Automatic Detection of High Vocal Effort in Telephone Speech

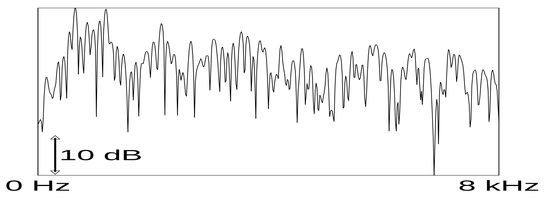
Jouni Pohjalainen, Tuomo Raitio, Hannu Pulakka, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Finland

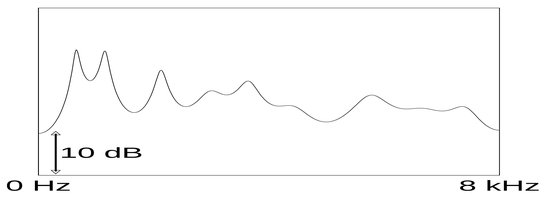
Focus

- High vocal effort increases F0, F1 and the spectral center of gravity
- The performance of automatic speech recognition (ASR) and speaker recognition (SR) systems is affected by **vocal effort mismatch** between the training and recognition phase
- In order to avoid performance degradation due to this mismatch, a **detection system** is needed to aid the recognizer in choosing acoustic models that are most appropriate for the changed conditions
- In a call center ASR setting, high vocal effort can arise due to the Lombard reflex or the customer's dissatisfaction with ASR performance

Feature Extraction

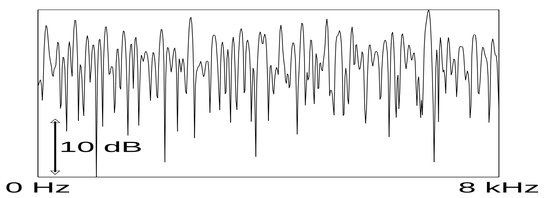


FFT

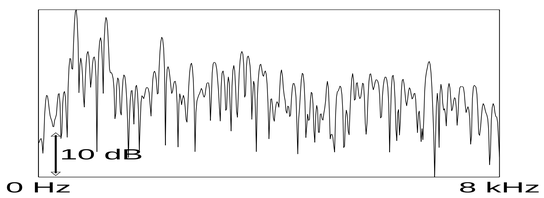


LP

- Using linear prediction (**LP**) as the spectrum estimation method in place of **FFT** in the **MFCC** computation chain has led to improved noise robustness in several earlier studies on ASR and SR
- The spectrum of the vocal tract excitation differs between normal and high-effort speech and can thus provide additional cues, especially with low signal-to-noise ratio (SNR)



Excitation



LP-CR

- To combine a robust envelope estimate with the excitation spectrum, **multiply the LP spectrum envelope by the spectral fine structure (excitation spectrum) obtained by cepstral processing**

Classification

- **Train Gaussian mixture models (GMMs) for two sound classes: normal and high-effort speech**
- In both the training and detection phases, k-means clustering is used to **select the high-energy frames** within an analysis block of two seconds (model and recognize only high-SNR frames)

Test Material

- 24 Finnish sentences spoken with normal and high vocal effort by 22 speakers, 22-fold cross-validation
- Noise from the NOISEX-92 database was added to simulate **far-end noise corruption**
- **Transmission over the GSM channel** was simulated

Results

*Equal error rates (%) for unprocessed and telephone-channel speech with **matched-condition** training and different spectrum estimation methods in MFCC feature extraction. Superscripts indicate, within each test condition, the spectrum estimation method pairs whose difference was not statistically significant.*

Spectrum estimation method	Test condition: Unprocessed speech clean (16kHz)	clean (8kHz)	Narrowband telephone speech *volvo* SNR=30	*volvo* SNR=0	*factory1* SNR=0	*babble* SNR=0
FFT	3.3 [1]	5.1	4.0	3.2 [1,2]	4.4 [1]	5.5
LP	2.0	3.5	2.9	3.0 [1]	5.1	4.7
LP-CR	3.2 [1]	4.2	3.5	3.5 [2]	4.0 [1]	4.2

*Equal error rates (%) for telephone-channel speech with the detector trained in high-SNR car noise condition and evaluated in **mismatched noise conditions**. Superscripts indicate, within each test condition, the spectrum estimation method pairs whose difference was not statistically significant.*

Spectrum estimation method	Test condition: *volvo* SNR=0	*factory1* SNR=0	*babble* SNR=0
FFT	3.9	4.8	5.1
LP	3.2 [1]	5.9	5.9
LP-CR	3.5 [1]	4.0	4.4

Conclusions

- A vocal effort detection system was developed and evaluated on unprocessed and telephone speech
- Concerning the spectrum analysis method used to obtain MFCCs, the spectral fine structure multiplied by a LP spectral envelope (LP-CR) led to improved performance compared to baseline FFT and LP
- Future work: use the methods to improve ASR and SR performance?