Frication and Voicing Classification

Luis M.T. Jesus[1] and Philip J.B. Jackson[2]

[1] Escola Superior de Saúde da Universidade de Aveiro, and
Instituto de Engenharia Electrónica e Telemática de Aveiro
Universidade de Aveiro, 3810 - 193 Aveiro, Portugal
lmtj@ua.pt
http://www.ieeta.pt/~lmtj/
[2] Centre for Vision, Speech & Signal Processing
University of Surrey, Guildford GU2 7XH, UK
p.jackson@surrey.ac.uk
http://personal.ee.surrey.ac.uk/Personal/P.Jackson/

Abstract. Phonetic detail of voiced and unvoiced fricatives was examined using speech analysis tools. Outputs of eight f0 trackers were combined to give reliable voicing and f0 values. Log - energy and Mel frequency cepstral features were used to train a Gaussian classifier that objectively labeled speech frames for frication. Duration statistics were derived from the voicing and frication labels for distinguishing between unvoiced and voiced fricatives in British English and European Portuguese.

1 Introduction

1.1 Background

The long term objectives of the work presented in this book chapter are to deliver novel analysis methods for characterizing speech. Parameters for describing frication and voicing in fricatives are used to facilitate analysis of phonation and frication interaction effects observed. In particular, we aim to develop a concise model of the duration of voice and frication sources in fricative consonants in British English (BE) and European Portuguese (EP). The present work incorporates the following tasks: (i) development of speech analysis methods; (ii) development of new measures of voicing and frication to extend the phonetic description of Portuguese and English speech data; (iii) application of these parameters to the automatic classification of speech sounds; (iv) application of techniques across English and Portuguese using selected measures most apt for analysis, classification and modelling of mixed source speech signals.

This study deals with sounds produced by the simultaneous combination of two aeroacoustic sources, which have very different natures (one is quasi-periodic and the other noiselike). To measure properties of sounds like fricatives, stops and affricates, we evaluated the feasibility of conventional temporal and spectral measures, to yield useful descriptions of speech events. Pre - recorded EP and BE corpora of contextually - balanced acoustic data were used (Jesus and Shadle 2002; Pincas 2004).

A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, pp. 11–20, 2008.

The accurate determination of voicing onset/offset and the extraction of the fundamental frequency are important for the quantification of differences between normal and pathological voices, and for the robust encoding of normal voicing information in speech analysis/synthesis systems, as well as automatic labeling and segmentation. Francis et al. (2003) compared acoustic measures of voicing onset and found methods based on the waveform and low-frequency "voicing bar" to be more accurate and consistent than methods based on formants. Time-domain (McCree et al. 2002; Droppo and Acero 2007) and frequency-domain (Quatieri 2001; Pelle and Estienne 2007) methods for fundamental frequency analysis, used for low bit rate speech coding, have typically aimed at delivering a binary voiced/unvoiced decision and very few researchers (Childers et al. 1989) have tried to identify three different voicing states, i.e., voiced, partially voiced and unvoiced. Estimation of fundamental frequency typically relies on the signal periodicity (Hess 1992), and some researchers have explicitly disregarded irregular voice segments (Cheveigné and Kawahara 2002).

Previous work on fricatives with mixed sources includes the identification of the unvoiced fricative duration (UFD) as an essential feature for voicing categorization in English fricatives (Stevens et al. 1992; Pincas 2004). One important interaction effect, the modulation of frication during voicing, has been studied (Jackson and Shadle 2000; Pincas and Jackson 2005), as have the voicing characteristics of Portuguese fricatives (Jesus and Shadle 2003).

1.2 Motivation

Here, we combine our knowledge about observable (in the acoustic signal) differences in production strategy between unvoiced, devoiced and voiced fricatives for the same place of articulation. Interactions between voicing and frication sources are characterized by relative timings of onsets and offsets of voicing and frication, the fundamental frequency (f0), and relative levels of voicing and frication.

We believe that a processing approach inspired by speech production (data driven and knowledge based) can contribute to the performance of speech technology systems.

In vowel production, the vocal tract is relatively unconstricted and vocal folds tend to vibrate easily. In voiced obstruent consonants (fricatives or stops), a strong simultaneous noise source can only be produced at the expense of weakened voicing or devoicing.

In a study of devoicing of Portuguese voiced fricatives (Jesus and Shadle 2003), a criterion based on the ratio of variances in the electroglottograph (EGG) signal was used, during the VF transition and during the fricative, to derive a two-way classification (voiced/devoiced). The EGG variance, calculated at the beginning, middle and end of the fricatives, can be compared to the present classification scheme based on the f0 tracks of the speech signal.

Although f0 trackers seek periodicity in ways often similar to those used for manual annotation, they tend to be least reliable at voice onset/offset. We decided to test a range of freely accessible algorithms and combine their outputs to achieve a more reliable set of measurements.

The aim of the work is in using statistical tools in the fine phonetic analysis of fricatives. We have devised experiments that use an HMM to automatically classify both voicing and frication.

2 Speech Data

2.1 European Portuguese

A speech corpus, containing 1304 words that included fricatives /f, v, s, z, ʃ, ʒ/ from two male and two female adult native EP speakers, was recorded in a sound treated room using a Bruel & Kjaer 4165 $\frac{1}{2}$ inch microphone located 1 m in front of the subject's mouth, connected to a B & K 2639 pre-amplifier, then amplified and filtered by a B & K 2636 measurement amplifier (22 Hz-22 kHz). Acoustic and EGG signals were recorded with a Sony TCD-D7 DAT (16 bits, 48 kHz sampling frequency) and digitally transferred to PC. The simultaneous EGG signal was not used in the present study. Corpora were devised that included Portuguese words containing fricatives in frame sentences (Corpus 3), and the same set of words in sentences (Corpus 4). The EP corpus has manual annotations of the fricative start and end times that mark the transitions into and out of each fricative. Phonetic and phonological details of the corpus are described in Jesus and Shadle (2002).

2.2 British English

Fricatives from eight subjects, four male and four female, were recorded, all native speakers of BE. Speech-like tokens were obtained using nonsense /VFə/ words, F=/f, v, θ, ð, s, z, ʃ, ʒ/, embedded in the phrase "What does /VFə/ mean?" with vowel V=/ɑ, i, u/. Mono recordings were made in an acoustically-sheltered cubicle by Beyerdynamic M59 dynamic microphone linked directly to PC with a Creative Audigy soundcard (16 bits, 44.1 kHz sampling frequency). Nine repetitions of each possible VF combination by each speaker made 1728 sentences. The BE corpus was manually annotated separately for voicing and frication (Pincas 2004).

2.3 Dividing the Data

The data was divided into eight sets, having equivalent dimensions, and an even distribution of fricatives according to their place of articulation and phonological voicing classification, as shown in Table 1. Each data - set also has approximately the same number of samples from each speaker, gender, and for EP data the same number of samples from Corpus 3 and Corpus 4 (Jesus and Shadle 2002).

We needed to divide the data up for jack - knife experiments, maintaining separation of the training and the test data, meanwhile providing the most informative test results from the limited total data. Given the fact, that the BE data are all in vowel context, any files in the EP corpus that contained consonantal contexts were excluded. This resulted in the loss of 9% of the data (a fairly small proportion overall).

Table 1. Number of fricatives in the BE and EP data-sets

	British English									European Portuguese						
Set	[f]	[v]	[θ]	[ð]	[s]	[z]	[ʃ]	[ʒ]	Total	[f]	[v]	[s]	[z]	[ʃ]	[ʒ]	Total
set1	38	8	56	30	32	16	24	32	236	22	33	32	26	26	27	166
set2	24	7	31	21	40	40	24	30	217	22	33	31	25	26	27	164
set3	32	37	32	30	24	24	22	31	232	22	33	31	26	27	27	166
set4	24	59	32	30	24	23	32	8	232	22	34	32	27	27	29	171
set5	24	22	23	14	40	40	32	24	219	22	37	33	27	27	27	173
set6	22	20	16	38	16	39	24	45	220	22	38	34	28	26	26	174
set7	8	32	16	18	16	8	32	24	154	22	39	32	27	26	28	174
set8	40	16	8	8	24	24	24	16	160	20	39	32	27	23	28	169

3 Extraction of Reference f0

Wave files were processed to give a set of eight f0 tracks each, from which a reference f0 track was calculated. These were analysed together to evaluate voicing and f0 errors, which were treated as either *gross* (e.g., halving or doubling octave errors) or *fine*.

3.1 f0 Determination Algorithms

Only open-source software was employed, which enabled investigation (and correction) of the algorithms and represented widely-used speech research tools. Our selection included a number of standard f0 determination algorithms available in the Speech Filing System (SFS v. 4.6), the Auditory Perception Toolbox by MARCS Auditory Laboratories (MARCS v. 1.01) and Praat (v. 5.0.02):

1. `fxrapt -isp ...` – autocorrelation algorithm similar to Secrest and Doddington (1983) and used in `get_f0` Entropics' ESPS/Waves.
2. `fxcep -isp ...` – cepstral algorithm by Whittaker, Howard and Huckvale using Noll (1967)'s rules .
3. `fxanal -isp ...` – autocorrelation algorithm similar to Secrest and Doddington (1983) and implemented by Huckvale.
4. `fxac -isp ...` – autocorrelation algorithm by Huckvale.
5. `extractfundamental(...,...,0.01,'threshamp',0.02)` – Matlab implementation by Morris of Yehia's LPC-based algorithm.
6. `To Pitch (ac)... 0.0 75.0 15 off 0.03 0.45 0.01 0.35 0.14 600.0` – autocorrelation method implemented by Boersma (1993).
7. `To Pitch (cc)... 0.0 75.0 15 off 0.03 0.45 0.01 0.35 0.14 600.0` – forward cross-correlation method (Boersma).
8. `To Pitch (shs)... 0.01 50.0 15 1250.0 15 0.84 600.0 48` – subharmonic summation algorithm (Hermes 1988).

3.2 Combining f0 Tracks

The output from each f0 tracker was treated as the product of two simultaneous tracks, a binary voicing decision and the estimated fundamental frequency. Gaps in the f0 data (i.e., during unvoiced segments) were filled by linear interpolation. Both pieces of information, typically provided every 10 ms, were upsampled to every 1 ms. Hence, each f0 track yielded a voicing state and f0 estimate at 1 kHz frame rate. The median[1] gave the majority voicing state and a robust f0 value (see Figure 1).

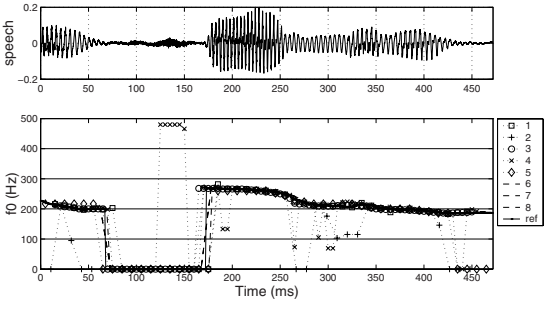

Fig. 1. Upper: acoustic signal of "a febra" [ɐˈfebɾɐ]. Lower: f0 tracks from 8 programs and the reference (ref).

The differences between the various f0 tracks and the reference track were analyzed to assess the consistency of the tracking methods, and hence an indication of the accuracy of the reference track. These differences fell into three broad categories: *voicing* errors, *gross* f0 errors and *fine* f0 errors. Voicing errors occurred when the voicing status of a given f0 track disagreed with that of the reference, and were classed as *false alarms* if the reference was unvoiced and as *false rejections* if it was voiced. With the same voicing status, a gross error indicated that the f0 track was closer (on a logarithmic scale) to either double or half of the current reference f0. The remaining voiced frames were considered *matched* and the fine errors were described for these by the RMS amplitude of the f0 difference (in Hz). A summary of the results of the error analysis is given in Table 2 for BE and EP data. The RAPT algorithm gave the best voicing decisions, while Boersma's methods provided most accurate f0.

4 Duration Analysis

In seeking an automatic and objective method for detecting and classifying the fine phonetic detail of fricatives, a series of hidden Markov models (HMMs) were built with Gaussian probability density functions. The MFCC and log-energy

[1] With eight values, the median was taken as mean of values ranked 4th and 5th; voicing status was rounded toward being voiced.

Table 2. f0 tracker (8 programs) error analysis (overall summary)

	1	2	3	4	5	6	7	8
Voicing error as proportion of entire corpus (%) – 69.8% voiced								
EP	4.7	30.0	6.7	9.5	12.0	6.0	6.2	14.0
BE	1.5	26.5	2.3	12.2	4.4	1.7	1.2	30.0
False alarm as proportion of unvoiced frames (%)								
EP	4.8	36.9	11.2	13.0	3.2	9.7	13.0	30.0
BE	1.3	24.3	1.9	13.5	0.7	0.5	0.5	36.9
False reject as proportion of voiced frames (%)								
EP	4.7	27.0	4.8	7.9	15.8	4.4	3.3	7.1
BE	2.3	34.7	3.5	7.2	18.5	6.2	4.0	4.3
Gross errors as proportion of voiced frames (%)								
EP	3.2	7.5	6.4	6.6	2.4	1.2	1.5	3.0
BE	3.1	8.5	9.6	11.2	2.8	1.4	3.4	3.9
Matched as proportion of voiced frames (%)								
EP	92.1	65.5	88.8	85.5	81.9	94.4	95.2	90.0
BE	94.7	56.8	86.9	81.5	78.7	92.4	92.6	91.9
RMS fine errors (Hz)								
EP	7.0	9.7	6.8	8.9	7.5	5.8	6.0	5.6
BE	7.2	10.1	5.9	10.5	9.3	6.3	6.2	7.0

features were obtained from the acoustic waveform (0.1 - 7.5 kHz) via HTK with 15 ms windows; only static features were used. The number of MFCCs was varied. The results of framewise classification accuracy against manual labels supported the use of 12 MFCCs plus log energy.

4.1 Method

Two experiments examined BE and EP respectively, using an HMM automatically to classify both voicing and frication. From the state alignment with respect to the acoustic features (i.e., the time spent in each state), we can derive an objective measure of devoicing, as well as other characteristics of the fricatives in our data sets.

Short audio clips containing one fricative plus 50 ms either side to give context and transitions, were extracted. Acoustic features (12 MFCCs and log energy) were computed with just 1 ms offset between frames, giving a 13 - D feature vector every 1 ms. Phonologically unvoiced fricatives typically start with a little or no overlap (<20 ms) between the voicing from the vowel to the onset of frication, then there is the main period of unvoiced frication until the onset of the following sound. For phonologically voiced fricatives, we expect there to be voicing throughout accompanied by the fricative source, although devoicing does sometimes occur. So, the state topology was defined to allow /V-uF-V/, /V-vF-uF-V/ or /V-vF-V/, where uF denotes unvoiced frication, vF denotes voiced frication, and V denotes the context of adjacent phonemes that were typically vowels (e.g., /ɑFə/ for BE). We have defined the topology to account for the state sequences

that occur in our data set, so we do not allow /V-uF-vF-V/ because it does not normally occur, whereas there is often a short period of overlap between voicing and frication at the start of phonologically unvoiced fricatives. The timing of these transitions is critical to their categorical perception, because it carries important cues to whether the fricative should be considered voiced or unvoiced.

Models for the BE data provided two states for the preceeding vowel, two for the fricative (one voiced, one unvoiced), and two for the following schwa. In order to balance the amount of training data used for each of the model states, and to accommodate the increased variability of the contexts in the EP database, six separate 2-state models were defined as follows: voiced frication (as with BE), unvoiced frication (as with BE), front, central and back vowels (and diphthongs starting with a front, central and back configuration), and silence. Nasalised and non-nasalised vowels were grouped together. This made a total of 12 states in the EP models, whereas the uniform context led to just 6 states in the BE models.

Initial state alignments were based on manual phone boundaries, dividing vowel segments, and using voicing decision from reference f0 for fricatives. One state was created for each of these with a 13-D Gaussian pdf. These initial definitions of state occupation were used to determine the mean and covariance for each state in Viterbi training. Training comprised of 10 further iterations in which the new state alignments were used to refine the models (allowing slight adjustments of the state boundaries for a better fit to the observed data).

The first set of multiple training iterations of jack-knife experiments, used *set2-8* for training and *set1* for testing. In the second set, we trained on *set1* and *set3-7* and tested on *set2*. The rest followed this pattern, i.e., the state alignment output from the HMMs were trained on 7/8 of the data and decoded on the remaining unseen files.

The final step consisted of using the trained models on the withheld test utterances to yield a completely automatic segmentation of the portion of the utterance around the fricative. This segmentation was then used to derive the duration statistics for final analysis of the data. The goal was a quantitative description of voiced and unvoiced periods during the phonological voiced and unvoiced fricatives.

4.2 Results

Manual annotations provided an initial alignment and the automatic ones were taken from the final alignment. These were used to extract the unvoiced frication duration (UFD) and the duration of frication with voicing, which we term the source overlap duration (SOD).

Figure 2 (top) shows the results of plotting SOD versus UFD for all eight English subjects, across all places of articulation. Voiced fricatives lie on the SOD axis, unvoiced lie on the UFD axis, and most of the data fall into the main area with some SOD and some UFD. The phonologically voiced and unvoiced fricatives tend to form two distinct clusters which are highlighted by the red and blue ellipses on those plots.

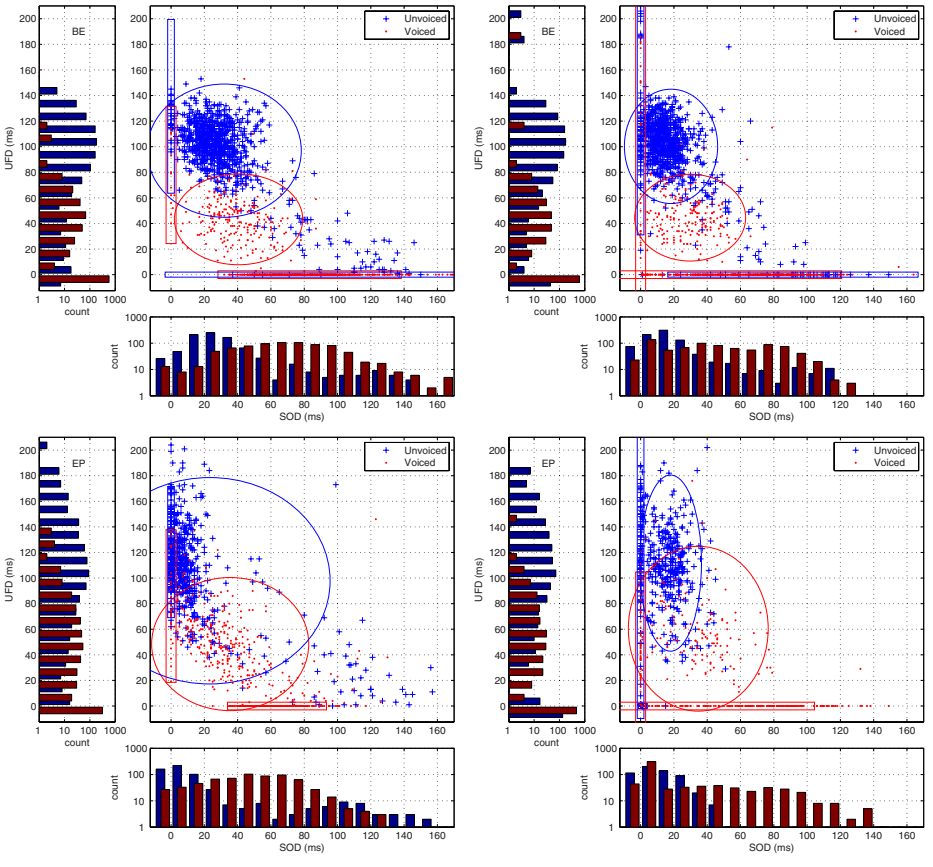

Fig. 2. Source overlap duration (SOD) and unvoiced frication duration (UFD) voicing classifications in BE (upper) and EP (lower) fricatives with manual (left) and HMM (right) alignments. Histograms show more clearly the distribution of data points.

Unvoiced fricatives cluster around (20, 100) ms, and a high classification accuracy of the phonological categories can be achieved by thresholding at UFD $\approx$ 60 ms (as reported previously by Pincas (2004)).

Considering the automatic voicing classification (Figure 2 top left), we see that the pattern is broadly consistent: SOD times have increased slightly at the expense of UFD. Figure 2 (top right) shows the output from the HMM annotation of states. The new clusters for unvoiced and voiced fricatives are centred at (10, 115) ms and (20, 50) ms respectively, suggesting a higher threshold UFD $\approx$ 70 ms.

Figure 2 (bottom) shows an analysis of Portuguese fricatives. As before, the left panel shows SOD versus UFD with manual frication annotation and voicing classification from the reference f0 track for the entire EP corpus. The distributions are similar to those from the BE corpus, however there is greater overlap

including a large number of phonologically voiced fricatives that were devoiced. This variability may be attributed to differences in annotation procedure and the more natural context of the EP tokens.

5 Conclusions

In this book chapter we have proposed the development of an automatic method for phonetic analysis of the durational characteristics of voicing and frication features. Our experiments consider both British English and European Portuguese fricatives recorded as nonsense and real words respectively. By combining the outputs of eight publicly - available f0 determination algorithms, we obtained a more reliable reference f0 track for each utterance which was used to evaluate the accuracy of each technique, with an emphasis on fricative speech. Together with manual annotation of phone boundaries, we used the voicing state of the reference f0 track to define initial regions of voiced and unvoiced frication. Jack - knife experiments were then conducted, training HMMs to recognize these states in unseen test utterances. The final output was an objective annotation of voiced and unvoiced frication to 1 ms resolution, from which duration statistics were obtained.

We have shown that the technique can be applied across languages. It is relevant both to English and Portuguese, and enables objective investigation of the duration characteristics observed in various contexts. Further work is needed to extend the results of this pilot study to a wider range of speech data, and to encapsulate our knowledge of fricative duration characteristics. Such duration models could be made context - dependent and incorporated into model - based speech synthesis and articulatory - feature based speech recognition.

Acknowledgements

This work was partially supported by Fundação para a Ciência e a Tecnologia, Portugal, Conselho de Reitores das Universidades Portuguesas, Portugal, and British Council, UK (Treaty of Windsor Programme).

References

Boersma, P.: Accurate short - term analysis of the fundamental frequency and the harmonics - to - noise ratio of a sampled sound. In: Proc. Institute of Phonetic Sciences, U. Amsterdam, vol. 17, pp. 97–110 (1993)

Cheveigné, A., Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. JASA 111(4), 1917–1930 (2002)

Childers, D., Hahn, M., Larar, J.: Silent and voiced/unvoiced/mixed excitation (four - way) classification of speech. IEEE Transactions on Acoustics, Speech and Signal Processing 31(11), 1771–1774 (1989)

Droppo, J., Acero, A.: A fine pitch model for speech. In: Proc. InterSpeech, pp. 2757–2760 (2007)

Francis, A., Ciocca, V., Yu, J.: Accuracy and variability of acoustic measures of voicing onset. JASA 113(2), 1025–1032 (2003)

Hermes, D.: Measurement of pitch by subharmonic summation. JASA 83(1), 257–264 (1988)

Hess, W.: Pitch and voicing determination. In: Furui, S., Sondhi, M. (eds.) Advances in Speech Signal Processing, pp. 3–48. Marcel Dekker, New York (1992)

Jackson, P., Shadle, C.: Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. JASA 108(4), 1421–1434 (2000)

Jesus, L., Shadle, C.: A parametric study of the spectral characteristics of European Portuguese fricatives. J. Phon. 30(3), 437–464 (2002)

Jesus, L., Shadle, C.: Devoicing measures of European Portuguese fricatives. In: Mamede, N., Baptista, J., Trancoso, I., Nunes, M. (eds.) Comp. Processing of the Portuguese Language, pp. 1–8. Springer, Heidelberg (2003)

McCree, A., Stachurski, J., Unno, T., Ertan, E., Paksoy, E., Viswanathan, V., Heikkinen, A., Ramo, A., Himanen, S., Blocher, P., Dressler, O.: A 4kb/s hybrid MELP/CELP speech coding candidate for ITU standardization. In: Proc. ICASSP, pp. 629–632 (2002)

Noll, A.: Cepstrum pitch determination. JASA 41(2), 293–309 (1967)

Pelle, P., Estienne, C.: A pitch extraction system based on phase locked loops and consensus decision. In: Proc. InterSpeech, pp. 1637–1640 (2007)

Pincas, J.: The interaction of voicing and frication sources in speech: An acoustic study. M.Res. Thesis, University of Surrey, Guildford, UK (2004)

Pincas, J., Jackson, P.: Amplitude modulation of frication noise by voicing saturates. In: Proc. InterSpeech, pp. 349–352 (2005)

Quatieri, T.: Discrete-time Speech Signal Processing: Principles and Practice. Prentice Hall, Englewood Cliffs (2001)

Secrest, B., Doddington, G.: An integrated pitch tracking algorithm for speech systems. In: Proc. ICASSP, pp. 1352–1355 (1983)

Stevens, K., Blumstein, S., Glicksman, L., Burton, M., Kurowski, K.: Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. JASA 91(5), 2979–3000 (1992)