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
e-mail: lmtj@ua.pt

[2]Centre for Vision, Speech & Signal Processing
University of Surrey, Guildford GU2 7XH, UK
e-mail: p.jackson@surrey.ac.uk

Acknowledgements: Fundação para a Ciência e a Tecnologia, Portugal, Conselho de Reitores das Universidades Portuguesas, Portugal, and British Council, UK .

PROPOR 2008, Curia, Portugal (SLP 1 - Speech Analysis 8/9/2008)

Background

- ”A preliminary evaluation of the automatic criterion for devoicing showed great potential for the use of this technique in future work.”
  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, pp. 1-8. Berlin: Springer-Verlag.
- ”Consonant detection in speech by a machine based on purely spectral features is always problematic due to a number of reasons like the unvoiced (no-energy) portions of stop consonants that can be confused with real silence, the high energy fricative noise that maybe confused with environmental or additive noise, and the vowel like spectrum of the liquids, the nasals and the semi-vowels that make them hard to distinguish from vowels.” Second IEEE Spoken Language Technology Workshop Goa, India 2008 - Special Session Call for Participation: Consonant Challenge for Indian Languages.

Our Work

- 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.
- Used an HMM to perform objective labeling of the voice and frication features.

Speech Data

European Portuguese (EP)

- 1304 words that included fricatives /f, v, s, z, ʃ, ʒ/.
- Two male and two female native EP speakers.
- Acoustic and EGG signals recorded (16 bits, 48 kHz).
- Manual annotations of the fricative start and end times that mark the transitions into and out of each fricative (Jesus and Shadle 2002).

British English (BE)

- 1728 words that included fricatives /f, v, θ, ð, s, z, ʃ, ʒ/.
- Four male and four female native speakers of BE.
- Mono acoustic recordings (16 bits, 44.1 kHz).
- Manually annotated separately for voicing and frication (Pincas 2004).

Number of fricatives in the BE and EP data - sets

- Data divided into eight sets, having equivalent dimensions, and an even distribution of fricatives according to their place of articulation and phonological voicing classification.
- Data divided for jack - knife experiments, maintaining separation of the training and the test data, whilst most informative test results were provided.

	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

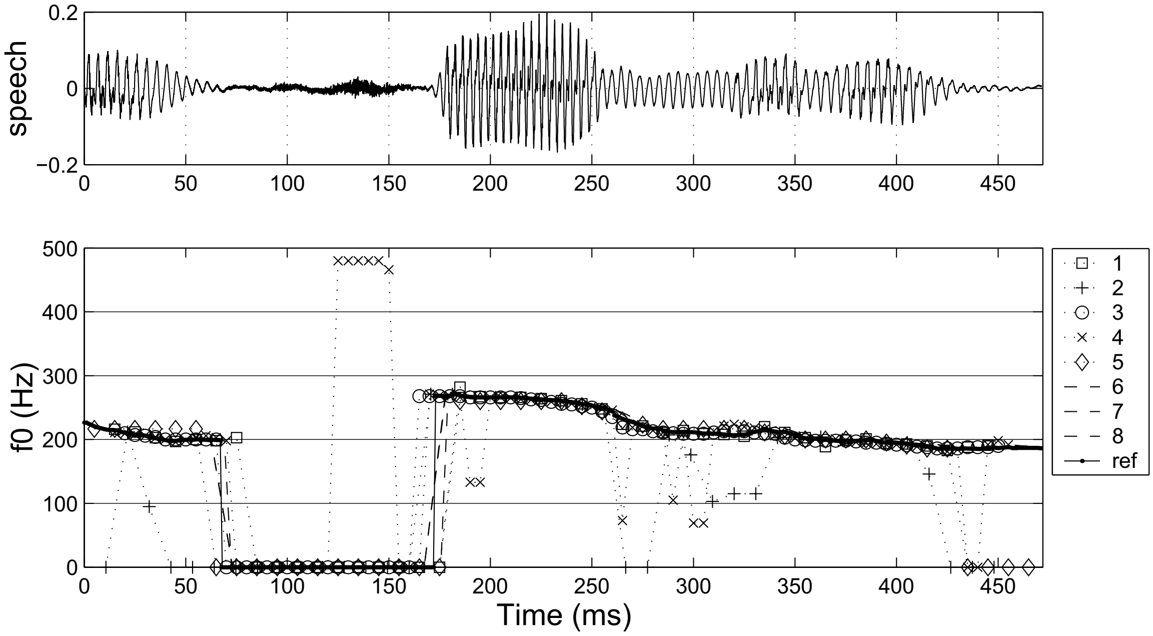
f0 Determination Algorithms

- Eight f0 tracks computed for each waveform by widely-used, open-source speech analysis tools:
  - Speech Filing System (SFS), v. 4.6
    1. Autocorrelation algorithm (Secrest and Doddington 1983).
    2. Cepstral algorithm (Noll 1967).
    3. Autocorrelation algorithm (Secrest and Doddington 1983).
    4. Autocorrelation algorithm by Huckvale.
  - Auditory Perception Toolbox (MARCS), v. 1.01
    5. Matlab implementation by Morris of Yehia's LPC-based algorithm.
  - Praat, v. 5.0.02
    6. Autocorrelation method (Boersma 1993).
    7. Forward cross-correlation method (Boersma).
    8. Subharmonic summation algorithm (Hermes 1988).
- Reference f0 track derived from 8 f0 tracks
- Analysed voicing and f0 errors, *gross* and *fine*

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 gave the majority voicing state and a robust f0 value.

Upper: acoustic signal of “a febra” [ɐ'febrɐ]
Lower: f0 tracks from 8 programs and the reference (ref)



f0 Tracker Error Analysis

- Differences between the various f0 tracks and the reference track were analysed to assess the consistency of the tracking methods, and hence an indication of the accuracy of the reference track.
- Differences fell into three broad categories:
    - Voicing errors - voicing status of a given f0 track disagreed with that of the reference.
        - *False alarms* if the reference was unvoiced.
        - *False rejections* if the reference was voiced.
    - Gross f0 errors - the f0 track was closer (on a logarithmic scale) to either double or half of the current reference f0.
    - Fine f0 errors - RMS amplitude of the f0 difference (in Hz) for the remaining voiced frames (considered *matched*).
- Boersma's methods (programs 6 and 7) provided most accurate f0.

f0 tracker (8 programs) error analysis

	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

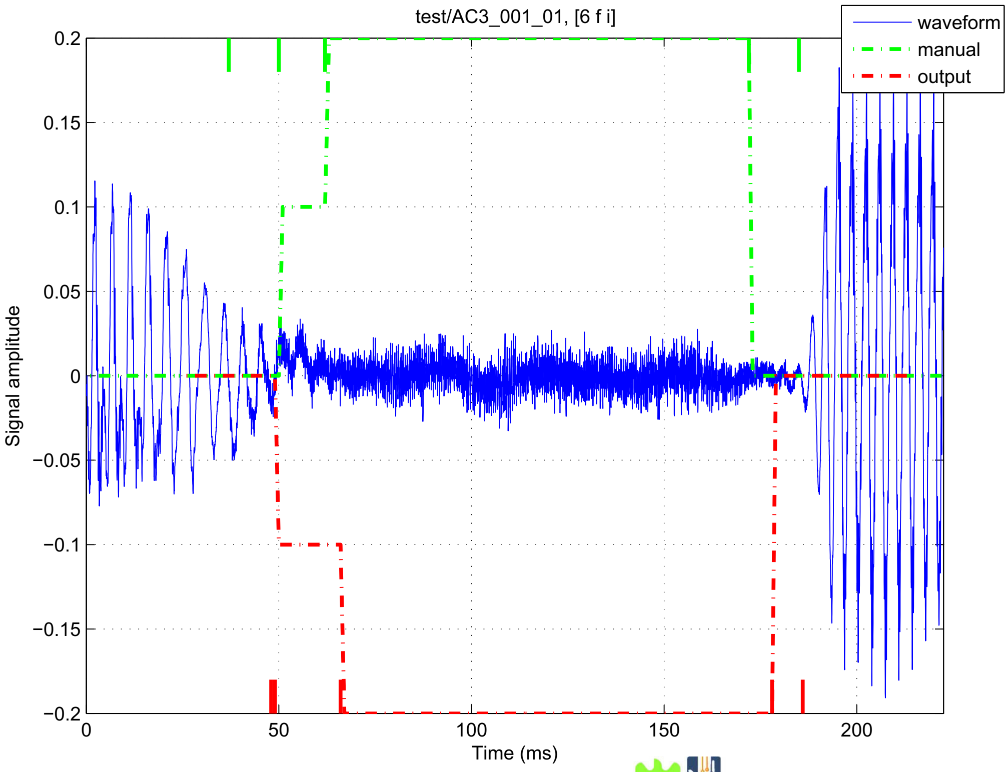
Duration Analysis

- 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.
- Two experiments examined BE and EP respectively, using an HMM automatically to classify both voicing and frication.

Method

- From the state alignment with respect to the acoustic features, we derived an objective measure of devoicing, as well as other characteristics of the fricatives in our data sets.
- Each fricative segment processed with 50 ms before and after.
- 12 MFCCs and log energy were computed from acoustic waveform (0.1–7.5 kHz) using 15 ms window and 1 ms frame offset.
- Only static features were used to identify frication and voicing irrespective of context.
- Unvoiced fricatives start with short overlap ($<$20 ms) between the voicing from preceding vowel and the onset of frication, followed by voiceless frication until the next phone.
- Voiced fricatives exhibit voicing throughout with frication, but devoicing can occur.

Initial labels and state alignment output from the HMM



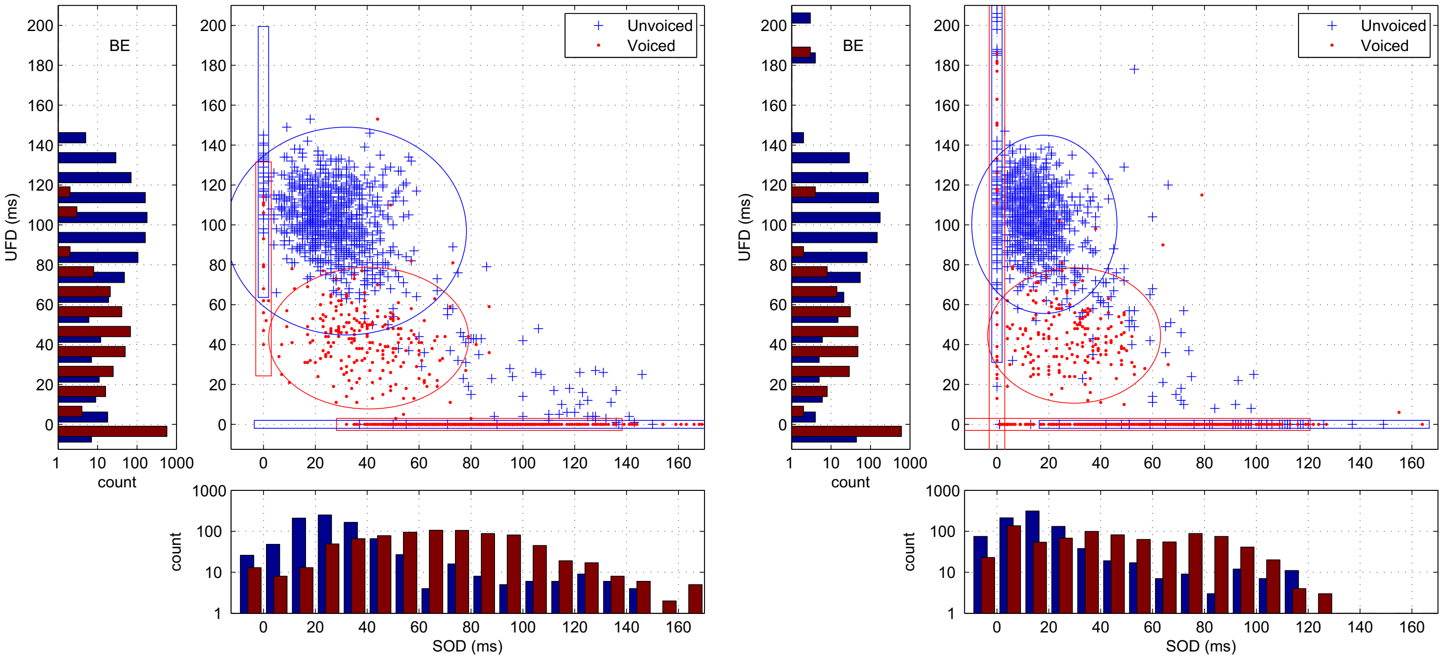
Method

- BE models – three separate 2 - state.
- EP models – six separate 2 - state.
- State alignment output from the HMMs were trained on 7/8 of the data and decoded on the remaining unseen files.
- 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.

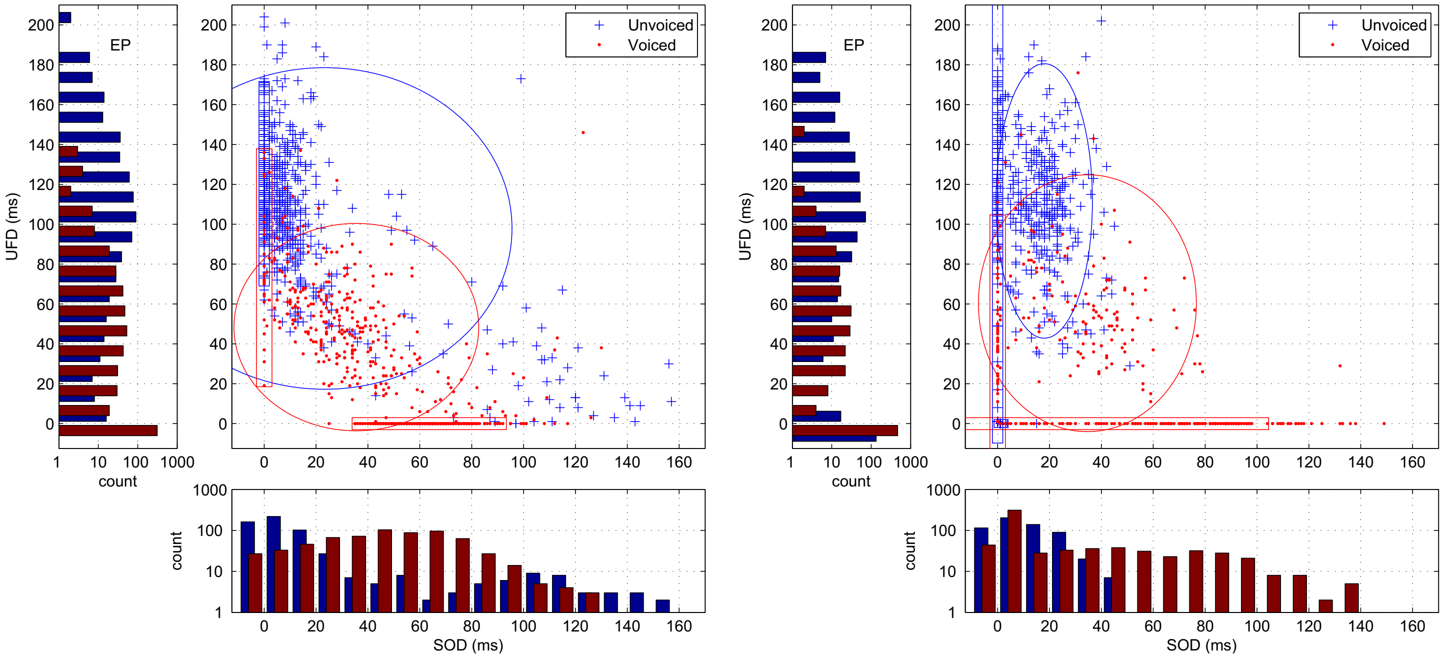
Results

- 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).

SOD and UFD voicing classifications in BE fricatives with manual (left) and HMM (right) alignments



SOD and UFD voicing classifications in EP fricatives with manual (left) and HMM (right) alignments



Conclusions

- Automatic method for phonetic analysis of the durational characteristics of voicing and frication features.
- Jack-knife experiments were conducted, training HMMs to recognise voicing states in unseen test utterances.
- Technique can be applied across languages.
- Relevant to EP and BE, and enables objective investigation of the duration characteristics observed in various contexts.
- Further work needed to extend the results to a wider range of speech data, and to encapsulate our knowledge of fricative duration characteristics.
- Duration models could be made context-dependent and incorporated into model-based speech synthesis and articulatory-feature based speech recognition.