PARAMETRIC HYBRID SOURCE MODELS FOR VOICED AND VOICELESS FRICATIVE CONSONANTS

Shrikanth Narayanan[1] *and Abeer Alwan*[2]

[1]AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974
[2]Department of Electrical Engineering, UCLA, 405 Hilgard Avenue, Los Angeles, CA 90095

ABSTRACT

Source models for fricative consonants are derived based on aerodynamic principles of sound generation, in conjunction with vocal-tract models obtained from MRI data. Results indicate that a linear source-filter model is adequate for capturing essential spectral characteristics of sustained fricatives below 10 kHz. The hybrid source models employ a combination of acoustic monopole and dipole sources, and a voiced source in the case of the voiced fricatives. The number of sources, source locations and spectral characteristics are chosen based on an analysis-by-synthesis approach and are motivated by aeroacoustic theory. The resulting model is computationally efficient and can be readily used for synthesis.

1. INTRODUCTION

Physiologically and physically motivated models for speech production are important for the development of high-quality speech synthesizers and articulatory-based coders and recognition systems. In this paper, the articulatory-to-acoustic mappings in fricative consonants, sounds characterized by turbulence sources in the vocal tract, are investigated. Two problems are addressed: (1) The lack of accurate vocal-tract models (including characterization of inter- and intra-speaker variabilities) was overcome by collecting and analyzing magnetic resonance images (MRI), dynamic electropalatography (EPG) data, high-quality acoustic recordings, and aerodynamic data from human subjects during speech production. (2) The lack of satisfactory source models for fricatives was tackled by the specification of physically-motivated parametric models of turbulent sources, leveraging off theoretical and experimental aeroacoustic studies of turbulence sound generation. The study resulted in the development of parametric hybrid source models for fricative production.

This research is a portion of SN's doctoral thesis [8] while at UCLA. Work supported in part by NSF.

2. METHODOLOGY

The modeling is based on the source-filter theory of speech production. Planar wave propagation in the vocal tract is assumed and the effect of the vocal-tract bend is ignored. Source models are derived based on an analysis-by-synthesis approach rather than inverse filtering due to the distributed nature of the turbulent sources and unknown exact source locations.
Vocal-tract models are based on area functions obtained from magnetic resonance imaging of four subjects. Physically-motivated fricative source models are then constructed to synthesize fricatives. The location, type, and the spectral characteristics of the sources are the parameters employed in the modeling.

Vocal-tract models: MR images of the vocal tract were obtained from four phonetically-trained native talkers of American English (2 male, 2 female) in all three anatomical planes (sagittal, coronal, and axial) while they sustained the fricative consonants of English [1]. All eight fricatives were considered: the stridents: /sh, s, zh, z/, and the nonstridents: /th, f, dh, v/. Vocal-tract lengths, area functions, and volumes (such as those of the sublingual cavities) were measured from these images.
As a first approximation, the vocal tract is modeled as a concatenation of uniform cylindrical tube-sections each section being 3 mm long. Depending on the subject's vocal-tract length, the total number of sections is 55-60. Sublingual cavities, such as those present in /sh/ and /zh/ are modeled as shunt branches specified in the anterior buccal cavity. Once the area function is known, several approaches may be used for simulating the acoustics in the vocal tract. In the current study, a time-domain simulation method proposed by Maeda [2] is used to determine the vocal-tract transfer function for each sound.

Source models: Sound sources in fricatives are generated by turbulence and/or random fluctuations in the

0-7803-3192-3/96 $5.00©1996 IEEE

airflow, primarily at, or in the vicinity of, a constriction and/or an obstacle such as the teeth. All sources of flow-induced sound are represented by some combination of the three canonical source types, namely, monopole, dipole, and quadrupole sources [3]. A *monopole* source results from a net unsteady mass injection into the fluid region (resulting in "random fluctuations in the flow volume velocity"). *Dipole* sound is emitted when there is no net mass injection into the fluid but there is a net distribution of fluctuating forces, whereas a *quadrupole* source exists in free turbulence. Depending upon the configuration of the vocal tract and the constriction geometry, which in turn depend on the fricative place of articulation, different turbulence-generating mechanisms emerge. Sound generation due to jets emerging from a constriction and impinging on an obstacle (such as the teeth) or on a surface (such as the vocal-tract walls) and causing a fluctuating force on the fluid medium is attributed predominantly to a distribution of dipole sources. Random velocity fluctuations within the constriction due to irregularities in the constriction geometry, on the other hand, constitute a monopole source. The sound generation due to turbulence in a free jet is attributed to a combination of monopole and quadrupole sources. For scenarios occurring in fricative production (predominantly impinging jets), the source effects are mainly due to the monopole and dipole sources which can be thought of, using the electrical network analogy, as series current sources and parallel voltage sources, respectively.

3. SIMULATIONS

The spectrum of the radiated sound pressure of the speech sound due to N multiple sources occurring in the vocal tract, is based on the linear frequency-domain relation:

$$P_r(\omega) = R(\omega) \sum_{i=1}^{N} T_i(\omega) S_i(\omega) \qquad (1)$$

The subscript "i" denotes a specific source location in the vocal tract, $T_i(\omega)$ is the transfer function between the volume velocity at the lips U_l and a source at the i^{th} section ($S_i(\omega)$), and $P_r(\omega)$ is the radiated sound pressure at a distance r cm in a far-field location. The effect of radiation from the lips to the far-field is accounted for by the term $R(\omega)$. The superposition in Eqn. 1 can be used to represent both the effect of multiple sources (such as voicing and turbulence) and the effect of distributed sources (approximated by lumped non-coherent entities). Note that the source $S_i(\omega)$ could be either volume velocity $U_i(\omega)$ or sound pressure $P_i(\omega)$. For a given $T_i(\omega)$ and $S_i(\omega)$, the product $T_i(\omega)S_i(\omega)$ in Eqn. 1 will always yield the corresponding volume velocity at the lips $U_l(\omega)$. For example, by superposition, $U_l(\omega)$ due to a pressure source $P_m(\omega)$ at location m and a volume velocity source $U_n(\omega)$ at location n is given by:

$$U_l(\omega) = T_m(\omega)P_m(\omega) + T_n(\omega)U_n(\omega) \qquad (2)$$

$$= [\frac{U_{l,m}(\omega)}{P_m(\omega)}]P_m(\omega) + [\frac{U_{l,n}(\omega)}{U_n(\omega)}]U_n(\omega) \qquad (3)$$

The boundary condition at the lips is based on the model of a piston in a spherical baffle. The boundary conditions at the glottis are specified by an appropriate subglottal pressure and a glottal impedance model [2]. The subglottal pressure was assumed to be 8 cm H_2O, corresponding to an open glottis condition, with the areas of glottal opening estimated from the MRI data. Synthesized spectra are then derived using *hybrid* source models in conjunction with the calculated transfer functions. The term *hybrid* implies the use of a combination of source **types** (such as monopole/dipole turbulent sources, and a voicing source) distributed at specific **locations** in the vocal tract where the aerodynamic generation of sound is thought to occur. In some cases, multiple sources of the same type may be specified to approximate the distributed nature of the source therein. The base-line (or, prototype) source models were derived from theoretical and experimental studies [4, 5, 6, 7] and used to specify parametric models for the turbulent sources. The dipole sources play a major role in determining the overall fricative spectrum. Hence, the emphasis was on 'optimizing' the dipole characteristics in order to achieve a closer match to the natural spectra. A three-parameter model for the dipole spectrum [Fig. 1] was proposed based on the results of previous aerodynamic studies on sources of turbulence generation due to jets impinging on an obstacle. The parameters are the frequency of the spectral peak, F_{peak} (in Hz), and the low- and high-frequency tilts T_{LF}, T_{HF} (in dB/octave). Based on experimental evidence [5], $F_{peak} = KUd/A_c$ (d = characteristic dimension of the jet, U = flow rate, A_c = area of jet constriction, K = scaling factor typically between 0.1-0.2) which implies a weak interaction between the source and filter. The nominal values for U were chosen from measured flow rates while those of d and A_c were provided by the MRI-data. Other base-line values for the model's parameters such as the spectral tilts were derived from various experimentally and empirically-derived prototype models. These base-line values were optimized to improve the match between the synthesized and natural fricative spectra.

The output spectra that correspond to the pressure sig-

nal at the microphone are calculated using Eqn. 1 and the performance of the hybrid sources is evaluated by comparing the output spectra from the models to those obtained from the natural utterances. In addition, the rms log-magnitude spectral distortion (L_2 norm) between the model and the natural speech spectra provided an objective performance measure.

4. RESULTS

A finite time-difference simulation [2] of the acoustic propagation in the vocal tract was employed to derive the various transfer functions using the MRI-derived area functions. Note that the MRI data were obtained from subjects in supine position. Analysis of EPG data, however, showed that the differences between supine and upright positions were insignificant. The losses required for synthesizing realistic fricative transfer functions were greater than the nominal values typically used for synthesizing vowels. Increased losses in the vocal tract during frication were specified by increasing the resistance due to viscous losses. Inclusion of a sublingual cavity resulted in lowering the resonant frequency associated with the cavity anterior to the supraglottal constriction.

The hybrid source models employed a combination of acoustic monopole and dipole sources, in addition to a voicing source for voiced fricatives, with the specification of the source location, number, and spectral characteristics motivated by aeroacoustic theory. Furthermore, multiple dipole sources (typically two) were used to approximate the distributed nature of the source. The results demonstrated good agreement between the natural and synthetic spectra. The number of sources, their location, and spectral characteristics for /s, sh/ of a male (MI) and female (PK) subject are summarized in Tables 1-3. The synthesized and natural spectra for both subjects are illustrated in Figures 2 and 3.

Results indicate that a source-filter type representation may be adequate for modeling fricative consonants in the frequency range below 10 kHz and that the sources can be represented by parameterizable-dipole and monopole models. The turbulence models comprised a combination of dipole sources at one or more obstacle locations and a monopole source at the constriction exit. The obstacle locations were estimated to be at the teeth and along the vocal-tract wall for /sh, zh/, at the teeth for /s, z/, and at the lip surface for the nonstrident fricatives. The distributed nature of the dipole sources was effectively approximated by two lumped sources placed at adjacent locations. The range of K values obtained through simulations was between 0.1 – 0.44 for the strident fricatives and 0.2 – 0.3 for the nonstridents; higher values were used in cases where obstacles were closer to the jet outlet than those farther away from it. The tilt T_{HF} values were in the range -3 to -14 dB/oct for the strident fricatives and 0 to -8 dB/oct for the nonstridents. There was a significant variability in the T_{HF} values across subjects which is primarily attributed to differences in flow rates. The roll-off T_{LF} in the frequency region below the peak F_{peak} was in the range 6-16 dB/oct.

In general, the match obtained between the spectra of the synthetic and natural utterances was found to be better for the strident fricatives than for the nonstridents (compare Figures 2 and 3). The spectra of the nonstrident fricatives are not, due to inter-subject variabilities, as well-defined as those of the stridents, and the resonance patterns are predominantly due to incomplete pole-zero cancelations resulting from the finite coupling between the front and back cavities. The 1D model used for the vocal-tract simulation is, perhaps, not quite adequate for capturing these effects.

The results for the voiced fricatives were found to be quite satisfactory through a superposition of the voicing and turbulence sources. The maximum voicing source strengths were 12 to 26 dB above the maximum dipole source strengths. This would imply a transglottal pressure in the range between 4.3 to 6.5-cm H_2O, assuming a subglottal pressure of 8-cm H_2O and A_c in the range 0.1 to 0.2 cm^2. Since the EPG data showed no significant effects of voicing on constriction location and width, the locations of the turbulent sources for the voiced and unvoiced fricative pairs were chosen to be similar. The role played by the monopole source in contributing to the overall spectra was found to be rather minimal while the dipole sources were found to play a significant role in defining the overall spectral characteristics.

The availability of parametric source models and reliable vocal-tract filter models provides insights into devising identification schemes for estimating the source parameters from the acoustic signal, i.e., the inverse problem or the 'acoustic-to-articulatory' mapping. Future work will focus 2D and 3D acoustic modeling and on examining the inverse problem for fricative consonants.

References: [1] S. S. Narayanan, A. A. Alwan, and K. Haker, "An articulatory study of fricative consonants using magnetic resonance imaging," *JASA*, vol. 98, pp. 1325-1347, Sept. 1995. [2] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, pp. 199–229, Oct. 1982. [3] K.N.Stevens, *Acoustic Phonetics*, in preparation. [4] G. Fant, *Acoustic Theory of Speech Production*. Mouton: The Hague, 1960. [5]

Source type	*monopole*	*dipole*	*dipole*
SUBJECT MI			
Number	1	2	2
Location	constriction	wall	teeth
Distance from lips (cm)	2.7	2.7,3.0	0.9,1.2
Maximum levels (dB) (rel. to monopole max)	0	21	17
SUBJECT PK			
Number	1	1	1
Location	constriction	wall	teeth
Distance from lips (cm)	1.8	1.8	0.9
Maximum levels (dB) (rel. to monopole max)	0	23	17

Table 1: Hybrid source models for /sh/ (MI and PK).

Source type	*monopole*	*dipole*
SUBJECT MI		
Number	1	2
Location	constriction	teeth
Distance from lips (cm)	1.8	0.9, 1.2
Maximum levels (dB) (rel. to monopole max.)	0	17
SUBJECT PK		
Number	1	2
Location	constriction	teeth
Distance from lips (cm)	1.5	0.9, 1.2
Maximum levels (dB) (rel. to monopole max.)	0	17

Table 2: Hybrid source models for /s/ (MI and PK).

K. N. Stevens, "Airflow and turbulence noise for fricative and stop consonants: Static considerations," *JASA*, vol. 50, pp. 1180–1192, May 1971. [6] C. H. Shadle, *The Acoustics of Fricative Consonants.* RLE Tech. Rep. 506, M.I.T, 1985. [7] L. M. P. Pastel, "Turbulent noise sources in vocal tract models," unpublished S.M. EE thesis, M.I.T., 1987. [8] S. S. Narayanan, "Fricative Consonants: An articulatory, acoustic, and systems study," Ph.D. thesis, EE Dept., UCLA, 1995.

Fricative	F_{peak} Hz	Scaling K	T_{LF} dB/oct	T_{HF} dB/oct
SUBJECT: MI /s/	2670	0.20	12	-6
/sh/ (wall)	5125	0.24	11	-8
(teeth)	3187	0.15	8	-5
SUBJECT: PK /s/	6993	0.30	8	-14
/sh/ (wall)	5211	0.20	9	-10
(teeth)	2584	0.10	6	-10

Table 3: Dipole spectral characteristics: /s/ and /sh/.

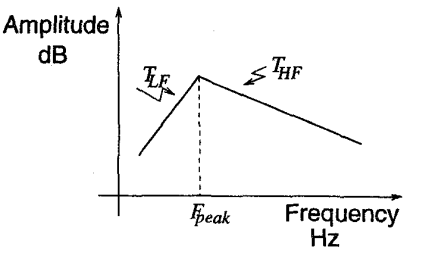

Figure 1: *Stylized Dipole Spectra*

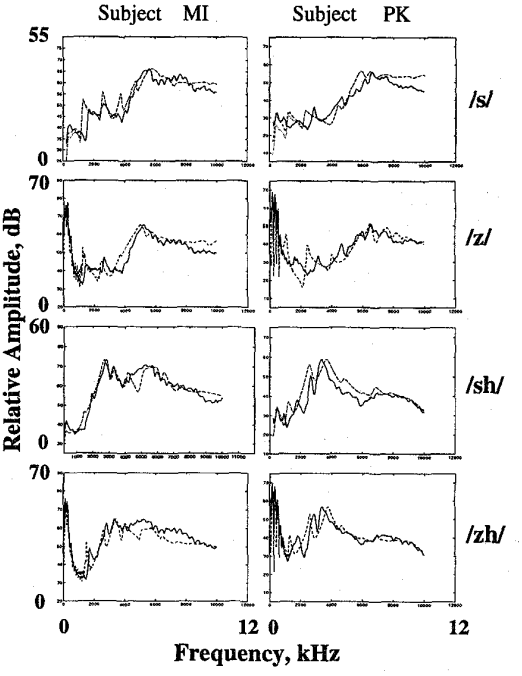

Figure 2: *Modeling results for strident fricatives: synthesizedspectra (dashed), natural spectra (solid).*

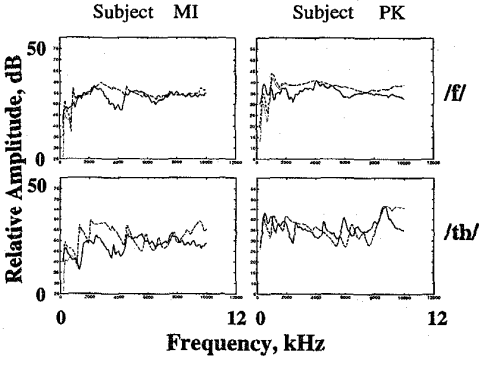

Figure 3: *Modeling results for nonstrident fricatives: synthesized spectra (dashed), natural spectra (solid).*