Enhanced area functions for noise source modeling in the vocal tract

Peter Birkholz

Department of Phoniatrics, Pedaudiology, and Communication Disorders
University Hospital Aachen and RWTH Aachen University
peterbirkholz@gmx.de

Abstract

The synthesis of natural-sounding fricatives based on a source-filter model is still a major challenge. While the filter is effectively modeled in terms of the vocal tract area function, noise sources are more difficult to model. Source properties critically depend on aerodynamic conditions and vocal tract geometry near the constriction in a way that is not fully understood. Therefore, noise source models usually assume different relations between noise source parameters and the aerodynamic state for different places of articulation. However, the place of articulation cannot be reliably determined from a dynamically changing area function, as in articulatory synthesis. Here we introduce the concept of an enhanced area function that adds the identity of the articulator that confines the vocal tract at the anterior-inferior side as a new layer of information to the classic area function. This allows to distinguish places of articulation and is therefore an effective representation of the vocal tract not only in terms of filter function but also for noise source modeling. A noise source model on this basis is presented.

Keywords: Area function, noise sources, fricatives, place of articulation

1. Introduction

Fricative production for articulatory speech synthesis is mostly modeled as a source-filter process (e.g., Shadle 1991; Narayanan and Alwan 2000). Hence, synthesis of fricatives requires the specification of source and filter parameters. The filter is usually specified in terms of the vocal tract area function $A(x)$, which describes the cross-sectional area A of the vocal tract normal to the longitudinal dimension x. The area function is an appropriate abstraction of the complex 3D shape of the vocal tract with regard to its acoustic filter effect for the relevant frequencies of up to about 4-5 kHz.

The more difficult problem is the prediction of the noise source parameters, namely the position, amplitude, and spectral shape of potential sources. They critically depend both on the aerodynamic conditions and vocal tract geometry in the vicinity of the turbulent jet, which may substantially differ from one fricative to the other (Shadle 1991; Ramsay and Shadle 2006). While it is not clear, which aspects of the complex shape of the vocal tract are relevant for this, it is sure that the abstract geometric information contained in the area function is not sufficient for predicting source parameters (Shadle et al. 2008). Therefore, the typical approach is to presume different noise source characteristics for different places of articulation. For example, Shadle (1991) identified significant differences between the noise source characteristics for /ʃ/ on the one hand, and for /ç, x/ on the other hand. She termed the corresponding sources *obstacle* and *wall* sources, respectively, "to indicate a critical difference in the geometry presented to the turbulent jet downstream of the constriction". To account for such differences in modeling experiments, Badin, Mawass, and Castelli (1995) modeled the noise source amplitude L (and analogously the spectral tilt) of fricatives with the general equation $L = k \cdot A_c^a \cdot \Delta p^b$, where A_c is the cross-sectional area of the constriction, Δp is the pressure drop across the constriction, and k, a, and b are parameters. The parameter values for the different places of articulation were experimentally determined.

This strategy – to use different values for noise source parameters depending on the place of articulation – currently seems to be the most promising way to synthesize high-quality fricatives. However, it is not straightforward to implement this idea for dynamic articulatory speech synthesis. If we assume the area function as a representation of the vocal tract shape, it is impossible to reliably discriminate places of articulation, and hence to predict position-dependent noise source parameters. For example, an apico-alveolar constriction cannot be distinguished from a labio-dental constriction based on the area function under all circumstances. Also when there are two or more constrictions, it is difficult to tell which of them produces a turbulent jet and hence a noise source. For example, /ʃ/ has two constrictions, one created with the anterior tongue, and one with the incisors (see the area function in Shadle 1991, p. 419), but only the lingual constriction gives rise to a significant turbulent jet. Furthermore, for /s/ and /ʃ/, the major source of noise is usually assumed at the incisors, whose position is not available from the pure area function.

In this paper, we propose the notion of an *enhanced area function* to represent the vocal tract shape. The purpose of the enhanced area function is to allow to determine the place of articulation of fricatives and so to support the optimal parametrization of noise sources. The basic idea is to add the identity of the articulator or structure that confines the vocal tract at the anterior-inferior side as a new layer of information to the area function. Therefore, each position along the longitudinal dimension of the vocal tract (or each tube section in the case of a discrete area function) is associated with a nominal value that specifies the articulator, i.e., the tongue, the lower incisors, or the lower lip. With the enhanced area function, the primary articulator forming a constriction can be identified and it becomes easily possible to distinguish places of articulation. The following section describes the extraction of an enhanced area function from a 3D vocal tract model (Birkholz 2013), and Section 3 presents a simple noise source model on this basis.

2. Model-based extraction of the enhanced area function

Area functions are mostly obtained from either MRI or X-ray images of the vocal tract (e.g., Narayanan, Alwan, and

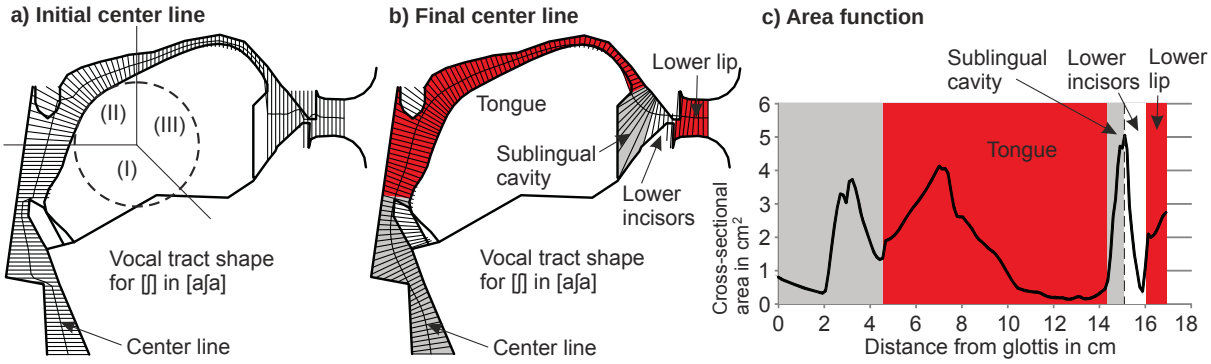

Figure 1: *a) Midsagittal view of the vocal tract model with the initial grid system and the initial center line superimposed on the outline. (I), (II), and (III) mark sectors with horizontal, radial, and vertical grid lines, respectively. The dashed circle represents the tongue body of the model. b) The same vocal tract shape with the final center line and projections of the cutting planes perpendicular to this center line. The vocal tract is segmented in multiple regions that represent different confining articulators or structures. c) Enhanced area function with the same segmentation as in b).*

Haker 1995), or from 2D or 3D models of the vocal tract (e.g., Birkholz 2013). Both the images and the models usually provide the information about the identity of the articulator or structure confining the vocal tract at the anterior-inferior side that is needed for enhanced area functions. Here we describe the calculation of the enhanced area function of the vocal tract model by Birkholz (2013) that is implemented in the articulatory synthesizer VocalTractLab 2.1 (`www.vocaltractlab.de`).

The first task is to find the center line of the vocal tract. Here, it is calculated in two steps (Birkholz 2005). In the first step, a grid system is superimposed on the vocal tract outline in the midsagittal plane, as shown in Figure 1a. This grid comprises a number of closely spaced horizontal grid lines in the pharyngeal region (sector I), vertical grid lines in the oral region (sector III) and radial grid lines in the velar region (sector II). The most inferior horizontal grid line represents the position of the glottis, and the most anterior vertical grid line represents the vocal tract termination at the mouth. The latter is positioned about half-way between the corners of the mouth and the most anterior points of the lips. In this way, the acoustic effect of the notch-shaped vocal tract termination at the lips is roughly accounted for according to the data by Lindblom et al. (2007). The boundaries between the three sectors intersect in the center point of a circle that represents the (movable) tongue body in the vocal tract model (cf. Figure 1a).

Each grid line intersects the posterior-superior outline and the anterior-inferior outline of the vocal tract as shown in Figure 1a. The first estimate of the center line is the sequence of straight-line segments joining the midpoints of the grid lines delimitated by the outlines. At positions where the cross-sectional area of the vocal tract changes abruptly, this center line exhibits sharp bends, which are unlikely to represent the true path of acoustic wave propagation. Therefore, this initial center line is smoothed in the second step with a 2 cm long moving average filter to obtain the final center line that is shown in Figure 1b. At each of 129 equally-spaced points along the final center line, the 3D wire-frame meshes that constitute the vocal tract model are intersected with a plane perpendicular to the center line in the respective point. For each cut, the cross-sectional area is obtained as one sample of the area function. Figure 1c shows the piece-wise linear area function corresponding to the vocal tract shape and center line in Figure 1b.

The idea of the *enhanced* area function was to associate each position x along the tube axis not only with the cross-sectional area $A(x)$, but also with the articulator $\alpha(x)$ that confines the vocal tract at the anterior-inferior side. In our model, this information is directly available from the identity of the wire-frame mesh that is intersected by each cutting plane at the anterior-inferior end. In the current implementation, the associated articulator α can assume one of four nominal values, namely "tongue", "lower incisors", "lower lip", and "other". In Figures 1b and c, the segments for tongue and lower lip are red, the segment for the lower incisors is white, and the remaining segments for the laryngeal region and for the sublingual cavity are gray.

For computational reasons, the vocal tract is mostly represented in terms of a *discrete* area function that corresponds to a sequence of incremental cylindrical tube sections. In our case, this requires to map the piece-wise linear area function $A(x)$ and the associated articulators $\alpha(x)$ to N cylindrical tube sections, each having a cross-sectional area A_i, a length l_i, and an associated articulator α_i ($0 \leq i < N$). Hence, if $x_i = \sum_{k=0}^{i-1} l_i$ (with $x_0 = 0$ and $1 \leq i < N$) denotes the position of the ith tube section, $A(x)$ and $\alpha(x)$ in the intervals $[x_i, x_i + l_i]$ have to be mapped to single values for A_i and α_i, respectively. With regard to the area, the obvious approach would be to assign A_i the average of $A(x)$ in the respective interval. However, in this way, very short closures in the vocal tract (shorter than a tube section length) may be "released" and become a narrow constriction instead, because the areas greater than zero directly next to the closure contribute to the average value. To prevent this problem, we propose to use the minimum of $A(x)$ in the interval, i.e., $A_i = \min_{x=x_i}^{x_i+l_i}\{A(x)\}$. With regard to the associated articulator, $\alpha(x)$ may have different nominal values in the interval $[x_i, x_i + l_i]$. Here, α_i should take the value of $\alpha(x)$ at the position of the minimal area in the interval, i.e., $\alpha_i = \alpha(x_0)$ with $x_0 = \arg\min_{x \in [x_i, x_i + l_i]} A(x)$, because this is the relevant articulator in the case that the tube section forms a constriction for noise generation.

As a compromise between low computational cost, which requires as few tube sections as possible, and high spatial detail, which requires as many sections as possible, we represent the area function with longer tube sections in the posterior part, and shorter tube sections in the anterior part, where most fricatives are produced and spatial accuracy is preferable. In fact, we

use 16 tube sections of equal length between the glottis and the velo-pharyngeal port position, and an additional 24 tube sections with decreasing length between the velo-pharyngeal port and the mouth opening (40 sections in total). As an example for the discretization, Figure 2a shows the discrete enhanced area function corresponding to the piece-wise linear area function in Figure 1c.

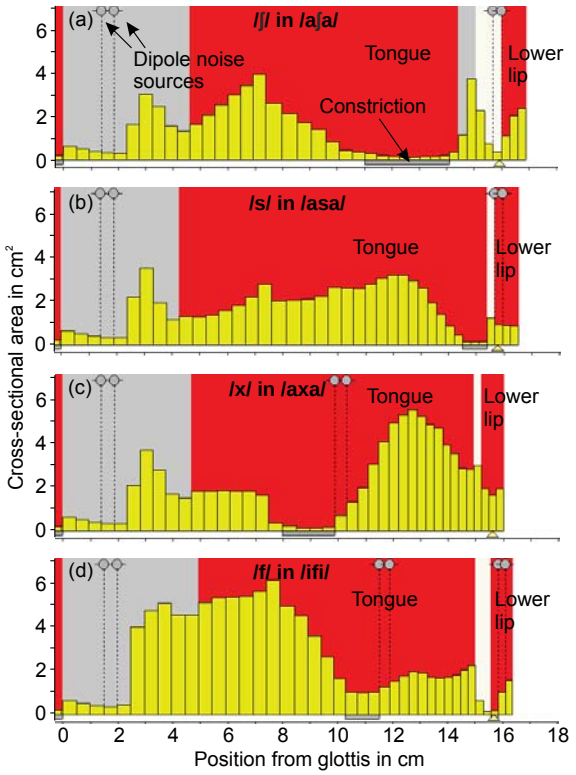

Figure 2: *Discrete enhanced area functions of four fricatives. Noise source locations are marked by gray circles, constricted regions are marked by gray horizontal bars, and the positions of the upper incisors are marked by yellow triangles.*

3. Noise source modeling

Based on the discrete enhanced area functions, we propose the following method for predicting noise sources. First of all, potential critical constrictions formed with the tongue or the lower lip are identified. Therefore, in each of the two regions where α_i ="tongue" and α_i ="lower lip", the tube section with the smallest area $A_{\min}$ is identified. If $A_{\min} < 1$ cm^2, all contiguous tube sections left and right of this section for which $A_i < A_{\min} + 0.2$ cm^2 are considered as one continuous constriction. In each example in Figure 2, there is one such constriction in the lingual region (marked by the gray bars below the area functions). In the example for /f/ in Figure 2d, there is also such a constriction formed with the posterior section of the lower lip region (the gray bar is hidden under the yellow triangle). In the lingual region, there may generally be a second constriction if it is not connected to the first one and satisfies $A_{\min} < 1$ cm^2. This is important to prevent acoustic artifacts when a constriction in the lingual region suddenly "jumps" from one position to another in running speech, for example from the tongue tip to the tongue back in /su/ when the tongue tip constriction is released so far that it becomes greater than the tongue back constriction of /u/. As Figure 2 shows, the area function has been extended with two short tube sections to the left that represent the glottis. These glottis sections are considered as a permanent additional constriction for aspiration noise with a variable gain.

Each of the identified constrictions gives rise to one noise source. All noise sources are modeled as localized turbulent sound pressure sources and characterized by their magnitude, spectral shape and location. According to Stevens (1998), p. 103, the source magnitude is calculated as $p_s = G \cdot |\overline{v}_c|^3 \sqrt{A_{\min}}$, where G is a position-dependent gain and $\overline{v}_c$ is the low-frequency part of the air particle velocity in the constriction. The latter was obtained by low-pass filtering the actual air particle velocity v_c with a first-order low-pass filter with a cutoff frequency of 500 Hz. The values of G for different constriction locations were determined in re-synthesis experiments as described further below.

Each noise source generates Gaussian white noise that is shaped with a first-order low-pass filter with a specific cutoff frequency f_c. The filter produces a high-frequency tilt of the source spectrum that is evident from a variety of experimental studies (Narayanan and Alwan 2000). For lingual constrictions, f_c is set to $0.15 \cdot v_c/d$ (following Stevens 1998, p. 104), where $d = \sqrt{4 \cdot A_{\min}/\pi}$ is the diameter of the constriction. For labiodental and glottal constrictions, f_c is set to 6000 Hz to generate an essentially flat source spectrum (Stevens 1998).

The position of a noise source in our model depends on the source type. When x_{sep} denotes the assumed point of flow separation at the anterior end of the respective constriction, and D is the distance between x_{sep} and the tip of the incisors, the noise source is assumed 0.15 cm downstream of x_{sep} for a labiodental source, 0.25 cm downstream of x_{sep} if $D > 2$ cm ("wall source" type), directly at the tip of the incisors when $D \leq 2$ cm, and 1.5 cm above the glottis for the glottal constriction. The information about the position of the upper incisors is currently not contained in the enhanced area function but obtained separately from the vocal tract model.

In the transmission-line model of the vocal tract underlying the acoustic simulation, actual noise sources can only be inserted at boundaries between tube sections (Birkholz, Jackèl, and Kröger 2007). Therefore, a predicted noise source located at some point within a tube section is always realized as *two* actual pressure sources in the transmission-line network – one at each end of the section, with the amplitudes linearly scaled according to the distance between the predicted and actual sources. The two actual sources generated for each constriction are shown as gray circles in Figure 2.

All source properties discussed above are summarized in Table 1. The relative gain G of the different types of sources was determined with analysis-by-synthesis. Therefore, a natural production of the utterance /afasaʃaçaχa/, spoken with a flat intonation, was re-synthesized with VocalTractLab 2.1, where the proposed noise source model was implemented. The values of G for the different noise source types were incrementally adjusted such that the noise levels of the synthetic fricatives approached the levels of the natural fricatives. Figure 3 shows the natural and (final) synthetic productions of the utterance one below the other for comparison. While the levels of the synthetic productions of /f,s,ʃ,χ/ are very similar to the original levels, it is somewhat too high for the synthetic /ç/ (which was realized with an "obstacle source" like /s, ʃ/).

Table 1: *Properties of modeled noise sources. v_c is air particle velocity in the constriction and d is constriction diameter.*

Articulator forming the constriction	Glottis	Tongue		Lower lip
Source type	Aspiration source	Wall source	Obstacle source at the incisors	Labiodental source
Typical phonemes	many	/χ, ʁ/	/s, z, ʃ, ʒ, ç, j/	/f,v/
Condition for source	none	Distance from flow separation point to tip of upper incisors greater than 2 cm	Distance from flow separation point to tip of upper incisors smaller than 2 cm	Constriction area must be smaller than that of a potential tongue tip constriction
Source position	1.5 cm downstr. from the flow separation point (glottal exit)	0.25 cm downstream from the flow separation point	Tip of the upper incisors	0.15 cm downstream from the flow separation point
Cuttoff freq. f_c in Hz	6000.0	$0.15 \cdot v_c/d$	$0.15 \cdot v_c/d$	6000.0
Relative gain G	$0.005 \ldots 0.5$	5.0	10.0	3.0

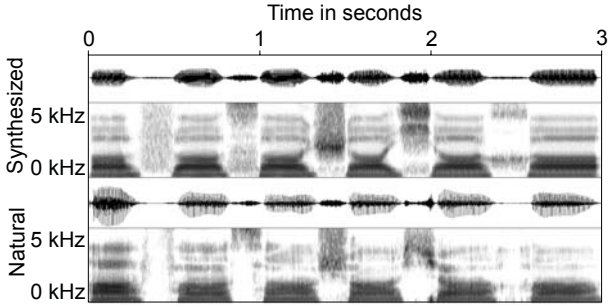

Figure 3: *Oscillograms and spectrograms of the utterance /afasaʃaçaχa/ - natural production (bottom) and re-synthesis (top).*

4. Discussion and conclusion

The main goal of this paper was to introduce the idea of the enhanced area function. It was devised as a coherent extension of the classic area function to support modeling of noise sources, without giving up the simplicity of the area function as an abstract representation of the complex 3D shape of the vocal tract for acoustic simulations. Based on the enhanced area function, a noise source model was presented, which was recently implemented in the articulatory speech synthesizer VocalTractLab 2.1 (`www.vocaltractlab.de`). According to informal listening tests, the model could generate all German fricatives in high quality in connected synthetic utterances.

The noise source model can still be improved in many ways. For example, not only dipole sources contribute to the spectrum of fricatives, but also monopole sources should be modeled. Furthermore, in voiced fricatives, the frication source is known to be modulated by voicing, where the phase of modulation depends on the distance between constriction and obstacle (Jackson and Shadle 2000). This delay in the modulation is perceptually relevant and should be considered. Also the fixed threshold of 2 cm between the exit of a lingual constriction and the incisors to discriminate between wall and obstacle sources should be replaced by a smooth "blending" of sources. Finally, a formal perceptual evaluation of the synthesized fricatives is needed.

5. References

Badin, P., K. Mawass, and E. Castelli (1995). "A model of frication noise source based on data from fricative consonants in vowel context". In: *13th International Congress of Phonetic Sciences (ICPhS 1995)*. Stockholm, Sweden, pp. 202–205.

Birkholz, P. (2005). *3D-Artikulatorische Sprachsynthese*. Logos Verlag Berlin.

— (2013). "Modeling consonant-vowel coarticulation for articulatory speech synthesis". In: *PLoS ONE* 8.4, e60603.

Birkholz, P., D. Jackèl, and B. J. Kröger (2007). "Simulation of Losses Due to Turbulence in the Time-Varying Vocal System". In: *IEEE Transactions on Audio, Speech and Language Processing* 15.4, pp. 1218–1226.

Jackson, P. J. B. and C. H. Shadle (2000). "Aero-acoustic modelling of voiced and unvoiced fricatives based on MRI data". In: *5th Seminar on Speech Production*. Seeon, Bavaria, pp. 185–188.

Lindblom, B., J. Sundberg, P. Branderud, and H. Djamshidpey (2007). "On the acoustics of spread lips". In: *Proceedings of Fonetik, TMH-QPSR* 50.1, pp. 13–16.

Narayanan, S. S. and A. A. Alwan (2000). "Noise source models for fricative consonants". In: *IEEE Transactions on Speech and Audio Processing* 8.3, pp. 328–344.

Narayanan, S. S., A. A. Alwan, and K. Haker (1995). "An articulatory study of fricative consonants using magnetic resonance imaging". In: *Journal of the Acoustical Society of America* 98.3, pp. 1325–1347.

Ramsay, G. and C. Shadle (2006). "The influence of geometry on the initiation of turbulence in the vocal tract during the production of fricatives". In: *7th International Seminar on Speech Production*. Ubatuba, Brazil, pp. 581–588.

Shadle, C. H. (1991). "The effect of geometry on source mechanisms of fricative consonants". In: *Journal of Phonetics* 19, pp. 409–424.

Shadle, C. H., M. Berezina, M. Proctor, and K. Iskarous (2008). "Mechanical models of fricatives based on MRI-derived vocal tract shapes". In: *8th International Seminar on Speech Production (ISSP 2008)*. Strasbourg, France, pp. 413–416.

Stevens, K. N. (1998). *Acoustic Phonetics*. The MIT Press, Cambridge, Massachusetts.