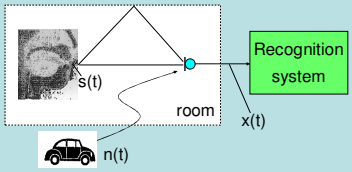
EXTRACTION OF ROBUST FEATURES BY COMBINING NOISE REDUCTION AND FDLP FOR THE RECOGNITION OF NOISY SPEECH SIGNALS IN HANDS-FREE MODE

Hans-Günter Hirsch, Institute for Pattern Recognition, Niederrhein University, Krefeld, Germany

Speech Input in Hands-free Mode

1



Two major effects

- Additive noise
- Reverberation

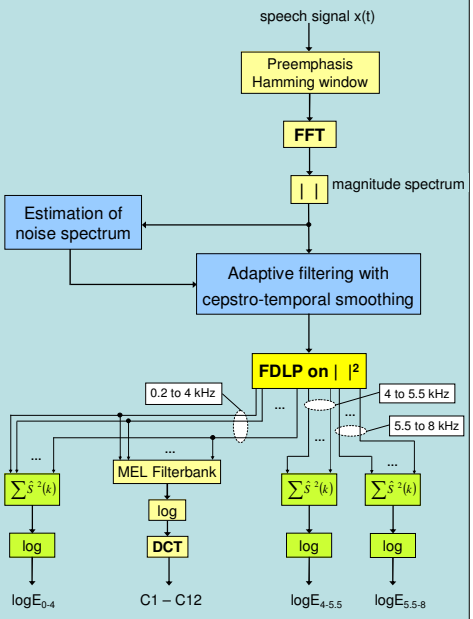
$$x(t) = s(t) * h_{rir}(t) + n(t)$$

Effects on acoustic features

- Stationary noise can be treated as additive component in the short-term spectrum
  ➔ Adaptive filtering in the spectral domain
- Reverberation causes a modification of the energy contour in each subband (can be roughly modeled as low pass filtering, Houtgast&Steeneken)
  ➔ Processing of subband energy contours (FDLP)

Robust Feature Extraction

Modified cepstral analysis scheme:

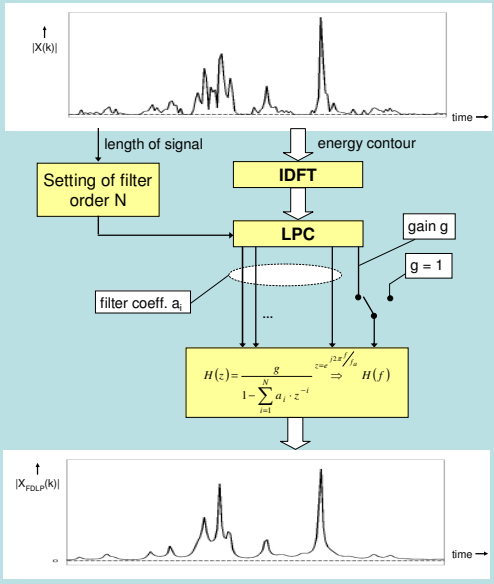


- Adaptive filtering in the DFT magnitude spectral domain
  - Cepstro temporal smoothing of filter characteristics (C. Breithaupt, T. Gerkmann, R. Martin) to reduce „musical tones“
- Frequency Domain Linear Prediction (FDLP) on contours of spectral magnitude in each DFT bin
- Calculation of 12 cepstral coefficients C1 to C12 and the logarithmic energy $\log E_{0\text{-}4}$ in the frequency range from 200 Hz up to 4 kHz
- For higher sampling frequencies of 11 or 16 kHz one or two additional energy parameters $\log E_{4\text{-}5.5}$ and $\log E_{5.5\text{-}8}$ are calculated.
- Adding Delta and Delta-Delta coefficients a feature vector consists of 39, 42 or 45 components dependent on the sampling frequency.
  - Training HMM parameters on the features of 16 kHz data, the HMMs can also be used for the recognition of signals at lower frequencies.

Frequency Domain Linear Prediction (FDLP)

2

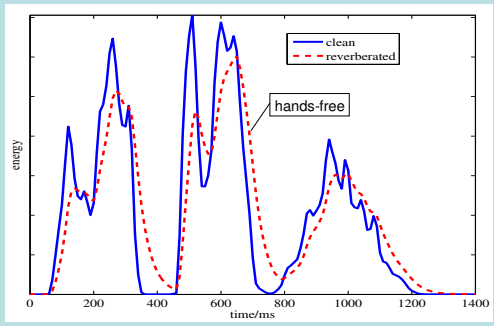
- Processing of energy contours in spectral subbands (here: temporal sequence of magnitude components in each DFT bin)



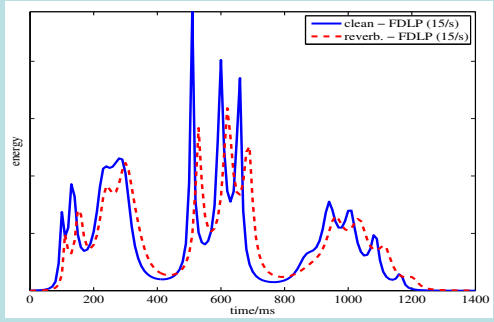
- The energy contour is considered as the filter characteristic of an all-pole model.
- FDLP leads to a smoothing of the energy contour in case of a low filter order.

Example

- The figure below contains the energy contours of a clean signal and a second hands-free version:



- Processing both contours by FDLP with a filter order of 15 per second:

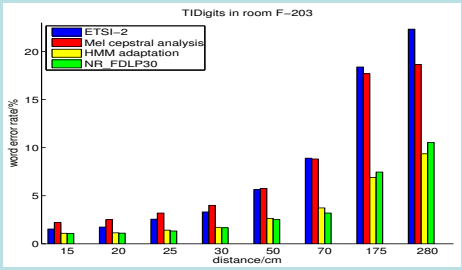


- The two contours after FDLP processing look similar.
- Indication that this processing step should create acoustic features that are fairly robust against the effects of reverberation.
- FDLP has been applied with a filter of order 30 per second and gain normalization in all listed recognition experiments. ➔ NR_FDLP30 in figures and tables

Recognition Experiments

3

1. **Recognition** of several versions of the **TIDigits** containing the effect of a hands-free recording in an office at a varying distance between speaker and microphone.



- Comparison with an alternative HMM adaptation scheme and with ETSI's robust front-end

2. Reverb Challenge Task

- Word error rates for the evaluation test sets:

analysis	SimData						
	Room1		Room2		Room3		Average
	Near	Far	Near	Far	Near	Far	
MFCC_0_D_A_Z	18.06	25.38	42.98	82.20	53.54	88.04	**51.68 %**
NR_FDLP30	21.08	25.50	27.72	58.76	31.72	69.73	**39.07 %**

analysis	RealData		
	Room1		Average
	Near	Far	
MFCC_0_D_A_Z	89.72	87.34	**88.53 %**
NR_FDLP30	75.53	72.48	**74.00 %**

- Considerable overall improvement
- Small degradation in case of clean speech or signals with low reverberation

3. Noisy WSJCAM0 data

- Only a small amount of stationary noise at a SNR of 20 dB is taken into account within the Reverb Challenge task. In a lot of applications noise signals will occur in the background at a lower SNR.
- Two further versions of the development test set have been created by adding car noise respectively interior noise to the 742 utterances at a SNR of 10 dB.

analysis	interior noise (SNR=10dB)			car noise (SNR=10dB)		
	set r1	set r2	set r3	set r1	set r2	set r3
MFCC_0_D_A_Z	44.5 %	45.5 %	44.4 %	40.8 %	41.9 %	39.6 %
NR_FDLP30	31.5 %	32.7 %	30.5 %	32.6 %	35.0 %	33.1 %

- High word error rates due to the noise
- Considerable improvements by applying the adaptive filtering technique

Discussion

- In most applications with a hands-free speech input additive noise and reverberation will occur together.
- Which one is the dominant effect?
- The author could assess within other experiments that noise becomes dominant at low SNR.

Acknowledgements

The author could experience the basics and the effects of FDLP processing during a research stay at the Center for Language and Speech Processing at the Johns Hopkins University in Baltimore, USA. The author would like to thank Hynek Hermansky as well as the whole speech group for the opportunity of doing research in a very stimulating environment.

The Center For Language and Speech Processing
at the Johns Hopkins University