Noise Source Models for Fricative Consonants

Shrikanth Narayanan, *Member, IEEE,* and Abeer Alwan, *Member, IEEE*

***Abstract*—Hybrid source models for fricative consonants are derived based on aeroacoustic principles of sound generation employing vocal tract area functions obtained from magnetic resonance imaging data of voiced and unvoiced English fricatives. Results based on data from a male and a female subject indicate that a linear source-filter model is fairly adequate for capturing essential spectral characteristics of sustained voiced and unvoiced strident fricatives below 10 kHz. The hybrid source models employ a combination of acoustic monopole and distributed dipole sources, and a voice source in the case of the voiced fricatives. The number of sources, source locations, and spectral characteristics and the relative source levels are chosen based on an analysis-by-synthesis approach and are motivated by aeroacoustic theory of speech production. The resulting model is computationally efficient and can be readily used for synthesis.**

***Index Terms*—Articulatory acoustics, fricatives, magnetic resonance images (MRI), noise source models, speech synthesis, turbulence, vocal tract area functions.**

I. Introduction

ARTICULATORY-TO-ACOUSTIC mapping 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 was overcome by collecting and analyzing magnetic resonance images (MRI) and high-quality acoustic recordings from human subjects during speech production.
2) The lack of satisfactory source models for fricatives was addressed by the specification of physically motivated parametric models of turbulent sources that leverages theoretical and experimental aeroacoustic studies of turbulence sound generation. The study resulted in the development of parametric hybrid source models for fricative production.

The flow problem in a complicated three-dimensional (3-D) vocal tract geometry is reduced to a one-dimensional (1-D) problem based on several simplifying assumptions including source–filter separability. Separability of the source(s) from the vocal tract assumes that the source properties are dependent only on the airflow and the constriction geometry where the airflow gets modulated (at the glottis and/or at a supraglottal location) and are independent of the rest of the tract. The linear acoustic behavior of the vocal tract filter can then be modeled as an electrical transmission-line model assuming planar wave propagation [4], [5]. Several approaches exist for simulating the acoustics in the vocal tract [4], [8], [23]. In this study, a time–domain simulation method of acoustic propagation in the vocal tract proposed by Maeda [8] is used to derive the vocal tract transfer functions. Simulations of such 1-D models, besides providing a convenient framework for studying speech production, are essential to validate the adequacy/inadequacy of source–filter type models.

A review of fricative production mechanisms, the articulatory and acoustic data and source models used in this study are given in Section II. The simulation methodology is described in Section III and the results are given in Section IV. A summary and discussion of the results are provided in Section V.

Manuscript received October 31, 1997; revised August 9, 1999. This work was supported in part by the National Science Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Douglas D. O'Shaughnessy.

S. Narayanan is with AT&T Labs—Research, Florham Park, NJ 07932 USA (e-mail: shri@research.att.com).

A. Alwan is with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: alwan@icsl.ucla.edu).

Publisher Item Identifier S 1063-6676(00)01718-1.

II. Background

A. Fricative Production Mechanisms

Fricatives are produced when a narrow supraglottal constriction is formed in the vocal tract, and air flowing through the tract, and constriction, generates turbulence in the region downstream from the constriction [4], [20], [24]. The generation of turbulence occurs near the vocal tract walls and/or the teeth which may act as an obstacle. In addition to turbulence, the vocal folds vibrate, at least for part of the frication period, for voiced fricatives. The eight fricative consonants in English, specified in terms of their place of articulation in unvoiced–voiced pairs, are: the labiodentals /**f**/ and /**v**/, interdentals /**θ**/ and /ð/, alveolars /**s**/ and /**z**/, and postalveolars /ʃ/ and /ʒ/. The alveolar and postalveolar fricatives are often referred to as sibilant or strident fricatives while the labiodentals and interdentals are referred to as nonsibilant or nonstrident fricatives.

In the past, several interesting theoretical and experimental studies of modeling the acoustics of fricative consonants have been made in the articulatory, aerodynamic, and/or perceptual domains. In one of the earliest works on modeling fricatives, Meyer-Epplere [13] used mechanical tube models to study the relationship between sound-pressure and the Reynolds number of flow and used those results to infer the articulatory parameters for fricative production. Fant [4], Heinz and Stevens [9], and Flanagan [6] all used a serial pressure source in a source-filter model for synthesizing fricative spectra. Acoustic mechanisms of fricatives were investigated using aerodynamic theory of flow induced sound by Stevens [24] and followed later by extensive experimental and theoretical work by Shadle [19]–[21] using mechanical models. A similar study by Pastel [17] on mechanical models focused on the aerodynamics and acoustics of /h/. Source model parameters estimated from the experiments were used to model the distributed nature of the noise source in a linear vocal tract (system) model. Some of the experimental findings of Shadle and Pastel will be further discussed in Section II-D2 and invoked in the simulations reported in this paper.

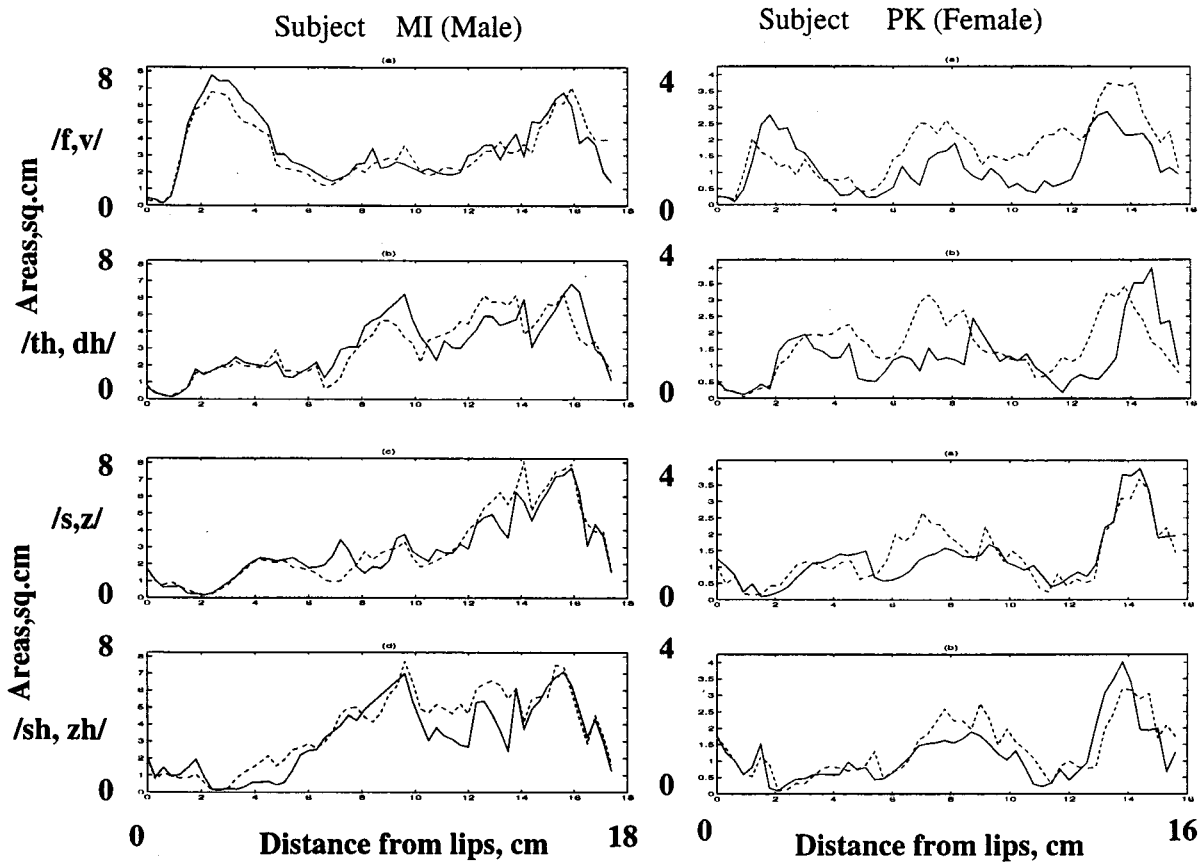


Fig. 1. Area functions measured from MRI data for subjects MI and PK. Solid line: unvoiced, dashed line: voiced. The last two laryngeal sections of the vocal tract are not shown.

Badin [1], [2] has contributed significantly to modeling the relation between the aerodynamic and articulatory parameters of fricative production such as the sound pressure, flow velocity, and vocal tract constriction areas.

B. Data Acquisition and Analysis

1) MRI Data: Area functions used in the modeling were derived from MRI data [14]. Two phonetically trained native talkers of American English (one male, one female), in supine position, sustained each consonant for about 13–16 s enabling four to five image slices to be recorded in a particular scanning plane (about 3.2 s/image). MRI data for all eight fricatives were obtained in the three anatomical planes: 28 to 35 images/sound/subject in the sagittal plane, and 40 to 45 images/sound/subject in the axial and coronal planes. These data were then used for 3-D reconstructions and measurements of vocal tract length, area, and volume. Note that the "effective" area of the airway was obtained by a simplification of the morphology: subtracting areas of the tissues (uvula, epiglottis, aryepiglottic and glossoepiglottic folds, false vocal folds, piriform sinuses, and valecullae) from the total pharyngeal cavity areas. Furthermore, since the teeth are not captured well by MRI, measurements in the dental region were aided by information provided by the subjects' dental casts. Area functions from the male (MI) and female subject (PK), shown in Fig. 1, were used for modeling. Calibration errors, overestimates, for area measurements were in the range of 2–8% [14]. A section length of about 3 mm was used, yielding a total of 55 to 60 sections depending on the length of the subject's vocal tract (16.5 and 18 cm for PK and MI, respectively). The alveolar fricatives of both MI and PK were apical (i.e., constriction formed with a raised tongue tip) while the postalveolars were laminal (i.e., constriction formed with a raised tongue blade). A sublingual space was consistently found in the postalveolars.

2) Acoustic Data: Speech data were collected in a soundproof facility with the subjects producing, from a supine position, eight consecutive repetitions of each sustained fricative, each lasting for about 5 s, with a short pause between each repetition. Each sustained fricative was produced beginning with the neutral vowel /ə/. Note that sound recordings could not be made during the MRI experiments due to the high level of ambient noise. The speech data were recorded at 44.1 kHz directly onto a Sun workstation and were later downsampled to 22.05 kHz. An omnidirectional microphone (Beyerdynamic M101) with a flat frequency response (within 4 dB) between 40–20 000 Hz was placed approximately 22 cm from the subject's mouth at about 15° angle off the midline. Fricative spectra were calculated using the Welch periodogram technique that used 100 FFT spectra obtained from overlapping Hanning window segments (23.22 ms length with 10 ms overlap). For each fricative, the natural spectrum used as a reference for acoustic modeling was an average of the eight repetitions. Fig. 2 shows sample unvoiced fricative spectra for subject MI. Note that the acoustic

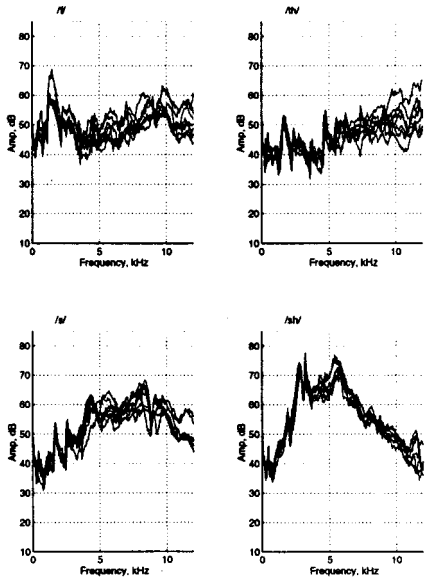


Fig. 2. Spectra of unvoiced fricatives (spoken by subject MI while in supine posture, overlay of 8 tokens/panel): top: left- /**f**/, right- /θ/; bottom: left- /**s**/, right- /ʃ/.

recordings were not calibrated to provide absolute sound pressure levels. Hence, any match of the overall amplitudes of the natural and synthesized spectra is arbitrary.

Spectral characteristics of the sustained fricatives agree well with the fricative spectra described in previous studies [1], [10], [22]. Strident fricatives exhibit a dynamic range of about 15–20 dB greater than nonstridents. Experiments on mechanical models have shown that obstacles such as the teeth in the path of an air jet emerging from a constriction contribute significantly to the relatively large values of flow-induced sound in stridents [19]. Strident fricatives exhibit significant spectral power in distinct frequency ranges: postalveolars, between 2.2–3.5 kHz and alveolars, 4.6–6.6 kHz. The significant spectral peak in this region is, in general, attributed to the lowest front-cavity resonance [9]. The broad peak in the high-frequency region, generally well above 5 kHz, found in the labiodentals and interdentals is attributed to the relatively short front cavity [9]. High-frequency spectral tilts (i.e., in the region above the main spectral peak) in stridents were found to be different across speakers, perhaps, due to differences in flow rates [21]. Overall spectral levels in the low frequency end of strident spectra, on the other hand, were anywhere between 15 to 30 dB below that of the main peak. This is primarily due to the presence of free zeros in the transconductance between the nonterminal pressure source excitation and the volume velocity at the lips [1].

The characteristic peaks observed in the low frequency end of the spectra, especially in the nonstridents, are attributed to the resonances of the cavity posterior to the supraglottal constriction (back cavity and subglottal structures) [1], [25]. The strength of the spectral power in these resonances is largely a function of the degree of coupling between the cavities in front and back of the constriction. For example, a prominent peak around 2500 Hz (associated with a back cavity resonance), seen close to the primary front cavity resonance in the postalveolars suggests that the front/back cavity coupling therein is appreciable. In addition, there may be an influence of other cavities such as the subglottal system, piriform sinuses, and epiglottic valleculae (see for example, [3]).

The high-frequency behavior of the voiced fricatives was similar to the unvoiced ones [1], [15]. The relative level of the main spectral peaks in the voiced fricatives, particularly the stridents, was lower than the unvoiced ones, presumably due to a smaller pressure drop across the supraglottal constriction [26]. The low-frequency region (about 1 kHz and below) of the voiced fricatives is dominated by the effects of voicing. The amplitude of "voicing" around the fundamental frequency was comparable to or greater than the amplitude of the main spectral peak associated with the lowest front cavity resonance.

C. Source Models

All sources of turbulent flow-induced sound can be represented by some combination of three canonical source types: monopole, dipole, and quadrupole [8], [11]. A flow monopole source results from a net unsteady mass injection into the fluid region. For example, flow monopole sources exist at the nozzle of a jet wherein a net fluctuation in mass occurs. If the dimensions of the source are smaller than the acoustic wavelength, then the magnitude of the radiated sound pressure is independent of both the sound speed in the medium and the angular orientation of the source with respect to the observer. A flow dipole sound is emitted when there is no net mass injection into the fluid but there is a net distribution of fluctuating forces. A flow dipole may be viewed as two flow monopole sources pulsating in opposite phase such that one monopole releases mass and the other absorbs it. Flow dipole sources exist at hard boundaries such as at an obstacle in the path of an impinging jet. When the unbounded medium has no net mass injection and no net forces, the source distribution is represented as a system of force couples and shear stresses resulting in a flow quadrupole source. Reference to monopole, dipole, and quadrupole sources in the rest of this paper implies flow-induced sources.

High-frequency components in turbulence are attenuated as the jet widens and the rate of decay depends on the presence of boundaries and the viscosity of the fluid. Experimental evidence has shown that spectra at locations farther away from the jet nozzle exhibit relatively smaller energy at high frequencies and greater energy in the low frequencies when compared to locations closer to the jet nozzle [8], [19]. As the jet widens, the particle velocity drops, affecting the strength of a potential dipole should the jet encounter an obstacle downstream.

Depending upon the configuration of the vocal tract and the constriction geometry (which vary depending on the place of articulation), different turbulence-generating mechanisms result. Stevens [25] summarizes at least three such mechanisms in fricative production.

1) Free jet emerging from a constriction with the turbulent velocity fluctuations distributed in the region downstream from the constriction. Such a sound generation mechanism is attributed to a combination of monopole and quadrupole sources.

2) Random velocity fluctuations within the constriction due to irregularities in the constriction geometry constitute a monopole source.
3) Jet emerging from a constriction impinging on an obstacle (e.g., teeth) or on a surface (e.g., vocal tract walls) resulting in a fluctuating force on the fluid medium.

Jets that impinge at a normal direction to the obstacle result in a greater fluctuating force on the medium than those that impinge at smaller angles (i.e., $<90^{\circ}$). The spectrum of a dipole source exhibits a broad peak at a frequency proportional to (V/d) where V is the air particle velocity and d is the characteristic dimension of the constriction (such as the width of a rectangular slit). In speech, the peak frequency is expected to be between $0.1(V/d)$ and $0.2(V/d)$, above and below which the spectrum decays monotonically [24], [25]. There is other evidence, for example, [16], that shows a general log-linear spectral roll-off without the presence of any such intermediate peak. Shadle's experimental results derived from realistic vocal tract-like mechanical models while showed no low frequency roll-off in some cases (/**s**, ʃ/) demonstrated a gentle low frequency roll-off in others (/ç, x/) [21]. It should be noted that these experiments also revealed that source spectra critically depend on the geometry of the vocal tract enclosing the noise source.

A monopole source representing volume velocity fluctuations can be modeled by an equivalent current source in a transmission-line model. Similarly, dipole sources at an obstacle normal to the surface that are uniformly distributed, can be modeled by an equivalent pressure voltage source in a transmission-line model. The role played by quadrupole sources, is relatively insignificant in fricative production when compared to dipole and monopole sources, and hence are not considered in the acoustic modeling [19], [25].

D. Prototype Spectral Models

In this section, some prototype spectral models for turbulence sources which are based on theoretical and experimental results from previous studies are presented. These models are used to investigate the acoustics of fricative consonants. A parametric voice source model, is used in conjunction with turbulent sources when modeling voiced fricatives.

1) Source Models Obtained Through Analysis-by-Synthesis Methods:

a) Fant's spectral models for turbulent sources: Fant suggested the use of a voltage source to model turbulent sound source in fricatives [4]. Based on transmission-line models specified by X-ray derived area functions, the following empirical source characteristics were suggested: 1) /**f**/: −3 dB/oct between 0.8 and 10 kHz. 2) /**s**/: 0 dB/oct between 0.8 and 4 kHz, and −6 dB/oct between 4 and 10 kHz. 3) /ʃ/: For an apical source, 0 dB/oct between 0.3 and 6 kHz, and −12 dB/oct between 6 and 10 kHz. For a dental source, 0 dB/oct between 0.3 and 2 kHz, and −12 dB/oct between 2 and 10 kHz.

b) Voice source model: A simplified parametric model for the glottal waveform [18] is used to derive the voice source spectrum for voiced fricatives. The open-quotient OQ, the duration within each pitch cycle where the glottis is open, was chosen to be $0.75N$, where N is the pitch period. The assumption of a relatively high OQ is based on the rationale that the glottal waveform during voiced fricatives is often more or less sinusoidal due to an active partial opening of the glottis to permit a greater airflow that supports turbulence generation at the supraglottal constriction. The rise and fall times of the glottal pulse were $0.7OQ$ and $0.3OQ$, respectively. The amplitude of the voicing source is proportional to $\Delta P_g^{1.5}$, where ΔP_g is the pressure drop across the glottal constriction, as long as the vocal folds continue to vibrate [25].

2) Experimentally Derived Fricative Source Models:

a) Monopole spectra: A monopole source spectrum, experimentally obtained by Pastel [17], is shown in Fig. 3(a). The spectrum was estimated from the sound pressure radiated from a uniform tube configuration with a jet (flow rate of 442 cm^3/s) impinging normally on the tube's wall such that the jet axis was perpendicular to the longitudinal axis of the tube (inset in Fig. 3(a) shows a schematic of the experimental configuration). The nozzle where the jet emerges was located at the end of the tube away from the mouth. Since the majority of dipole sources, due to the wall obstacle in this configuration, are perpendicular to the longitudinal axis of the tube, they couple poorly to the sound field. Moreover, the free jet region in this configuration is relatively small implying that the contribution of the quadrupole sources to the radiated sound pressure is negligible as well. At subsonic speeds, the acoustic efficiency of the quadrupole sources is the smallest when compared to those of the other source types [8]. Hence, it is reasonable to assume that the radiated sound pressure for this configuration is primarily due to the monopole source at the jet inlet. The actual source spectrum was estimated by inverse filtering the radiated sound-pressure signal using the transfer function of the tube.

b) Dipole spectra: Dipole source spectra estimated from mechanical models with configurations, dimensions, and flow rates relevant to fricative production can be found in [17] and [19]. In Pastel's study, the mechanical tube model was 17 cm in length and 2.85 cm^2 in cross sectional area. Turbulence source spectra for sound generated due to jets ($A = 0.114$ cm^2) impinging on a wall obstacle (a movable annular ring in this case) at various angles were estimated through inverse filtering for sources located at various points along the tube. Fig. 3(b) shows sample source spectra estimated for a 27° jet (flow rate of 442 cm^3/s) at distances 1, 2, 3, and 4 cm from the obstacle. Note that the spectrum at the location closest to the obstacle (1 cm) shows greater energy at high frequencies when compared to those located farther away. A comparison of these dipole spectra [Fig. 3(b)] with the monopole spectrum of Fig. 3(a) reveals a faster roll-off at high frequencies for the monopole case. The monopole sources contribute toward greater spectral amplitudes in the low frequency range.

Fig. 3(c) shows the sound source spectra obtained by Shadle [19] for a jet emerging from a circular constriction ($l = 1$ cm, $A = 0.08$ cm^2) at the mouth of a cylindrical tube model ($l = 17$ cm, $A = 2.54$ cm^2) and impinging perpendicularly on a semicircular obstacle at 3 cm downstream from the constriction (for two flow rates: 160 and 420 cm^3/s). These spectra were approximated by exponential fits of the form $P_s(f) = ae^{(fb)}$ where f is the frequency in hertz. For $U = 420$ cm^3/s, $a = 95$ and $b = -0.0004$ and for $U = 160$ cm^3/s, $a = 80$ and $b = -0.0007$.

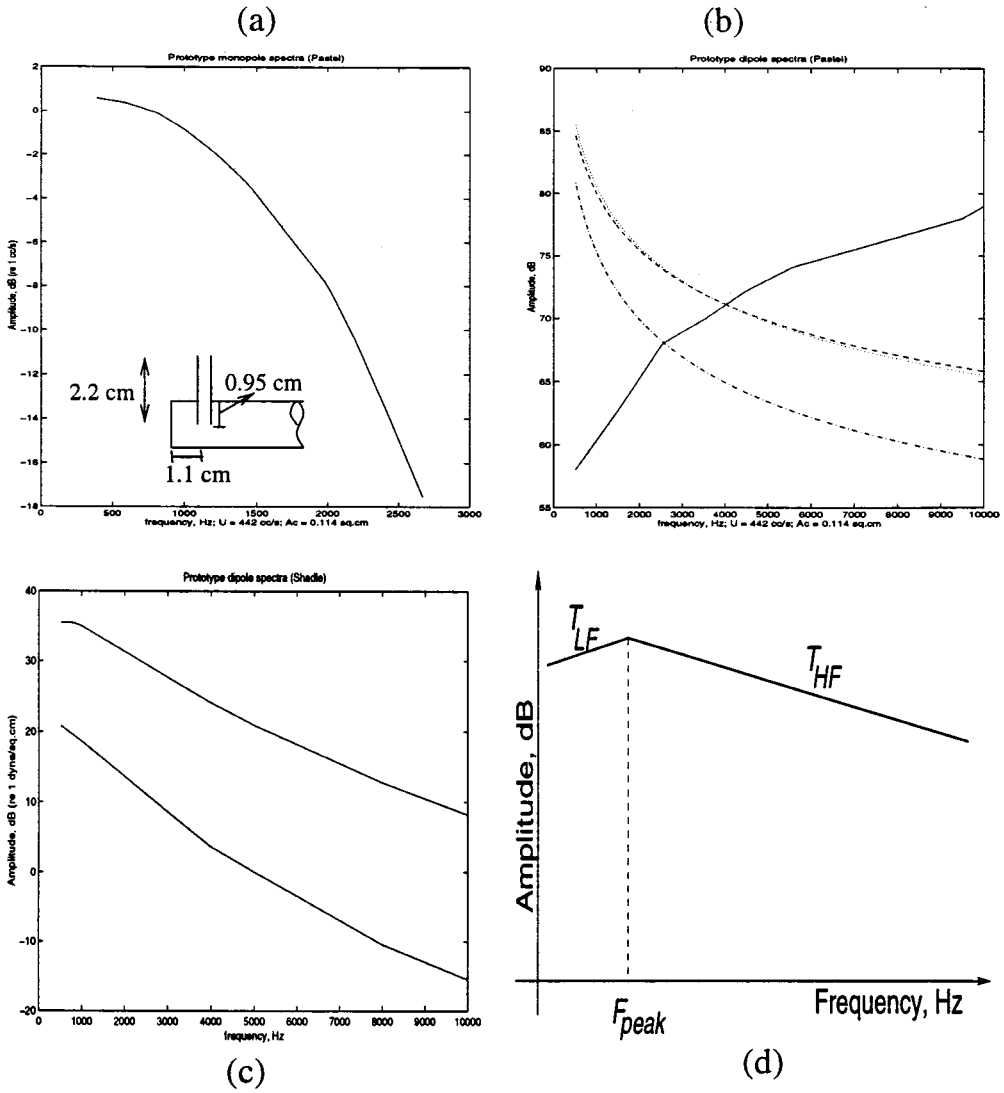


Fig. 3. (a) Spectrum of a monopole source estimated from a uniform tube ($l = 17$ cm, $a = 2.85$ cm^2) with a nozzle area of 0.114 cm^2 (after Pastel [17]). The inset shows the schematic of the experimental configuration with the jet inlet at 90° to the longitudinal axis of the main tube. (b) Dipole spectrum estimated from experiments on mechanical models (after Pastel [17]). The various curves shown correspond to different constriction to obstacle distances: 1 cm (solid), 2 cm (dashed), 3 cm (dotted), and 4 cm (dot-dashed). The flow velocity $U = 442$ cm^3 and the jet angle = 27°. (c) Dipole spectrum estimated from experiments on mechanical models (after Shadle [19]). The top curve corresponds to $U = 420$ cm^3/s and the bottom curve corresponds to $U = 160$ cm^3/s. (d) A stylized three-parameter model for the dipole spectrum. F_{peak} denotes the peak frequency while T_{LF} and T_{HF} denote the low-frequency and high-frequency spectral tilts, respectively.

The adaptation of experimentally derived source data to newer conditions (flow rates and constriction areas) requires adjustments in the magnitude levels of the baseline, or prototype, source spectra. Based on experimental evidence relating sound power for turbulent flows at a constriction of a given cross sectional area to the sixth power of air flow velocity, Stevens [25] states that the magnitude of the turbulent sound pressure source is proportional to $U^3 A_c^{-2.5}$, where U is the volume velocity and A_c is the constriction area. Assuming such a relation, the relative levels of the dipole source spectra can be scaled to reflect new U and A_c values. It should be noted that other recent experimental evidence indicate variations in the dependency relation between the sound pressure magnitude and U and A_c. Badin *et al.* [2] report a dependency of the source SPL on $U^2 A_c^{-1.6}$ based on aerodynamic and acoustic data obtained from sustained fricatives of one subject. In this paper, we will use the proportionality relation $U^3 A_c^{-2.5}$. Note that this assumption, however, does not account for the experimental evidence indicating decreased source strength for obstacles located farther away from the jet constriction [8], [17], [19].

III. Simulation Methodology

The spectrum of the radiated sound pressure specifying the output (fricative) speech is derived, in the frequency domain, according to the relation $P_{r,i}(\omega) = R(\omega) T_i(\omega) S_i(\omega)$. If the subscript "$i$" denotes a specific source location in the vocal tract, then $T_i(\omega)$ specifies the transfer function between a source (volume velocity U_i or sound pressure P_i) at the ith location and the volume velocity at the lips U_l. $S_i(\omega)$ represents the spectrum of the source at the ith location and $P_{r,i}(\omega)$ is the contribution to the radiated sound pressure due to this source at a distance of r cm in a far-field location. $R(\omega)$ is the relation of the sound pressure at (r, θ) to U_l. The magnitude

of $R(\omega)$ for frequencies less than 4 kHz is approximated by $|R(\omega)| \triangleq |P_r(\omega)/U_l(\omega)| = (\rho\omega/4\pi r)$ based on a point source approximation for the radiating surface at the lips. The dependence of $|R(\omega)|$ on ω yields a 6 dB/octave slope in the frequency response of the radiation transfer function. In the simulations, deviations from this approximation at high frequencies are assumed to be small and hence, are ignored. Notice that, for a given $T_i(\omega)$ and $S_i(\omega)$, the product $T_i(\omega)S_i(\omega)$ will always yield the corresponding volume velocity at the lips. Since the acoustic signal was measured in the far-field, the pressure signal at the microphone placed at a distance r cm from the lips is obtained by multiplying the volume velocity at the lips, $U_l(\omega)$, by the radiation transfer function $R(\omega)$.

The volume velocity $U_l(\omega)$ and the radiated pressure $P_l(\omega)$ at the lips are related by the radiation impedance $Z_l(\omega) \triangleq P_l(\omega)/U_l(\omega)$. A realistic model for $Z_l(\omega)$ can be obtained by approximating the radiation load at the mouth as a piston set in a spherical baffle [5]. Although the criterion of the radius of the spherical baffle (head) being greater than the piston radius (mouth) will be always met, Flanagan's two-term approximation for the piston-in-sphere model, which is primarily limited to low frequencies and/or small mouth openings will be violated, especially at high frequencies, for condition of $ka \ll 1$ ($k = \omega/c, a$ is the piston radius).

The boundary conditions at the glottis are specified by an appropriate subglottal pressure and a glottal impedance, comprising of a glottal resistance and a glottal inductance [(4) and (15)] [12]. The glottal impedance depends on the glottal dimensions: the nominal length and thickness of the glottis were assumed to be 1.4 cm and 0.3 cm, respectively [12]. The subglottal pressure was assumed to be 8 cm H_2O, corresponding to an open glottis condition, with the nominal areas of glottal opening assumed to be 1 cm^2 for the unvoiced fricatives and 0.5 cm^2 for the voiced ones.

If N input sources are involved, assuming superposition holds, we can express the output as $P_r(\omega) = R(\omega)\sum_{i=1}^{N} T_i(\omega)S_i(\omega)$. Such an approximation may be used to represent the effect of the *distributed* nature of the sound sources in terms of lumped, noninteracting entities. Furthermore, it allows us to consider the effect of different source types, such as the monopole and dipole sources, in the acoustic model.

In the current study, the simulation method of Maeda [12] was used to generate the vocal tract transfer functions $T_i(\omega)$, using the area functions obtained from MRI data [15]. The values of the various constants used in the simulations are given in Table I; the sampling frequency was 48 kHz. For each area function, a set of transfer functions $T_i(\omega)$, was derived for different source locations in the vocal tract.

1) Transfer functions between a series pressure source $P_i(\omega)$ and the output volume velocity at the lips $U_l(\omega)$, for different source locations in the cavity anterior (i.e., downstream) to the supraglottal constriction (*front* cavity).
2) Transfer function between a volume velocity source at the location of the maximum constriction U_c and the volume velocity at the lips U_l.

TABLE I
PHYSICAL CONSTANTS AND OTHER SIMULATION PARAMETERS

Symbol	Description	Value	Units
ρ	air density	1.14×10^{-3}	gm/cm^3
c	speed of sound in air	34480	cm/s
μ	viscosity coefficient	1.86×10^{-4}	dyne-s/cm^2
λ	heat conduction coefficient	5.5×10^{-5}	cal/cm-s-deg C
η	adiabatic constant of air	1.4	-
c_p	specific heat of air at constant pressure	0.24	cal/gm-deg C
R_w	wall resistance	1600	gm/s/cm^2
L_w	wall mass	1.5	gm/cm^2
C_w	wall compliance	3.0×10^5	s^2/gm/cm^2
l	length of a tube section	0.3	cm
S	circumference of tube	variable	cm
d	characteristic jet dimension	variable	cm
A	cross-sectional area of tube	variable	cm^2
A_c	area of maximum constriction	variable	cm^2
U	flow volume velocity	variable	cm^3/s

3) Transfer function between a volume velocity source at the glottis U_g and the volume velocity at the lips U_l.

Next, output spectra are derived using *hybrid* source models in conjunction with the calculated transfer functions. The term *hybrid* in the current context implies the use of a combination of various source *types* (such as monopole/dipole turbulent sources and a voice source) distributed at specific *locations* in the vocal tract where aerodynamic generation of sound is believed to occur. In some cases, multiple sources of the same type may be specified to approximate the distributed nature of the source therein. Prototype source models are derived from results of previous empirical and experimental studies. Specifically, source models considered here are those of Fant [4], Stevens [24], Shadle [19], and Pastel [17]. Based on simulation results with the prototype dipole sources, a parametric dipole source model is proposed. Finally, improved hybrid source models for fricative consonants using this three parameter dipole source model are derived based on seeking an optimal match between the synthesized and natural spectra. The $L2$ log-magnitude spectral distortion between the synthetic spectrum $P_{m,\alpha}(\omega)$ and natural spectrum $P_s(\omega)$

$$d_2(P_s, P_{m,\alpha})^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\log(P_s(\omega)) - \log(P_{m,\alpha}(\omega))|^2 \, d\omega \tag{1}$$

provides an objective performance measure for parameter optimization: $\hat{\alpha} = \arg\min_{\alpha \in S} d_2(P_s, P_{m,\alpha})$ where α is an index vector of source model parameters. The strategy adopted to investigate source modeling of fricatives in terms of source type, source levels, and spectral characteristics is described below.

1) Source Type, Location, and Number: Derivation of the source location and type is largely based on the articulatory and acoustic/aerodynamic theory of sound generation in fricatives. For a given place of articulation, these source model parameters are relatively fixed. Specifically, we consider the following.

1) A monopole source is specified at the constriction exit.
2) A voice source is specified at the glottis for voiced fricatives.

3) Dipole sources are specified in the vicinity of obstacles in the path of the air flow emerging from the constriction.

The obstacles are the teeth and vocal tract walls for the postalveolars, teeth for alveolars and lips for nonstridents. Moreover, since in reality such turbulent sources are distributed in nature, multiple, lumped noninteracting dipoles may be used as an approximation. Where multiple dipole sources were used in this paper, their number was limited to one wall source and two obstacle sources (specified in adjacent vocal tract sections). These approximations need to be revisited while modeling surface sources in uvular and pharyngeal fricatives.

2) Spectral Characteristics: Prototype spectral characteristics are adopted from previous theoretical and experimental studies (Section II-D). The following strategy was used.

1) The spectral characteristics of the monopole source were deemed "fixed" and results from [17] were used.
2) The spectral characteristics of the voice source were obtained from the glottal model given in Section II-D1. The fundamental frequency was subject dependent.
3) The dipole source spectral characteristics greatly influence the shaping of the fricative spectra over most of the frequency range.

First, the performance of prototype dipole source models (Fant, Shadle, Pastel) are investigated. Second, parametric dipole source models that provide a better match to the natural speech spectra are presented. The baseline parametric dipole spectrum is defined by three parameters. As illustrated in Fig. 3(d), the dipole spectrum is characterized by a broad peak at a frequency F_{peak} with a roll-off below this peak, specified by the tilt T_{LF}, and a roll-off above this peak, specified by the tilt T_{HF}.

Experimental results for a free jet spectrum have predicted a broad peak around $0.2V/d$ Hz (where V is the jet velocity at the constriction and d is the diameter of the constriction) [8], [24]. In fricative production, however, the jet emerging from a constriction often is influenced by the presence of an obstacle, such as the teeth, vocal tract walls, or the lips, in the jet's path. The extent of the obstacle's influence on the sound power generated depends to a great extent on the geometry involved, which varies depending on the place of articulation. Specifically, the peak frequencies of the spectra of jets impinging on an obstacle are likely to depend on the obstacle to constriction distance (l_o) and the obstacle size, in addition to the jet dimension and velocity. Unfortunately, experimental results on scaling relations for F_{peak} values are not readily available for cases involving obstacles. Based on the results for the characteristic oscillation frequency of an edge-tone configuration and her own experimental data for impinging jets, Shadle [19] suggested that the empirical scaling for F_{peak} perhaps lies between $0.2V/d$ Hz and $0.2V/l_o$ Hz. Results of Pastel [17] seem to indicate that higher values of F_{peak} are associated with smaller values of l_o. It is possible, however, that the presence of a relatively high-frequency zero in the transfer function, resulting from the small l_o, was not properly accounted for during inverse filtering used to obtain the experimental source spectra; in Pastel's study this may have caused the apparent high-frequency energy in the source spectra. The results of Fant [4] also suggest a similar idea (Section II-D1): The break frequency (frequency at which the slope of the spectrum changes, interpreted here as F_{peak}) for the "apical" source for /ʃ/ is 6 kHz compared to a "dental" source which is at 2 kHz (note that this apical source has a smaller l_o than the dental one). Nevertheless, currently there is no satisfactory scaling relation for F_{peak}. It is quite likely that the scaling relation would depend on both d and l_o values, V, and the dimensions of the obstacle. Determination of a scaling relation that incorporates these factors is not possible with the available data. Hence, in the current study, F_{peak} values were empirically adjusted as described below.

The current study relies on selecting the values of the parameters F_{peak}, T_{LF}, and T_{HF} using an analysis-by-synthesis parameter optimization scheme. The (broad) peak frequency of the dipole spectrum is chosen in accordance to the relation $F_{peak} = KU/A_c d$, with the scaling parameter K being adjustable. Although experiments on mechanical models suggest that K lies between 0.1 and 0.2 for speech-like configurations [21], [24], the current modeling allowed $0 < K < 0.5$ (in steps of 0.05). The characteristic dimension d was $\sqrt{4A_c/\pi}$. The range for low- and high-frequency tilts used for the optimization was 0–24 dB/oct (minimum step size 1 dB/oct). The source parameter set that provided the smallest log-spectral distortion was deemed optimal.

3) Relative Source Levels: The exact nature of scaling relations between the amplitude levels of various source types (monopole, dipole, voice) and the aerodynamic state of the vocal tract (in terms of pressure, volume velocity and dimensions) is not well understood. Furthermore, the dipole source strength value is known to vary depending on factors such as the obstacle-constriction distance and the configuration of jet impingement [19], [21], [25]. This problem is further complicated by the complex geometry of the vocal tract enclosing the noise sources and inter/intra-speaker articulatory and aerodynamic variability.

Approximate guidelines for adjusting relative levels of the three source types are given in [25]. For example, the maximum strength of the voice source for voiced fricatives is estimated to be greater than the maximum turbulent dipole strength by 5 to 25 dB [26]. These estimates assume that the voice source amplitude is proportional to $\Delta P_g^{1.5}$, where ΔP_g is the transglottal pressure, and the noise source amplitude is proportional to $\Delta P_c^{1.5} A_c^{0.5}$ where ΔP_c is the pressure drop across the supraglottal constriction with area A_c. For subglottal pressure of 8-cm H_2O and A_c, between 0.1 and 0.3 cm^2, ΔP_c is between 4 and 6-cm H_2O. Similarly, monopole strength is assumed to be 10–15 dB lower than the maximum dipole strength in strident fricatives [17].

In our first set of simulations with the various prototype dipole source models, the relative levels of the different source types were kept fixed in accordance to the guidelines outlined above. Based on the results of these experiments, it was determined that the relative source levels need to be varied in order to provide a better match with the natural spectra. Hence, in our second simulations with the parametric dipole models relative source levels (of monopole and voice source spectra peaks with respect to dipole spectrum peak) were also included as variable parameters in the optimization scheme. Knowledge

from previous studies provided rough guidelines in specifying constraints on the range of values used for the optimization.

IV. Simulation Results

In this section, results of modeling fricative speech spectra based on a source-filter approach are presented. The simulations use acoustic data and MRI-derived area functions from a male (MI) and a female (PK) subject.

A. Vocal Tract Transfer Functions

Vocal tract transfer functions were obtained using a finite time-difference simulation of acoustic propagation [12]. Both volume velocity and pressure sources could be specified in any section of the vocal tract. Here, we consider two issues.

1) The influence of the constriction-to-obstacle distance (l_o) on the transfer function [1]. We illustrate this effect of varying l_o on the transfer function of the stridents /**s**/ and /ʃ/ of MI.
2) The effect of including a sublingual cavity on the vocal tract transfer function. This effect is illustrated using MI's /ʃ/ data.

1) Effects of Changing Dipole Source Location: Since the location of the dipole source in the vocal tract, relative to the constriction and the obstacles, is crucial in fricative production [1], [19], it is instructive to study the effect of changing the dipole source location on the vocal tract transfer functions. Fig. 4(a) shows transfer functions corresponding to various dipole source locations between the teeth and the constriction for MI's /**s**/ (the transfer functions are spaced by 5 dB for clarity). The most striking effect is the change in the frequency of the zero from about 2000 Hz (source near the teeth, 0.9 cm from lips) to about 3700 Hz (source anterior to oral constriction, 2.1 cm). This zero is attributed to the free zero arising from the cavity between the source and the oral constriction [1]. Additional zeros in the high-frequency end of the spectrum can be observed for particularly large constriction-to-source distances such as for the source located in the vicinity of the teeth. Similar effects were observed in the high-frequency region of MI's /ʃ/ transfer functions which can also be attributed to changing the dipole source location [Fig. 4(b)]. However, in this case, there is an additional influence from the sublingual cavity, as will be further explained in the following section.

2) Effects of the Sublingual Cavity: The simulations were modified to allow for the inclusion of sublingual cavities; sublingual cavities were observed in postalveolar fricatives and some labiodental fricatives [14]. A parallel side-branch coupled to the main oral tract at one end and terminated by a "hard wall" at the other, was used to specify the sublingual cavity in the anterior part of the vocal tract. In Fig. 4(b), transfer functions calculated for dipole sources located at four different vocal tract sections (0.9, 1.2, 1.5, and 1.8 cm from the lip opening) for the cases with and without a sublingual branch are shown for MI's /ʃ/. The sublingual branch is about 1.2 cm long and is coupled to the oral cavity at 2.1 cm from the lip opening. A 100–350 Hz decrease in the frequency of the lowest front cavity resonance is observed when the sublingual branch is added (the smaller decreases are for larger values of l_o). This effect is expected since there is an increase in the effective length of the cavity anterior to the oral constriction. The shifts in the other pole frequencies of the transfer function as a result of the additional sublingual branch are relatively small. As the source location is moved away from the constriction, changes in the zeros at the high-frequency end of the spectrum (>6 kHz) can be observed. The addition of the sublingual branch lowers the frequency of the free zero arising from the cavity between the constriction and the source, with the effect most prominent for locations farther away from the constriction (a drop of about 2 kHz is observed for the dipole located near the teeth). There is also a less prominent zero (around 10 kHz) whose exact cavity affiliation is unclear.

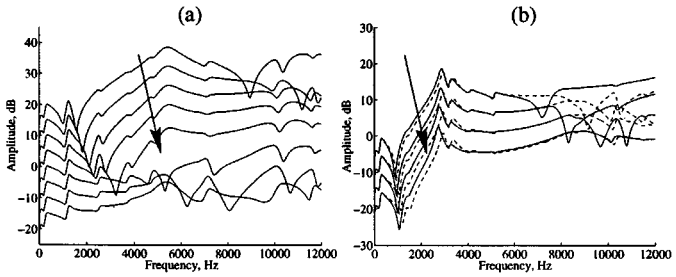


Fig. 4. (a) Effect of varying the dipole source location on the transfer function for subject MI's /**s**/. Arrow points to location changes away from the lips and toward the oral constriction (in steps of 0.3 cm). (b) The effect of including a sublingual cavity on the transfer functions for subject MI's /ʃ/. Transfer functions shown correspond to dipole sources placed at four contiguous vocal tract sections, 0.3 cm apart, starting at the vicinity of the teeth (0.9 cm from the lip opening) and moved away from the lips (direction of the arrow): dashed lines (without sublingual branch), solid lines (with a sublingual branch of 1.2 cm).

B. Investigation of Source Models

A two-step process was taken in investigating the source models: first, the best approximations to noise source models were made using the available aeroacoustic modeling data, and these models were used with transfer functions derived from MRI data to predict speech spectra; second, the source parameters were then optimized to produce the best fit of predicted to measured natural speech spectra. This relegates to the second stage the process of independently varying parameters that were not actually physically independent. The main result of the first step was thus a measure of the suitability of existing experimental data for synthesis. The main result of the second step was a set of source models that provide the best fits possible for these transfer functions and these two subjects.

Optimization experiments for the second step mentioned above were done in two stages. First, simulations with prototype dipole source models wherein the number of dipole sources and their locations were selected by parameter optimization. Second, simulations with parametric dipole spectral models wherein the optimization parameters included the three parameters of the stylized dipole spectrum (F_{peak}, T_{LF}, and T_{TF}), the dipole source location, and the relative source levels. Note that results from the first set of simulations were used in the second set of simulations. Multiparameter optimization was done as explained in Section III.

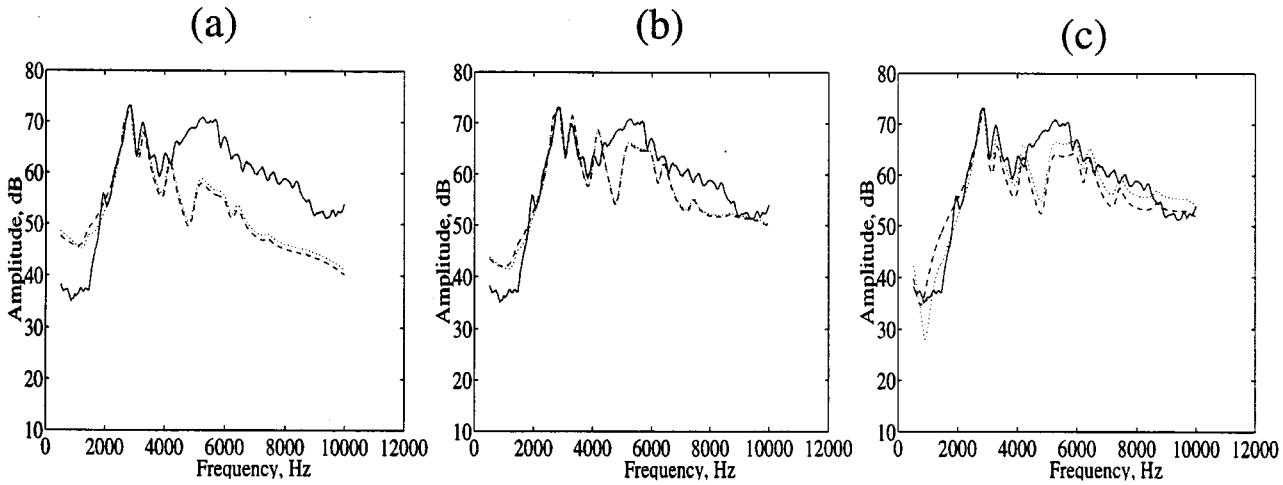


Fig. 5. Model output spectra with prototype dipole source models for subject MI's /ʃ/: (a) Shadle, (b) Fant, and (c) Pastel. Monopole spectrum is from Pastel's data. Dipole sources only (dotted), in conjunction with a constriction monopole (dashed). Dipole sources were located as indicated in Table IV. Spectrum of a natural /ʃ/ is shown (solid) for comparison.

1) Strident Fricatives (Postalveolar Fricatives) /ʃ/ and /ʒ/ : A set of transfer functions corresponding to dipole sources located at different sections anterior (i.e., downstream) to the minimum constriction was first calculated. In addition, a transfer function for a monopole source at the constriction was calculated and used in conjunction with the monopole source spectrum from Pastel [17]. A nominal volume velocity of 450 cm^3/s was used in the simulations. Recall that F_{peak} of the dipole spectrum is dependent on the volume velocity. The transfer functions for dipole sources located in the vicinity of the teeth are found to provide the closest match to the overall spectra, particularly below 5 kHz. The effect of the dipole sources at the wall obstacle (in the vicinity of the constriction) accounts for frequency characteristics between 6 and 8 kHz. A wall-type obstacle is reasonable for a postalveolar strident wherein the anterior tongue body is significantly raised, and "domed" in shape [14]. This tongue shape causes the turbulent flow to impinge on the upper region of the vocal tract and the teeth. The spectral characteristics and relative levels of the dipoles at the wall and teeth are, however, likely to be different [17], [19]. The hybrid model for an unvoiced postalveolar strident uses a monopole at the constriction exit, and dipoles in the immediate vicinity of the constriction (wall-obstacle) and at the teeth (teeth-obstacle). Preliminary simulations with the prototype source models indicated the optimal source locations for MI and PK.

Results using the dipole characteristics of Shadle are shown in Fig. 5(a) for MI. For this simulation, the dipole spectral characteristics at the different source locations were assumed to be identical. The overall spectral shapes of the output do not match the natural spectra well, particularly in the region above 5 kHz. Compare these to the results obtained using Fant and Pastel's models [Fig. 5(b) and (c)] wherein spectral characteristics for the wall and teeth obstacles were assumed to be distinct. The results for Fant's model used the "apical" source spectrum for the wall-obstacle dipole and the "dental" source spectrum for the teeth dipole (Section II-D1). For Pastel's case, the spectra corresponding to $l_o = 1$ cm and 2 cm [Fig. 3(b)] were used for the wall-obstacle dipole and teeth obstacle dipoles, respectively (Section II-D2). Note that a smaller T_{HF} for the wall source, i.e., giving it more high-frequency energy, results in a better fit to the spectrum between 5 and 7 kHz. In fact, the experimental results of both Shadle [19] and Pastel [17] indicate that spectra of sources appearing closer to the origination of jet—such as the wall obstacle source here—tend to have a relatively greater high-frequency power than those that are located farther away. In summary, the source spectrum shape is influenced by the obstacle-to-constriction distance. It should be noted that l_o values also influence the frequency of the "free" zero in the transfer function (Fig. 4) which in turn influences the peak spectral amplitude [1]. In the low-frequency region ($<$2.5 kHz), the match is further improved when the low-frequency spectral tilt stays flat or has a gentle roll-off [compare Fig. 5(a) to Fig. 5(b)]. Then, any influence of the monopole source also becomes more apparent (Fig. 5(c) illustrates this effect although similar results for subject PK, not shown, were somewhat more pronounced). However, to fully investigate the extent of the monopole source's influence, it was deemed necessary that relative source levels should also be varied.

The log-spectral distortions for the experiments with the prototype dipole source models are summarized in Table II. These results, in general, indicated the following.

1) High-frequency tilts of the dipole spectrum (in general, $>$2.5 kHz) need to be different, particularly across sounds and subjects.
2) Low-frequency tilt of the dipole spectrum (in general, $<$2.5 kHz) also needs to be adjusted, although to a lesser degree than the high-frequency tilt.
3) Relative levels of the various source spectra need to be adjusted.

With insights obtained from simulations using prototype dipole models, an improved hybrid source model was developed. The spectral characteristics of the dipole sources and their relative levels were used as parameters in an optimization scheme in order to obtain a better match with the natural spectra (in the sense of minimizing the $L2$ log-spectral distortion). The corresponding distortion values obtained by the optimization are

TABLE II
SUMMARY OF LOG SPECTRAL DISTORTIONS (IN DECIBELS) OBTAINED FOR THE PROTOTYPE MODELS AND THE PARAMETRIC DIPOLE MODEL. SOURCE LOCATION AND NUMBER USED IN THE SIMULATIONS ARE LISTED IN TABLES IV–VII. DIPOLE SOURCE SPECTRAL CHARACTERISTICS AND RELATIVE SOURCE LEVELS WERE FIXED FOR ALL SIMULATIONS WITH THE PROTOTYPE MODELS, WHILE THEY WERE CHOSEN AUTOMATICALLY BY THE ANALYSIS-BY-SYNTHESIS SCHEME FOR THE PARAMETRIC MODEL. SEE SECTIONS II AND III FOR FURTHER DETAILS

SUBJECT	Fricative	Fant	Shadle	Pastel	Parametric
MI	/ʃ/	28	57	22	19
	/ʒ/	36	61	47	28
PK	/ʃ/	66	38	99	26
	/ʒ/	60	39	83	28
MI	/s/	41	32	40	24
	/z/	49	60	76	25
PK	/s/	38	64	48	24
	/z/	44	40	43	28
MI	/θ/	40	47	41	23
	/ð/	35	43	42	25
PK	/θ/	26	55	23	23
	/ð/	32	26	34	24
MI	/f/	27	23	26	13
	/v/	33	30	23	20
PK	/f/	41	79	46	37
	/v/	26	31	23	23

TABLE III
SUMMARY OF SPECTRAL CHARACTERISTICS OF PARAMETRIC DIPOLE SOURCE MODELS FOR SUBJECTS MI AND PK. THE PARAMETERS F_{peak}, T_{LF}, AND T_{HF} ARE DEFINED IN FIG. 3(d). THE SCALING FACTOR K FOR THE PEAK FREQUENCY IS WITH RESPECT TO U/dA_c. SEE TEXT FOR FURTHER DETAILS

SUBJECT	Fricative	F_{peak} Hz	Scaling K	T_{LF} dB/oct	T_{HF} dB/oct
MI	/ʃ/ (wall)	4056	0.2	2	0
	(teeth)	2128	0.1	1	-8
	/ʒ/ (wall)	2841	0.2	0	-3
	(teeth)	1420	0.1	0	-3
PK	/ʃ/ (wall)	3900	0.15	0	-14
	(teeth)	2600	0.1	0	-16
	/ʒ/ (wall)	2250	0.15	0	-14
	(teeth)	1500	0.1	0	-16
MI	/s/	2650	0.2	3	-12
	/z/	1957	0.2	3	-16
PK	/s/	4622	0.2	1	0
	/z/	2475	0.2	0	0
MI	/θ/	4066	0.4	1	-14
	/ð/	3419	0.4	2	-8
PK	/θ/	3270	0.2	0	-6
	/ð/	1810	0.2	0	-8
MI	/f/	2807	0.2	3	-12
	/v/	2590	0.3	3	-12
PK	/f/	3923	0.2	0	-4
	/v/	3363	0.2	0	-10

listed in Table II. Details regarding the hybrid source models thus derived are given in Table III and the dipole spectral characteristics, in Table IV.

The synthesized /ʃ/ spectra for MI and PK are shown in Fig. 6(a) and (c), respectively. A good match between the synthesized and natural spectra is found both in the overall shape and the values of the significant resonances. Significant reduction in distortion was obtained for both subjects compared to simulations with prototype dipole models. The peak of the dipole spectra for the teeth-obstacle case was higher than suggested by previous models. The peak of the wall-obstacle spectra was lower than that found in [17]. The estimated low-frequency spectral tilts indicated a flat response, except for MI's wall-dipole spectrum, similar to Fant's prototype model. The high-frequency tilts, while in general comparable to previous experimental data, appear to be speaker dependent: T_{HF} values for PK are much higher than for MI. The maximum level of the wall-obstacle dipole had to be maintained about 5 dB higher than that of the teeth-obstacle. The contribution from the monopole component was found to be relatively insignificant for MI while it resulted in minor improvements for PK.

The strategy adopted for modeling /ʒ/ was similar to /ʃ/ except for the inclusion of a voice source at the glottis. The fundamental frequency used in the simulations was 127 Hz for MI and 172 Hz for PK, and the nominal value of U was 337.5 cm^3/s which is 25% less than that used for the unvoiced cases (note that flow rates for voiced fricatives are typically lower than the unvoiced ones). The agreement of the model's output spectrum with the natural spectra, in general, is very good [Fig. 6(b) and (d)]. Voicing effects dominate in the low-frequency region (below 800 Hz). The influence of the monopole source in the region between 800 Hz and 3500 Hz results in a slight improvement in the spectral match. The high-frequency match for subject MI is relatively poor compared to that of PK.

2) Alveolar Fricatives: /s/ and /z/: The output spectrum for /s/ was a result of two identical dipoles in the vicinity of the teeth and one monopole at the constriction exit (Table V). Among the prototype models, Shadle's model provided the best match in the high-frequency region for /s/ of both subjects. At low frequencies, on the other hand, the best results were provided by Fant's model. Shadle's and Pastel's models tend to overempha-

TABLE IV
HYBRID SOURCE MODELS FOR POSTALVEOLAR FRICATIVES

Fricative	/ʃ/			/ʒ/			
Source type	*monopole*	*dipole*	*dipole*	*monopole*	*dipole*	*dipole*	*voicing*
SUBJECT MI							
Number	1	1	2	1	1	2	1
Location	constriction	wall	teeth	constriction	wall	teeth	glottis
Distance from lips (cm)	2.7	2.4	0.9,1.2	2.7	2.4	0.9, 1.2	17.7
Max. levels (dB) (rel. to monopole max)	0	25	20	0	25	20	5
SUBJECT PK							
Number	1	1	2	1	1	2	1
Location	constriction	wall	teeth	constriction	wall	teeth	glottis
Distance from lips (cm)	1.8	1.8	0.9,1.2	1.8	1.8	0.9,1.2	15.9
Relative Levels (dB) (rel. to monopole max)	0	23	17	0	26.6	17	53

size the low-frequency region of the synthesized spectra, especially in PK (note that Fant's spectral model is flat in the frequency region between 0.8–4 kHz). These results suggest that the peak of the dipole spectrum is likely to be somewhat higher than suggested by these two models. Further, these results indicate that the monopole source influence could be more effective if its level relative to those of the dipole sources was adjusted properly (rather than holding it fixed). Next, improved hybrid source models were derived in conjunction with parametric dipole source models. Details are given in Tables III and V and the corresponding output spectra are shown in Fig. 6(e) and (g). Although the general agreement in the spectral shapes of the model's output with that of natural speech is good for both subjects, there are slight discrepancies in the resonance frequency values below 4 kHz. A small effect of the monopole source can be noticed in the spectra below 3 kHz.

The results for /**z**/ [Fig. 6(f) and (h)] also show good agreement with the overall natural spectra. However, in the frequency region around 1 kHz where there is a "transition" in the dominance of the dipole source over the voice source, the match is poor for PK. This, perhaps, reflects the inadequacy in assuming a simple superposition of the voicing and turbulence sources. The optimal source model parameters for /**s**/ and /**z**/ are similar. It should be, however, noted that the spectral tilts predicted by the analysis-synthesis scheme for subjects MI and PK are vastly different: the dipole source spectrum is relatively flat for PK, while T_{HF} was relatively high for MI. Further, the results for PK's /**s**/ and /**z**/ being somewhat worse than for MI, especially in the low-frequency region (<3 kHz), is perhaps due to greater mismatches in vocal tract conditions between MRI data gathering and acoustic recordings.

*3) Nonstrident Fricatives (Interdental Fricatives: /**θ**/ and /ð/) :* Examination of the transfer functions for dipole sources located in the teeth/lip region (equivalent to using a dipole source with white-noise spectrum) showed only moderate agreement with the poles and zeros of the natural speech spectra. The acoustic theory of speech production predicts that most of the poles and zeros seen in the 0–10 kHz frequency range can be attributed to the cavity posterior to the supraglottal constriction. The back cavity poles and zeros are, however, only approximately cancelled due to finite coupling between the front and back cavities and the front cavity source location. These resonances are not significant and are variable across the two subjects (due to greater articulatory variability in the nonstrident fricatives such as /**θ**/). Moreover the dynamic range of the amplitudes of these resonances in the natural spectra of /**θ**/ is within 20 dB.

The hybrid source model used for /**θ**/ includes a monopole source at the constriction and a single dipole source at the teeth (Tables III and VI). Simulations using the prototype dipole spectral models indicated that while models proposed by Fant (the one suggested for /**f**/ was used here) and Pastel were inadequate in capturing the high-frequency region of MI's interdentals, Shadle's models tend to somewhat overemphasize the low-frequency region in both subjects.

The output spectra using parametric dipole source models are shown in Fig. 7(a)–(d) for the interdental fricatives. Details of the parametric dipole sources are summarized in Table III and those of the hybrid source models in Table VI. Although the estimated T_{HF} values for both subjects are in the range of the prototype models, the F_{peak} scaling required for MI is particularly high. The modeling details for /ð/ were similar to /**θ**/ except for the inclusion of the voicing source at the glottis. It is clear that the match between the model's output spectra and natural spectra is relatively poor when compared to the strident fricatives' results especially in the spectral peaks rather than in the overall spectral shapes. Several reasons could account for these results. First, unlike strident fricatives, the spectral structure for a nonstrident fricative such as /**θ**/ (and even more so for the labiodentals such as /**f**/ discussed in the next section) does not possess characteristic peaks of significant spectral energy and is subject to greater intra- and inter-speaker variability. Hence, finding a match for such spectra is prone to be difficult and inexact. Moreover, limitations in the vocal tract simulation, including errors due to the limited spatial resolution with which the area functions are represented (the resolution

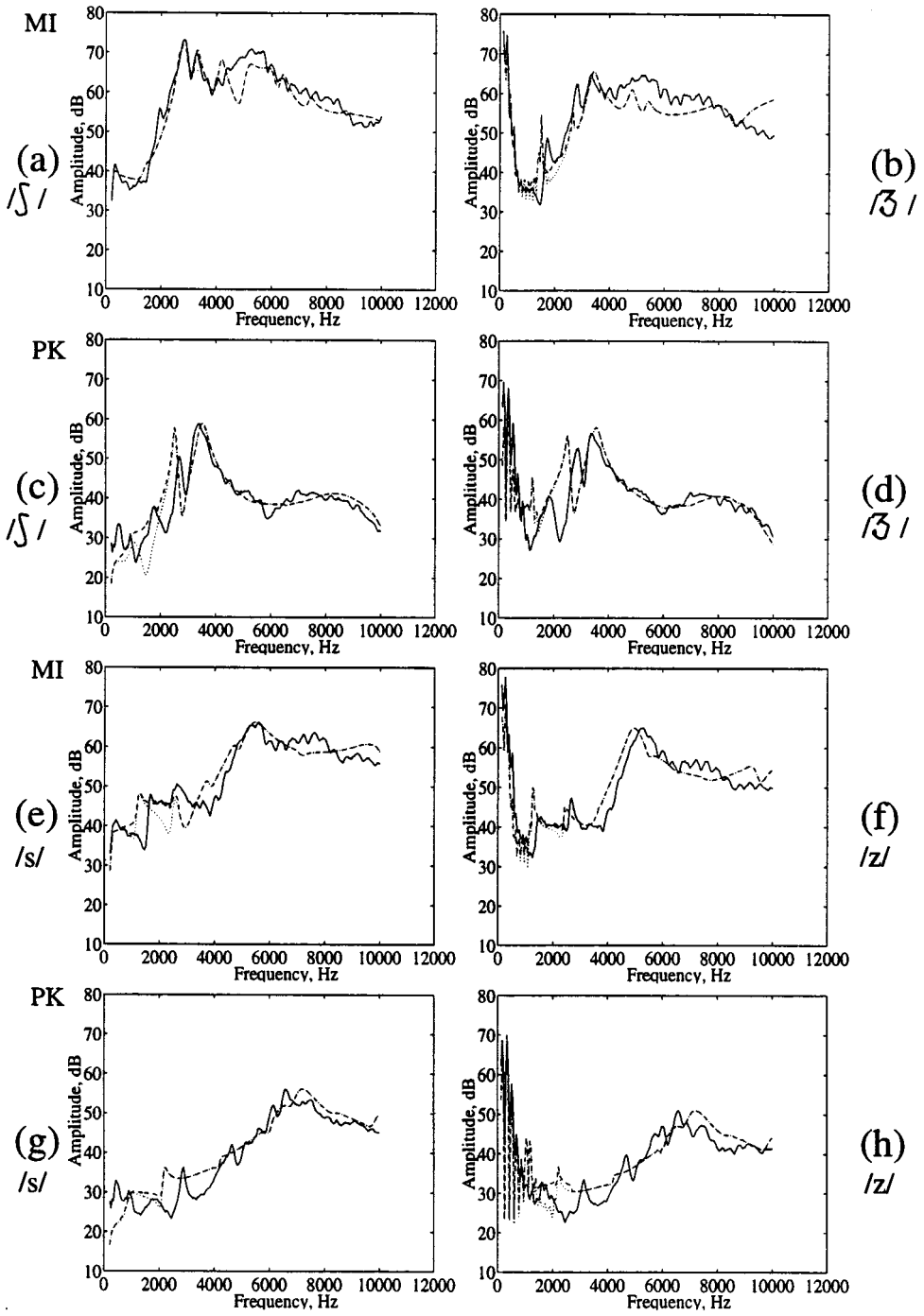


Fig. 6. Strident fricative spectra derived from optimal hybrid source model inputs using the parametric dipole spectra listed in Table II. For subject MI: (a) /ʃ/ (b) /ʒ/ (e) /**s**/ (f) /**z**/. For subject PK: (c) /ʃ/ (d) /ʒ/ (g) /**s**/ (h) /**z**/. The input source type and location are given in Tables IV and V for the postalveolar and the alveolar fricatives, respectively. Output spectrum from the model as a result of superposition of outputs from different sources: Dotted line—dipole sources only (/ʃ, **s**/) or dipole source plus voice source (/ʒ, **z**/), Dashed line—with an additional monopole source. The natural fricative spectrum is shown in each of the panels for comparison (solid).

effect becoming more appreciable for smaller front-cavity dimensions) are likely to affect the estimated transfer functions. In particular, the transfer functions (and possibly source functions) are more sensitive to lip shape, which cannot be measured any more accurately than for the stridents, and perhaps is not captured adequately by a 1-D sound propagation model.

4) Labiodental Fricatives: */***f***/ and /***v***/:* The results for MI and PK are shown in Fig. 7(e)–(h) while the details of the parametric dipole models and the hybrid source models are summarized in Tables III and VII, respectively. As expected, the results for the labiodentals are also poor, especially for PK. The same explanations offered for the interdentals hold: The articulatory and acoustic variabilities in labiodentals are quite high. The tongue is unconstrained, making articulatory variations more likely. The natural spectrum of the labiodental fricatives is relatively flat (dynamic range of about 15 dB) in the frequency range considered (below 10 kHz) characterized only by incompletely canceled back cavity resonances, depending on the degree of

TABLE V
HYBRID SOURCE MODELS FOR ALVEOLAR FRICATIVES

Fricative	/s/		/z/		
Source type	*monopole*	*dipole*	*monopole*	*dipole*	*voicing*
SUBJECT MI					
Number	1	2	1	2	1
Location	constriction	teeth	constriction	teeth	glottis
Distance from lips (cm)	1.8	0.9, 1.2	1.8	0.9, 1.2	17.7
Maximum levels (dB) (rel. to monopole max.)	0	30	0	30	15
SUBJECT PK					
Number	1	2	1	2	1
Location	constriction	teeth	constriction	teeth	glottis
Distance from lips (cm)	1.5	0.9, 1.2	1.5	0.9, 1.2	15.9
Maximum Levels (dB) (rel. to monopole max.)	0	35	0	35	20

TABLE VI
HYBRID SOURCE MODELS FOR INTERDENTAL FRICATIVES

Fricative	/θ/		/ð/		
Source type	*monopole*	*dipole*	*monopole*	*dipole*	*voicing*
SUBJECT MI					
Number	1	1	1	1	1
Location	constriction	lips	constriction	lips	glottis
Distance from lips (cm)	0.9	0.9	0.9	0.9	17.7
Maximum Levels (dB) (rel. to monopole max.)	0	35	0	35	15
SUBJECT PK					
Number	1	1	1	2	1
Location	constriction	lips	constriction	lips	glottis
Distance from lips (cm)	0.9	0.9	0.9	0.9	15.9
Maximum Levels (dB) (rel. to monopole max.)	0	12	0	35	20

coupling between the back and front cavities, a variable factor. Hence, to begin with, the spectral details of the natural speech template, to which the model seeks a match, is inherently variable.

V. SUMMARY AND DISCUSSION

In this paper, the results of acoustic modeling of sustained unvoiced and voiced fricatives are presented. The modeling is based on the source-filter theory of speech production wherein the vocal tract is specified as a concatenation of cylindrical tubes and assuming planar wave propagation. The areas for these cylindrical sections were derived from MRI data obtained from human subjects during the production of sustained fricatives. A time-difference method of simulating linear acoustic propagation in the tract, based on a program by Maeda [12], was used to derive the vocal tract transfer functions.

The modeling was performed in the frequency domain based on linear systems theory. This enabled the use of the superposition principle, as a result of which the output (speech spectrum) due to multiple sources exciting the vocal tract could be calculated with ease. A hybrid source modeling strategy was adopted and the sound sources occurring in the tract were assumed to be independent. The sound generation in the vocal tract due to turbulence effects in the cavity anterior to the supraglottal constriction was modeled in terms of a combination of flow monopole (at the constriction) and dipole sources (in the vicinity of an obstacle to the airflow). Furthermore, multiple dipole sources (at most two) were used to approximate the distributed nature of the source. A voice source at the glottis was included for voiced fricatives. The source type, number of sources, their location in the vocal tract, and their relative levels and spectral characteristics were the variables involved in the modeling. The first set of simulations

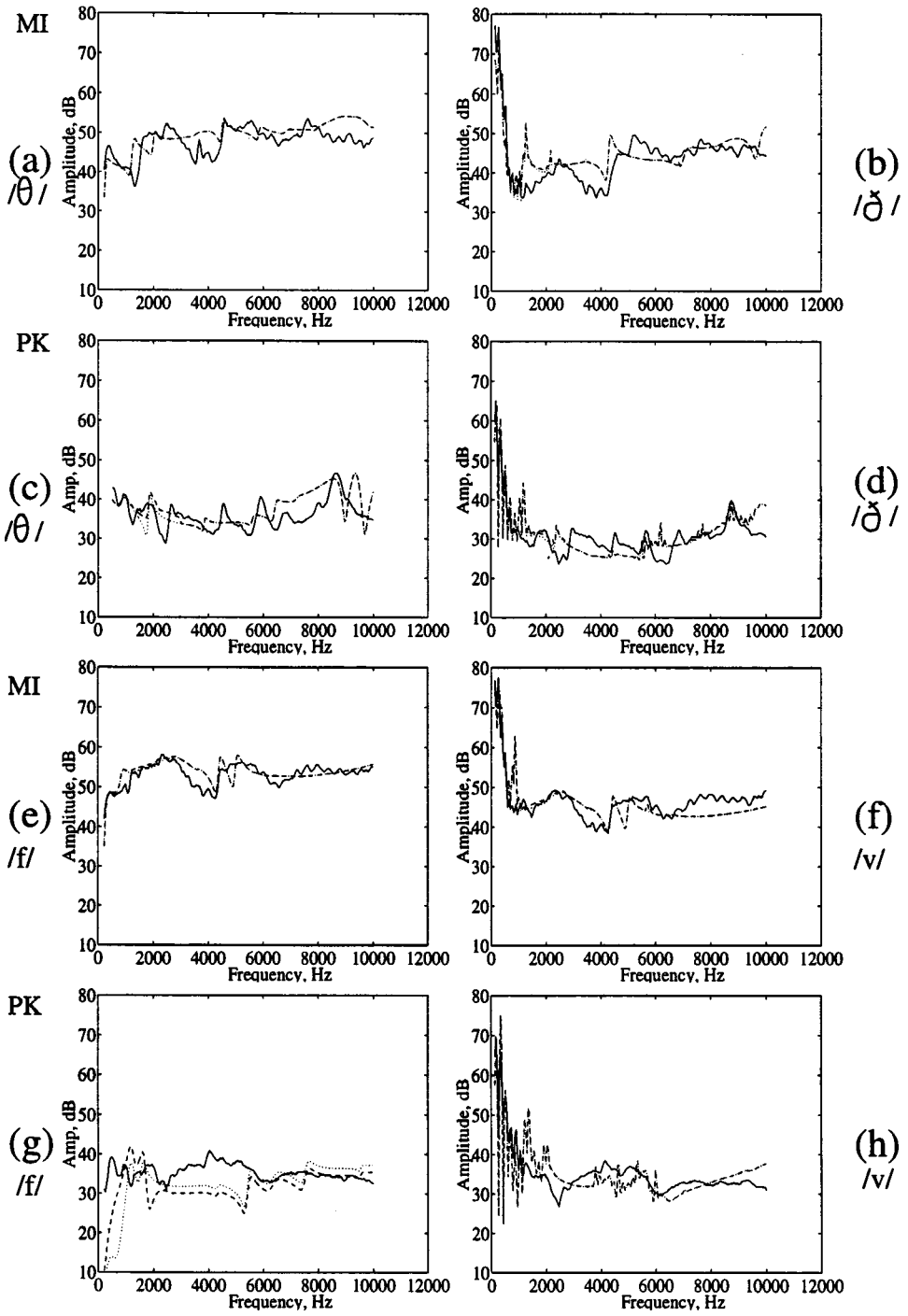


Fig. 7. Nonstrident fricative spectra derived from optimal hybrid source model inputs using the parametric dipole spectra listed in Table II. For subject MI: (a) /θ/ (b) /ð/ (e) /**f**/ (f) /**v**/. For subject PK: (c) /θ/ (d) /ð/ (g) /**f**/ (h) /**v**/.The input source type and location are given in Tables VI and VII for the interdental and the labiodental fricatives, respectively. Output spectrum from the model as a result of superposition of outputs from different sources: Dotted—dipole sources only (/θ, **f**/) or dipole source plus voice source (/ð, **v**/), Dashed—with an additional monopole source. The natural fricative spectrum is shown in each of the panels for comparison (solid).

investigated the effect of the baseline, or prototype, dipole source characteristics obtained from published experimental and empirical studies. The source location and number were the parameters used in the optimization to minimize the $L2$ log-spectral distortion between the synthesized and natural speech spectra. It was found that dipole sources played a major role in determining the overall fricative spectra whereas the monopole source's effect was minimal. Hence, further improvement of source models concentrated on optimizing the characteristics of a parameterized dipole spectral model. A simple three-parameter model for the dipole spectra was proposed based on both the results of previous aerodynamic studies on sources of turbulence generation and results of our first set of simulations. In addition to source location and number, the optimization parameters for the second set of simulations included the three dipole spectral parameters and

TABLE VII
HYBRID SOURCE MODELS FOR LABIODENTAL FRICATIVES

Fricative	/f/		/v/		
Source type	*monopole*	*dipole*	*monopole*	*dipole*	*voicing*
SUBJECT MI					
Number	1	2	1	2	1
Location	constriction	lips	constriction	lips	glottis
Distance from lips (cm)	0.6	0.6	0.6	0.6	17.7
Maximum Levels (dB) (rel. monopole max.)	0	25	0	25	10
SUBJECT PK					
Number	1	2	1	2	1
Location	constriction	lips	constriction	lips	glottis
Distance from lips (cm)	0.6	0.6	0.6	0.6	15.9
Maximum Levels (dB) (rel. monopole max.)	0	20	0	20	10

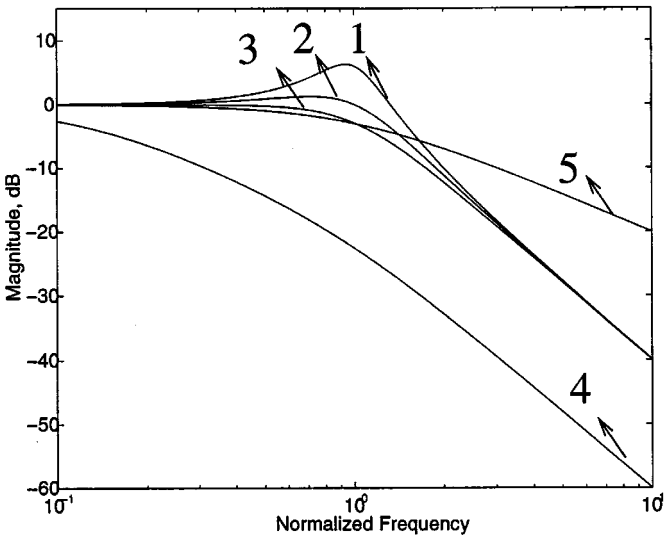


Fig. 8. Magnitude of a second-order transfer function: (1 and 2) underdamped case, −12 dB/oct; 1 represents a system with smaller damping factor than 2; 3: critically damped; 4 same as 3, but with a smaller break frequency; and 5: first-order case, −6 dB/oct slope.

the relative source levels. The simulation results demonstrated fairly good agreement between natural and synthetic spectra.

Improvements in hybrid models entailed adjusting the peak frequency and the high- and low-frequency tilts of the dipole source spectrum. First, the peak values, F_{peak}, are considered. The stylized dipole spectra assume a value of $KU/(A_c d)$, where the characteristic dimension was chosen to be $\sqrt{4A_c/\pi}$. The value of K was adjusted to provide an improved match to the natural spectra. The K values obtained through simulation were between 0.1–0.2 for the stridents and 0.2–0.4 for the nonstridents. For stridents, the scaling required for obstacles closer to the jet outlet (i.e., the minimum constriction) is much higher than for obstacles located farther away from it. For example, for /ʃ/ and /ʒ/, the dipole source at the vocal tract wall in the constrictions vicinity has a higher value of K than for the source located near the teeth. Similarly, the value of F_{peak} for the dipole near the teeth for /**s**/ is higher than that for the dipole at a similar location for /ʃ/. For nonstrident fricatives, the lip surface which acts as an obstacle is located relatively close to where the jet originates, and the values of K are higher. It should, however, be noted that although some strident fricatives have comparable obstacle-to-constriction distances (l_o) to those of nonstridents (e.g., wall-obstacle source in postalveolars versus lip-obstacle in interdentals), the F_{peak} values in the dipole spectra for the strident fricatives are higher than in nonstridents suggesting the influence of other factors such as the obstacle dimensions and angle of jet impingement in determining the source spectral characteristics.

The values of the tilt T_{HF} were between 0 and −16 dB/oct for both strident and nonstrident fricatives. There is a significant variability in the T_{HF} values of the two subjects. In general, T_{HF} values for PK tend to be lower than MI's, except for the postalveolars. It is not clear why the T_{HF} values are especially high for PK's postalveolars and especially low for PK's alveolars. For postalveolars, the empirical value for the high-frequency tilt suggested by Fant (i.e., −12 dB/oct) is within the range of values obtained in this study. For alveolars, the −6 dB/oct high-frequency roll-off suggested by Fant is somewhat lower than those obtained for MI but higher than those for PK. For labiodentals, the −3 dB/oct high-frequency slope suggested by Fant is on the lower side of values obtained in this study. Experimental results from Shadle's obstacle model indicate an approximate high-frequency tilt of −9 dB/oct (for $U =$ 420 cm^3/s) while those of Pastel show a tilt of −4 dB/oct ($U =$ 442 cm^3/s, $l_o = 2$ cm). It should be emphasized that these results pertain to specific mechanical model configurations and aerodynamic conditions. Nevertheless, T_{HF} values in Table III show a general agreement with the above-mentioned experimental and empirical values. Differences in T_{HF} values across subjects, among other things, may be due to variability in flow rates.

The roll-off T_{LF} ranged between 0 and 3 dB/oct and was, in general, subject-dependent. These values are in the range suggested by some previous studies [4], [24]. Other more recent experimental evidence such as [16] do not support the existence

of a T_{LF} component in the dipole spectrum. Experimental results of Shadle [21], directly relevant to fricative aerodynamics, however, do not offer compelling evidence for ignoring T_{LF} in the parametric dipole spectrum.

A direct comparison between the turbulent (dipole) sources for the voiced and unvoiced fricatives cannot be made due to the lack of information regarding the actual source SPL's and flow rates. Nevertheless, the levels of the dipole sources for voiced fricatives are expected to be about 5 to 10 dB lower than the unvoiced ones due to a smaller pressure drop across the supraglottal constriction. For a particular subglottal pressure, since the (time-averaged) glottal area is smaller for the voiced fricatives, a relatively large pressure drop across the glottis and a relatively small pressure drop across the oral constriction would result when compared to the unvoiced fricatives.

Results for voiced fricatives were quite satisfactory even though a simplistic superposition of the effects of the voicing and turbulence sources was adopted. The frequency region in the output spectrum where a cross-over occurs in the dominant source type, from voicing to turbulence, is, however, not always captured well by the model. A model that takes into account the interaction between the turbulence and voicing sources would perhaps provide a satisfactory solution.

Overall, the match obtained between the synthesized and natural spectra was better for strident fricatives than for nonstridents. This is primarily because the spectra of the nonstrident fricatives are not characterized by consistent, well-defined spectral peaks as do the stridents. The resonance patterns observed in the nonstrident spectra are predominantly due to incompletely cancelled back cavity pole-zero pairs resulting from the finite coupling present between the front and back cavities. The transfer functions are more sensitive to lip shape in nonstridents but cannot be measured any more accurately than for the stridents. The 1-D model used for the vocal tract simulation, perhaps, is not adequate for capturing these effects.

In general, there are a few other potential reasons, mainly arising due to various limitations in data acquisition and modeling used, that could have contributed to the less than perfect match between the synthesized and natural fricative spectra. These include the following.

1) The strategy for accounting for the teeth which are not captured by MRI may have been somewhat inaccurate explaining the generally poor match for the labiodentals and interdentals, and perhaps the need for more than one teeth-source for the stridents.
2) Inability in obtaining precisely the same articulatory position for both MRI and audio-recording sessions.
3) Limitations due to the various approximations used in the acoustic modeling including those used to obtain the transfer functions, the radiation impedance model, the far-field transfer characteristic, and the low-frequency piston-in-sphere approximation.

Finally, an attempt at unifying the empirical noise source models for fricatives can be made. The parameters of the hybrid source model used in our experiments are source location, type, number, dipole spectral characteristics and relative source levels. First, let us consider source location, type and number. Since dipole sources were the most influential, and since the effect of the monopole sources was relatively minimal, the noise source model for these fricatives in practice can be approximated just by dipole (pressure) source models, similar to the strategy employed in [4], [6]. Further the dipole source location can be held fixed for a given fricative articulation class as dictated by the obstacle locations, an observation that is consistent with the findings of Shadle [19], [21]. This is different from the approach taken in some articulatory synthesis schemes wherein the noise sources were primarily located in the vicinity of or at constriction locations [7], [23].

Next, let us consider the spectral characteristics of the dipole source. The various experimental and empirical results agree on the inclusion of a high-frequency tilt in the dipole spectrum although there is considerable variability in its values across these various studies (and, in their experimental conditions). Hence, it is not straightforward to pin down a set of empirical values for the spectral tilts in defining a general dipole noise source spectrum, even at the level of each articulation class. Instead, we propose a general noise-shaping filter interpretation toward a unified analysis and modeling of the dipole spectrum. Consider a two-port network driven by a white Gaussian noise (WGN) voltage source (pressure) and whose voltage output represents the dipole source. For example, the -6 dB/oct spectrum (Fant's /**s**/ model) with a break frequency at 4 kHz ($\omega_b = 8000\,\pi$ rad/s) can be simply modeled as the output of a first-order filter with a transfer function of the form $H_{dipole}(\omega) = 1/(1 + j\omega/\omega_b)$ that is excited by a white Gaussian noise source. A spectral tilt of -12 dB/oct with a break frequency of ω_b and damping factor ζ requires a quadratic term, such as $(1+2\zeta(j\omega/\omega_b)+(j\omega/\omega_b)^2)$ in the transfer function's denominator. Since the zeroth-order and first-order systems can be considered as special cases of a second order system, a preliminary model for the dipole source can be a WGN-excited all-pole second order system; a simple analog realization of this would be a series RLC circuit (or equivalently a mass-spring system) where the voltage across the capacitor represents the dipole voltage. If such a model is assumed, the small low-frequency roll-off in the dipole spectrum corresponds to moderate under-damping as illustrated in Fig. 8 (or equivalently in circuit terminology, a relatively high Q value due to a somewhat larger energy storage in the acoustic inertance relative to that dissipated in the resistance). If the effect of the inertance is minimal, then the dipole spectrum corresponds to first-order effects (tilts in the order of –6 dB/oct) and in addition if resistive effects are minimal, then the effects are zeroth-order (only a gain term with a flat spectrum). Since the low-frequency tilt T_{LF} appears to be subject dependent (Table III), the damping factor ζ perhaps can be assumed to be subject dependent. The break frequency ω_b can be related to the aerodynamic-articulatory state of the vocal tract. While a scaling relation similar to that used for F_{peak} might be a good starting point for ω_b, deriving relations between the (acoustic) elements of such a system and the aerodynamic state of the vocal tract is not possible with the data currently available and is beyond the scope of this paper.

In the future, acoustics in more realistic 3-D geometries should be investigated. This can be achieved either through numerical simulations and/or experimental investigations in

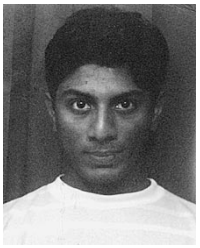
mechanical models. Furthermore, *in vivo* measurements of flow and pressure in the human vocal tract would provide further insights toward constructing improved source models for fricatives. The current study indicates that a source-filter type representation may be adequate for frequencies below 10 kHz and that the sources may be adequately represented by a parameterizable-dipole model. Significant work, however, is required for a detailed investigation of the source parameters (peak, tilts) through aerodynamic experiments. Such experiments are crucial for validating the results of the current study.

ACKNOWLEDGMENT

The authors are grateful to S. Maeda for providing his simulation program. They also thank four anonymous reviewers and the editor for their insightful comments and suggestions.

REFERENCES

[1] P. Badin, "Acoustics of voiceless fricatives: Production theory and data," *Speech Technol. Lett.*, pp. 45–52, Apr.–Sept. 1989.
[2] P. Badin, C. Shadle, Y. P. T. Ngoc, J. N. Carter, W. S. C. Chiu, C. Scully, and K. Stromberg, "Frication and aspiration noise sources: Contribution of experimental data to articulatory synthesis," in *Int. Conf. Spoken Language Processing*, Yokohama, Japan, 1994, pp. 163–166.
[3] J. Dang and K. Honda, "Acoustic characteristics of the piriform fossa in models and humans," *J. Acoust. Soc. Amer.*, vol. 101, pp. 456–465, 1997.
[4] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.
[5] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*. New York: Springer-Verlag, 1972.
[6] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *Bell Syst. Tech. J.*, vol. 54, no. 3, pp. 485–506, 1975.
[7] J. L. Flanagan and K. Ishizaka, "Automatic generation of voiceless excitation in a vocal cord- vocal tract speech synthesizer," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 163–170, Feb. 1976.
[8] M. E. Goldstein, *Aeroacoustics*. New York: McGraw-Hill, 1976.
[9] J. M. Heinz and K. N. Stevens, "On the properties of voiceless fricative consonants," *J. Acoust. Soc. Amer.*, vol. 33, no. 5, pp. 589–596, 1961.
[10] G. W. Hughes and M. Hale, "Spectral properties of fricative consonants," *J. Acoust. Soc. Amer.*, vol. 28, no. 2, pp. 303–310, 1956.
[11] M. J. Lighthill, "On sound generated aerodynamically I. General Theory," *Proc. R. Soc. A*, vol. 211, pp. 564–587, 1952.
[12] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Commun.*, vol. 1, pp. 199–229, 1982.
[13] W. Meyer-Epplere, "Zum erzeugungsmehanismus der gerauschlaute," *Z. Phonetik*, vol. 7, pp. 196–212, 1953.
[14] S. Narayanan, A. Alwan, and K. Haker, "An articulatory study of fricative consonants using magnetic resonance imaging," *J. Acoust. Soc. Amer.*, vol. 98, no. 3, pp. 1325–1347, 1995.
[15] S. S. Narayanan, "Fricative consonants: An articulatory, acoustic, and systems study," Ph.D. dissertation, Dept. Elect. Eng., Univ. Calif., Los Angeles, 1995.
[16] P. A. Nelson and C. L. Morfey, "Aerodynamic sound production in low speed flow ducts," *J. Sound Vibr.*, vol. 79, no. 2, pp. 263–289, 1981.
[17] L. M. P. Pastel, "Turbulent noise sources in vocal tract models," M.S. thesis, Mass. Inst. Technol., Cambridge, MA, 1987.
[18] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, no. 2, pp. 583–590, 1971.
[19] C. H. Shadle, "the acoustics of fricative consonants," Mass. Inst. Technol., Cambridge, Tech. Rep. 506, 1985.
[20] ——, "Articulatory-acoustic relationships in fricative consonants," in *Speech Production and Speech Modeling*, W. Hardcastle and A. Marchal, Eds. Norwell, MA: Kluwer, 1990, pp. 187–209.
[21] ——, "The effect of geometry on source mechanisms of fricative consonants," *J. Phonetics*, vol. 19, pp. 409–424, 1991.
[22] C. H. Shadle, P. Badin, and A. Moulinier, "Toward the spectral characteristics of the fricative consonants," in *Proc. XII Int. Conf. Phonetic Sciences*, vol. 3, Aix-en-Provence, France, 1991, pp. 58–61.
[23] M. M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 955–967, July 1987.
[24] K. N. Stevens, "Airflow and turbulence noise for fricative and stop consonants: Static considerations," *J. Acoust. Soc. Amer.*, vol. 50, no. 4 (pt. 2), pp. 1180–1192, 1971.
[25] ——, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.
[26] K. N. Stevens, S. E. Blumstein, L. Glicksman, M. Burton, and K. Kurowski, "Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters," *J. Acoust. Soc. Amer.*, vol. 91, no. 5, pp. 2979–3000, 1992.



Shrikanth Narayanan (S'88–M'95) received the M.S., Eng., and Ph.D. degrees, all in electrical engineering, from the University of California, Los Angeles, in 1990, 1992, and 1995, respectively.

Since 1995, he has been with AT&T, first with AT&T Bell Labs and later with AT&T Shannon Laboratory. His research interests include several areas of speech processing–speech production modeling, speech recognition, spoken language and multimodal dialog systems.

Dr. Narayanan is a member of Tau Beta Pi and Eta Kappa Nu.

Abeer Alwan (S'82–M'85) received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1992.

Since then, she has been with the Electrical Engineering Department, University of California, Los Angeles (UCLA), as an Assistant Professor (1992–1996) and Associate Professor (1996–present). She established and directs the Speech Processing and Auditory Perception Laboratory at UCLA. Her research interests include modeling human speech production and perception mechanisms and applying these models to speech-processing applications.

Dr. Alwan received the NSF Research Initiation Award in 1993, the NIH FIRST Career Development Award in 1994, the UCLA-TRW Excellence in Teaching Award in 1994, the NSF Career Development Award in 1995, and the Okawa Foundation Award in Telecommunications in 1997. She is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She is an elected member of the Acoustical Society of America Technical Committee on Speech Communication and the IEEE Signal Processing Technical Committees on Audio and Electroacoustics and on Speech Processing.