Detection of Shouted Speech in the Presence of Ambient Noise

Jouni Pohjalainen, Tuomo Raitio, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland

Background

- Detection of shouted speech is a central research problem in audio event detection for surveillance and can also benefit automatic speech recognition and speaker recognition
- Robustness in noisy environments is required
- Accuracy of machine vs. human in detecting shouted speech?

Topic

- A system based on mel frequency cepstral coefficient (MFCC) feature extraction, unsupervised frame selection and Gaussian mixture model (GMM) classification is developed and evaluated
- The shout detection capability of human listeners is measured by a listening test and the results are used as reference values to evaluate the automatic system

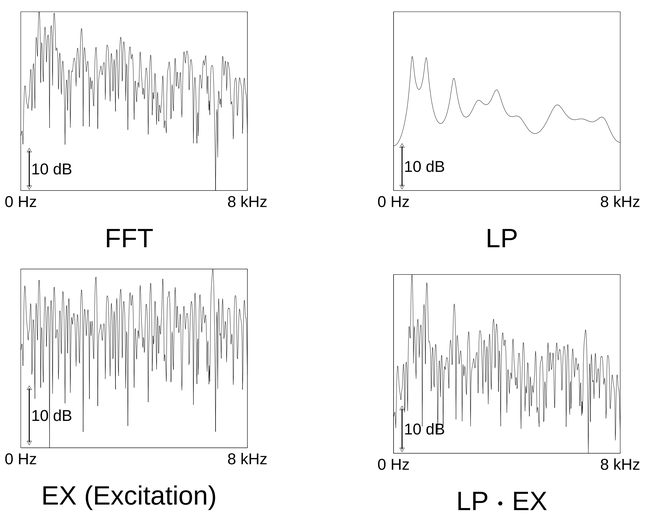
Test Material

- Speech material: 24 short Finnish sentences spoken normally and shouted by 22 speakers
- *factory1* and *babble* noise from NOISEX-92 database used to artificially corrupt the test material with additive noise as well as to simulate a noise-alone condition

Feature Extraction

- Typically, MFCC computation begins by FFT spectrum analysis
- FFT can be replaced, e.g., by linear predictive methods for improved robustness in many applications
- In an earlier study [1], it was found useful to multiply together
  - spectrum envelope obtained from linear prediction (LP) analysis
  - excitation spectrum/spectral fine structure obtained by cepstral analysis

  and use this spectrum (LP · EX) to obtain MFCCs



- RESULT: LP · EX outperforms FFT as the MFCC spectrum analysis method in shout detection
- RESULT: According to the analysis of MFCC coefficient distributions for normal and shouted speech, as well as shout detection experiments with 12, 18, 24, 30 and 36 MFCCs, the best MFCC vector form appears to be 30 MFCCs without deltas

Classification

- GMM modeling of different sound classes: shouted speech (target class), normal speech, ambient noise and non-shouting (combination of normal speech and noise)
- In both the training and detection phases, a suitably initialized unsupervised method (k-means or HMM re-estimation) is used to select the high-energy frames within an analysis block of 2 s (i.e., model and recognize only high SNR frames)
- Decision for each block is based on the GMM likelihoods averaged over the selected frames
- RESULT: Decision rule

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

  outperforms

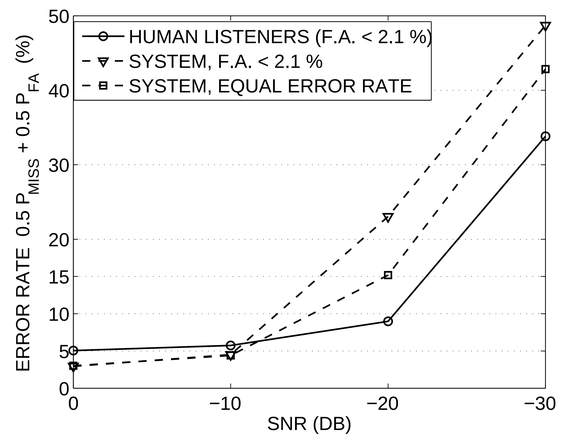
$$L = L_{\text{shout}} - L_{\text{nonshout}}$$

  with likelihoods given by 8-component GMMs
- RESULT: The frame selection approach can have a noticeable effect on performance. The best performance was obtained using the HMM and k-means based methods in training and detection, respectively

Main Results

Listening test (*babble* noise)		
SNR (dB)	Miss (%)	False alarm (%)
0	8.7	1.4
-10	10.0	1.5
-20	16.0	1.9
-30	65.6	2.1

- For comparison, the operating point of the system was set in two different ways, corresponding both to the EER criterion and to the low false alarm rate exhibited by listeners



Conclusions

- The automatic system outperforms human listeners at moderate “babble” noise levels
- Detection by humans is better in severely noisy conditions

References

[1] Pohjalainen, J., Alku, P. and Kinnunen, T. “Shout Detection in Noise”, in Proc. ICASSP, Prague, Czech Republic, May 2011.