FROM MRI AND ACOUSTIC DATA TO ARTICULATORY SYNTHESIS: A CASE STUDY OF THE LATERAL APPROXIMANTS IN AMERICAN ENGLISH

*Philbert Bangayan, Abeer Alwan, and Shrikanth Narayanan**

Department of Electrical Engineering, UCLA
Los Angeles, CA 90095
* AT&T Research Labs., Murray Hill, NJ

ABSTRACT

Articulatory patterns of the lateral approximant /l/, both dark and light allophones, in American English were analyzed through magnetic resonance imaging (MRI) and electropalatography (EPG). Vocal tract lengths, area functions, and cavity volumes, were measured from the MR images, while EPG data were used for studying inter- and intra-speaker variabilities in lingua-palatal contact patterns. Acoustic modeling was based on the MRI-derived vocal-tract area functions and the acoustic spectra of these sounds. Acoustic modeling utilized the analog circuit simulator HSPICE.

1. INTRODUCTION

Laterals are typically produced with a lingual contact along the midsagittal line such that air flows along one or both sides of the tongue. The contact is made with the anterior tongue tip/blade in the anterior region of the roof of the oral cavity. /l/ is voiced and has been broadly classified into two canonical allophones, namely the light and dark varieties, which will be referred to by the symbols [l] and [ɫ], respectively, in this paper. The light allophone [l] occurs in prevocalic contexts and the dark allophone [ɫ] occurs in post-vocalic and syllabic cases Acoustically, [ɫ] is characterized by a relatively lower F_2 and higher F_1 when compared to the F_2 and F_1 values of [l] [Espy-Wilson, 1992].

Due to the lack of articulatory data, the articulatory-to-acoustic mapping of these sounds is not well understood. Fant [1960] and Stevens [in preparation] propose a model of lateral production in which a side branch is created in the oral region. The main effect of the side branch is to introduce a pole-zero pair in the vocal-tract transfer function. In this paper, we investigate the articulatory-acoustic mapping for /l/ based on MRI and acoustic data, and a transmission-line model of the vocal tract.

2. ARTICULATORY AND ACOUSTIC DATA

Articulatory patterns of the laterals in American English were analyzed using magnetic resonance imaging (MRI) and electropalatography (EPG). MR images of the vocal tract during sustained production of dark [ɫ] and light [l] were used for length, volume, and area measurements and for the analysis of the 3D vocal tract and tongue shapes. EPG contact profiles were used for a detailed analysis of inter- and intra-speaker variabilities.

2.1. Methods

A detailed description of the MRI acquisition and analysis procedures is provided in Narayanan *et al.* [1995a]. Magnetic Resonance (MR) images were collected using a GE 1.5 Tesla SIGNA machine with a fast SPGR protocol in the coronal, axial, and sagittal planes. The image slice thickness was 3 mm with no inter-scan spacing. Image resolution was 0.0081 cm^2 per pixel for an FOV = 24 *cm*.

Four phonetically-trained, native American English speakers [2 males (MI, SC) and 2 females (AK, PK)] served as subjects. The subjects, in supine position, sustained each consonant for about 13-16 s enabling four to five image slices to be recorded in a particular plane (about 3.2 sec/image). Subjects repeated each sound six to nine times, with a pause of three to ten seconds between repetitions, to enable the entire vocal tract to be scanned. The data set comprised 28 to 35 images/sound/subject in the sagittal plane, and 40 to 45 images/sound/subject in the axial and coronal planes.

Stability of the articulators could not be monitored during scanning. Instead, analysis results of EPG data, which were collected on a different day, suggest that our phonetically-trained subjects maintained stable gestures while sustaining these sounds. EPG data were recorded using *Kay Elemetrics Palatometer*. Each subject has a custom-fitted acrylic palate with 96 sensing electrodes. The sweep rate of this system is 1.7 ms and the sampling period is 10 ms. The subjects assumed a supine position, similar to that assumed inside the MRI machine, while phonating the sustained utterances.

Acoustic recordings of the sustained sounds were made in a sound-proof chamber using an omni-directional microphone (Beyerdynamic M101). The data were digitized on to a SUN SPARC station at a sampling rate of 44 kHz (16 bits/sample) using Ariel ProPort (model 656).

2.2. Results

MR images for both the light allophone [l] and the dark allophone [ɫ] indicate that the *midsagittal* tongue contours

can be different across subjects. Common characteristics, however, were revealed in cross-sectional and 3D tongue shapes, area functions, and linguopalatal contact profiles. Moreover, for each subject, the tongue shapes for the dark and light allophones showed many similar characteristics, particularly in the oral region but certain systematic differences were also found such as the degree of tongue-root retraction in the pharyngeal region.

The laterals were characterized by a complete linguo-alveolar contact or, just a constriction as observed in some cases of the [ɬ] of one subject. The contact location was about 1.5 cm away from the lip opening and the contact length, 0.3-1.0 cm in the alveolar region with relatively small openings around both sides of the contact. The *'lateral channels'* along the sides of the tongue began appearing from where the alveolar contact/constriction was seen and continued posteriorly until lingua-velar contact was established. The right and left channels appear to be, in general, unequal and their areas start increasing behind the alveolar contact (due to inward lateral compression of the tongue body) and start decreasing again as the region of lingua-velar contact is approached.

Area functions were similar in their patterns across the four subjects, particularly up to about 4 cm from the lips in the front region. Increasing area values are observed in the palatal region posterior to the alveolar contact. Post-contact lateral channels (between the tongue body and the teeth) contribute to increased areas in the vicinity of the palatal region; the decrease in the areas in the vicinity of the velar region is attributed to the disappearance of the lateral channel due to lingual bracing with the roof of the oropharynx. Presence of grooving along the midsagittal line immediately behind the alveolar contact contributes to an abrupt increase in the area functions. Sample area functions measured along the midsagittal line for [l] are shown in Figure 1. The cross-sectional areas of the side branches in the vicinity of the maximum alveolar contact ranged between .1-.5 cm^2. In addition, sublingual cavities, of length 1-1.5 cm and cross-sectional areas between .1-.2 cm^2 were observed for both subjects.

It appears that the primary tongue-shaping mechanisms for laterals are responsible for the alveolar contact, inward-lateral compression, and convex shaping of the dorsum and the posterior tongue body. These features seem to be invariant across subjects. Flattening or medial grooving of the tongue body immediately behind the alveolar occlusion, appears to be a *secondary* feature and is likely to be affected by anatomical differences. These secondary features are influenced by the extent and force of front-region linguopalatal contact and the muscular activity of the dorsum and posterior tongue body [Narayanan et al, 1995b]. Analysis results of the EPG data were consistent with those of the MRI study.

Acoustic analysis consisted of spectrographic analysis, and short-time DFT and LPC spectra (using a 25 ms Hamming window). LPC spectra were used to estimate the locations of the first four formant frequencies, while DFT spectra were useful in estimating the pole-zero pairs, if present, in the natural spectra. DFT spectra of light [l] for 2 speakers are shown in Figure 2. In the following section, we describe our modeling effort in synthesizing the laterals using the MRI-derived area functions and the acoustic spectra of these sounds.

3. MODELING METHODOLOGY

The vocal tract is modeled as a concatenation of uniform cylindrical-tube sections. The length of each section is 3 mm (similar to the MR image resolution). Depending on the subject's vocal-tract length, the total number of sections is between 55-60. Each section is then modeled as a lumped circuit. Figure 3 shows a lumped circuit approximation for one section of the vocal tract; the values of the components (R, L, C, and G) can be derived from the the vocal tract's area function [Flanagan, 1972]. Using the analog circuit simulator HSPICE, a transmission-line model of the vocal tract was constructed. The transfer function of the volume velocity (or current, using the impedance analogy) at the lips to that at the glottis was calculated for each configuration. As was shown earlier [Rael et al., 1995], there are several advantages of using HSPICE over other articulatory synthesizers: 1) side branches (needed for modeling nasals and /l/, for example) can be easily simulated by additional transmission lines in parallel, 2) drive-dependent sources, at any location, could be added, and 3) the number of sections can be varied without changing the sampling rate as is the case with most discrete-time synthesizers. The lumped circuit approximation is valid as long as the cross dimensions are small when compared to the wavelength of the sound. For the configurations studied here, the approximation is valid up to about 4-5 kHz. Small-signal analysis is used to determine the formant frequencies from the frequency response of the circuit.

A computer interface, using MATLAB, was developed such that the input to HSPICE can be specified in terms of the area function of the vocal tract and the type and location of dependent or independent sources; voltage or current sources can be specified. The transmission line(s) were terminated by a radiation impedance model proposed by Stevens et al. [1953].

To calibrate the performance of the synthesizer, vowel area functions from [Fant, 1960] were used, and the formant frequencies of the synthesized waveforms, calculated from HSPICE, were similar to those reported by Fant.

In this paper, we focus on modeling light [l] for 2 speakers (AK and MI.) When modeling [l], two topologies were studied: 1) a 'lumped area function' topology in which the cross-sectional areas of the two side branches were summed with the area of the main cavity to produce an 'effective area function'. This topology can be represented by a single transmission line, and 2) a 'branched' topology, in which one side branch was present. The transmission-line model corresponding to the second topology is shown in Figure 4. The branching point (*point a*) was about 10 cm away from the glottis and the side branch had the same cross-sectional

area function as that measured along the midsagittal line, with the exception of the region corresponding to the alveolar occlusion. In that region, the measured areas of the side branches were used. The main cavity is effectively terminated at the point of occlusion, resulting in the side branch (*ab*) being 2-3 cm longer than *ac*.

4. RESULTS AND DISCUSSION

Synthesized spectra using the lumped and branched topology for MI and AK are shown in Figure 5, also shown in that figure are the locations of the first four formant frequencies of the natural utterances. The branched topology resulted in a better estimate of the formant frequencies of the natural tokens, in a pole-zero pair in the vicinity of 2000 Hz, and a change in the average spacing of the formant frequencies. The pole-zero pair was evident in the natural spectra of these sounds. These results are in agreement with Stevens' side-branch model for lateral production. It should be noted that simulations with 2 side branches, instead of one, did not result in a better match to the acoustic spectra. Similarly, the acoustic effect of adding a sublingual cavity, for the configurations studied here, was minimal.

The transmission-line model, with circuit components calculated from MRI-derived area functions, resulted in a reasonable prediction of the formant frequencies for [l]. The relative formant *amplitudes*, however, do not exactly match those of the natural spectra. This mismatch may be due to the difficulty in incorporating frequency-dependent losses into HSPICE. Future work will examine the effects of such losses, contrast acoustic models for dark and light laterals, and examine the effect of inter-speaker articulatory variabilities, such as shapes and sizes of the different cavities and lateral channels, on the acoustic modeling of these sounds.

This work was supported in part by NSF.

5. REFERENCES

1. Espy-Wilson, C. Y. (1992). "Acoustic measures for linguistic features distinguishing the semivowels in American English," *JASA*, 92 (2), 736–757.
2. G. Fant, *Acoustic Theory of Speech Production* (1960). Mouton: The Hague.
3. Flanagan, J. L. *Speech Analysis, Synthesis, and Perception*, (1972). Academic Press, New York.
4. HSPICE 95.2 release, Meta-Software, Campbell, CA.
5. S. Narayanan, A. Alwan, and K. Haker (1995a), "An articulatory study of fricative consonants using magnetic resonance imaging," *JASA*, Vol. 98, pp. 1325-1347, Sept. 1995.
6. S. Narayanan, A. Alwan, and K. Haker (1995b), "An Articulatory Study of Liquid Consonants in American English," ICPhS Proc., Stockholm, Sweden. August 1995, Vol. 3, 576-579.
7. J. Rael, J. Chang, and A. Alwan (1995). "A Computationally-Efficient Articulatory Synthesizer," *JASA*, May 1995, Vol. 97, (5), 3245.
8. Stevens,K. N. *Acoustic Phonetics*, in preparation.
9. Stevens, K. N., Kasowski, S., and Fant, G. (1953), "An Electrical Analog of the Vocal Tract," *JASA*, Vol. 25, Number 4, pp. 734-742.

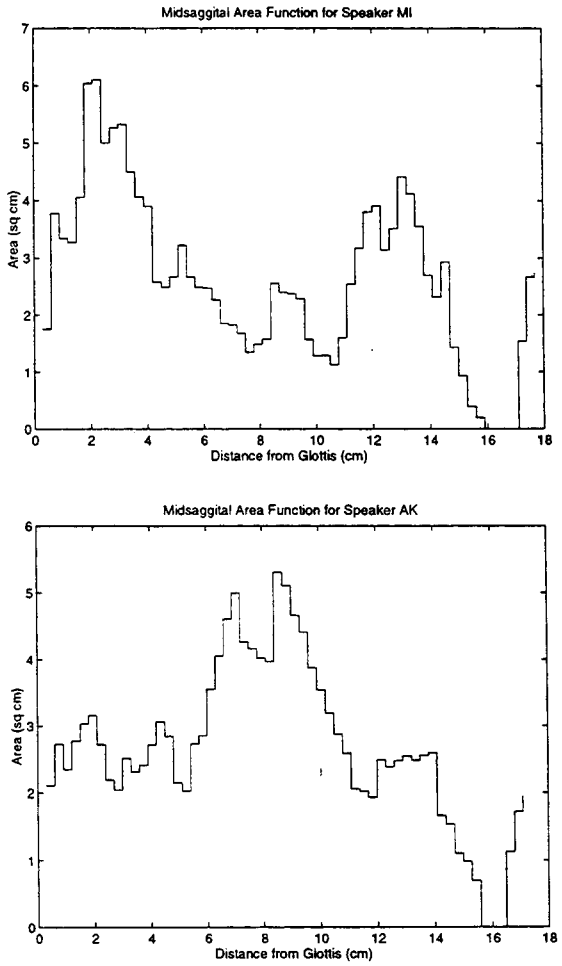

Figure 1: **Area functions** for two speakers (MI and AK) measured along the midsagittal plane. At the point of maximum contact (15.9 cm - 17.2 cm for MI, 15.6 cm - 16.5 cm for AK), cross sectional areas for the side branches ranged between .1 - .5 cm^2.

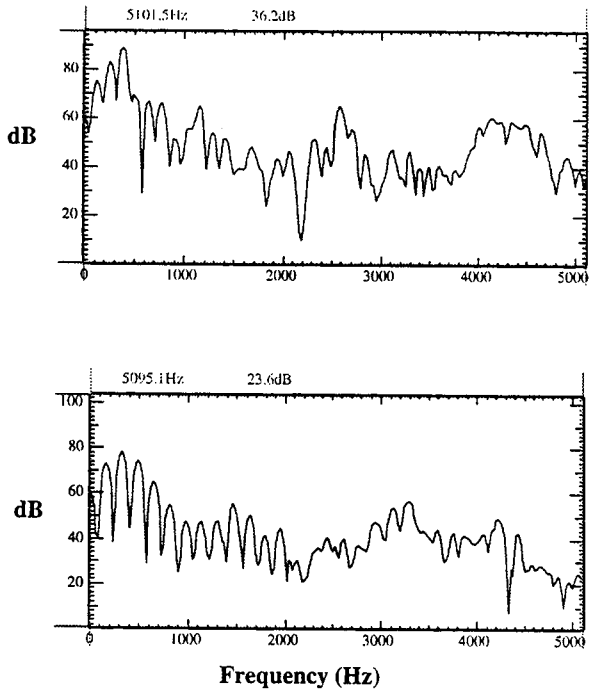

Figure 2: **DFT spectra of natural [I] tokens:** subject MI (top panel) and subject AK (bottom panel). A pole/zero pair appears near 2 kHz for both samples. Analysis was done with a 25 ms Hamming window and a preemphasis coefficient of 0.9.

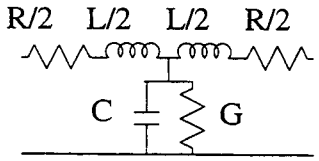

Figure 3: **T-section.** The numerical values for the elements are:

$$L = \frac{\rho l}{A}, C = \frac{Al}{\rho c^2}, R = \frac{Sl}{A^2}\sqrt{\frac{\omega\rho\mu}{2}}, \text{and } G = Sl\frac{\eta - 1}{\rho c^2}\sqrt{\frac{\lambda\omega}{2c_p\rho}}.$$

l, A, and S are the length, cross-sectional area, and circumference of the section, respectively. and w is frequency. Frequency was fixed at 100 Hz in all sumulations. The other parameters are constants related to air density, viscosity, heat conduction, sound velocity, specific heat of air at constant pressure, and the adiabatic constant (after Flanagan, 1972).

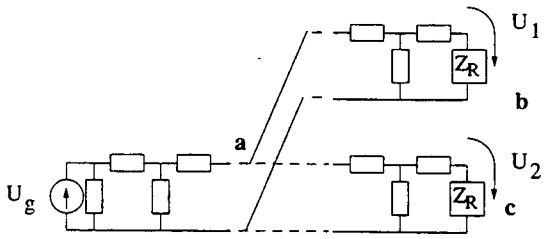

Figure 4: **Branched topology:** the vocal tract modeled as a transmission line with a side branch. U_g is the volume velocity at the glottis and Z_R is the radiation impedance at the lips. The length *ab* is about 3 cm longer than *ac*.

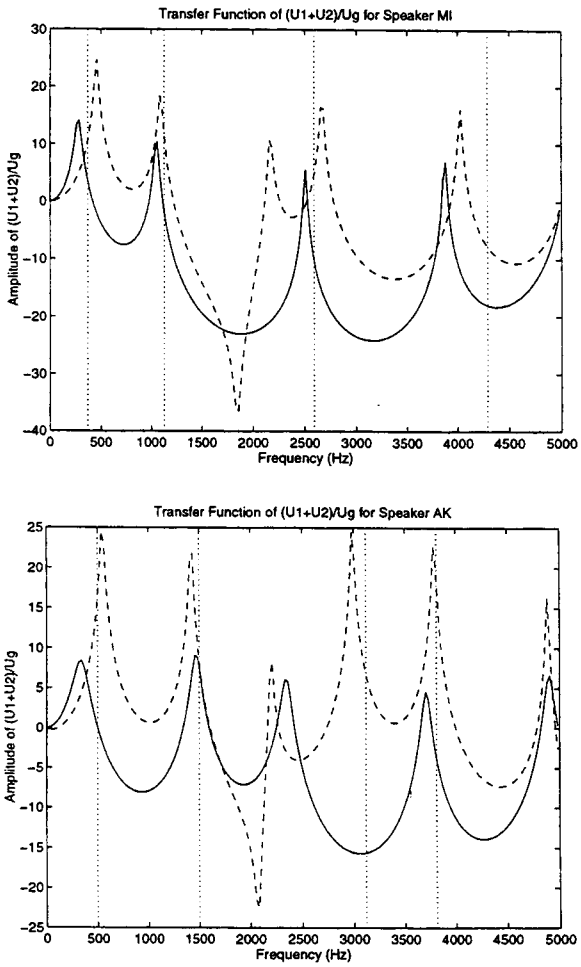

Figure 5: **Simulated transfer functions,** and the first four formant frequencies of the natural spectra. The solid line represents the transfer function from the lumped topology while the dashed line is calculated from the branched topology. The vertical dotted lines correspond to the first four formant frequencies measured from the LPC spectra of the natural tokens. The branched topology resulted in a pole/zero pair, and an improved prediction of the natural formant frequencies, especially for AK's F3.