Acoustic Modeling of American English Lateral Approximants

Zhaoyan Zhang[†], *Carol Espy-Wilson*[†], *Mark Tiede*[‡]

[†]University of Maryland, College Park, USA

[‡]Haskins Laboratories, College Park, USA

zhaoyan@glue.umd.edu, espy@glue.umd.edu, tiede@haskins.yale.edu

Abstract

A vocal tract model for an American English /l/ production with lateral channels and a supralingual side branch has been developed. Acoustic modeling of an /l/ production using MRI-derived vocal tract dimensions shows that both the lateral channels and the supralingual side branch contribute to the production of zeros in the F3 to F5 frequency range, thereby resulting in pole-zero clusters around 2-5 kHz in the spectrum of the /l/ sound.

1. Introduction

Research on vocal tract acoustics requires accurate models of the complex geometry formed by the articulators. In this paper, a vocal tract model with lateral channels is described. Lateral channels appear mostly in the production of /l/ sounds. They are airflow paths around the tongue that are produced by the laterally inward movement of the tongue toward the midsagittal plane [1]. These two airflow paths join again anterior to the lingual-alveolar contact along the midsagittal line. In most cases, the main flow above the tongue terminates at the lingual-alveolar contact, giving rise to a supralingual side branch (space enclosed by the roof of the oral cavity and the tongue posterior to the lingual closure) to the lateral channels.

The effects of these geometric features on the acoustics of the vocal tract are not clear. Particularly, the /l/ sound is generally characterized by a pole-zero cluster around 2-3 kHz in its spectrum. The source of the zero(s) still remains unclear. A zero may be generated by the supralingual side branch [1-3]. Additional zeros may also be produced by two lateral channels of different length [4]. In this paper, a vocal tract model with both a supralingual side branch and lateral channels is presented. Possible sources of the pole-zero clusters were investigated using MRI-derived vocal tract dimensions for an American English syllabic /l/ production.

2. Model Description

Models of the vocal tract have already been discussed in many studies [5-7]. However, there are few studies on lateral channel modeling for /l/ sound production. In this study, a frequency-domain model for vocal tract acoustic response function calculation (VTAR) was developed based on the transmission-line model of the vocal tract. The vocal tract is simulated as a concatenation of cylindrical sections with lengths far less than the acoustic wavelength. Each cylindrical section is represented by an analog circuit as shown in Fig. 1. The transmission-line model has been discussed extensively in many studies (cf. [8]). The exact expressions for each circuit element used in this study are given as follows,

$$R_a = \frac{lS}{2A^2}\sqrt{\frac{\omega\rho\mu}{2}}, \qquad L_a = \frac{\rho l}{2A}\omega$$
$$C_a = \frac{lA}{\rho c^2}\omega, \quad G_a = Sl\frac{\eta - 1}{\rho c^2}\sqrt{\frac{\lambda\omega}{2C_p\rho}}$$
$$L_w = \frac{m}{lS}\omega, \quad R_w = \frac{b}{lS}, \quad C_w = \frac{lS}{k}\omega. \tag{1}$$

where *m*, *b*, and *k* are the mass, mechanical resistance, and the stiffness of the wall per unit area of the tube, respectively.

The vocal tract is decomposed into various modules (single tubes, branching, and lateral channels). For each module the input and output pressures and volume velocities are related by a transfer matrix, as:

$$\begin{bmatrix} p_{in} \\ U_{in} \end{bmatrix} = K \begin{bmatrix} p_{out} \\ U_{out} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} p_{out} \\ U_{out} \end{bmatrix} \tag{2}$$

where *A*, *B*, *C*, and *D* are determined by the properties of the air and vocal tract wall, and can be calculated using the transmission-line model as discussed above.

At the point of branching (Fig. 2a), such as the coupling of the nasal tract, sublingual, or supralingual space to the vocal tract, an extra branch coupling matrix is used to relate the state variables across the branching point,

$$\begin{bmatrix} p_1 \\ U_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/Z_3 & 1 \end{bmatrix} \begin{bmatrix} p_2 \\ U_2 \end{bmatrix},$$
$$\text{or} \begin{bmatrix} p_1 \\ U_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/Z_2 & 1 \end{bmatrix} \begin{bmatrix} p_3 \\ U_3 \end{bmatrix} \tag{3}$$

where Z_2 and Z_3 are the input impedances of the side branches 2 and 3, respectively.

For the two lateral channels in the vocal tract, the two channels have the same input and output pressure (Fig. 2b). We assume for each lateral channel that the input and output state variables are related by

$$\begin{bmatrix} p_{in,i} \\ U_{in,i} \end{bmatrix} = K_i \begin{bmatrix} p_{out,i} \\ U_{out,i} \end{bmatrix} \tag{4}$$

Applying boundary conditions and flow continuity, simple algebraic manipulation leads to a relationship between the input and output of the lateral channels,

$$\begin{bmatrix} p_{in} \\ U_{in} \end{bmatrix} = \begin{bmatrix} \dfrac{A_1B_2 + A_2B_1}{B_1 + B_2} & \dfrac{B_1B_2}{B_1 + B_2} \\ C_1 + C_2 - \dfrac{(D_1 - D_2)(A_1 - A_2)}{B_1 + B_2} & \dfrac{D_1B_2 + D_2B_1}{B_1 + B_2} \end{bmatrix} \begin{bmatrix} p_{out} \\ U_{out} \end{bmatrix} \tag{5}$$

The entire vocal tract can be modeled by combining the appropriate modules and multiplying the individual transfer matrices in an order corresponding to their geometric location. This results in the following single equation relating the pressures and volume velocities at the glottis and the lips:

$$\begin{bmatrix} p_g \\ U_g \end{bmatrix} = K \begin{bmatrix} p_l \\ U_l \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} p_l \\ U_l \end{bmatrix} \quad (6)$$

The transfer function can be calculated as

$$20\log_{10}\left(U_l / U_g\right) = 20\log_{10}(1/(CZ_l + D)) \quad (7)$$

where Z_l is the radiation impedance at the lips.

3. Model validation

The model was first validated against Maeda's model [5] for vowels, and Jackson's model [7] for the side-branching module. Good agreements were obtained for both cases.

The vocal tract geometry in /l/ production is very complex and multiple features play a role in affecting the vocal tract acoustic response. This makes it impossible to directly validate the model using the vocal tract geometry data for /l/ production. Instead, nasalized vowels were chosen for this purpose. Nasalized vowels are similar in production to the /l/ sound since they are produced with two airflow paths (the oral tract and nasal tract) and the posterior end of the two paths have the same pressure, assuming plane wave propagation. The difference between the production of nasalized vowels and the /l/ sound is that, for the nasalized vowels, the two paths generally have different radiation impedances and therefore different termination pressures. The two paths in lateral channels have the same termination pressure. Therefore, if the radiation impedances are neglected, the two paths in the nasalized vowels would have the same termination pressures, as in the case of lateral channels. The nasalized vowels can then be modeled using the lateral channel model as well.

A nasalized vowel /iy/ was selected for the validation. Maeda's simulation code was used as a validation reference. Fig. 3 shows the comparison between the vocal tract acoustic response functions predicted using Maeda's code and the VTAR program. Zero radiation was imposed at both the nose and mouth. The agreement is very good, except some mismatch in the amplitude of the first zero (around 600 Hz) in the spectrum.

4. Results

Magnetic Resonance Images were collected for a single male native speaker of American English producing a sustained /el/ sound. A semi-polar grid superimposed on a midsagittal projection through the MRI volume dataset was first aligned using anatomical landmarks [9]. The centroids of the intersection of this grid with the anterior and posterior tract walls were then used to define the approximate airflow path through the tract. Area functions were then measured from the cross-section projections orthogonal to this path. The measured area function is shown in Fig. 4. The lingual-alveolar contact was observed in the region 17-18 cm from the glottis along the midsagittal line.

Figure 5 shows MRI-derived cross-sectional coronal slices in the oral cavity. For this particular case, the length of the lateral channels is not clear. From the figure, the lateral channels first appear at 13.9 cm from the glottis (the rightmost picture of second row in Fig. 5), and become apparent about 15.2 cm from the glottis (second picture of third row in Fig. 5). The two lateral channels are not separated from each other and the main flow channel until the lingual-alveolar contact is made, and rejoin immediately after the lingual closure. Extra efforts have been made to split these three channels and measure their individual areas for the purpose of acoustic modeling. The boundaries between the channels were chosen as the location of minimum areas of inter-connection. These areas are shown in Fig. 4 as well.

Acoustic data were recorded from the same speaker separately from the MRI session. The measured power spectrum is shown in Fig. 6. The spectrum shows the first four formants at 390, 780, 3670, and 4610 Hz, respectively. In the frequency range between F2 and F3, the spectrum is relatively flat. Separate acoustic recordings of vowels spoken by the same speaker show that, for this speaker, the range for F3 is from 2.2 to 2.8 kHz, 3.2-3.9 kHz for F4, and 4.3-4.5 kHz for F5. This indicates that the speaker's F3 may be canceled out by one or more zeros in the range 1-3.5 kHz in the spectrum of Fig. 6.

The measured area function was used as input to the VTAR simulation program. The lateral channels were modeled as starting from the midsagittal lingual closure and of a length of about 1 cm. The vocal tract otherwise was modeled as two single tubes, one anterior and one posterior to the lateral channels, with varying cross-sectional areas derived from MRI data. The calculated vocal tract acoustic response function is shown in Fig. 6. This model gives fairly accurate predictions for the formant structure between 1 kHz and above 3 kHz. However, it gives a prominent F3 peak in the frequency range in between where the measured power spectrum is relatively flat.

Prahler [4] has shown that uniform lateral channels of about 8 cm long (or a combined length of 16 cm) are needed to produce zeros in the region around 2 kHz. The lateral channels used in this study are obviously not long enough. It has been speculated that the zeros (around 2 kHz) may be due to the quarter-wave length resonance of the supralingual cavity with an effective length of about 4.4 cm [1-2]. To cancel out the formant in the range of 1-3 kHz, the length of the lateral channels and the supralingual cavity need to be increased. As a first attempt, the starting point of the lateral channels was moved back to the location where they first become apparent, about 15.2 cm from the glottis. The section of the main flow channel from this location to the point of lingual closure was modeled as a (supralingual) side branch to the lateral channels. The length of the lateral channels and supralingual side branch is approximately 1.8 cm and 2.8 cm, respectively. The calculated vocal tract acoustic response function is shown in Fig. 6 for comparison. A zero was observed around 5.5 kHz, still too high to cancel out the F3 formant. In the next step, the starting point of the lateral channel was moved further back to the location where they first appear, about 13.9 cm from the glottis. This results in both a longer lateral channel (about 4.1 cm) and supralingual side branch (about 3.1 cm). The resulting vocal tract acoustic response function is shown in Fig. 7. Two zeros appear at 2390 and 4160 Hz, giving rise to pole-zero clusters in that frequency range. The general agreement between the spectral shapes of the prediction and the measurement is much improved. If the effects of glottal spectrum and radiation characteristics are taken into account, the agreements in both the spectral shape and amplitude are further improved (Fig. 8).

5. Discussion

The first zero is caused by the lateral channels. This can be verified by replacing the lateral channels module in the model with a single-tube module having an area of the sum of the two lateral channels. The acoustic response function predicted in this way is also shown in Fig. 7 for comparison. The first zero disappears in this case, indicating the first zero is due to the lateral channels and the second one is due to the supralingual side branch. For lateral channels of uniform areas, the zero it produces is always canceled out by a pole at the same location if the two channels have the same length [4]. However, this study has shown that for lateral channels of varying areas, which is the case in real speech production, a zero-pole pair can be produced even for lateral channels of the same length, as shown in Fig. 7.

The predicted acoustic response function in Fig. 7 shows multiple peaks in the 2-4 kHz region, which is due to the incomplete cancellation of the speaker's F3 formant by the first zero. The location of the first zero is very sensitive to the areas of the lateral channels. Due to the low resolution of the MRI data, it is difficult to precisely locate the starting point of the lateral channels and measure the areas of the individual channels. Since the speaker's teeth have not been superimposed on the MRI-slices, the cross-sectional areas of the lateral channels may be overestimated. Smaller areas of the lateral channels will result in more loss, which could reduce the prominence of the peaks in the cluster region. Also, the asymmetric lateral channels may have different length. Future work will aim to collect MRI data of higher quality and perform detailed area measurement to further explore these issues.

6. Acknowledgement

This work was supported by NIH grant 1 R01 DC05250-01.

7. References

[1]. Stevens, K. N., 1998, *Acoustic phonetics*, The MIT Press, Cambridge, Massachusetts.

[2]. Fant, G., 1970, *Acoustic theory of speech production*, The Hague, Netherlands: Mouton.

[3]. Bangayan, P., Alwan, A., Narayanan, S., 1999, "From MRI and acoustic data to articulatory synthesis: a case study of the lateral approximants in American English," *Proceedings of ICSLP*, Philadelphia, PA.

[4]. Prahler, A., 1998, *Analysis and synthesis of the American English Lateral Consonant*, MIT Thesis, Cambridge, Massachusetts.

[5]. Maeda, S., 1982, "A digital simulation method of the vocal-tract system," *Speech Commun.* 1, 199-229.

[6]. Sondhi, M. M., Schroeter, J., 1987, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoustics, Speech, And Signal Processing*, 35, 955-967.

[7]. Jackson, M. T., Espy-Wilson, C., Boyce, S., 2001, "Verifying a vocal tract model with a closed side-branch," *J. Acoust. Soc. Am.*, 109, 2983-2987.

[8]. Flanagan, J.L., 1972, *Speech Analysis, Synthesis, and Perception*, Academic Press, New York.

[9]. Tiede, M. Yehia, H., and Vatikiotis-Bateson, E., 1996, "A shape-based approach to vocal tract area function estimation," *Proceedings ETRW-SPM96* (Autrans), 41-44.

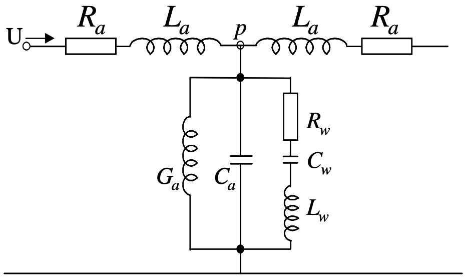

Figure 1: Transmission line representation of the vocal tract.

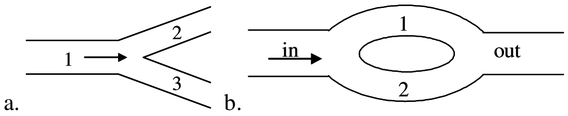

Figure 2: Models for a) tube branching and b) lateral channels.

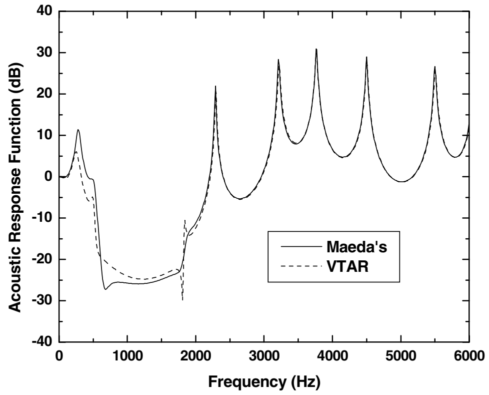

Figure 3: Comparison between the vocal tract acoustic response function predicted from Maeda's code and VTAR, for nasalized vowel /iy/, with zero radiation impedance.

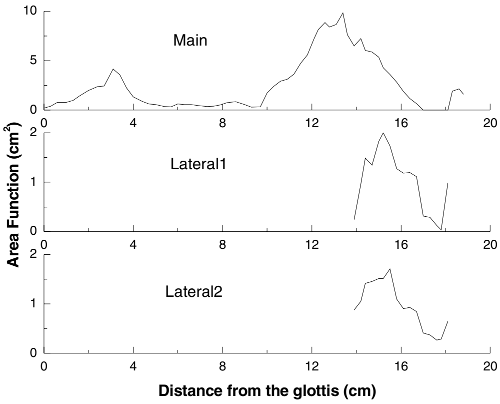

Figure 4: MRI-derived cross-sectional area functions along the midsagittal line (main flow channel) and of the two lateral channels.

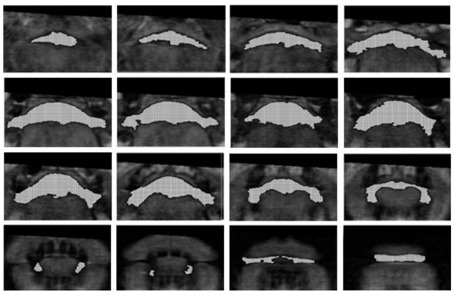

Figure 5: Cross-sectional images of the oral cavity showing the thresholded airway, from 10.3 cm (first picture in first row) to 18.6 cm (fourth picture in fourth row) from the glottis with approximate 6 mm spacing.

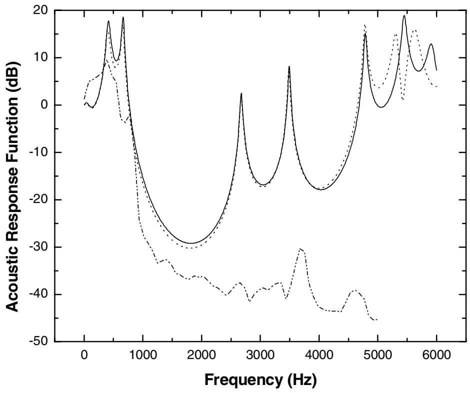

Figure 6: Predicted vocal tract acoustic response functions using MRI-derived vocal tract geometry data. Lateral channels begin from the midsagittal lingual closure (solid line), and 15.2 cm from the glottis (dotted line). Dash-dotted line is the power spectrum measured from the acoustical data.

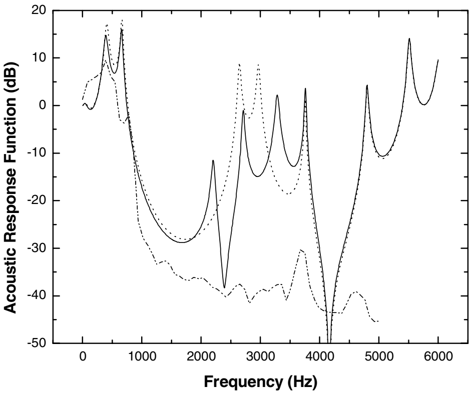

Figure 7: Predicted vocal tract acoustic response functions, with both lateral channels and supralingual side branch (solid line), and with only supralingual side branch (dotted line). Lateral channels begin from 13.9 cm from the glottis. Dash-dotted line is the power spectrum measured from the acoustical data.

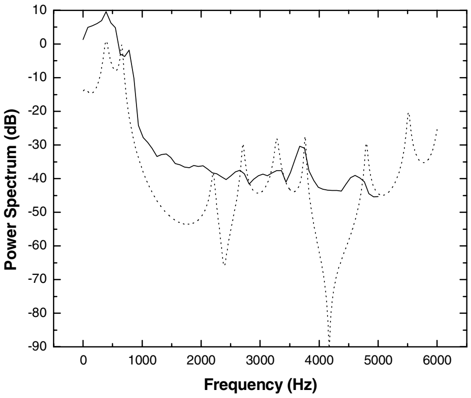

Figure 8: Measured power spectrum (solid line) and the prediction from VTAR (dashed line) with corrections for the glottal spectrum and lip radiation.