ESTIMATION OF VOCAL TRACT FRONT CAVITY RESONANCE IN UNVOICED FRICATIVE SPEECH

Minkyu Lee and *Donald G. Childers*
Department of Electrical and Computer Engineering
University of Florida
Gainesville, FL 32611, U.S.A.

ABSTRACT

The purpose of this paper is to study the effect of the front cavity resonance and the vocal tract area function on the quality of synthesized unvoiced speech. From prior experiments, it has been determined that unvoiced speech is highly related to the vocal tract front cavity resonance. The noise source is located near the vocal tract constriction and the front cavity serves as a spectral shaping filter.

An algorithm is proposed to estimate front cavity resonances, from which effective length of the vocal tract front cavity can be calculated. The parameters are used to construct a simple vocal tract area function. Unvoiced speech is generated using an articulatory synthesizer. And effects of the front cavity length, back cavity shape on the perception of unvoiced fricatives are investigated.

1 INTRODUCTION

Research in linguistics and psychoacoustics suggests that the human auditory system tends to simplify the formant information. It is observed, in the study of auditory perception, that two spectral peaks are all that are needed for simulating front vowels and that one spectral peak is sufficient for simulating back vowels (Delattre *et al.*, 1952). Carlson *et al.* (1975) and Bladon and Fant (1978) also found that vowels can be simulated perceptually by using only two spectral peaks. While the first peak is held at the first formant F1, the second spectral peak has to be put between F_3 and F_4 for simulating front vowels, and into the vicinity of the second formant F_2 when simulating back vowels. Fant called the second spectral peak the effective second formant, F_2', because it does not correspond to any of the formants (Fant and Risberg, 1962). It seems that the F_2' has some correlation with speech production as well as speech perception. Fant showed that, when simulating a vowel by a single harmonic signal, the listeners responded to a resonance frequency of the uncoupled front cavity of the vocal tract. He suggested that the F_2' might be equivalent to this resonance frequency of the uncoupled front cavity.

According to the basic acoustic theory of speech production, the fundamental resonance frequency of the front cavity may be associated with any of the first four formants. An experiment by G. M. Kuhn (1975) using spectrographic data from two types of speech, one from normal speech and the other from fricative speech, suggested that it may be possible to estimate the frequency of the front cavity resonance from information in the speech signal. It is also suggested that the front cavity resonance could be estimated from speech data that is of a form more like that found in the auditory system (Kuhn, 1975).

In this study, we attempt to estimate the length of vocal tract front cavity for unvoiced fricatives using the front cavity resonance frequency estimates. The analysis method adopted in this research is based on the psychoacoustics of human hearing. Perceptual linear predictive (PLP) analysis, originally suggested by Hynek Hermansky (Hermansky, 1990), is used to approximate the auditory spectrum of speech by an all−pole model. The auditory spectrum is obtained from the speech signal by filtering using a critical−band filter bank followed by an equal loudness curve pre−emphasis and an intensity to loudness conversion by the intensity−loudness power law. Then, the auditory spectrum is modelled by an autoregressive (all pole) model. The estimates are, then, verified by informal listening test of synthetic fricatives generated by an articulatory synthesizer.

2 Analysis of unvoiced fricatives

Analysis data used are sustained fricatives of /s/, /ʃ/, /f/, and /h/. The FFT spectra for each of the four fricative classes are shown in Figure 1. The spectra are computed using a 512 point FFT for Hamming windowed speech segments. The spectrum of /s/ shows one major spectral peak near 5000 Hz, while for /ʃ/, there is a spectral peak at about 1900 Hz. The spectrum of /f/ does not have a prominent peak and the spectrum of /h/ has a peak at a low frequency below 1000 Hz. There is only a small difference in the frequency region below 1000 Hz for the fricatives /s/, /ʃ/, and /f/.

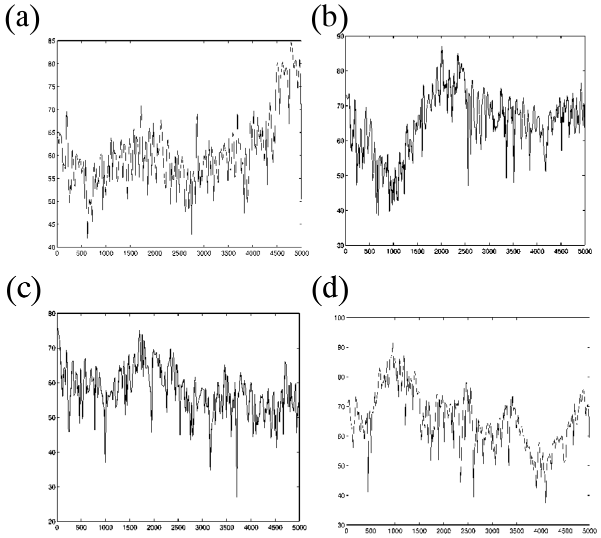

Figure 1. Example of measured spectra for (a) /s/, (b) /ʃ/, (c) /f/, and (d) /h/.

2.1 Analysis procedure

This section explaines analysis procedure. First, speech signal is segmented into frames and windowed and the PLP analysis was performed for each frame (Hermansky, 1990). The pole frequency and bandwidth are transformed back to the linear scale (Hz) from the Bark scale. Poles that have large bandwidth are eliminated. A pole frequency that corresponds to a minimum bandwidth is chosen as a major spectral peak frequency. The pole frequency is our estimate of the vocal tract front cavity resonance frequency. The front cavity resonance can be chosen by a pole frequency with a bandwidth less than a threshold value. From this experiment, a threshold of about 400 Hz seems to be a reasonable value. If there are no poles that meet the bandwidth requirement, it is assumed that the place of articulation is at the mouth opening, as in the example of /f/, and /θ/. The spectrum of /s/ shows one major spectral peak near 5000 Hz, while for /ʃ/, there is a spectral peak at about 1900 Hz. The spectrum of /f/ does not have a prominent peak and the spectrum of /h/ has a peak at a low frequency below 1000 Hz.

The front cavity resonance corresponds to the quarter−wavelength resonance frequency of the front cavity (Kuhn, 1975). Once the front cavity resonance frequency is estimated, the functional length of the front cavity, which is the length from the lips to a supraglottal constriction, can be calculated using the formula, $l=c/(4*f)$, where l is the front cavity length, c is the speed of sound (353m/sec for 35° C) and f is a quarter−wave resonance. For sibilant fricative sounds, the estimated average front cavity length are quite reasonable, i.e., 1.74 cm for /s/ and 4.1 cm for /ʃ/. On the other hand, there are no front cavity and back cavity for /θ/ and /f/ sounds because the constriction is at the lips, and, therefore, the analysis algorithm could not find a front cavity resonance frequency that has a bandwidth of less than the threshold 400 Hz. In this case, the place of articulation is at the mouth opening, and thus, the effective length front cavity is zero. Finally, the effective front cavity length of 9.94 cm for aspiration noise /h/ is shorter than we expected. This error is probably because the analysis detected a peak from the tracheal pole−zero pair (Klatt and Klatt, 1990).

	/s/−sat	/ʃ/−ship	/θ/−thin	/f/−fix	/h/−hat
Front cavity length (cm)	1.74	4.1	0	0	9.94*

From this experiment, we can conclude that the front cavity resonance frequency moves upward in frequency as the distance from the consonant constriction to the lips decrease. The analysis algorithm based on the PLP algorithm could provide a resonable estimate of front cavity resonance frequency. When the constriction is even more anterior, there is no front cavity and the analysis algorithm could not produce any reasonable estimate of the front cavity resonance, in which case we can assume that the constriction is at the mouth opening.

3 Unvoiced Fricative Speech Synthesis

3.1 Unvoiced Fricative Model

A schematized model of the vocal tract for an unvoiced fricative consonant is shown in Figure 2. A_g and A_c are the cross−sectional areas of the glottis and the constriction, respectively. Likewise, L_g and L_c are length of the glottis and the vocal tract constriction, respectively. If $A_g > A_c$, the supraglottal constriction plays a major roll in the generation of fricative speech. Aspiration noise is generated when $A_g < A_c$.

3.2 Articulatory synthesis

In order to verify the validity of the parameters estimated in the previous section, unvoiced sounds are generated using an articulatory synthesizer, which is based on the time−domain approach suggested by Sondhi and Schroeter (1987) and improved by Hsieh (1994). Using the estimates of the front cavity length with

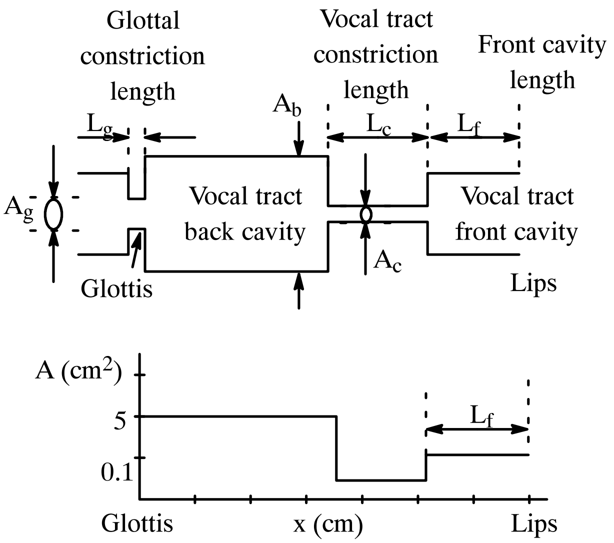

Figure 2. A model of the vocal tract and area function for fricative consonants. The vocal tract constrict divides the vocal tract into front and back cavity. A_c and L_c determine the characteristic of turbulence noise source.

nominal values of supraglottal constriction length and back cavity shape, a vocal tract area function can be constructed. The input area function is transformed to the equivalent RLC−network. Using the noise source excitation from the spectral shaping filter as a voltage source and by applying Kirchoff's and Ohm's law to the network, the discrete−time acoustic matrix equations are formed. The pressure at the midpoint of each section and volume velocity are calculated as solutions using the elimination procedure and a backward substitution. The synthetic speech is the backward difference between the sum of the volume velocities at the nostrils and lips at the current time and the sum of the volume velocities at the nostrils and lips at the previous time instant (Hsieh, 1994). A noise generator is implemented based on the model proposed by Sondhi and Schroeter (1986, 1987). The noise source model defines the characteristics of the noise source as a function of the airflow through the constriction and of the constriction cross-sectional area A_c. The turbulence gain and critical Reynolds number can also be specified. The noise spectrum is white.

3.2.1 Experiment I– Effect of the front cavity length

With the noise source parameters fixed, the effect of the front cavity length on the synthetic fricative sound can be determined by comparing the spectra of the synthetic sounds generated by changing the front cavity length. With the back cavity width fixed and the cross sectional area for the constriction is set to 0.1 cm^2, the front cavity length is gradually increased from 0 cm to 6.4 cm. Figure 3 shows the FFT spectra for synthetic unvoiced sounds. When the front cavity length is zero, i.e., the place of articulation is at the lips, the spectrum of the synthetic sound is similar to the spectrum of the /f/ sound (Figure 3−a). As the front cavity length increases to one to two centimeters, the spectrum becomes close to the spectrum of the /s/ sound (Figure 3−b). When the length of the front cavity is around three to six centimeters, the spectrum resembles the spectrum of /ʃ/ (Figure 3−c). Finally, Figure 3−d is when the front cavity length is greater than 9 centimeter and the spectrum is similar to the aspiration /h/. The observations in this experiment are confirmed by informal listening tests.

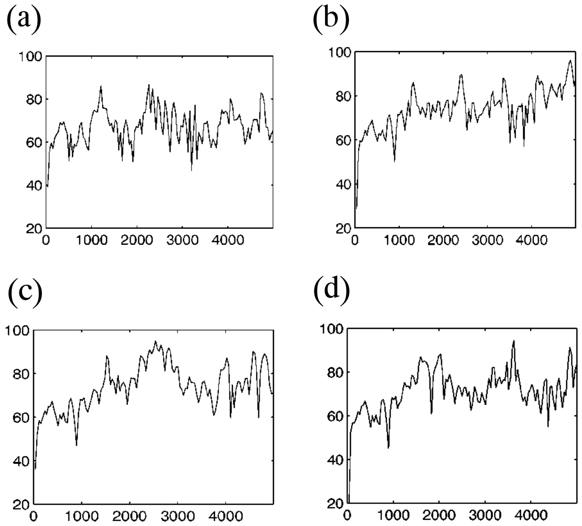

Figure 3. The spectra of unvoiced synthetic speech varying the front cavity length. (a) L_f=0, (b) L_f=1.5, (c) $L_{f=}6$, (d) L_f=6 cm.

3.2.2 Experiment II– Effect of the back cavity on the synthetic fricative sounds

This experiment is to study the effect of the back cavity resonance on the synthetic fricative sound. Given noise source parameters, a front cavity length, and a constriction length/width, the shape of the area function corresponding to the back cavity is changed as in Figure 4−a and Figure 4−c. The front cavity and constriction part of the two area functions are almost the same, i.e., the cross sectional area of the supraglottal constriction is 0.4 cm^2, the front cavity length is 1.9 cm, and the front cavity area is 5 cm^2. These numerical value are measured area function using X−rays (Figure 4−c) (Badin, 1991). Vocal tract shape for fricative /ʃ/ was chosen for this experiment because it has distinct front and back cavity.

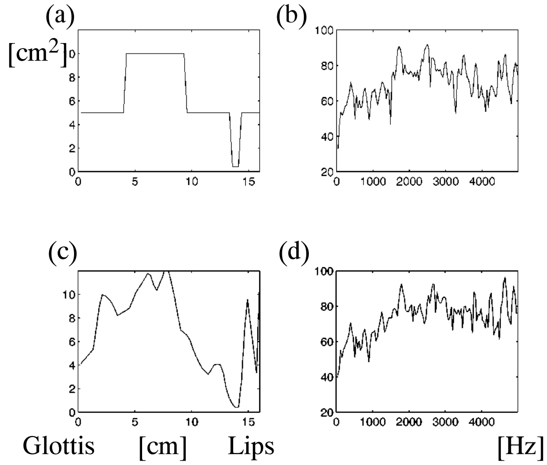

Figure 4. Area function and FFT spectrum of synthetic speech.

It can be observed that the formant structure of the spectra are very similar. This is probably because the back cavity resonance is decoupled from the constriction and the front cavity by narrow and long supraglottal constriction, thus, having little influence on the overall speech spectrum (Heinz and Stevens, 1961).

4 CONCLUSION

Using the front cavity resonance estimates from the algorithm based on an human auditory model, effective length of the vocal tract front cavity can be calculated. The parameters obtained from the analysis are used to construct a simple vocal tract area function. Using an articulatory synthesizer, which consists of a turbulent noise source generator and a vocal tract filter, unvoiced speech is generated. Informal listening test proves that the length of the front cavity determines which sound is pronounced while the shape and size of the back cavity is less important for fricative sound generation.

The main point of the experiments is the importance of the front cavity resonance in the production of unvoiced fricative speech. The length of the front cavity could be estimated from acoustic speech signal using the analysis method developed here. In general, a simple vocal tract area function is enough to generate an intelligible unvoiced sound. From the experiments on unvoiced fricatives speech, we can summarize the production of unvoiced speech as follows:

$\boxed{1}$ For sibilant fricatives (/s/ and /ʃ/), the supraglottal constrictions are narrow and long enough to decouple the back cavity resonance with the front cavity. The length of the front cavity determines which sound is pronounced. And the length can be estimated using the analysis methods developed here.

$\boxed{2}$ The fricative /f/ is generated at the lower lip and upper incisor, and /θ/ is generated by the tongue tip and incisors. The turbulence noise source for these fricatives is located at the mouth opening and, therefore, there is no front or back cavity. The FFT spectrum of speech sound is similar to the spectrum of turbulence noise source.

5 REFERENCES

Badin, P. (1991). “Fricative consonants: Acoustic and X-ray measurements,” J. of Phonetics, 19, 397–408.

Bladon, R. A. W., and Fant, G. (1978). “A two–formant model and cardinal vowels,” STL–QPRS, RIT, Stockholm, Sweden, 1, 1–8.

Carlson, R., Fant, G., and Granstrom, B. (1975). “Two–formant models, pitch and vowel perception,” *Auditory Analysis and Perception of Speech,* Academic Press, London, 55–82.

Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, L. J. (1952). “An experimental study of the acoustic determinants of vowel color,” Word, 8, 195–210.

Fant, G., and Risberg, A. (1962). “Auditory matching of vowels with two formant synthetic sounds,” STL–QPSR, RIT, Stockholm, Sweden, 4, 7–11, .

Heinz, J. M., and Stevens, K. N. (1961). “On the properties of voiceless fricatives,” J. Acoust. Soc. Am., 33, 589–596.

Hermansky, H. (1990). “Perceptual linear predictive (PLP) analysis of speech,” J. Acoust. Soc. Am., 87(4), 1738–1752.

Hsieh, Y. F. (1994). “A flexible and high quality articulatory speech synthesizer,” Ph.D. Dissertation, University of Florida.

Klatt, D. H., and Klatt, L. C. (1990). “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” J. Acoust. Soc. Am., 87(2), 820–857.

Kuhn, G.M. (1975). “On the front cavity resonance and its possible role in speech perception,” J. Acoust. Soc. Am., 58, No. 2, 428–433.

Sondhi, M. M., and Schroeter, J. (1986). “A nonlinear articulatory speech synthesizer using both time- and frequency-domain elements,” Proc. IEEE Int. Conf. on ASSP, 1999–2002.

Sondhi, M. M., and Schroeter, J. (1987). “A hybrid time-frequency domain articulatory speech synthesizer,” IEEE Trans. on ASSP, 35(7), 955–967.

Stevens, K. N. (1971). “Airflow and turbulence noise for fricative and stop consonants: Static considerations,” J. Acoust. Soc. Am, 50(4), 1180–1192.