Measures of voiced frication for automatic classification

Philip JB Jackson[1], Luis MT Jesus[2],
Christine H Shadle[3], Jonathan Pincas[1]

[1] Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford, GU2 7XH, UK.

[2] Escola Superior de Saúde da Universidade de Aveiro, and
Instituto de Engenharia Electrónica e Telemática de Aveiro,
Universidade de Aveiro, 3810-193 Aveiro, Portugal.

[3] School of Electronics and Computer Science,
University of Southampton, Southampton, SO17 1BJ, UK.

UniS

Abstract

To characterize acoustic sources in voiced fricatives, it seems apt to use vowel and voiceless-fricative results. However, having both phonation and frication in these mixed-source sounds creates *interaction effects*, that vary across place of articulation. Acoustic and articulatory interactions are examined and automatic statistical techniques employed to describe these phenomena quantitatively, which could help in speech recognition and phone classification. The study focuses on three types of information:

1. spectral measures of the acoustic signal during the fricative,
2. durations of acoustic and laryngeal events related to sources,
3. modulation of frication noise by glottal vibration.

Observed interaction effects are analyzed for British English and European Portuguese speech recordings.

Topic: speech analysis.

Acknowledgement

This collaborative project was supported by the participating institutions.

Presented at 147$^{\text{th}}$ meeting of the Acoustical Society of America, New York NY, May 2004.

Voicing and Frication

Linear superposition of individual sources fails to account for:

1. relative strengths of voicing and frication;
2. corresponding changes in amplitude and spectral tilt;
3. periods of devoicing in voiced fricatives [2,3];
4. periods of source overlap in voiceless fricatives [6,7];
5. modulation of frication by voicing [2].

We refer to these deviations from the traditional single-source theory as mixed-source *interactions*.

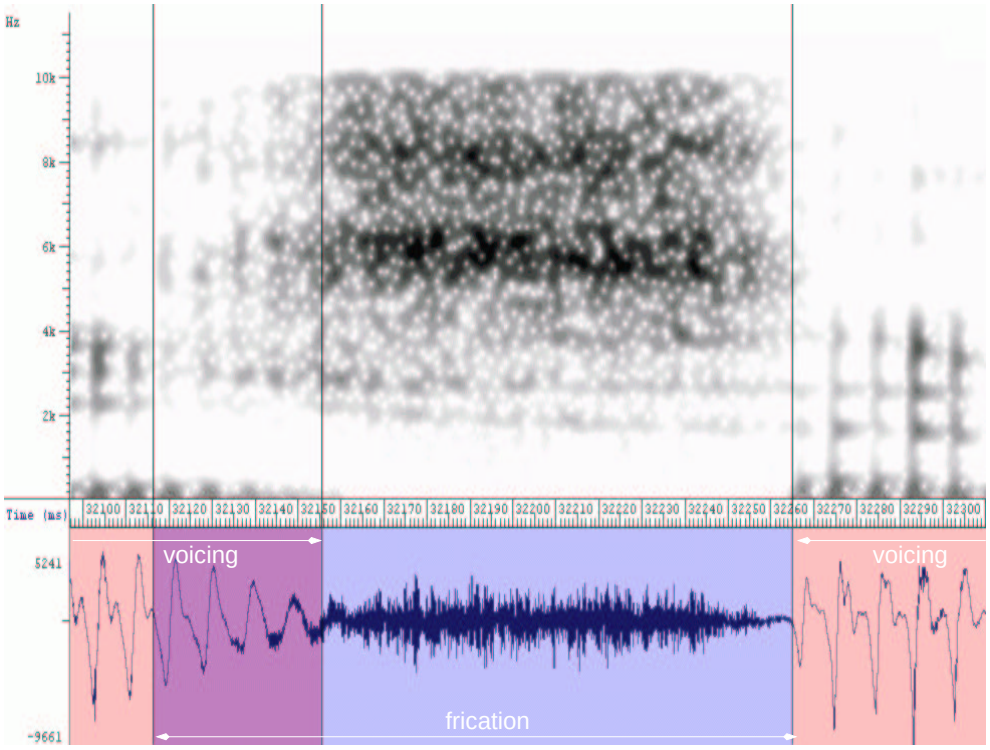


Figure 1: Waveform and spectrogram of a /VFV/ syllable showing voicing (red), frication (blue) and voicing-frication source overlap (purple).

Speech data

European Portuguese corpus. Two male (LMTJ, CFGA) and two female (ACC, ISSS) adult native speakers were digitally recorded at 48 kHz in a sound-treated room: B & K microphne at 1 m and Laryngograph EGG.

"Diga /FV_1FV_2/, por favor." where F and/or V may be deleted for words with an initial vowel or final fricative, vowel context included V=/a, u, i/, and fricative F=/f, v, s, z, sh, zh/ (approx. 160 sentences/speaker).

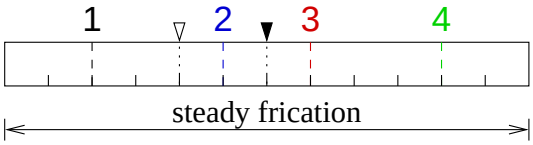
British English corpus. Two male (JP, PJ) and two female (AT, RG) adult native speakers were digitally recorded at 44.1 kHz in an acoustically-screened cubicle: Beyerdynamic microphone at 0.3 m.

"What does /VFe/ mean?", where vowel context V=/a, u, i/, fricative F=/f, v, th, dh, s, z, sh, zh/ and /e/ denotes a neutral schwa vowel (9 reps of each VF combination = 216 sentences/speaker).

Speaking rate was controlled by playing (through a single-ear headphone) prompt sentences in time to a beat followed by a pause.

Spectra

Multitaper spectra (PSD Thomson estimates) were calculated with 11-ms windows centered at four positions (1-4) within each fricative, in twelfths:



where empty and solid triangles show limit of *devoiced* and *partially devoiced*.

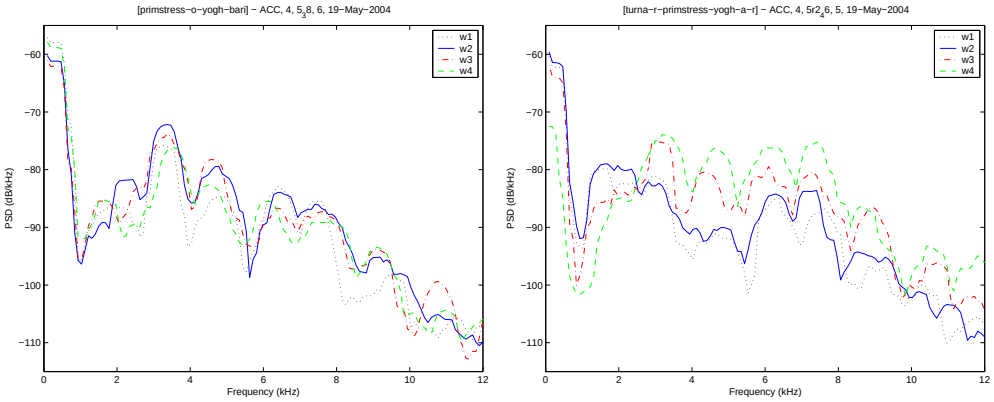


Figure 2: Multitaper spectra of (left) fully-voiced and (right) partially-devoiced postalveolar fricatives /zh/ (spkr. ACC), at the four window positions: w1–w4.

Fricative spectra were parameterised, as shown in fig. 2 [3]:

- A_d is amplitude difference between the minimum 0–2 kHz and the maximum 0.5–12 kHz, which is maximized for a localized source, and for a higher relative strength of the frication source;
- S_p is slope of regression line fit to spectrum from the frequency of the main peak (mean value for that place, $\bar{F}$) and maximum spectral frequency (20 kHz), which increases with respect to the source strength;
- $S_p{}'$ is slope of regression line fit from the low-frequency minimum up to $\bar{F}$, which behaves similarly to A_d.

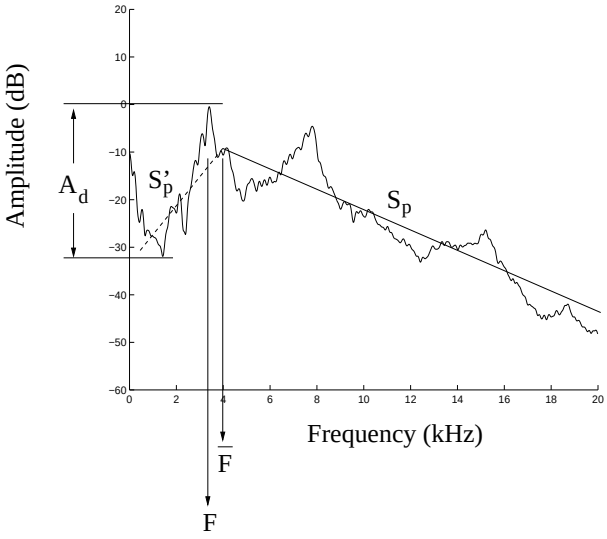


Figure 3: Dynamic amplitude A_d, and regression lines used to calculate low frequency (500 Hz to $\overline{F}$ kHz) slope $S_p{}'$ (dashed) and high frequency ($\overline{F}$ kHz to 20 kHz) slope S_p (solid). Sustained fricative /sh/ (spkr. ISSS).

A_d results for labio-dental fricatives /f, v/ revealed four different settings (with decreasing HNR of the sources): voiced /v/ (V /v/); partially devoiced /v/ (PDEV /v/); devoiced /v/ (DEV /v/); voiceless /f/.

$S_p{}'$ results for sibilant fricatives /s, z, sh, zh/ enabled classification into the same four categories; results for /sh, zh/ are shown in figure 4.

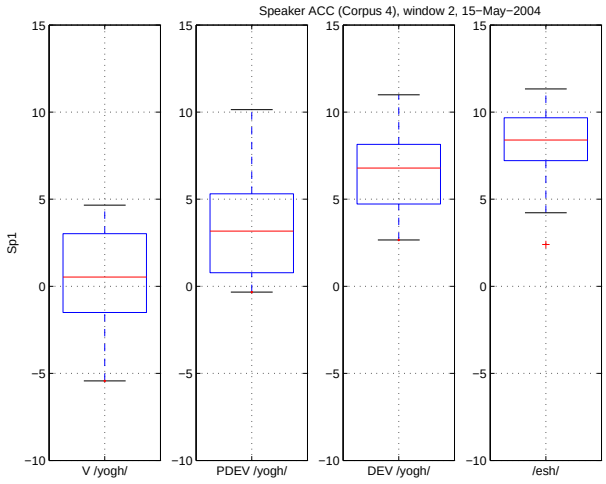


Figure 4: $S_p{}'$ parameter (spkr. ACC) for postalveolar fricatives (from left): voiced, partially devoiced and devoiced /zh/, and voiceless /sh/.

Durations

To provide an objective labeling of voiced and unvoiced regions, a procedure for automatic classification was adopted. Such a technique improves the consistency for larger data sets, from which to derive more reliable statistics.

First, fricatives for spkr. ISSS were manually annotated for voicing using the EGG signal, which was band-pass filtered (to reduce the effects of dc drift and electrical noise). After smoothing and conversion to decibels, the resultant laryngeal signal (Lx level) was sampled to give values at 5-ms intervals, as shown below in figure 6.

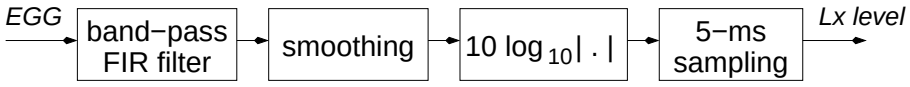


Figure 5: Processing stages applied to EGG signal to extract Lx level.

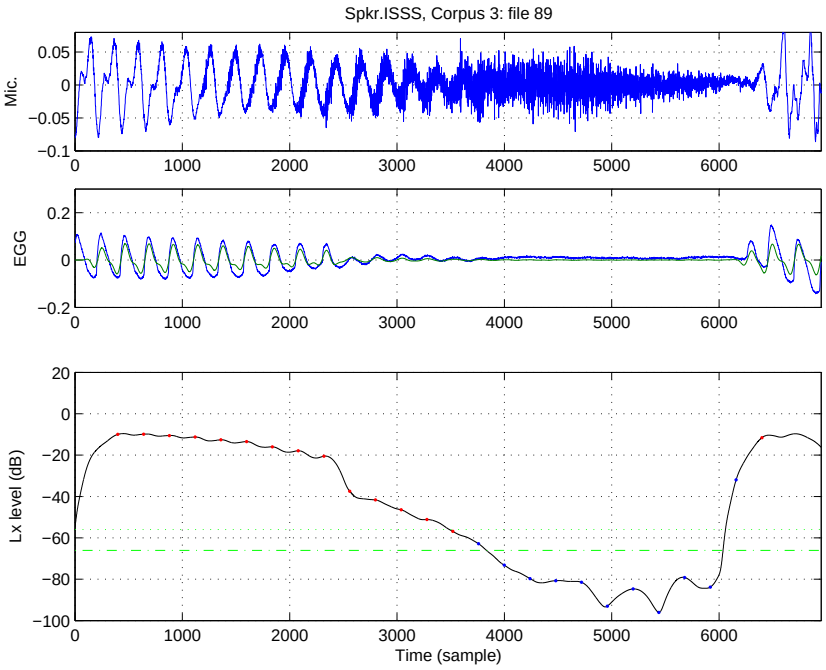


Figure 6: Acoustic waveform (top) of a /VFV/ utterance with devoiced /zh/ (spkr. ISSS). Glottal waveform (middle) with band-pass filtered EGG (in green). Derived Lx level curve (bottom) and samples (dots) with their manual classification (red, voiced; blue, unvoiced). Green dotted/dash-dot lines indicate initial/final classification boundaries.

We can explore speaker differences in devoicing from the Lx-level distribution:

- some have three modes (strong voicing, weak voicing and unvoiced);
- other speakers only two (voiced and unvoiced).

These features were used with the manual annotations to initialise a hidden Markov model (HMM), whose two states indicated the absence/presence of glottal oscillation, respectively. The model was trained through 10 iterations of Baum-Welch re-estimation, using all 160 fricative tokens.

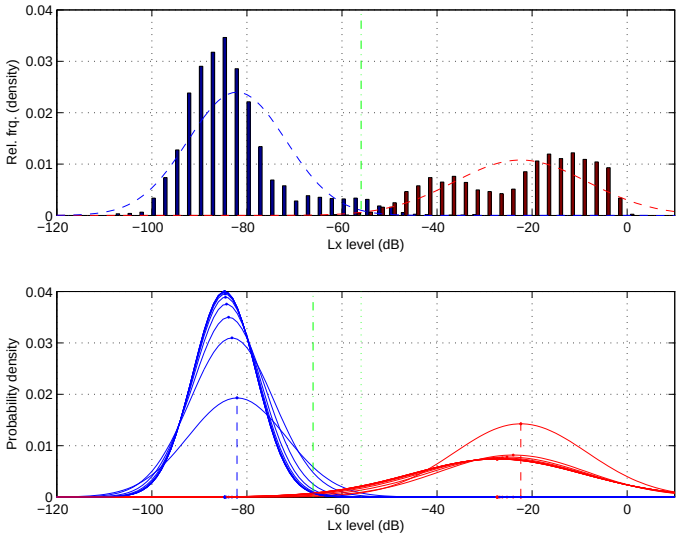


Figure 7: Histograms of Lx levels (top) classified manually into voiced (red) and unvoiced (blue) with fitted Gaussians (dashed). Effect of training (bottom), starting from Gaussians with dashed means (spkr. ISSS).

The state alignments produced by Viterbi decoding with this HMM enabled absence-of-glottal-oscillation durations to be extracted for each fricative. These can be used, in conjunction with fricative onset and offset times to investigate the effects of place of articulation and vowel context.

The results correspond well to the manual annotations with small adjustments to the transition times, and give us confidence in the model parameters obtained for characterizing vocal-fold vibration. It is anticipated that extension of these automatic techniques will allow us to examine further the timing relationships between stages in fricatives and their transitions.

Duration measures extracted from BE corpus

Trication onset t_F^+, voicing offset t_V^-, frication offset t_F^-, voicing onset t_V^+, and

- *total frication duration* (TFD $= t_F^- - t_F^+$),
- *source overlap duration* (SOD $= t_V^- - t_F^+$).

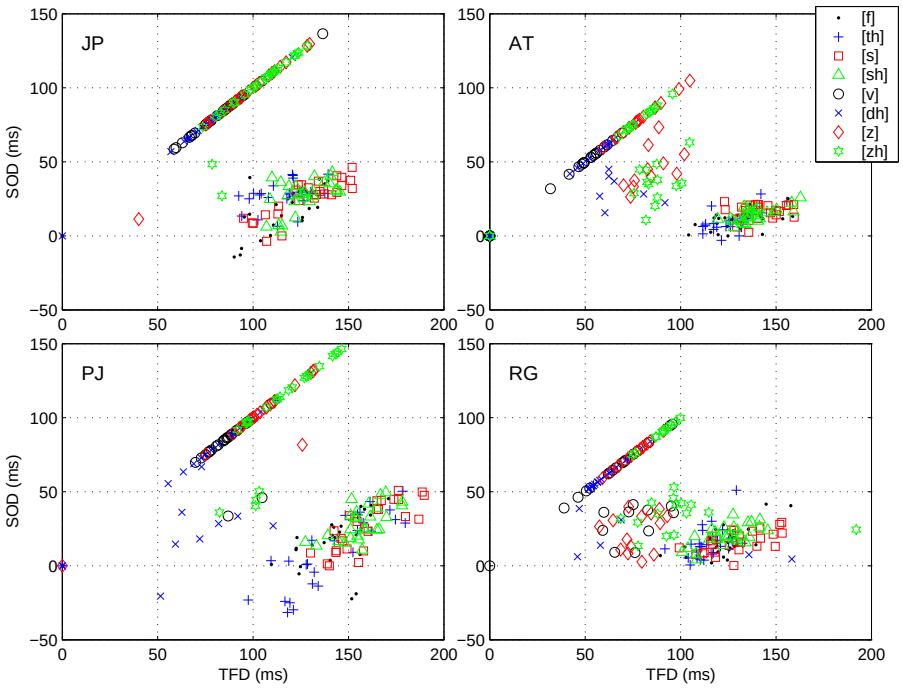


Figure 8: SOD versus TFD for (left) male and (right) female speakers.

- total fricative duration (TFD) is greater for unvoiced than voiced
- male TFD is longer than female
- voiced: females tend to devoice more than males, especially /z, zh/
- unvoiced: stronger correlation between SOD and TFD for males
- despite significant speaker variations, there are consistent patterns with place for unvoiced and voiced fricatives across speakers

Modulations

An algorithm was designed to measure the amplitude of any voicing-modulation of frication (AVM) in areas of source overlap, which characterizes the degree of fluctuation around the mean aperiodic power (see fig. 9 and example in fig. 10). Thus, it was independent of overall frication power and allowed for comparison across place of articulation.

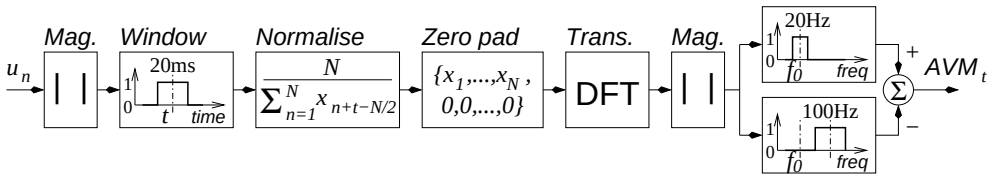


Figure 9: Processing stages applied to the high-pass filtered (3 kHz, order 4) speech signal, u_n, to extract the amplitude of voicing modulation (AVM) via discrete Fourier transform.

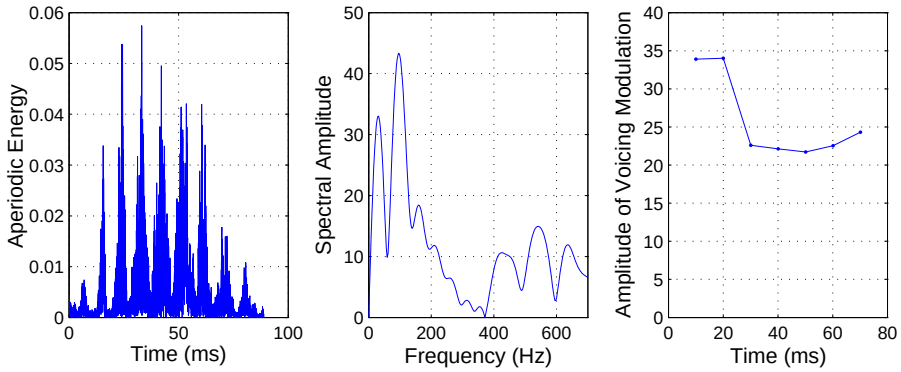


Figure 10(from left): Aperiodic energy, its spectrum and extracted AVM.

A measure of the aperiodic energy (AE), x_n, was obtained by averaging u_n over 20-ms frames (at 10 ms intervals).

Voicing strength (VS) was taken as peak-peak amplitude of the low-pass filtered waveform for each frame, normalized against the preceeding vowel to compensate for variations in speaking and recording levels.

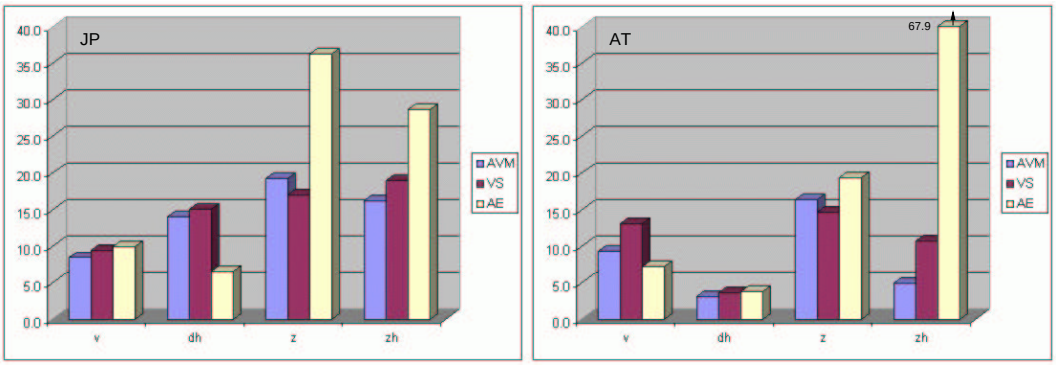


Figure 11: Mean AVM, VS and AE values: (left) male and (right) female.

- Aperiodic Energy proportionately greater for "sibilants"
- Amplitude of Voicing Modulation depended on Voicing Strength
- AVM greater than VS only for the alveolar fricative /z/, as in [2]

Conclusion

Summary:

- broad, multi-domain analysis of voicing-frication interaction
- spectral measures A_d and $S_p{}'$ used to distinguish V/PDev/Dev/U frics
- TFD and SOD pattern with voicing and place of articulation
- fricative modulation mainly reflects voicing strength

Further work:

- How can we model frication development within a fricative?
- Which of these cues is perceptually relevant?
- Use of parameters and automatic methods for speech synthesis.

References

[1] Blacklock, O. S. and C. H. Shadle (2003). Spectral moments and alternative methods of characterizing fricatives. *Journal of the Acoustical Society of America* 113(4):2199.

[2] Jackson, P. J. B. and C. H. Shadle (2000). Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. *Journal of the Acoustical Society of America* 108(4):1421–1434.

[3] Jesus, L. M. T. and C. H. Shadle (2002). A parametric study of the spectral characteristics of European Portuguese fricatives. *Journal of Phonetics* 30(3):437–464.

[4] Jesus, L. M. T. and C. H. Shadle (2003). Devoicing measures of European Portuguese fricatives. In N. J. Mamede, J. Baptista, I. Trancoso, and M. G. V. Nunes (Eds.), *Computational Processing of the Portuguese Language*, Chapter 1, pp. 1–8. Berlin: Springer-Verlag.

[5] Percival, D. B. and A. T. Walden (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge: Cambridge University Press.

[6] Pincas, J. and P. J. B. Jackson (2004). Quantifying voicing-frication interaction effects in voiced and voiceless fricatives. In *Proc. One Day Meeting for Young Speech Researchers*, p. 27, London, UK.

[7] Pincas, J. and P. J. B. Jackson (2004). Acoustic correlates of voicing-frication interaction in fricatives. In *Proc. From Sound to Sense*, Cambridge MA, USA.