Noise Sources and Area Functions for the Synthesis of Fricative Consonants

Peter Birkholz and Dietmar Jackél

Institut für Informatik, Universität Rostock, Email: piet@informatik.uni-rostock.de

Abstract

In this study, we characterize the noise sources and the critical parts of vocal tract area functions for the synthesis of voiceless fricatives. We derive these characteristics indirectly by fitting synthetic to natural fricative spectra in an interactive procedure. The adjustable parameters are the number, location, type, amplitude, and spectral shape of the noise sources as well as the cross-sectional areas of the vocal tract. From the results of these experiments we derive a model for the calculation of noise source parameters for an articulatory speech synthesizer.

1 Introduction

Fricative consonants are produced, when the air from the lungs is forced through a narrow passage in the vocal tract such that the airflow becomes turbulent in the region downstream from the constriction [1, 2]. The turbulence can generate noise by different mechanisms. One source of sound is generated by turbulent velocity fluctuations in the jet of air leaving the constriction. These fluctuations constitute a *monopole* sound source. When the jet impinges on an obstacle like the teeth or the vocal tract walls, the surface of the obstacle generates a fluctuating force on the medium. The fluctuating force is acting as an acoustic *dipole* source. A third mechanism of turbulence noise generation is a consequence of random velocity fluctuations in the airstream *within* the constrictions [1], constituting a monopole source.

Nearly all existing fricative models (see [3] for a review) are based on a linear source-filter model of the vocal tract and insert one or more lumped noise sources that add random fluctuations to either sound pressure or volume velocity. In these models, pressure sources are equivalent to acoustic dipoles and volume velocity sources correspond to acoustic monopoles. The success of such models hinges on the ability to characterize the noise source(s) in terms of its/their location, spectrum, and strength. Another difficulty is the accurate representation of the vocal tract area function downstream from the constriction. This part of the area function largely determines the filter properties.

The aim of this study was to characterize both the noise sources and the area functions for the fricatives /f/, /s/, /ʃ/ (hard and soft), /ç/ and /x/ by means of an analysis-by-synthesis approach. In other words, we adjusted the area functions and noise sources of synthetic fricatives such that the synthetic output spectra closely matched the corresponding natural fricative spectra. Theoretically, different area functions and source configurations can produce one and the same synthetic spectrum (non-uniqueness of the articulatory-to-acoustic mapping). To avoid physiologically unrealistic area functions or source properties that contradict the theory of noise production, we decided to adjust the

parameters interactively instead automatically. The area functions were initialized with the measurements published by Fant [4] and were modified only when it was strictly necessary. Both monopole and dipole sources could be introduced at different places within the area functions and modified in terms of amplitude and spectral shape.
The following section describes the way we synthesized the fricative spectra. Section 3 gives details about the fitting procedure and its results. In Sec. 4 we derive a general noise source model and draw conclusions in Sec. 5.

2 Acoustic Simulation

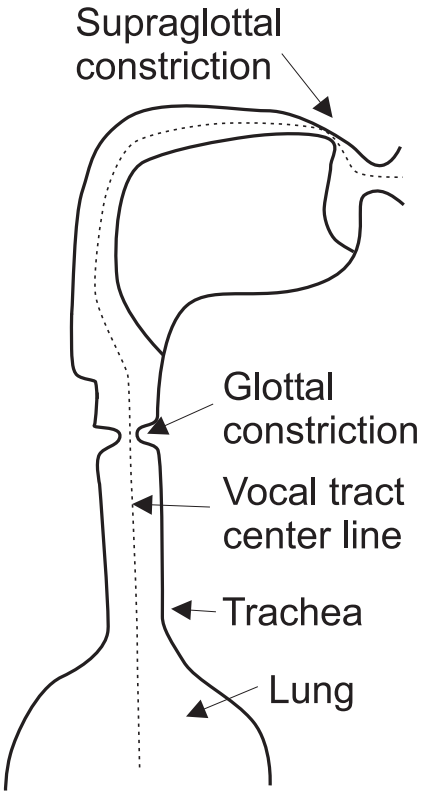

Figure 1: Vocal tract model.

The vocal tract model that we used for the simulations is shown in Fig. 1 and consits of the subglottal system, the glottis, and the pharyngeal and oral cavity. Assuming plane wave propagation, we approximated the vocal tract pipe by incremental abutting cylindrical tube sections. Each cylindrical section was represented by its equivalent electrical T-network as shown in Fig. 2. The whole vocal tract was thus treated in analogy to a nonuniform electrical transmission line with lumped elements. The derivation of the lumped transmission line model can be found, e.g., in [4, 5]. The network elements that we used in this study are summarized in [6].
The subglottal system was modeled according to [7], and the glottis was represented by a tube section with a fixed area of 0.3 cm^2 and a length of 7 mm. This corresponds to a typical glottal configuration for the production of voiceless fricatives [1]. The area function between the glottis and the lips was represented by 40 equally spaced tube sections. The cross-sectional area of these tube sections and the vocal tract length between glottis and lips could be interactively varied.
Besides the vocal tract area function we had to specify the noise sources in order to synthesize the fricative spectra. Both monopole and dipole sources could be inserted in the transmission line model as indicated by the dashed parts in Fig. 2. Acoustic monopoles were represented as shunt volume velocity sources (U_n) and dipoles as series pressure sources (P_n). Each tube section could contain at most one pressure source (located at the left end of the section) and one volume velocity source (located in the center of the section). In order to specify the spectral characteristics of the noise sources we followed the proposal by Narayanan and Alwan [3] and used for each source a low-pass filter of either first or second order. These filters offer the possibility to choose an arbitrary cutoff frequency and gain, a high-frequency slope of either -6 dB/oct (first order) or -12 dB/oct (second order), and a quality factor Q between 0.5 and infinity for second order filters.
The generation of turbulence and noise in the vicinity of a constriction is generally accompanied by a pressure loss. This pressure loss is mainly caused by flow separation from the

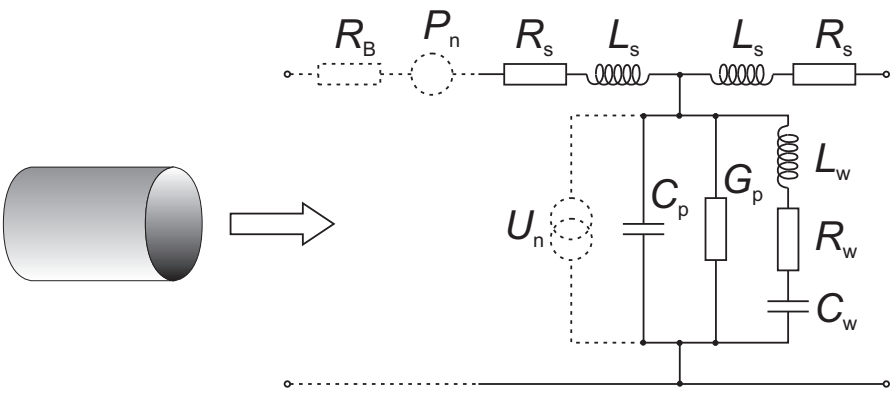

Figure 2: Equivalent electrical T-network corresponding to one tube section.

vocal tract walls when a jet of air is generated at the outlet of a constriction. The loss can be quantified in terms of the dynamic pressure in the constriction as $\Delta p_c = k\rho u^2/2A_c^2$, where ρ is the ambient density, u is the volume velocity and A_c is the cross-sectional area of the constriction [1, p. 30]. As a first approximation, the constant k can be set to 1 (complete loss of the dynamic pressure in the constricted tube). The same relation is given for the pressure loss at the narrow glottal passage with the area A_g and the pressure loss $\Delta p_g = k\rho u^2/2A_g^2$. We took into account both of these losses in our simulation by means of two additional resistances in the transmission line (labeled R_B in Fig 2): one at the inlet of the glottis (R_g) and one at the left end of the most constricted supraglottal tube section (R_c). In order to get *linear* resistances, they were approximated for the small-signal (or ac) case as

$$R_c = \left.\frac{\partial \Delta p_c}{\partial u}\right|_{u=\overline{u}} = \rho\frac{\overline{u}}{A_c^2} \quad \text{and} \quad R_g = \left.\frac{\partial \Delta p_g}{\partial u}\right|_{u=\overline{u}} = \rho\frac{\overline{u}}{A_g^2},$$

where $\overline{u}$ is the mean volume velocity through the vocal tract. Assuming that Δp_c and Δp_g are the dominating pressure losses in the vocal tract, $\overline{u}$ can be calculated as [1, p.108]

$$\overline{u} = \sqrt{\frac{2P_{\text{lung}}}{\rho(1/A_g^2 + 1/A_c^2)}}.$$

The pressure P_{lung} produced by the lung was held constant at 8 cm H_2O in all simulations. The spectrum of the radiated sound pressure was calculated in the frequency domain as

$$P_{\text{rad}}(\omega) = R(\omega)\sum_{i=1}^{N} S_i(\omega)H_i(\omega),$$

where $R(\omega)$ is the radiation characteristic [1, p. 127], N is the number of noise sources, $S_i(\omega)$ is the source spectrum of the ith source (either volume velocity or pressure), and $H_i(\omega)$ is the transfer function between the ith source and the volume velocity at the lips. For details about the calculation of the transfer functions we refer to standard textbooks about circuit theory.

3 Spectral Fitting

For each fricative, we interactively modified the area function and the noise sources such that the synthetic spectra closely fitted to the spectra of the corresponding natural fricatives.

The natural fricatives were recorded from four subjects (two male and two female) at a sampling rate of 22 kHz and 16 bits per sample. The subjects were asked to produce the fricatives /f/, /s/, /ʃ/ (hard and soft), /ç/ and /x/ and to sustain each of them for at least 500 ms. The computer calculated the spectrum of each fricative by averaging 40 short-time magnitude spectra of overlapping Hamming windowed segments (23.2 ms length and 11 ms overlap). From all of the calculated spectra we chose one per fricative for the fitting procedure.

For initial estimates of the area functions we used the measurements provided by Fant [4, p. 172] and approximated the curves by 40 discrete area values/tube sections. During the spectral fitting we restricted our modifications of the area functions to the overall vocal tract length and the cross-sectional areas of the tube sections downstream from the main constriction. These area values are most significant for the filter properties of fricatives. In order to make the interactive experiments tractable, we limited the allowed areas – in accordance with Fant [4] – to the following set of values: 0.16–0.32–0.65–1–1.3–1.6–2–2.6–3.2–4–5–6.5–8–10.5–13–16 cm^2. The area values and the source properties were modified in incremental steps and the similarity of the synthetic and natural spectra in terms of pole and zero locations as well as spectral slopes was checked after each modification.

The fitting results are summarized in Fig. 3. The left side of the figure shows the area functions, where the dark gray areas indicate the tube sections downstream from the main constriction. The assumed position of the incisors is marked by a small triangle. Noise sources are indicated by arrows. The letter **M** denotes a monopole (volume velocity) source and **D** denotes a dipole (pressure) source. The right side of the figure shows the corresponding fricative spectra of the recordings (top curve) and the simulation (bottom curve). The curves demonstrate that we could achieve a fairly good agreement between natural and synthetic spectra with the analysis-by-synthesis approach.

It was possible to synthesize each fricative with one monopole noise source in the main constriction and one dipole noise source at an obstacle (vocal tract walls, teeth or lips) in the region of the turbulent jet. The source spectra of all monopole and dipole sources could be shaped with a *second order* low-pass filter with a fixed Q-value of $1/\sqrt{2}$. Due to this Q-value they belong to the class of Butterworth-filters that have a maximally flat passband without peaking in the frequency response. The cutoff frequency for all monopole sources was furthermore set to 1100 Hz, which gives them a magnitude spectrum that agrees closely with the data experimentally found by Pastel (cited in [3]). The cutoff frequencies f_c for the dipole sources vary between 2500 Hz and 6000 Hz (underneath the letter **D**).

The amplitudes of the dipole sources were set to a fixed constant value for all fricatives, and the amplitudes of the monopole sources were adjusted with respect to them. The relative amplitudes are written underneath the letters **M**. A value of $a = 0.2$ means, that the real amplitude $\hat{u}$ of the volume velocity source equals $\hat{u} = 0.2\hat{p}/R_\mathrm{B}$, where R_B is the additional supraglottal resistance and $\hat{p}$ is the amplitude of the pressure source. The relative amplitudes of the monopole sources vary between 0.1 and 0.5. In general, their contribution to the fricative spectra is small compared to the dipole sources. Nevertheless, the good fit for the fricatives /ç/ and the hard /ʃ/ in the low-frequency region was only possible with the contribution of the monopole sources.

Besides the source characteristics, the cross-sectional areas in the front part of the vocal tract were critical for good spectral fits. For /ç/ and /ʃ/, the area functions have a second striking constriction at the location of the teeth, and the relatively large cross-sectional areas between the tongue constriction and the teeth constriction can be interpreted as

a sublingual space, which was consistently found in MRI-scans by Narayanan et al. [3]. We also want to point out, that a realistic value for the cross-sectional area of the glottis (we used 0.3 cm^2) and the resulting coupling between the vocal tract and the subglottal system is essential for the proper damping of the formants and antiformants in the fricative spectra.

Let us now turn to an interpretation of the source locations. The monopole sources were set into the main constriction for each fricative, which is formed with the upper incisors and the lower lip for /f/, and with the tongue and the palate for /s/, /ʃ/, /ç/, and /x/. The dipole sources were set at the position of the teeth for /s/, /ʃ/, and /ç/, at the position of the lips for /f/, and at the position of the vocal tract walls a short distance downstream of the constriction for /x/. The teeth, lips and vocal tract walls can be interpreted as the dominant obstacles in the region of the turbulent jet for these fricatives.

4 Noise Source Model

Based on the results of this study and data from previous studies about turbulence in the vocal tract, we propose the following noise source model for articulatory speech synthesis. It is based on three principles, that define the position, spectrum, and strength of the noise sources:

(1) Two noise sources are used in the case of turbulence: A monopole source in the most constricted tube section and a dipole source at the downstream end of the tube section corresponding to an obstacle. When the constriction is located at the incisors as in /f/, the lips (one section downstream from the constriction) are assumed to act as the obstacle. When the constriction is located up to 3 cm behind the incisors as in /s/, /ʃ/ and /ç/, the incisors are assumed to act as the obstacle. When the constriction is further upstream, as in /x/, the vocal tract walls (one section downstream from the constriction) act as the obstacle.

(2) The sources produce Gaussian white noise, filtered with a low-pass Butterworth filter of second order. The cutoff frequency f_c for the monopole source is 1100 Hz, and for the dipole source, f_c is a function of the velocity v_c in the constriction and its cross-dimension d_c [1, p. 103]: $f_c = \gamma v_c / d_c$. The constant γ was empirically determined as 0.4.

(3) The amplitude of the dipole source is calculated according to [8] as $\hat{p} = \max\{0, \alpha \cdot \eta(Re^2 - Re_{\text{crit}}^2)\}$, where $Re = v_c d_c / \nu$ is the dimensionless Reynold's number in the constriction, $Re_{\text{crit}} = 1800$ is the critical Reynolds number, ν is the kinematic viscosity of the fluid and α ($\approx 4 \cdot 10^{-6}$ N/m^2) is an empirically determined gain. The additional factor η is a modification to the original formula and is supposed to reflect the efficiency of the obstacle for the generation of noise. The sound generated by an obstacle is greatest, when the airflow impinges on the obstacle perpendiculary, and smaller, when the angle of incidence decreases [1]. As a rough approximation, we suggest to set $\eta = 1$ when the incisors act as obstacle, and $\eta = 0.5$ when the lips or vocal tract walls act as obstacle. The amplitude $\hat{u}$ of the monopole source is calculated analog to $\hat{p}$ as $\hat{u} = \max\{0, \beta(Re^2 - Re_{\text{crit}}^2)\}$, but with a different gain β ($\approx 2 \cdot 10^{-13}$ m^3/s) and without the factor η.

We have integrated this noise source model in an articulatory speech synthesizer that is based on a discrete-time simulation of the vocal tract system [6]. From the area functions in Fig. 3, we synthesized the fricatives both in isolation and in vowel context (/a:/, /i:/ and /u:/). Preliminary results of listening tests confirm that the synthetic fricatives are

both highly intelligible and distinguishable. Sound files with some examples can be found in the internet under the URL [9].

5 Conclusions

In the first part of this study, we have interactively fitted synthetic to natural fricative spectra to find characteristic properties of the area functions and noise sources of fricatives. In the second part, we used these findings, together with information from other studies about frication noise, to derive a noise source model that can be used for articulatory speech synthesis. The application of this model in a discrete-time simulation of the vocal tract system resulted in the production of very natural fricative sounds. In the context of an articulatory synthesizer, the noise source model requires not only the pure area function of the vocal tract, but also the location of the teeth. An important aspect for high quality synthetic fricatives is furthermore the coupling to the subglottal system through the open glottis. Finally, we propose a finer spatial resolution of the area function in the front part of the vocal tract in articulatory synthesizers, because the fricative spectra are very sensitive to area changes in this region.

6 Acknowledgments

This research was funded by the DFG (German Research Foundation).

References

[1] K. N. Stevens, *Acoustic Phonetics*. The MIT Press, 1998.

[2] P. Badin, "Acoustics of voiceless fricatives: Production theory and data," *STL-QPSR*, vol. 3, pp. 33–55, 1989.

[3] S. Narayanan and A. Alwan, "Noise source models for fricative consonants," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 328–344, 2000.

[4] G. Fant, *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

[5] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. Springer-Verlag, Berlin, 1965.

[6] P. Birkholz and D. Jackèl, "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system," in *Interspeech 2004-ICSLP*, Jeju, Korea, 2004.

[7] K. Ishizaka, M. Matsudaira, and T. Kaneko, "Input acoustic-impedance measurement of the subglottal system," *Journal of the Acoustical Society of America*, vol. 60, no. 1, pp. 190–197, 1976.

[8] W. Meyer-Eppler, "Zum Erzeugungsmechanismus der Geräuschlaute," *Zeitschrift für Phonetik und allgemeine Sprachwissenschaft*, vol. 7, pp. 196–212, 1953.

[9] P. Birkholz, *http://wwwicg.informatik.uni-rostock.de/~ piet/new_ vocal_ tract.html*. Project page, 2005.

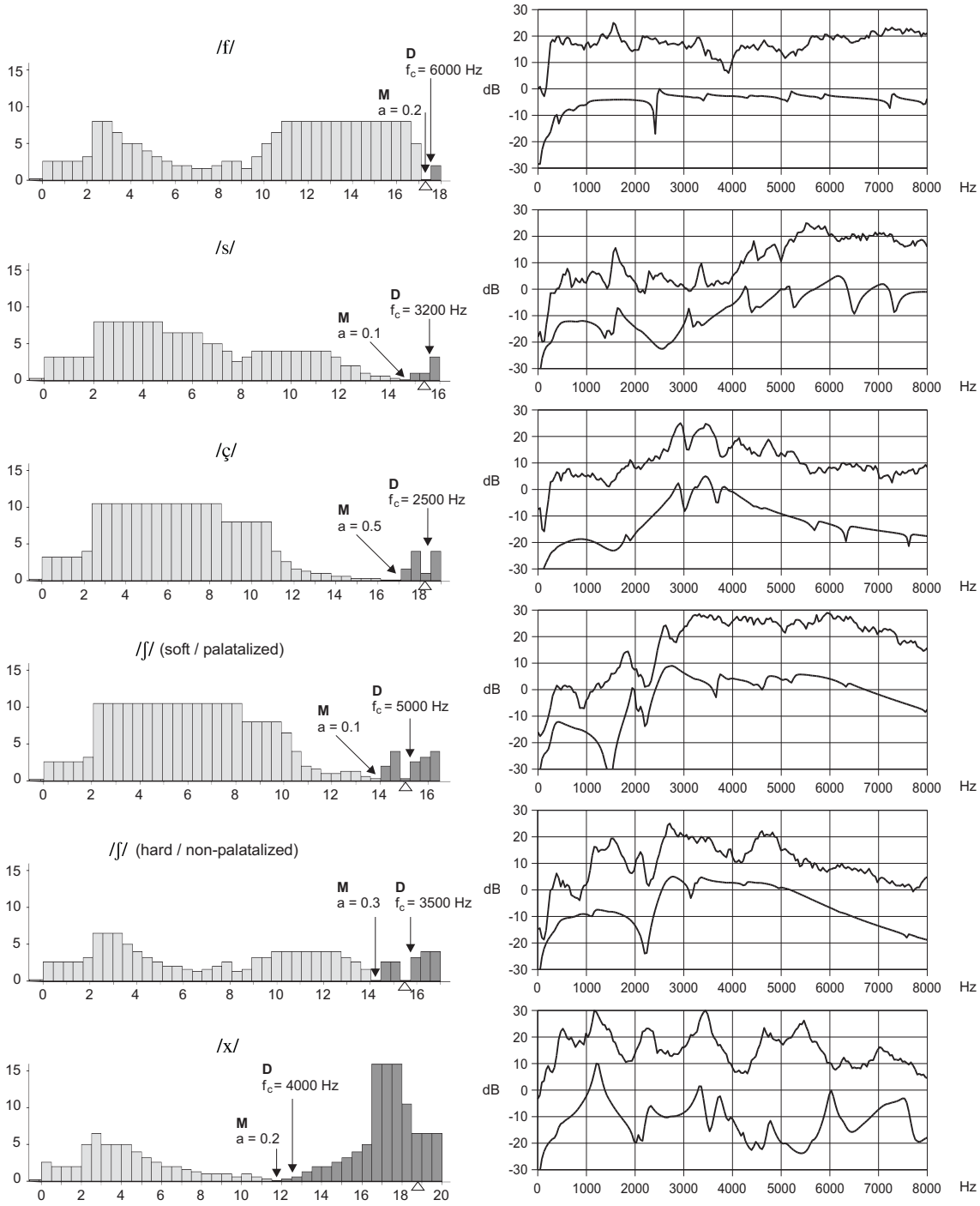

Figure 3: Spectral fitting results. The left side shows the area functions, where dark gray areas indicate tube sections downstream from the main constriction. The positions of the noise sources are marked by arrows. Dipole sources are labeled by the letter **D** and the cutoff frequency of the spectral shaping filter. Monopole sources are labeled by the letter **M** and their amplitude relative to the dipole source amplitude. The position of the teeth is marked by a white triangle. The right side of the picture shows the corresponding fricative spectra. The upper spectra originate from natural fricatives and the lower spectra are synthesized.