

SIMULATION OF DISORDERED SPEECH USING A FREQUENCY-DOMAIN VOCAL TRACT MODEL*

L. Deng¹, X. Shen¹, D. Jamieson², and J. Till³

¹ Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada

² Hearing Health Care Research Unit, University of Western Ontario, London, Ontario, Canada

³ Department of Speech Pathology, Veterans Administration Hospital, Santa Barbara, CA.

ABSTRACT

In this paper, we address the issue of how the perception of disorderness in selected types of speech disorders may be correlated with the abnormal articulatory structure and with the related acoustic properties. As a first step towards this end we have developed an articulatory synthesizer based on frequency-domain simulation of vocal-tract wave propagation. The synthesizer has been implemented by three numerical methods — Runge-Kutta, ABCD matrix, and Finite difference, which provide frequency-domain solutions to the transmission-line equation characterizing a lossy vocal tract. The synthesizer is applied in preliminary experiments where the synthesizer's outputs are used to match samples from a corpus of steady-state speech sound, obtained from a dysarthric speaker, uttered in the /hV/ context.

1. INTRODUCTION

The object of this study is to improve understanding of the articulatory correlate(s) and the related acoustic properties associated with selected speech disorders. While listeners may distinguish disordered speech from normal speech quite readily, studies have as yet failed to identify the acoustical factors underlying these perceived distinctions. We hypothesize that an improved understanding of the perceived differences between samples of some types of disordered speech and normal speech may come from studies of differences in the articulatory structures and in the articulatory-to-acoustic relations.

Reported in this paper is a first step towards examining the above hypothesis by developing a specialized articulatory synthesizer capable of simulating certain aspects of disordered (as well as normal) speech. The design of the synthesizer has been based on frequency-domain simulation of vocal-tract wave propagation. The articulatory synthesizer is implemented by three numerical methods — Runge-Kutta approach, ABCD matrix approach, and finite difference approach, which provide frequency-domain solutions for transmission-line equations that characterize a lossy vocal tract. Both vocal-tract shaping (filter) and source excitation are modeled

using well-established data, and the radiation effect is taken into account. This synthesizer is then applied to model samples from a corpus of speech data from persons with a range of speech production disorders. By manipulating the vocal-tract area function and the voice-source characteristics, the synthesized speech can be judged as being perceptually similar to the disordered speech samples (“target” sounds), thereby enabling a better understanding of what aspects of the synthesizer parameter manipulations and their acoustic consequences are correlated with the perception of disorderness.

Previous work on disordered speech includes replacing the voicing sources of tracheoesophageal speech using LPC synthesis [6], analysis of vocal tract area function in Parkinsonian speech [3], acoustic analysis of dysarthric speech [4]. In the present paper, we focus on studying steady-state speech sounds in the /hV/ environment, obtained from a dysarthric speaker. The applications of our results are primarily to understanding the effects of speech disorders on speech acoustics and on the perceived quality of speech. The work also has implications for understanding the interaction between speech production disorders and speech coding methods which have been designed and evaluated using only normal speech. As such, these results are expected to assist in the design of future systems for converting disordered speech samples into speech which is perceived as being more like normal speech.

2. VOCAL-TRACT WAVE PROPAGATION

Sound waves are created by vibration and are propagated in air through the transmission system. The vocal tract is a non-uniform acoustic tube which begins at the opening between the vocal cords, or glottis, and ends at the lips. A set of linear partial-differential equations characterizing the acoustic wave propagation in a nonuniform vocal tract system can be found in [5, 8]. Let $p = p(x, t)$ represent the variation in sound pressure in the tube at position x and time t , $u = u(x, t)$ the variation in volume velocity flow, ρ the equilibrium density of air in the tube, c the corresponding velocity of sound, and $A(x, t)$ the “area function” of the tube. If $A(x, t) = A$ is time invariant, then sound waves in the vocal tract tube satisfy the following equations

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A(x)} \frac{\partial u}{\partial t}, \quad -\frac{\partial u}{\partial x} = \frac{A(x)}{\rho c^2} \frac{\partial p}{\partial t}. \quad (1)$$

* Support of this work by the Natural Science and Engineering Research Council (NSERC) of Canada and useful discussions with G. Ramsay and P. Mermelstein on the synthesizer implementation are gratefully acknowledged.

Since the vocal/nasal tracts constitute lossy tubes of non-uniform cross-sectional area, it is conventional to break up the tracts into a number of contiguous cylindrical sections so that the sound waves in each section satisfy Eqn. (1).

If the vocal tract configuration is constrained to be quasi-static (over a short period of time which we call a frame), then the steady-state solution to the vocal tract transmission equations can be obtained by transforming the equations to the frequency domain. Express the variation in sound pressure $p(x, t)$ and the variation in volume velocity flow $u(x, t)$ for a complex exponential excitation directly as

$$p(x, t) = P(x, \omega)e^{j\omega t}, \quad u(x, t) = U(x, \omega)e^{j\omega t}, \quad (2)$$

where ω is angular frequency and j is the imaginary unit. Substituting these solutions into Eqn.(1) gives the ordinary differential equations relating the complex amplitudes

$$\frac{dP(x, \omega)}{dx} = -ZU(x, \omega), \quad \frac{dU(x, \omega)}{dx} = -YP(x, \omega), \quad (3)$$

where

$$Z = \frac{j\omega\rho}{A(x)}, \quad Y = \frac{j\omega A(x)}{\rho c^2}.$$

The boundary conditions are (in an ideal case)

$$P(l, \omega) = 0, \quad U(0, \omega) = U_g(\omega) \quad (4)$$

Eliminating $U(x, \omega)$ or $P(x, \omega)$ from the two equations (3), the set of equations for sound propagation reduces to the familiar Webster Horn equation:

$$\frac{d^2 P(x, \omega)}{dx^2} = ZYP(x, \omega) \quad (5)$$

or

$$\frac{d^2 U(x, \omega)}{dx^2} = ZYU(x, \omega). \quad (6)$$

This pair of equations have been derived under the assumption of no energy loss in the tube. In reality, energy loss exists as a result of viscous friction between the air and the walls of the tube, heat conduction through the walls of the tube, and vibration of the tube walls. To include these effects, Z should be changed into $\bar{Z} = Z + R_a$, and Y into $\bar{Y} = Y + Y_w + G_a$ where R_a characterizes the effect of viscous friction at the tube wall, Y_w the effect of the vibration of the tube wall (yielding wall), and G_a the effect of heat conduction. Since the vocal tract tube terminates with the opening between the lips, the effect of speech radiation at the lips can be represented by

$$P_{mouth}(l, \omega) - Z_{rad}U_{mouth}(l, \omega) = 0 \quad (7)$$

where Z_{rad} is the frequency-domain radiation load:

$$Z_{rad} = \frac{\rho c}{A(x)} \frac{j\omega R_l L_l}{R_l + j\omega L_l}, \quad (8)$$

with R_l being the radiation resistance of the vocal-tract and L_l being the radiation inductance of the vocal tract. For the purpose of studying the frequency-domain characteristics of the vocal-tract, the relation between pressure and volume velocity at the glottis is given by [2]:

$$Y_g P(0, \omega) + U(0, \omega) = U_g(\omega), \quad (9)$$

where Y_g is the admittance of the glottis and $U_g(\omega)$ is the equivalent volume-velocity source.

3. FREQUENCY-DOMAIN SOLUTIONS

In the following, we will present three numerical integration approaches, with tradeoffs between computation efficiency and simulation accuracy, to solve Eqn.(3) with boundary conditions given by Eqns.(4) and (9) for arbitrary vocal tract area functions. Such solutions provide considerable insights into the nature of the speech production process and of the spectral properties of the speech signal. For purposes of simplifying the writing, we will use the following notations: $P_0 = P(0, \omega)$, $U_0 = U(0, \omega)$, $P = P(x, \omega)$, $U = U(x, \omega)$, $P_l = P(l, \omega)$ and $U_l = U(l, \omega)$.

3.1. Runge-Kutta approach

Solving Eqn.(3) with boundary conditions given by (4) and (9) is a two-point-boundary-value (*TPBV*) problem. Since initial value U_0 and boundary value P_l are known, we first transform the *TPBV* problem to an initial-value problem, i.e., to obtain the initial value P_0 . The boundary conditions (4) and (9) can be expressed in the standard form as

$$M \begin{bmatrix} P_0 \\ U_0 \end{bmatrix} + N \begin{bmatrix} P_l \\ U_l \end{bmatrix} = \begin{bmatrix} 0 \\ U_g \end{bmatrix}, \quad (10)$$

and Eqn.(3) as

$$\begin{bmatrix} dP/dx \\ dU/dx \end{bmatrix} = L \begin{bmatrix} P \\ U \end{bmatrix}, \quad (11)$$

where

$$M = \begin{bmatrix} 0 & 0 \\ Y_g & 1 \end{bmatrix}, \quad N = \begin{bmatrix} 1 & -Z_{rad} \\ 0 & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 0 & -\bar{Z} \\ -\bar{Y} & 0 \end{bmatrix}. \quad (12)$$

Since the solution to Eqn.(11) is given by

$$\begin{bmatrix} P_x \\ U_x \end{bmatrix} = e^{Lx} \begin{bmatrix} P_0 \\ U_0 \end{bmatrix}. \quad (13)$$

At $x = l$, we have

$$\begin{bmatrix} P_l \\ U_l \end{bmatrix} = e^{Ll} \begin{bmatrix} P_0 \\ U_0 \end{bmatrix} \quad (14)$$

Substituting Eqn.(14) into Eqn.(10), we can eliminate P_l and U_l from Eqn.(10) to obtain the initial condition

$$(M + Ne^{Ll}) \begin{bmatrix} P_0 \\ U_0 \end{bmatrix} = \begin{bmatrix} 0 \\ U_g \end{bmatrix}, \quad (15)$$

or

$$\begin{bmatrix} P_0 \\ U_0 \end{bmatrix} = (M + Ne^{Ll})^{-1} \begin{bmatrix} 0 \\ U_g \end{bmatrix}. \quad (16)$$

Solution to Eqn.(3) with the initial condition Eqn.(16) can then be easily obtained.

For notational convenience, we note that in Eqn.(14), e^{Ll} is a 2×2 matrix. Therefore, let

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = e^{Ll}. \quad (17)$$

Then by

$$Y_g P_0 + U_0 = U_g \quad (18)$$

$$\begin{bmatrix} P_l \\ U_l \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} P_0 \\ U_0 \end{bmatrix} \quad (19)$$

$$P_l = Z_{rad} U_l, \quad (20)$$

we obtain the following relations

$$H = \frac{P_l}{U_0} = \frac{Z_{rad}(AD - BC)}{A - CZ_{rad}} \quad (21)$$

$$Z_{in} = \frac{P_0}{U_0} = \frac{DZ_{rad} - B}{A - CZ_{rad}} \quad (22)$$

$$\frac{U_l}{U_0} = \frac{AD - BC}{A - CZ_{rad}}. \quad (23)$$

Note that the matrix e^{Lx} can represent any portion of the tract, of any length or variable cross-sectional area. Therefore any portion of the tract is obtained by multiplying the matrices $\prod e^{L_i x_i}$, $i = 1, \dots, N$ for the sequence of elementary homogeneous segments comprising it.

3.2. ABCD matrix approach

The characteristics of sound propagation in the vocal tract tube can be described by drawing upon the elementary electrical theory. For a vocal tract with l in length, sending-end(glottis) pressure and velocity P_0 and U_0 , the receiving-end(mouth) pressure and velocity P_l and U_l are given by [8]

$$P_l = \cosh(\gamma l) P_0 - Z_0 \sinh(\gamma l) U_0 \quad (24)$$

$$U_l = -Y_0 \sinh(\gamma l) P_0 + \cosh(\gamma l) U_0, \quad (25)$$

where $Z_0 = (\bar{Z}/\bar{Y})^{1/2}$, $Y_0 = (\bar{Y}/\bar{Z})^{1/2}$ and $\gamma = (\bar{Z}\bar{Y})^{1/2}$. The above equation can be represented by the following $ABCD$ matrix

$$\begin{bmatrix} P_l \\ U_l \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} P_0 \\ U_0 \end{bmatrix} \quad (26)$$

with obvious A, B, C and D elements. Eqn.(26) has the same form as Eqn. (18). Again, the $ABCD$ matrix can represent any portion of the tract, of any length or variable cross-sectional area. Therefore the cross-sectional area function is computed at some finite number of points N and the tube approximated as a concatenation of piecewise constant segments. The chain matrix for any portion of the tract is then obtained by multiplying the 2×2 matrices for the sequence of elementary homogeneous segments comprising it.

3.3. Finite difference approach

The differential equations

$$\frac{dP}{dx} = -\bar{Z}U, \quad \frac{dU}{dx} = -\bar{Y}P \quad (27)$$

are first transformed to a set of finite-difference equations. Given the length l of the vocal tract, the glottis located at $x = 0$, and the mouth located at $x = l$, we seek the acoustic pressure and volume velocity at N equally spaced points ($k = 0, 1, \dots, N - 1$) on the

interval $[0, l]$. The differential equations are transformed to finite-difference equations¹ according to the rule

$$\frac{df}{dx} = (f_k - f_{k-1})/\Delta x \quad (28)$$

$$g = (g_k + g_{k+1})/2 \quad (29)$$

where $\Delta x = l/(N - 1)$, $f_k = f(k\Delta x)$, $g_k = g(k\Delta x)$. We obtain

$$\frac{P_k - P_{k-1}}{\Delta x} = -\bar{Z}_k \frac{U_k + U_{k+1}}{2} \quad (30)$$

$$\frac{U_k - U_{k-1}}{\Delta x} = -\bar{Y}_k \frac{P_k + P_{k+1}}{2} \quad (31)$$

or

$$P_k + \frac{\Delta x}{2} \bar{Z}_k U_k - P_{k-1} + \frac{\Delta x}{2} \bar{Z}_{k-1} U_{k-1} = 0 \quad (32)$$

$$\frac{\Delta x}{2} \bar{Y}_k P_k + U_k + \frac{\Delta x}{2} \bar{Y}_{k-1} P_{k-1} - U_{k-1} = 0 \quad (33)$$

for $k = 1, 2, \dots, N - 1$. The complete system of the $2N$ equations determines the acoustic pressure and volume velocity distributions. Since

$$\begin{bmatrix} 1 & Z_k^* \\ Y_k^* & 1 \end{bmatrix} \begin{bmatrix} P_k \\ U_k \end{bmatrix} = \begin{bmatrix} 1 & -Z_{k-1}^* \\ -Y_{k-1}^* & 1 \end{bmatrix} \begin{bmatrix} P_{k-1} \\ U_{k-1} \end{bmatrix} \quad (34)$$

$$\begin{bmatrix} P_k \\ U_k \end{bmatrix} = \begin{bmatrix} 1 & Z_k^* \\ Y_k^* & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -Z_{k-1}^* \\ -Y_{k-1}^* & 1 \end{bmatrix} \begin{bmatrix} P_{k-1} \\ U_{k-1} \end{bmatrix}, \quad (35)$$

where

$$Z_k^* = \frac{\Delta x}{2} \bar{Z}_k, \quad Y_k^* = \frac{\Delta x}{2} \bar{Y}_k, \quad (36)$$

we have

$$P_{N-1} = Z_{rad} U_{N-1} \quad (37)$$

$$\begin{bmatrix} P_{N-1} \\ U_{N-1} \end{bmatrix} = A_{N-1}^{-1} B_{N-1} A_{N-2}^{-1} B_{N-2} \dots A_1^{-1} B_1 \begin{bmatrix} P_0 \\ U_0 \end{bmatrix} \quad (38)$$

$$Y_g P_0 + U_0 = U_g \quad (39)$$

where

$$A_k = \begin{bmatrix} 1 & Z_k^* \\ Y_k^* & 1 \end{bmatrix}, \quad B_k = \begin{bmatrix} 1 & -Z_{k-1}^* \\ -Y_{k-1}^* & 1 \end{bmatrix} \quad (40)$$

Let

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = A_{N-1}^{-1} B_{N-1} A_{N-2}^{-1} B_{N-2} \dots A_1^{-1} B_1, \quad (41)$$

then Eqn.(38) can be seen to have the same form as Eqn.(18).

4. SIMULATION RESULTS

The vocal system is characterized by a set of resonances (formants) that depend primarily upon the vocal tract area function. When the area function $A(x)$, the wall impedance, and the loss parameters of the vocal tract are specified, we can compute the speech signal for a variety of sounds with a source of either a quasi-periodic pulse

¹This approach is identical to that used to solve the basilar membrane vibration model published in [1].

waveform or a random noise waveform. In our current early stage of the study, we used the synthetic pulse waveform of the form [7]

$$\begin{aligned}
 & - \\
 & \text{otherwise} \quad (42)
 \end{aligned}$$

The synthesized speech signal in the frequency domain is obtained by

$$\text{—————} \quad (43)$$

where \hat{X} is the Laplace transform of x . The inverse Laplace transform of \hat{X} gives the time-domain synthesized speech signal. In our simulation experiments, according to detailed analysis of the “target” samples of the disordered speech utterances, the spacings and shapes of the glottal pulses are perturbed and some irregularities or noise are added. This potentially allows us to approximate the synthesizer’s output to the jittering or shimmering types of the speech disorder. Fig. 1 shows one simple example of the target utterance with speech disorder. The utterance is [hi:d] *heed* from a speaker². A comparison between the spectrogram of the utterance shown in Fig.1 and that of some normal speech utterances suggests that the intensity variation and lack of smoothness in high-frequency resonances (F2 and F3) appear to be associated with consistent disorderness perception clear to the authors’ informal listening. We have manipulated both the source and the area function of the synthesizer. The preliminary results showed that, by carefully manipulating the vocal-tract area function and of excitation waveform characteristics, the synthesized speech can be made to progressively approach the perceptual quality of the target disordered speech utterances. But ultimate judgment on and quantification of the quality will need future work on extensive human perceptual testing.

5. SUMMARY AND CONCLUSIONS

This paper reports our preliminary efforts devoted to understanding the possible articulatory and acoustic correlates of the perception of the disordered speech quality. We have described in detail the tools of specialized articulatory synthesis already developed that would enable us to tackle the problem. Among the three methods (Runge-Kutta, ABCD matrix, and Finite-difference) for implementing the frequency-domain synthesizer by finding numerical solutions to the transmission-line equation characterizing the vocal tract, the finite-difference solution appears to best balance the tradeoffs between computation efficiency and simulation accuracy.

While at the early stage of the study, we have run a version of the synthesizer aimed at matching samples from a corpus of steady-state speech sounds recored from a dysarthric speaker. By performing detailed acoustic analysis on the disordered speech, we have been able to manipulate the vocal-tract area function and the excitation waveform characteristics so as to gradually improve the similarity between the synthetic speech and the target speech, both in the quality of perception and in the spectrographic display. Further

²The speaker is a 78-year old male. His dysarthria is characterized by slurring, omission of phones and by a rapid but unstable rate of speaking.

progresses of this research will lie in refinement of the articulatory synthesis tool and in detailed examination of the synthesizer parameters responsible for producing sounds of the disordered quality. Our experiences also suggest that there is a strong need to develop semi-automatic approaches to aid selection of the synthesizer parameters, from which further fine modifications of the parameters will quickly lead to synthesis of the sounds close to the target one containing a quality of disorderness.

6. REFERENCES

1. L. Deng and I. Kheirallah, “Numerical Property and Efficient Solution of a Transmission-Line Model for Basilar Membrane Wave Motions,” *Signal Processing*, Vol. 33, 269-285, 1993.
2. J.L. Flanagan, *Speech Analysis Synthesis and Perception*, Springer-Verlag, New York, 1972.
3. I. Gath and E. Yair, “Analysis of Vocal Tract Parameters in Parkinsonian Speech,” *J. Acoust. Soc. Am.*, Vol. 84(5), 1628-1634, 1988.
4. C.L. Ludlow and C.J. Bassich, “The Results of Acoustic and Perceptual Assessment of Two Types of Dysarthria,” in W.R. Berry (ed), *Clinical Dysarthria*, San Diego: College-Hill Press, 121-153, 1982.
5. M.R. Portnoff, *A Quasi-One-Dimensional Digital Simulation for the Time-Varying Vocal-Tract*, S.B./S.M. thesis, MIT, Cambridge, Mass., 1973.
6. Y. Qi, “Replacing Tracheoesophageal Voicing Source Using LPC Synthesis,” *J. Acoust. Soc. Am.*, Vol. 88, 1228-1235, 1990.
7. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, N.J., 1978.
8. J. Schroeter and M. M. Sondhi, “Techniques for Estimating Vocal-Tract Shapes from the Speech Signal,” *IEEE Trans. Speech Audio Proc.*, Vol. 2(1), 133-150, 1994.