

# Adding Fricatives to the Portuguese Articulatory Synthesiser

António Teixeira, Luis M. T. Jesus, Roberto Martinez

Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA)

Universidade de Aveiro, 3810-193 Aveiro, Portugal

{ajst, lmtj, martinezrs}@ieeta.pt

## Abstract

First attempts at incorporating models of friction into an articulatory synthesizer, with a modular and flexible design, are presented. Although the synthesizer allows the user to choose different combinations of source types, noise volume velocity sources have been used to generate turbulence. Preliminary results indicate that the model is capturing essential characteristics of the transfer functions and spectral characteristics of fricatives. Results also show the potential of performing synthesis based on broad articulatory configurations of fricatives.

## 1. Introduction

### 1.1. Fricative Production Mechanisms

When a vowel is being uttered, the vocal tract is relatively unconstricted ( $\sim 1 \text{ cm}^2$  cross-sectional area at the most constricted region) and the vocal folds vibrate periodically, causing the volume of air flowing through the glottis to fluctuate periodically as well. Fricative consonants are produced when the vocal tract is constricted ( $\sim 0.1 \text{ cm}^2$  at most constricted region) somewhere along its length, enough to produce turbulence noise when air is forced through the constriction. The place of constriction affects the tract resonances (filter properties), but also affects the shape of the tract downstream of the constriction and thus the source properties: where the turbulent jet will impinge on tract walls, generating more noise, and the particular spectral characteristics of that noise.

It is known from studies of jet noise and mechanical models that when a particular configuration is held constant, and only the air velocity is increased, the turbulence noise increases (i.e. sound pressure and power), and increases more at higher frequencies. Though it is not easy to control nor measure parameters so precisely in the vocal tract, the same phenomenon appears to occur for fricatives.

The acoustic mechanism for production of fricatives is thus not as well understood as for vowels because:

1. turbulence noise defies an analytic formulation, requiring empirical studies;
2. turbulence noise sources are much more sensitive to changes in the surrounding geometry than are acoustic resonances;
3. given the small constriction dimensions and the dependence of all aeroacoustic sources on flow velocities, it is much more difficult and more important to get sufficiently accurate vocal tract *shape* and simultaneous *aero-dynamic* and *acoustic* data for fricative configurations.

These difficulties have been reflected in the relatively poor quality of fricative and affricate synthesis.

### 1.2. Previous Studies of Fricatives

Our understanding of fricative production has been improved by the use of existing expertise in the production of speech corpora, the extraction of MRI data [1], fricative aeroacoustics analysis methods, and the incorporation of three dimensional vocal tract data in speech synthesis. The study of relations between articulatory, acoustic and perceptual cues provides crucial information for the articulatory synthesis of fricative consonants [2]. The study of the nature of the interaction between acoustic sources and vocal tract shapes for constricted consonantal configurations, and the study of mechanical models by Shadle [3], has supplied important data to drive various parametric multi-tube acoustic models [4, 5, 6, 7, 8].

European Portuguese fricatives have been previously analysed by Jesus and Shadle [9] in ways designed to enhance our description of the language, and to use and increase our understanding of the production of fricatives. The research presented by Jesus and Shadle [9] aimed to investigate the acoustic features that characterise the production of fricative consonants [9]. Their work focused on the analysis of friction in the Portuguese language, describing a novel methodology of corpus design, and temporal and spectral analysis techniques. Knowledge accumulated from their data could be used for improved speech synthesis. The peak frequencies, spectral amplitude characteristics, and temporal information could be useful for synthesis and the parameterisation of the spectra allows us to deduce the behaviour of sources for articulatory synthesis models such as the one proposed by Narayanan and Alwan [8]. The quantified spectral characteristics of Portuguese fricatives [9] can be related to specific properties of the transfer function and source spectrum during the production of these sounds, although using only the far-field acoustic signal will always present a limitation to source-filter separation.

### 1.3. Previous Fricative Production Models

Flanagan and Ishizaka [4, 5] modelled voiceless excitation, for a speech synthesizer that incorporated the two mass vocal fold model, assuming that the sound sources interacted with the resonant system. The model included turbulent excitation generated by a serial-random pressure source within the vocal fold model (produced by vorticity in the flow through the vocal fold opening, essentially during the time that the vocal folds do not vibrate) and/or turbulent flow (modelled as a series pressure source) that occurred at constricted points along the vocal tract.

Sondhi and Schroeter [6] developed a hybrid articulatory synthesiser that modelled the glottis in the time domain because of its nonlinear nature, and modelled the vocal and nasal tracts in the frequency domain taking advantage of the more convenient representation of losses and radiation using a product of  $2 \times 2$  chain matrices (also called ABCD matrices). Frication was

generated using only one series pressure source at the point of maximum constriction, or alternatively using a parallel volume velocity source downstream of the main constriction.

Badin [10] investigated the properties of the vocal tract transfer function in the frequency domain and its relation to the source location and impedance functions of a pseudo-static aerodynamic model that defined boundary conditions (glottis and constriction resistances) from aerodynamic parameters (subglottal pressure, glottis opening and constriction area). The spectra of natural speech were replicated using a model of the vocal tract area functions taking into account the glottal and subglottal impedances.

Shadle [3] studied the acoustic mechanism of fricative consonants in the context of three domains: theoretical models, mechanical models and speech. She described four sets of experiments with mechanical models of increasing realism, which were used to determine source characteristics such as location, degree of distribution and spectrum shape.

Scully, et al. [2] combined the analysis of real speech with analysis-by-synthesis using the Leeds model of speech production. The input of the model specified a succession of targets for each articulator and the output defined strengths and time domains of the voice, quasi-periodic, aspiration noise and frication noise sources. The articulatory descriptions obtained from analysis of natural speech provided components for the aerodynamic description, and each acoustic source of the model depended upon an appropriate combination of articulatory and aerodynamic conditions.

Narayanan and Alwan [8] used MRI, dynamic EPG, high quality acoustic recordings and aerodynamic studies to derive data for a parametric hybrid source model and vocal tract model. The source characteristics were derived based on an analysis-by-synthesis method and the vocal tract area functions were obtained from MRI of the fricatives. The vocal tract was modelled as a concatenation of 3 mm long uniform cylindrical tube sections, and the sublingual cavities were modelled as shunt branches specified in the anterior oral cavity. Their hybrid source model used a combination of acoustic monopole and dipole sources and a voiced source in the case of voiced fricatives.

## 2. SAPWindows

SAPWindows is the name given to the University of Aveiro's articulatory synthesizer and it stands for "Sintetizador Articulatorio de Português" for Windows. It consists of articulatory, source, and acoustic models. Different sounds are produced when the acoustic model is excited by various sources. The synthesizer was implemented using object oriented programming, therefore several abstract classes, parameter transfer protocol rules and data structures were designed. The main abstract classes, known as base classes, define only the criteria and methods used. Implementation of the different models available for a base class is performed in derived classes.

Fig. 1 shows how the main synthesis base classes interact with each other. The acoustic model produces the sound. Base classes could be used in other applications besides SAPWindows; with the proper interface those classes can be reused.

The anatomic model adopted in this work is based on the MMIRC (Mind Machine Interaction Research Center) model, which in turn is a modified version of the Mermelstein model [11]. The model assumes midsagittal plane symmetry, and the output is an estimate of the vocal tract cross-sectional area.

The acoustic model is responsible for speech wave gener-

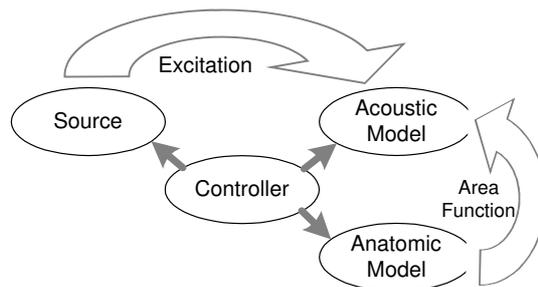


Figure 1: SAPWindows main synthesis base classes.

ation. The output of main synthesis base classes is a sound wave. The impulse response is given by the inverse Fourier Transform (IFFT) of the acoustic transfer function of a given vocal tract configuration, obtained by calculating the resulting transfer function at  $N$  frequencies. The FFTW fast implementation of the Fourier Transform was used. The convolution of the impulse response with the glottal excitation signal produces the sound waveform. A frequency domain analysis and time domain synthesis method, called hybrid or frequency analysis time synthesis method, was used.

The acoustic model must know the configuration of the anatomic model before any kind of computation is performed, so it can retrieve areas, length, sinus data and the section mapping. Information about different sound sources (oral or nasal) is provided for the model by setting the appropriate class parameter.

SAPWindows graphical user interface (GUI), shown in Fig. 3, displays detailed information during synthesis, including, the time trajectory of the vocal tract configuration, the glottal source volume velocity, the speech waveform and spectrogram. Detailed information about SAPWindows has already been presented [12].

## 3. Fricative Modelling

The synthesizer was developed in a way that it could incorporate the following features: to deal with several noise sources with different characteristics; to automatically detect conditions for the existence of noise sources; to use information produced by the articulatory model; to use as much as possible the existing code and data structures.

A model of frication was added to the synthesizer maintaining most of the existing modules, control processes and parameters. This system was used in previous experiments [12] to synthesize voiced sounds. In this investigation [12] the value of  $F_0$  was directly controlled by changing the parameterised glottal area of a two-mass vocal fold model [12]. Unvoiced sounds are now produced in the acoustic model by setting the  $F_0$  value below a certain threshold, as indicator of no-oscillation. The existing parameters **Agmax** (maximum glottal aperture) and **slope** (difference between masses aperture) are used to set the vocal folds opening, controlling  $A_{g1}$  and  $A_{g2}$  in the two-mass model. When  $F_0$  goes below the mentioned threshold "fake" periods of fixed duration are created as a result of having adopted a pitch synchronous synthesis method.

In order to obtain the radiated pressure at the lips due to glottal volume velocity [12], transfer functions are calculated

at the beginning and end of each period and linear interpolation of the impulse responses is used to obtain each sample [12]. The flow, pressure and resistance of noise sources, and the transfer functions from noise sources to the lips, are calculated several times, to allow the activation and deactivation of noise sources during a period. For each tube where a noise source was inserted, past values of noise source volume velocity were stored to calculate the convolution with the impulse response and therefore obtain the speech sound pressure waveform.

In this implementation, noise sources are part of the acoustic model, to model turbulence generated inside the tract which depends on volume velocity. This differs from the glottal source model, which is considered as a separate module of the synthesizer. We have a new acoustic model that can include several sources, extending the existing synthesizer to support noise sources.

In the current version of our synthesizer the volume flow at the constriction is assumed to be equal to the flow at the glottis. Additional work is being developed to test an improved model of volume flow through a constriction.

### 3.1. Noise sources

Fluctuations in the velocity of airflow emerging from a constriction (at an abrupt termination of a tube) create monopole sources and fluctuations of forces exerted by an obstacle (e.g. teeth, lips) or surface (e.g. palate) oriented normal to the flow generate dipole sources. Since dipole sources have been shown to be the most influential in the fricative spectra [8], the noise source of the fricatives has only been approximated by equivalent pressure voltage (dipole) sources in the transmission-line model. Nevertheless, it is also possible to insert the appropriate monopole sources, which contribute to the low-frequency amplitude and can be modelled by an equivalent current volume velocity source.

Frication noise is generated at the vocal tract according to the suggestions of Flanagan [4], and Sondhi and Schroeter [6]. A noise source can be introduced automatically at any T-section of the vocal tract network, between the velum and the lips. The synthesizer's articulatory module registers which vocal tract tube cross sectional areas are below a certain threshold ( $A < 1\text{ cm}^2$ ), producing a list of tube sections that might be part of an oral constriction that generates turbulence.

The acoustic module calculates the Reynolds number at the sections selected by the articulatory module and activates noise sources at tube sections where the Reynolds number is above a critical value ( $Re_{crit} = 2000$  according to [6]). Noise sources can also be inserted at any location in the vocal tract, based on additional information about the distribution and characteristics of sources [3, 8]. This is a different source placement strategy from that usually used in articulatory synthesis [6] where the sources are primarily located in the vicinity of the constriction. The distributed nature of some noise sources can be modelled by inserting several sources located in consecutive vocal tract sections. This will allow us to try combinations of the canonical source types (monopole, dipole and quadrupole).

A pressure source with an amplitude proportional to the squared Reynolds number ( $P_{noise} = 2 \times 10^{-6} \times \text{random}(Re^2 - Re_{crit}^2)$ , for  $Re > Re_{crit}$  and  $P_{noise} = 0$ , for  $Re \leq Re_{crit}$ ) is activated at the correct place in the tract [4, 6]. The internal resistance of the noise pressure source is proportional to the volume velocity at the constriction:  $R_{noise} = \frac{\rho |U_c|}{2A_c^2}$ , where  $\rho$  is the density of the air,  $U_c$  is the flow at the constriction, and  $A_c$  is the constriction cross-sectional

area. The turbulent flow can be calculated by dividing the noise pressure by the source resistance. This noise flow could also be filtered in the time domain to shape the noise spectrum [8] and test various experimentally derived dipole spectra.

### 3.2. Propagation and radiation

The general problem associated with having  $N$  noise sources is decomposed in  $N$  simple problems by using the superposition principle. In order to calculate the radiated pressure at the lips, the vocal tract is divided into three sections: pharyngeal, region between velum coupling point and noise source and region after the source. Data structures based on the area function of each section are defined and ABCD matrices calculated [6]. The ABCD matrices were then used to calculate downstream ( $Z_1$ ) and upstream ( $Z_2$ ) input impedances, as well as the transfer function ( $H$ )

$$H = \frac{Z_1}{Z_1 + Z_2} \frac{1}{CZ_{rad} + D},$$

where  $C$  and  $D$  are parameters from the ABCD matrix (from noise source to lips), and  $Z_{rad}$  is the lip radiation impedance.

The radiated pressure at the lips due to a specific source is given by:  $p_{radiated}(n) = h(n) * u_{noise}(n)$ , where  $h(n) = IFFT(H)$ . The output sound pressure due to the different noise sources are added together. The output sound pressure resulting from the excitation of the vocal tract by a glottal source is also added when there is voicing.

## 4. Results

The main goal of this work was to synthesize unvoiced fricatives. In a first experiment the synthesizer was used to produce sustained unvoiced fricatives. The vocal tract configuration derived from a high vowel was adjusted by raising the tongue tip in order to produce a sequence of reduced vocal tract cross-sectional areas. The lung pressure was linearly increased and decreased at the beginning and end of the utterance, to produce a gradual onset and offset of the glottal flow.

Fig. 2 shows the glottis volume velocity waveform, the speech waveform and spectrogram of a synthesized /j/. The synthesizer activated only one noise source at the onset and offset of frication, and used five sources during the steady state of the fricative.

The second goal was to synthesize fricatives in VCV sequences. Articulatory configurations for vowels obtained by an inversion method were used and during the fricative interval the tongue tip articulatory parameter was adjusted to a postalveolar fricative configuration. A  $F_0$  value of  $100\text{ Hz}$  and a maximum glottal opening of  $0.3\text{ cm}^2$  were used to synthesize the vowels. The time trajectory of the glottal source parameter **Agmax** starts at  $1.5\text{ cm}^2$ , rises to  $2\text{ cm}^2$  at the fricative middle point and returns to  $1.5\text{ cm}^2$  near the end, before assuming the value used during vowel production. Synthesis results for the non-sense word /jfi/ are presented in Fig. 3 using the actual SAP- Windows GUI after synthesis.

In Fig. 4 we compare the Power Spectral Density (PSD) estimate of the synthesized fricative with the PSD of a natural [j] from the corpus used in [9]. A reasonable fit between the two signals was obtained.

## 5. Conclusions

With the addition of noise source models and modifications to the acoustic model, our articulatory synthesizer is capable of

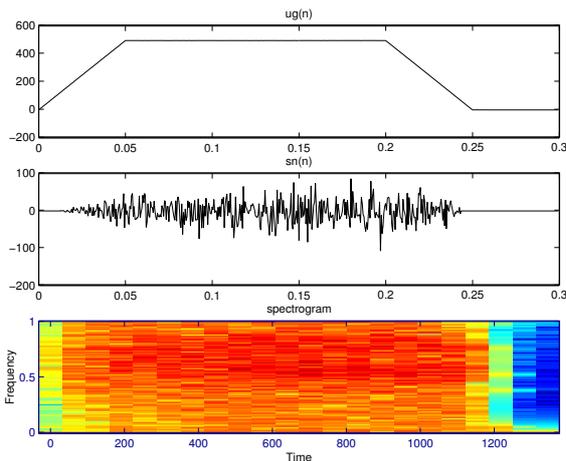


Figure 2: Synthesis results for an unvoiced fricative showing the glottal flow, the speech waveform and spectrogram.

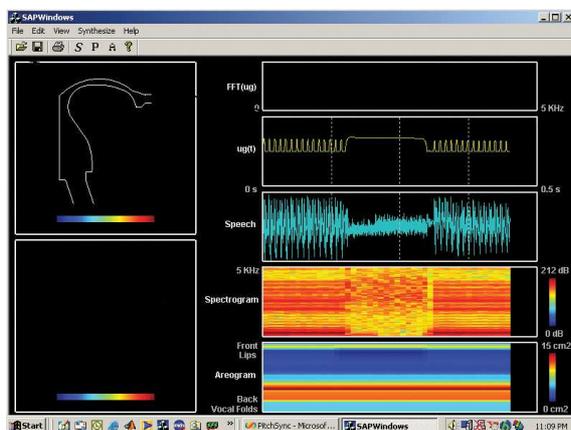


Figure 3: Dump of synthesizer GUI after /fi/ synthesis.

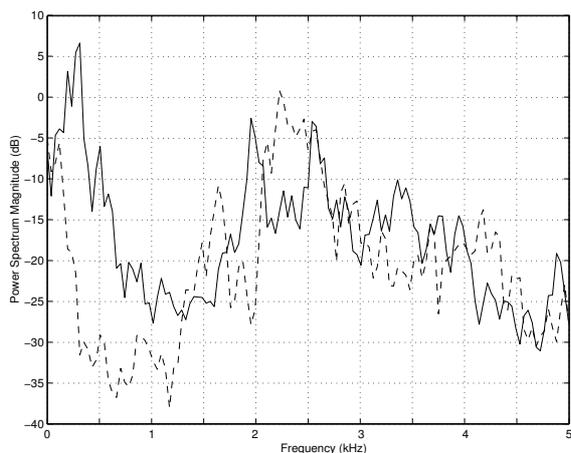


Figure 4: PSD of a synthesized fricative /f/ (solid line) and a natural /f/ (dashed line).

producing sustained fricatives and fricatives in VCV sequences. First results were judged in informal listening tests as being highly intelligible.

Preliminary results of an ongoing work were presented, so further validation and checking of the models is still required. Nevertheless, this is an important new step towards a complete articulatory synthesizer for Portuguese. Our model of fricatives is comprehensive and flexible, making the new version of SAP-Windows a value tool for trying out new or improved source models, and running production and perceptual studies of European Portuguese fricatives. The possibility of automatically inserting and removing noise sources along the oral tract is a feature we regard as having great potential.

## 6. References

- [1] S. S. Narayanan, A. A. H. Alwan, and K. Haker, "An articulatory study of fricative consonants using magnetic resonance imaging," *Journal of the Acoustical Society of America*, vol. 98, no. 3, pp. 1325–1347, 1995.
- [2] C. Scully, E. Castelli, E. Brearley, and M. Shirt, "Analysis and simulation of a speaker's aerodynamic and acoustic patterns for fricatives," *Journal of Phonetics*, vol. 20, no. 1, pp. 39–51, 1992.
- [3] C. H. Shadle, "Articulatory-acoustic relationships in fricative consonants," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds., pp. 187–209. Kluwer Academic, Dordrecht, 1990.
- [4] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer - Verlag, Berlin, second edition, 1972.
- [5] J. L. Flanagan and K. Ishizaka, "Automatic generation of voiceless excitation in a vocal cord - vocal tract speech synthesizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, no. 2, pp. 163–170, 1976.
- [6] M. M. Sondhi and J. Schroeter, "A hybrid time - frequency domain articulatory speech synthesizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, no. 7, pp. 955–967, 1987.
- [7] E. L. Riegelsberger, *The Acoustic-to-Articulatory Mapping of Voiced and Fricated Speech*, Ph.D., Department of Electrical Engineering, The Ohio State University, Ohio, USA, 1997.
- [8] S. S. Narayanan and A. A. H. Alwan, "Noise source models for fricative consonants," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 328–344, 2000.
- [9] L. M. T. Jesus and C. H. Shadle, "A parametric study of the spectral characteristics of European Portuguese fricatives," *Journal of Phonetics*, vol. 30, no. 3, pp. 437–464, 2002.
- [10] P. Badin, "Acoustics of voiceless fricatives: Production theory and data," Quarterly Progress and Status Report 3/1989, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, 1989.
- [11] D. G. Childers, *Speech Processing and Synthesis Toolboxes*, John Wiley & Sons, New York, 2000.
- [12] A. Teixeira, L. Silva, R. Martinez, and F. Vaz, "SAP-Windows - towards a versatile modular articulatory synthesizer," in *Proceedings of the IEEE-SP Workshop on Speech Synthesis*, USA, 2002.