Automatic Detection of Anger in Telephone Speech with Robust Autoregressive Modulation Filtering

Jouni Pohjalainen and Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Finland

Focus

- Automatic detection of a caller's angry emotional state can improve the user experience with telephone dialog systems
- In this study, speech material from the Berlin database of emotional speech is processed in order to simulate noisy telephone channels
  - Earlier studies on anger detection have mainly used non-public spontaneous speech material
  - Acted speech databases emphasize hot-anger and confusable emotion classes rather than the discrimination between angry and neutral speech
- The temporal dynamics of angry speech are modeled by autoregressive prediction of MFCC features computed using either Fourier or linear predictive spectrum analysis

Test Material

- The 535 utterances of the **Berlin database**
  - Seven emotion classes: *anger*, *boredom*, *disgust*, *fear*, *joy*, *sadness* and *neutral*
  - The goal is to separate the *anger* class from the other six classes
- Noise from the NOISEX-92 database is added to simulate **far-end noise corruption**
- **Transmission over the GSM channel** is simulated

Feature Extraction

- **39 features** extracted every 10 ms using a 25 ms analysis frame:
  12 **MFCCs** based on either FFT or LP + log energy normalized over each utterance + **Δ**'s and **ΔΔ**'s
  - MFCC chain: 1) **power spectrum estimate** 2) mel filterbank 3) logarithm 4) discrete cosine transform
  - Using linear prediction (**LP**) as the spectrum analysis method in place of **FFT** in MFCC computation has led to improved noise robustness in several earlier studies on automatic speech and speaker recognition

Frame Processing

- In the training phase, an **autoregressive** (AR) model is trained for each feature to represent its time dynamics in the target class (angry speech)
  - **Modulation filtering:** after initial feature extraction in both training and detection, the AR models are used to predict the feature values based on the unprocessed features and the features are replaced by their predicted values
  - If $x_{t,j}$ is the value of the jth original feature in the tth frame, it is replaced by the prediction $y_{t,j} = c_j + \sum_{k=1}^{r} b_{j,k} x_{t-sk,j}$, where r is the AR order, s is the *frame skip* parameter and c_j is a constant term

Classification

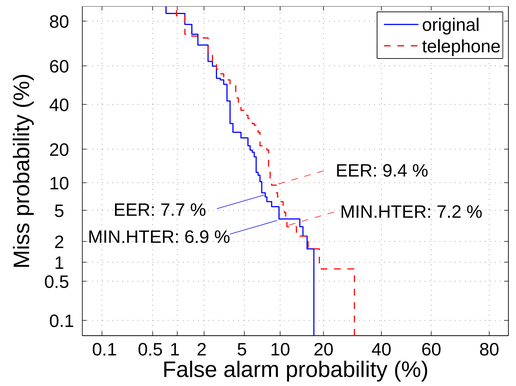
- Bayesian classification using Gaussian mixture models (**GMMs**)
  - Use two GMMs to model the distributions of the AR-processed feature vector in anger and non-anger

Results

- The best configuration in feature filtering was found to be $r = 8$ and $s = 4$, i.e., **AR(8)** with feature vectors observed at a **40 ms interval**, giving a prediction history of **320 ms** (see the paper for further results)

Setup				Number of GMM components					
				16				64	
Training		Testing		AR filtering of MFCC features					
				none		$r = 8, s = 4$			
Noise type	SNR (dB)	Noise type	SNR (dB)	Spectrum analysis in MFCC					
				FFT	LP	FFT	LP	FFT	LP
Car	30	Car	30	12.5	12.4	11.8	10.9	**9.4**	**9.4**
			0	14.8	10.7	12.4	10.8	10.1	**9.3**
		Factory	10	15.0	10.9	12.4	11.8	12.5	**10.1**
			0	18.0	17.1	**10.9**	11.4	17.2	13.9
			-10	33.8	35.4	22.6	25.1	29.7	**21.8**
		Babble	10	13.3	12.4	12.4	**10.9**	11.4	**10.9**
			0	14.1	**11.8**	12.5	**11.8**	13.3	13.3
			-10	17.2	16.5	18.0	**16.2**	28.3	20.4
Original non-telephone				12.5	*10.1*	9.9	*8.4*	*10.1*	**7.7**

- Equal error rates (EER, %) for matched and noise-mismatched detection of anger in narrowband telephone speech and matched detection in original, uncorrupted 16 kHz speech



- Detection-error-tradeoff (DET) curves for the system using LP, AR modulation filtering ($r = 8, s = 4$) and 64-component GMMs, evaluated on the original 16 kHz and high-SNR telephone data

Conclusions

- An angry speech detection system was developed and evaluated on unprocessed and telephone speech
- Linear prediction (LP) was more robust than FFT as the spectrum analysis method in MFCC computation
- Modulation filtering of feature vectors by autoregressive prediction across frames, using a maximum AR lag of approximately 300 ms, improved the detection performance overall
- Increasing the number of GMM components from 16 to 64 improved the performance in matched high-SNR conditions but also decreased the robustness in several mismatched conditions
- Future work: use the autoregressive modulation filtering approach in other paralinguistic classification problems