AERO-ACOUSTIC MODELLING OF VOICED AND UNVOICED FRICATIVES BASED ON MRI DATA

Philip J.B. Jackson* and Christine H. Shadle†
*School of Electronics and Electrical Engineering, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [p.jackson@bham.ac.uk][1]
†Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK. [chs@ecs.soton.ac.uk]

ABSTRACT

Area functions generated from dynamic MRI data were used with the program VOAC to generate transfer functions for [ɑ, i, s] from the repeated word [pʰɑsi]. A decomposition technique was applied to acoustic recordings of the same subject to derive modulated noise sources for voiced fricatives. Using a standard glottal source and both constant and modulated noise sources, [ɑ, i, s, z] were synthesized.

1. INTRODUCTION

We would like to develop a more realistic production model of unvoiced speech sounds, namely fricatives, plosives and aspiration noise. All three involve turbulence noise generation, with place-dependent source characteristics that vary with time (rapidly, in plosives). In this study, we aimed to produce, using an aero-acoustic model of the vocal-tract filter and source, voiced as well as unvoiced fricatives that provide a good match to analyses of speech recordings.

Our approach builds upon many aspects of our research. The vocal-tract acoustics program, VOAC (Davies et al. 1993), has been shown to provide predictions that closely match measurements in flow duct experiments (Jackson and Shadle 1999b). Our geometrical information comes from a dynamic magnetic resonance imaging (dMRI) study (Mohammad 1999), whose vowel data have already been used to produce area functions (Shadle et al. 1999). Here, for the first time, we report the use of VOAC to compute the vocal-tract transfer function (VTTF) for a fricative, for which its aero-acoustic formulation is most advantageous.

The development of an algorithm to decompose a speech signal into its harmonic and anharmonic components (Jackson and Shadle 1998; 2000) has offered the opportunity to analyse these estimates of the voiced and unvoiced signals separately. Indeed, such decomposition enables the interaction of sources to be examined, as in (Jackson and Shadle 1999a), which revealed a relationship between the place of articulation and the phase lag of amplitude modulation of the noise component, in relation to voicing.

[1]This research was conducted while the first author was at Dept. Electronics and Comp. Sci., Univ. of Southampton, UK.

In this paper, we combine our new knowledge about the source and filter to synthesize a voiced and unvoiced fricative pair /z, s/, and apply the observed phase delay to the frication noise envelope. The results of this hybrid synthesis are discussed and compared with recordings of the same phonemes spoken by the same subject.

2. VOCAL-TRACT FILTER

Our synthesis model was initially based on the assumption that the acoustic source and vocal-tract filter are independent. VOAC computes the VTTF from geometrical data, in the form of cross-sectional area and hydraulic radius functions, along the length of the tract. Although linear in sound pressure, the effects of flow may be highly nonlinear and exhibit strong source interactions. It has been noted (Scully 1990; Narayanan et al. 1995) that flow engenders additional losses. For our model to embrace these features, VOAC incorporates the effects of net flow into the transmission of plane waves through a tubular representation of the tract, and relaxes assumptions of rigid walls and isentropic propagation. The geometry functions were derived from multiple-slice dMRI data (Mohammad 1999; Shadle et al. 1999), using a method of converting from the pixel outlines that was improved over earlier efforts on vowels (Jackson and Shadle 1999b).

Vocal-tract images were acquired by dMRI in three sagittal planes using a 0.5 T SIGNA scanner (fast RF-spoiled GE), as part of an associated project (Mohammad 1999). The subject, adult male PJ, is a native speaker of British English RP. The corpus consisted of repetitions of the nonsense word [pʰɑsi]. Connected vocal-tract outlines were extracted from the images by linking manually-determined boundaries of the upper lip, upper teeth, hard palate, soft palate, velum, the back of the pharynx, the larynx, epiglottis, the tongue from root to tip, the lower teeth and lower lip. A grid was superimposed on the outline from each image slice, comprising a set of horizontal lines $1\frac{1}{2}$ pixels apart (2.8 mm), 18 radial lines separated by $\pi/32$ rad and another set of parallel lines declining at $\pi/16$ rad, as shown in Figure 1. The origin was located at the tongue centroid. The intercepts were found, from which the cross-sectional distance was calculated for each grid line. The length along the vocal tract was defined as the perpendicular distance between parallel grid lines and

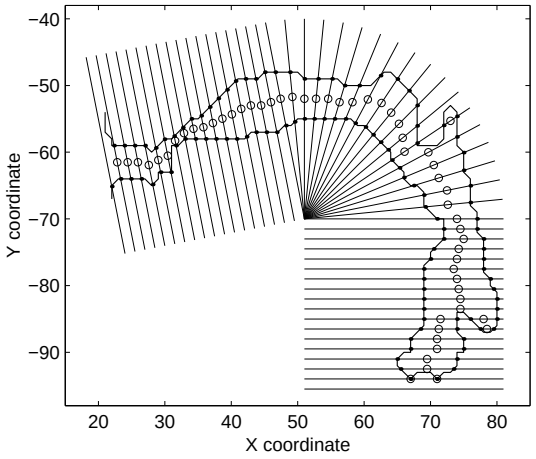

Figure 1: Grid lines (thin) superimposed on the outline from the mid-sagittal slice (thick) for [s] spoken by PJ (frame 21). The intercepts are marked by dots, and circles indicate the mid-point of each vocal tract section.

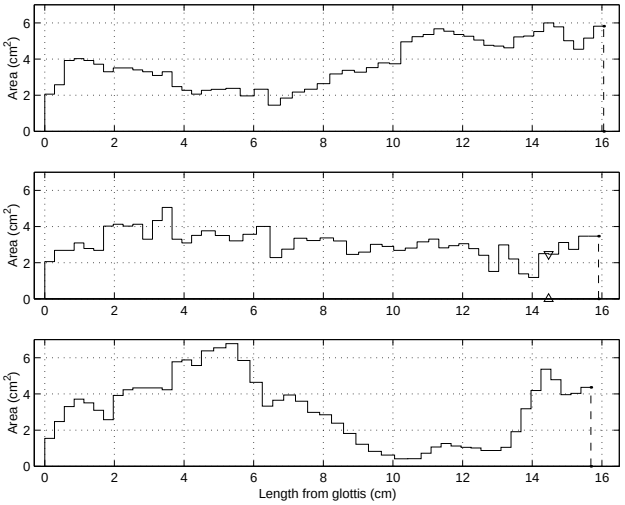

Figure 2: Area functions combining sagittal slices from the mid-points of three phones from the dynamic MRI data of [p^{h}ɑsi] by PJ: (top) vowel [ɑ], (middle) fricative [s], and (bottom) vowel [i]. The radiation surface is shown as a dashed line and, for the fricative [s], the location of the pressure source was at the teeth, as indicated by triangles.

as the length of the arc for radial lines, using the mid-point as the effective radius. Side branches were also identified and their details stored with the place and sense of attachment to the main tract.

The total cross-sectional area at each grid line was calculated as the sum of the distance measurements, multiplied by the inter-slice spacing of 11 mm. The hydraulic radius, the ratio of twice the area to the perimeter, was estimated by assuming an elliptical cross section, and using the mid-sagittal distance as one of its axes.

Three area functions obtained by this technique are plotted in Figure 2, which correspond to the frames midway through the phones in the nonsense word: [ɑ], [s] and [i]. In these examples, the side branches were discarded. The area functions for the vowels look about as we would expect. That for [s] has an atypically large constriction, of approximately 1 cm^2, at 14 cm from the glottis. Narayanan et al. (1995) report minimum constriction areas for [s] of 0.1–0.3 cm^2 for their four subjects. Their subjects sustained the fricatives, which would tend to result in smaller constrictions, but the discrepancy is still large.

The resolution of the dMRI images is 1 pixel within the plane of a slice, which corresponds for these images to 1.875 mm, and an area of 0.2 cm^2. If each of the three slices has a sagittal distance from tongue to palate of one pixel, the minimum constriction area is therefore 0.6 cm^2. In the mid-fricative frame, the minimum distance across the constriction in each slice was one pixel, but these points were not precisely the same length from the glottis. The rapidity of the area change for [s] also acts to decrease the dMRI resolution. Finally, the position of the teeth within the image was estimated manually by adjusting the brightness and contrast of the images and making a judgement, introducing additional uncertainty in the vicinity of the constriction.

We have chosen to synthesize with this area function rather than to attempt to correct it at this stage. Further, we use this area function to synthesize /z/ as well as /s/. First, we cannot justify decreasing the area in the constriction simply because we expect it to be smaller and know that that is an acoustically sensitive dimension. Second, we are interested in the performance of the entire chain, of which nearly every component has a novel aspect; it is more valid to compare results from varying source functions (e.g. from constant to two types of modulated noise) than to attempt to optimize the area function derived from dMRI. Third, Narayanan et al. (1995) note only small differences between [s] and [z] for their subjects; the range of constriction areas is similar (from 0.12 to 0.25 cm^2 for [z]).

For a volume-velocity source at the glottis U_G, such as voicing, the volume velocity at the lips was calculated from the volume-velocity VTTF:

$$H^V_{GL}(f) = \frac{U_L(f)}{U_G(f)}\,. \qquad (1)$$

The radiation from the lips was approximated by a piston in an infinite baffle (Beranek 1954), which was used to determine the reflection coefficient. Modelled as a pressure source within the tract, the frication source p_Q induces a response to the waves travelling both upstream (towards the glottis) and downstream (towards the lips). In linear models, the overall VTTF from the source to the lips is equal to the product of two transfer functions:

$$H^P_{QL}(f) = \frac{U_G(f)}{p_Q(f)}\,\frac{U_L(f)}{U_G(f)} = H^P_{QG}(f)\,H^V_{GL}(f)\,, \qquad (2)$$

where H^P_{QG} is the pressure transfer function of the rear-tract, that part upstream of the source, which uses a reflection coefficient $R = 1$ at Q. To predict the far-field sound p_{ff} radiated from U_L at $r = 0.3$ m, the VTTFs were

multiplied by the radiation factor $\rho f/r$, where ρ is the density of air.

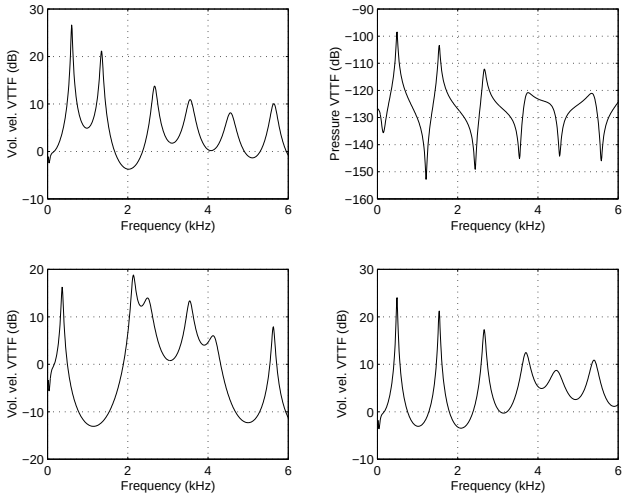

Figure 3: Transfer functions predicted by VOAC for the three phones, [ɑ], [i] and [s] for volume velocity sources at the glottis (top left, bottom left, and bottom right, respectively), and a pressure source downstream of the constriction for the fricative (top right).

The VTTFs for [ɑ], [i] and [s] are given in Figure 3. The formants differ between [ɑ] and [i] as expected, and approximate those observed for the same vowel and subject (Jackson and Shadle 1999b): (F1, F2, F3) = 0.60, 1.34, 2.67 kHz for [ɑ] and 0.36, 2.14, 2.49 kHz for [i]. However, neither VTTF is as expected for [s], though they do correspond well to the area function in Fig. 3. The pressure VTTF for the fricative, $H^P_{QL}(f)$, shows the spectral zeros introduced by the rear-tract transfer function, which is a characteristic of localised supraglottal sources. Note that the area functions from which these VTTFs were calculated were derived directly from the dMRI data, and no attempt has been made to modify them in relation to observations, e.g. (Beautemps et al. 1995).

3. SOURCE MODELS

To synthesize the unvoiced fricative /s/, white noise spectrum $N(f)$ was coloured, according to Shadle's (1985) estimates:

$$D(f) = N(f)\, a \exp(bf)\ , \qquad (3)$$

where $a = 95$ and $b = -0.0004$ are constants, and $D(f)$ is the source spectrum. The frication source was convolved with the impulse response of the pressure VTTF from source to lips and the radiation characteristic, yielding a stationary noise signal, /s/. For its voiced counterpart /z/, the voiced source was generated using a cubic waveform, as in (Klatt 1987), with an open quotient of 0.5 and fundamental frequency $f_0 \approx 130$ Hz. Filtering it by the volume-velocity VTTF from the glottis to the lips gave the voiced component.

Many researchers have noted that the frication source appears to be modulated by voicing, e.g. (Fant 1960), and the phase of the modulation has been shown to be perceptually significant (Hermes 1991). Our analysis (Jackson and Shadle 1999a) confirmed that the noise component varied periodically according to fluctuations in the flow velocity at the constriction exit. Moreover, the modulation phase appeared to be governed by the convection time for the flow perturbation to travel from the constriction to the obstacle. Therefore, to synthesize the voiced fricative, the frication source $d(n)$ was modulated by the voice source $g(n)$ and the phase of its envelope delayed to match our empirical observations:

$$\hat{d}(n) = d(n)\, g(n - \tau)\,, \qquad (4)$$

where the delay time τ was in the range 2.8–3.8 ms.

4. RESULTS

Figure 4 shows a portion of the synthesized voiced fricative /z/ with its constituent components, and a sustained [z] recorded by speaker PJ. The pitch-scaled harmonic filter (PSHF, Jackson and Shadle 2000) was used to decompose the speech recording into its harmonic and anharmonic components, which are estimates of the voiced and unvoiced signals, respectively. It is clear from the anharmonic signal that the noise has been modulated and that the peak amplitude is not synchronous with the glottal pulses, seen in the harmonic signal, although they share the same periodicity. In the synthetic components, the first formant dominates, yet the pulse-like excitation of the voiced components at glottal closure and the modulated envelope of the frication noise are characteristics echoed in the real signals.

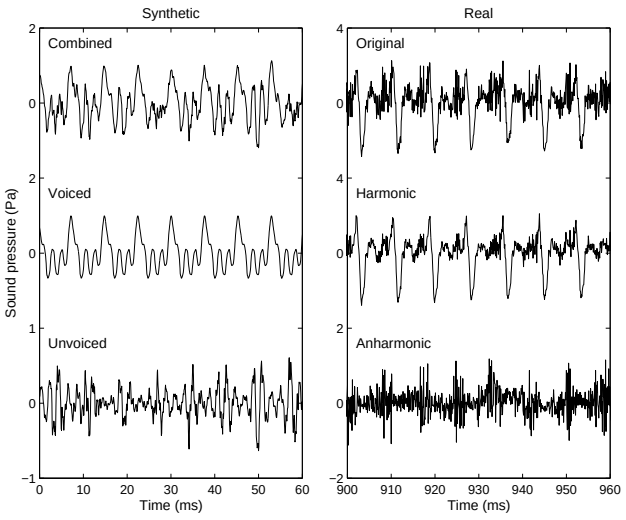

Figure 4: Synthetic (left) and real (right) signals for a sustained /z/ sound: (bottom, double amplitude scale) fricative component from a modulated noise source, (middle) voiced component, and (top) the combined signal.

Examples without the phase lag and with no modulation were also prepared for subjective assessment. These three synthetic examples of /z/ were all given the same

harmonics-to-noise ratio (HNR = 6 dB): no modulation, modulation in-phase with the glottal waveform, and delayed modulation.[2] Simple listening tests of the synthetic /z/ examples gave the following subjective impressions. None of the examples sounded like a /z/, in part because the synthetic fricatives were presented without any transitions, and probably also because of the problems with the area function already noted. With the constant noise source, the noise seemed detached and the example unnatural. The two modulated noise source examples did not give this detached impression, and differed perceptibly from each other. The modulated noise source with the delayed phase relation, as found in real speech, sounded the most natural of the three.

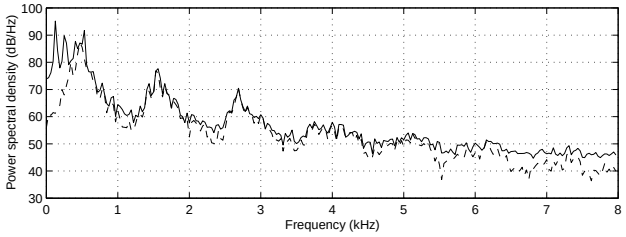

Figure 5: Power spectra of the synthetic voiced and unvoiced fricative pair: (solid) [z] and (dashed) [s].

The spectra of the voiced and unvoiced fricatives, shown in Figure 5, illustrate the effect of the voiced component. The voiceless spectrum has three prominent formant peaks (0.49, 1.54, 2.66 kHz), which give way to broader humps at higher frequencies (> 3 kHz), as seen in the VTTF (Fig. 3, top right). These formants are sometimes evident in measured spectra, but they usually peak in the 4–7 kHz band, and have a positive spectral tilt (Shadle 1990; Jackson and Shadle 2000). The presence of voicing has the effect of smoothing the spectrum at higher frequencies, as well as adding peaks at the first few harmonics of the fundamental frequency, f_0.

5. DISCUSSION

Our interpretation of the dMRI vocal-tract outlines resulted in reasonable area functions for [ɑ], [i] and [s], although we lacked resolution for the region near the fricative constriction. Differences in this region may also be attributed to the dynamic context in which the [s] was produced. Future attempts might seek to resolve this deficiency by incorporating data from electropalatography or static MRI. By incorporating the delay observed between peaks in the glottal waveform and the envelope of the turbulence noise, we have created a source model that is more realistic from a physical and aero-acoustic perspective. In future, we plan to include transitions within phonemes (based on the adjacent dMRI frames), and between phonemes; these and improvements to other aspects of the synthesis will justify the use of more formal listening tests. Finally, we plan to include other speech sounds, e.g. stop consonants, and data for other subjects.

[2] The sound (`.wav`) files can be found on the project web site: http://www.isis.ecs.soton.ac.uk/research/projects/nephthys/.

References

Beautemps, D., P. Badin, and R. Laboissière (1995). Deriving vocal-tract functions from mid-sagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data. *Speech Comm. 16*, 27–47.

Beranek, L. (1954). *Acoustics* (1st ed.). New York, NY: McGraw-Hill.

Davies, P., R. McGowan, and C. Shadle (1993). Practical flow duct acoustics applied to the vocal tract. *Vocal Fold Physiology: Frontiers in Basic Science,* ed. I.R. Titze, Singular Pub., San Diego, CA, 93–142.

Fant, G. (1960). *Acoustic Theory of Speech Production.* The Hague, Netherlands: Mouton.

Hermes, D. (1991). Synthesis of breathy vowels: some research methods. *Speech Comm. 10*(5-6), 497–502.

Jackson, P. and C. Shadle (1998). Pitch-synchronous decomposition of mixed-source speech signals. *Proc. Joint Int. Cong. on Acoust. and Acoust. Soc. Am.,* Seattle, WA *1*, 263–264.

Jackson, P. and C. Shadle (1999a). Analysis of mixed-source speech sounds: aspiration, voiced fricatives and breathiness. *Int. Conf. on Voice Phys. and Biomechanics,* Berlin, FRG.

Jackson, P. and C. Shadle (1999b). Modelling vocal-tract acoustics validated by flow experiments (abstract). *J. Acoust. Soc. Am. 105*(2, Pt. 2), 1161.

Jackson, P. and C. Shadle (2000). Performance of the pitch-scaled harmonic filter and applications in speech analysis. *Proc. IEEE-ICASSP*, Istanbul.

Klatt, D. (1987). Review of text-to-speech conversion for English. *J. Acoust. Soc. Am. 82*(3), 737–793.

Mohammad, M. (1999). *Dynamic measurements of speech articulators using Magnetic Resonance Imaging.* Ph. D. thesis, Dept. Electronics & Comp. Sci., Univ. of Southampton, UK.

Narayanan, S., A. Alwan, and K. Haker (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *J. Acoust. Soc. Am. 98*(3), 1325–1347.

Scully, C. (1990). Articulatory synthesis. In W. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modelling*, pp. 151–186. Kluwer Academic.

Shadle, C. (1985). *The acoustics of fricative consonants.* Ph. D. thesis, RLE Tech. Rep. 506, MIT, Cambridge, MA.

Shadle, C. (1990). Articulatory-acoustic relationships in fricative consonants. *Speech Production and Speech Modelling,* eds. W.J. Hardcastle and A. Marchal, Kluwer Academic, Netherlands, 187–209.

Shadle, C., M. Mohammad, J. Carter, and P. Jackson (1999). Multi-planar dynamic Magnetic Resonance Imaging: New tools for speech research. *Proc. ICPhS, San Francisco, CA 1*, 623–626.