Loma Linda University Medical Center as a staff speech pathologist and clinical faculty for three years before moving to Chicago to study clinical medical ethics. Helen completed a year-long fellowship training program in the Department of Medicine's Center for Clinical Medical Ethics at the University of Chicago in 1995. She stayed in Chicago for an additional year to participate in teaching, research and ethics consultation at the university and at a county long-term care facility. Helen is working with Dr. Jerry Moon and plans to develop her research and teaching interests in the interface between clinical ethics and speech pathology. For example, she would like to continue to develop some early work in the area of patient preferences about treatment decisions for head and neck cancer, which influence communication and swallowing functions.

Postdoctoral Trainees

Mike Edgerton, Ph.D.

Mike received his Doctor of Musical Arts from the University of Illinois in composition. He taught composition at Yonsei University and was a Visiting Scholar at Hanyang University in Seoul, South Korea. As a Postdoctoral Fellow at the University of Wisconsin-Madison, Mike is currently working with Dr. Diane Bless. He is researching the production of two simultaneously-voiced oscillators in the vocal tract, as well as the production of both voiced and unvoiced pharyngeal fricatives in order to align the voice more closely with the contemporary timbral explorations of other acoustic-based instruments.