Shout Detection in Noise

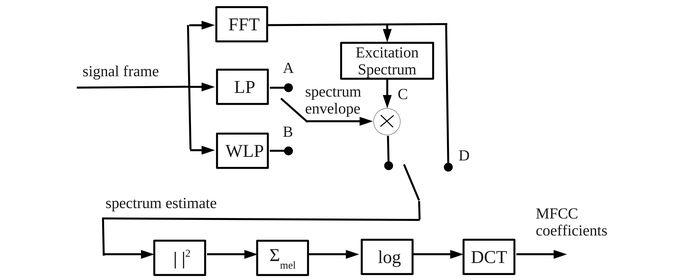
Jouni Pohjalainen[1], Paavo Alku[1], Tomi Kinnunen[2]

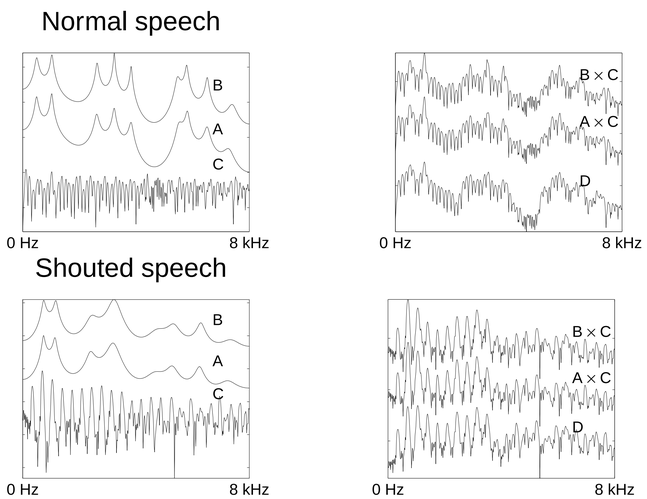
[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]School of Computing, University of Eastern Finland, Finland

Background

- Detection of shouted speech is an important research problem in audio event detection for surveillance
- In noisy environments, the distance between the shouter and the microphone determines the SNR
- Shouting is reflected in the vocal tract excitation by, e.g., the use of high F0
- A system based on mel frequency cepstral coefficient (**MFCC**) feature extraction and Gaussian mixture model (**GMM**) classification is described and evaluated

Feature Extraction





- Typically, MFCC analysis is based on FFT spectrum analysis as the first step
- Replacing FFT spectrum by magnitude spectrum envelope obtained by linear prediction (LP) or weighted linear prediction (WLP) [1] has improved the noise robustness of MFCCs in earlier studies [2]
- To combine a robust envelope estimate with the excitation spectrum, LP or WLP spectrum envelope can be multiplied by spectral fine structure obtained by cepstral analysis
- Analysis of MFCC coefficient distributions showed that normal speech and shouted speech differed mainly in the MFCC coefficients with index less than 30

Classification

- GMM modeling of three sound classes: shouted speech, normal speech and environmental noise
- In both the training and detection phases, unsupervised training of a two-state hidden Markov model is used to select the high-energy frames within an analysis block of 2 s (model only high SNR frames)
- Decision for each block is based on the GMM likelihoods averaged over the selected frames:

$$L = L_{\text{shout}} - \max(L_{\text{speech}}, L_{\text{noise}})$$

Results

- Speech material: 24 short Finnish sentences spoken normally and shouted by 22 speakers, one speaker in turn chosen as the test speaker and the other 21 speakers' material used to train the shouting and speech GMMs (22-fold cross validation)
- Noise from NOISEX-92 database used to artificially corrupt the test material with additive noise as well as to simulate a noise-alone condition

Equal error rate (EER) %							
12 MFCCs	Spectrum estimation	Signal-to-noise ratio (dB) 20	10	0	-10	-20	-30
FACTORY NOISE	FFT	2.4	2.5	3.2	12.2	28.1	50.2
	LP	3.9	4.3	5.6	10.3	22.0	45.4
	LP + ex.	2.7	2.3	3.3	6.6	21.2	46.4
BABBLE NOISE	FFT	2.6	2.9	3.2	9.5	22.8	46.0
	LP	3.5	4.3	4.6	7.7	22.9	45.6
	LP + ex.	2.8	2.7	2.9	4.7	17.0	43.8

- Combining the LP/WLP spectrum envelope with the cepstrally separated excitation spectrum ("ex.") gave better results than the envelope alone

Equal error rate (EER) %							
30 MFCCs	Spectrum estimation	Signal-to-noise ratio (dB) 20	10	0	-10	-20	-30
FACTORY NOISE	FFT	2.5	2.7	2.9	10.1	20.2	46.0
	LP + ex.	2.9	3.1	3.0	6.1	17.0	45.9
	WLP + ex.	2.6	2.5	3.3	6.8	18.4	46.6
BABBLE NOISE	FFT	2.6	2.3	2.1	5.1	19.8	45.0
	LP + ex.	3.2	2.9	3.5	4.6	15.6	42.9
	WLP + ex.	2.2	2.4	2.2	4.7	15.2	43.8

- In terms of robustness, i.e., at low SNR levels, the LP/WLP spectrum analysis approach outperformed FFT and 30 MFCCs performed better than 8, 12 or 20 MFCCs

Conclusions

- A system for shout detection was developed, giving low error rates for high and moderate SNRs
- The importance of the vocal tract excitation, manifested as the spectral fine structure, was observed
- **Improved performance in shout detection resulted from:**
  - spectral fine structure multiplied by the LP/WLP envelope
  - increasing the length of the MFCC vector, preserving more information about the fine structure

References

[1] Ma, C., Kamp, Y. and Willems, L. F., "Robust signal selection for linear prediction analysis of voiced speech", Speech Communication, 12(2):69–81, 1993.

[2] Pohjalainen, J., Kallasjoki, H., Palomäki, K. J., Kurimo, M. and Alku, P., "Weighted Linear Prediction for Speech Analysis in Noisy Conditions", in Proc. Interspeech, Brighton, UK, 2009.