Robust Speech Recognition by Combining a Robust Feature Extraction with an Adaptation of HMMs

Hans-Günter Hirsch, Andreas Kitzig

Department of Electrical Engineering and Computer Science, Niederrhein University of Applied Sciences, Krefeld
E-Mail: hans-guenter.hirsch@hs-niederrhein.de
Web: http://dnt.kr.hsnr.de/

Abstract

A method is presented to extract robust features from a noisy speech signal with the intention to improve the performance of an automatic speech recognition system. The processing is based on an adaptive filtering of the short-term spectra where the frequency response of the filter is smoothed with a cepstro-temporal approach [1], [2]. It turns out that the recognition performance is comparable with the performance that can be achieved with a robust feature extraction scheme standardized by ETSI [3]. Looking at the case of a hands-free speech input in a noisy and reverberant environment the recognition rates can be improved further by additionally adapting the HMMs to the acoustic conditions [4].

1 Extraction of robust features

A lot of investigations have been carried out to develop processing schemes for the extraction of robust acoustic features in the presence of background noise. A well known one is the algorithm that has been standardized by ETSI [3]. It is based on a two stage Wiener filtering where the filter characteristic is estimated in the frequency domain but the filtering itself is done in the time domain. The noise reduced time signal is again transformed to the frequency domain to determine the cepstral coefficients as acoustic parameters that describe the envelope of the short-term spectrum.

In this paper, we focus on an approach where the whole processing is done in the frequency domain without the need of a back transformation to the time domain. The processing scheme is shown in figure 1.

The filtering is based on a cepstral smoothing technique that has been developed by the authors of [1], [2]. They designed the algorithm mainly with the intention of generating an enhanced speech signal. Even though, they applied it also as preprocessing on the data base and the recognition experiments that are known under the abbreviation “Aurora-2” [5]. For their studies, they generated the noise reduced time signals and used them as input for the “Aurora-2” experiments. This version of the noise reduction scheme that was especially parameterised and optimised for the application to speech recognition is taken as basis for our investigations.

A preemphasis filtering is applied first on the speech signal sampled at a rate of 8 kHz. The samples of short segments with 25 ms duration are weighted with a Hamming window and transformed with a DFT of length 256.

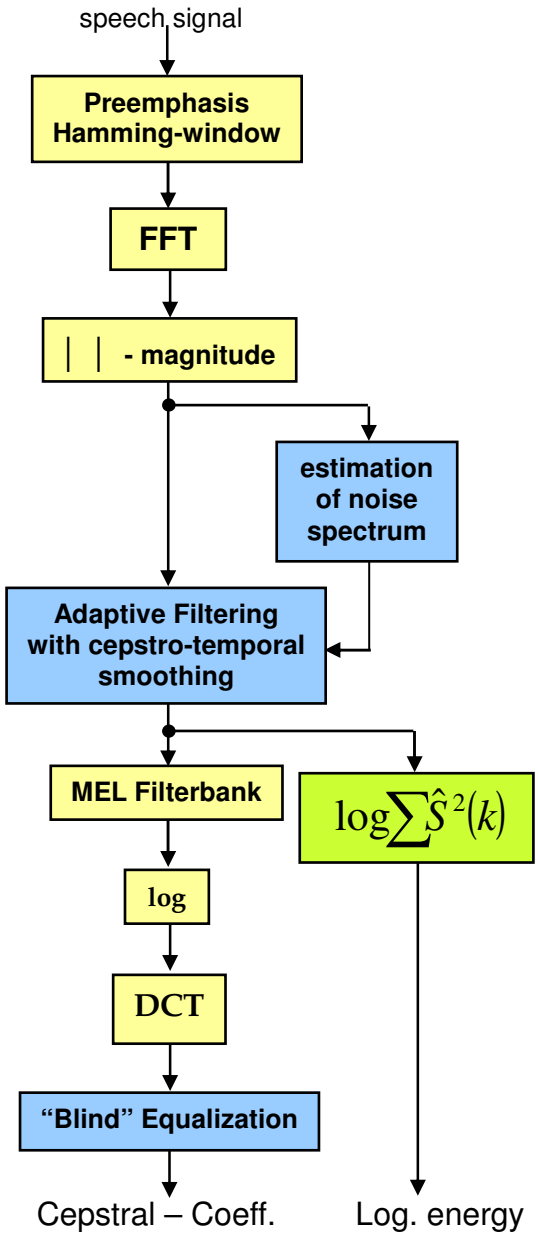

Figure 1: Scheme for the extraction of robust features

Consecutive spectra are calculated every 10 ms. Thus, complex spectra $Y(k,\ell)$ are determined with k as index of the DFT bin (0≤k≤128) and ℓ as index of the speech segment. The power density spectrum of the noise $P_n(k)$ is estimated based on an approach presented in [6].

ISBN 978-3-8007-3300-2

The power density spectrum of the clean speech

$$\hat{P}_S^{ml}(k,\ell) = (\gamma(k,\ell) - 1) \cdot P_n(k)$$

is estimated at the beginning based on the aposteriori SNR $\gamma(k,\ell) = \frac{|Y(k,\ell)|^2}{P_n(k)}$. The estimated logarithmic spectrum is transformed to the cepstral domain.

$$\hat{P}_S^{ceps}(q,\ell) = IDFT\left\{\ln\left(\hat{P}_S^{ml}(k,\ell)\right)\right\}$$

q is the index of the cepstral bin. The consecutive cepstral coefficients of each bin with index q are filtered along time. Especially the cepstral coefficients of higher order are filtered with a low pass filter with a very low cut-off frequency. This leads to the term "cepstro-temporal" smoothing to describe the smoothing effect of the low-pass filtering. In case of voiced speech segments the filtering with the extremely low cut-off frequency is not applied to cepstral bins that contain the corresponding values of the pitch frequency. Voiced segments contain a peak in the cepstrum at the bin corresponding to the pitch frequency. The intention of not extremely filtering these cepstral coefficients is to keep the harmonic structure in the logarithmic spectrum of voiced sounds. A detection of voiced segments in combination with an estimation of the pitch frequency is needed. Further details of the cepstro-temporal filtering can be found in [1], [2].

The smoothed cepstra $\overline{P}_S^{ceps}(q,\ell)$ are transformed again to the linear spectral domain.

$$\hat{P}_S^{ct}(k,\ell) = \exp\left(DFT\left\{\overline{P}_S^{ceps}(q,\ell)\right\}\right)$$

The estimated power density spectrum is taken to define an apriori SNR.

$$\hat{\xi}^{ct}(k,\ell) = \max\left\{\frac{\hat{P}_S^{ct}(k,\ell)}{P_n(k)}, \xi_{\min}\right\}$$

$$with\ 10\log_{10}\xi_{\min} = -25dB$$

The logarithmic SNR is limited to values above -25 dB. With the goal of estimating the logarithmic spectral amplitude a filter function is defined.

$$G(k,\ell) = \frac{\hat{\xi}^{ct}(k,\ell)}{1+\hat{\xi}^{ct}(k,\ell)} \cdot e^{0{,}5 \cdot \exp\mathrm{int}\left(\frac{\hat{\xi}^{ct}(k,\ell)}{1+\hat{\xi}^{ct}(k,\ell)} \cdot \gamma(k,\ell)\right)}$$

The final estimate of the clean speech spectrum is calculated by multiplying the noisy speech spectrum with the filter function.

$$\hat{S}(k,\ell) = G(k,\ell) \cdot Y(k,\ell)$$

The logarithmic energy logE is calculated as sum of the components of the power density spectrum $\left|\hat{S}(k,\ell)\right|^2$ for each frame with index ℓ. LogE is taken as one component of the feature vector. The Mel spectrum is determined from the estimated magnitude spectrum of the clean speech. A Discrete Cosine Transformation is applied on the logarithmic Mel spectrum to determine the Mel frequency cepstral coefficients C1 to C12. A "blind" spectral equalization scheme [3] is applied on the cepstral coefficients to compensate the influence of an unknown frequency characteristic, e.g. due to the frequency response of a microphone or a transmission channel.

The 13 acoustic parameters are extended by their first and second derivatives so that each feature vector contains 39 components in total. The calculation of the derivatives is done as defined in [3].

2 Recognition of Aurora-2 data

The robust feature extraction scheme is applied on the speech data of the Aurora-2 experiment. Aurora-2 contains the utterances of English digit sequences ("TI-digits") where different noises have been artificially added at several signal-to-noise ratios (SNRs). Comparing the achieved recognition results with the results presented in [1] is done as verification that the implementation of the filtering approach has been done correctly. The average word error rates are shown in table 1 for the noisy speech data of Set-A where the rates represent the average results for 4 different noise conditions and 5 SNRs in the range from 0 to 20 dB.

	Results of [1]	Own implementation	ETSI-2
word error rate	14,08 %	15,09 %	12,25 %

Table 1: Average word error rates for Set-A

The recognition performance that is achieved with the own implementation of a robust feature extraction is slightly worse than the results presented in [1]. The additional "blind" equalization scheme causes a small degradation. Disabling the equalization an average word error rate of 14,82 % is achieved. But the equalization is helpful in conditions where the speech is modified by unknown frequency responses. This condition was not investigated in [1]. The difference between the error rate of 14,08 % that has been achieved by the authors of [2] and the error rate of 14,82 % may be explained by an additional smoothing effect in the processing of [1] due to the back transformation of the estimated clean speech spectra to the time domain before applying again a cepstral analysis scheme.

Comparing the results with the robust feature extraction as defined by ETSI [3] slightly better average results are achieved with the ETSI scheme. Analysing the different word error rates a bit more in detail it turns out that the degradation is mainly due to worse results for low SNR conditions.

This is also the reason for the higher error rates in table 2 when comparing the recognition results for all subsets of the Aurora-2 task. Set-B contains the same number of noise conditions but different noise signals. Set-C contains one noise condition of Set-A and one condition of set-B but including an additional bandpass filtering of the noisy data to simulate the transmission over a telephone line.

ISBN 978-3-8007-3300-2

	Set-A	Set-B	Set-C
Cepstro-temporal smoothing	15,09 %	15,34 %	18,81 %
ETSI-2	12,25 %	12,90 %	13,97 %

Table 2: Average word error rates for all subsets

The ETSI noise reduction scheme creates especially lower error rates for noise conditions with a low SNR. Furthermore it has to be considered that the ETSI front-end has been developed over years to reach an optimal recognition performance on the Aurora-2 task. In the next section results will be presented for other tasks of recognizing noisy speech data. It will be shown that the highest performance is not always achieved by applying the ETSI robust feature extraction scheme.

3 Recognition of Aurora-5 data

The data base and recognition experiments known under the abbreviation “Aurora-5” contain speech data that also take into account the recording in hands-free mode in a noisy environment [5]. As for Aurora-2 the TI-digits have been taken as basis to artificially create the noisy and reverberant versions. But in comparison to Aurora-2 the whole set of 8700 test utterances from adults are used to create the noisy data for each noise condition. Only 1000 utterances have been randomly selected to create the noisy data in each condition for the Aurora-2 experiments. Furthermore a whole set of noise recordings representing a desired noise condition is taken to create the noisy speech data. For Aurora-2 only a single recording has been used to create the data for a specific noise condition. Thus, Aurora-5 includes a higher variance of noise conditions than Aurora-2.

Some recognition results are presented in figures 2 and 3. The word error rates are plotted for applying

- a cepstral analysis scheme without noise reduction,
- the ETSI front-end and
- the feature extraction scheme based on a filtering with cepstro-temporal smoothing.

A feature vector consists of 39 parameters (12 MFCCs, logE, Deltas and Delta-Deltas) in all cases. Gender dependent HMMs for the 11 English digits are used with 16 states and a mixture of 2 Gaussians per state.

Word error rates are shown in figure 2 for the condition of a hands-free speech input in a noisy car environment. A considerable improvement can be seen when comparing the recognition results that are achieved with the robust front-ends against the results for a feature extraction without any noise reduction. The error rates for the noise reduction with cepstro-temporal smoothing are slightly lower than the ones for the ETSI front-end in the SNR range from 5 to 15 dB. Only for the SNR of 0 dB the ETSI front-end has a higher performance. This corresponds to the results on the Aurora-2 data where the lower average error rates are due to a higher performance at low SNR conditions.

Looking at the condition of a hands-free speech input in a noisy office (figure 3) the error rates for the ETSI front-end are slightly worse in all cases. The reason is the additional modification of the speech signals by the reverberation of the room. It turns out that the ETSI front-end can not handle this effect as good as the robust feature extraction based on cepstro-temporal smoothing.

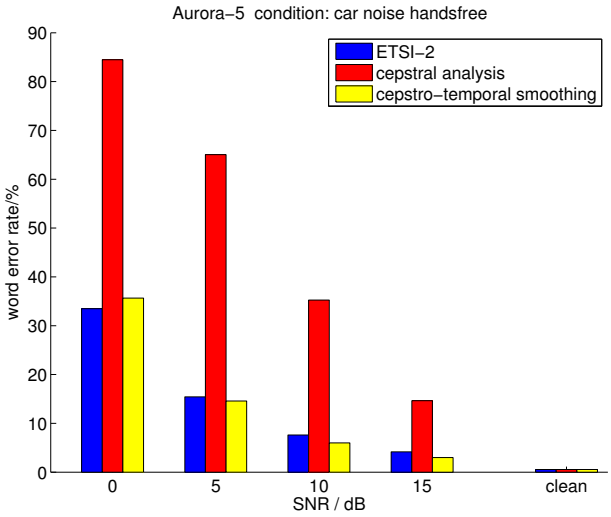

Figure 2: Word error rates for the hands-free speech input inside a noisy car

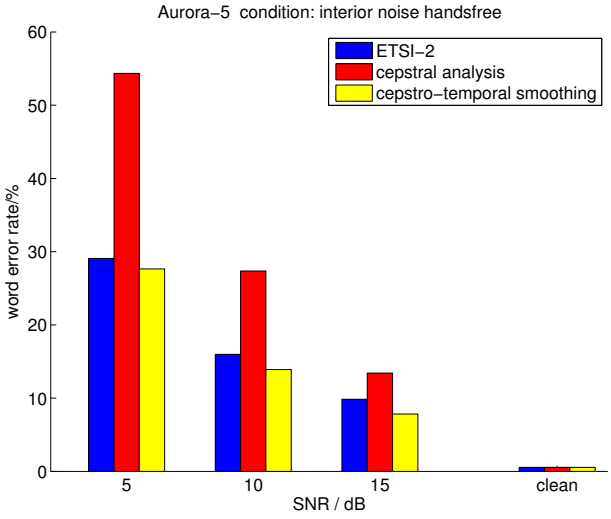

Figure 3: Word error rates for the hands-free speech input inside an noisy office

4 Recognition results with an additional HMM adaptation

Looking at the hands-free speech input in a room the performance of a speech recognition system decreases due to the reverberation in a room. The term reverberation describes the effect that the recording of a speech signal in a room contains not only the signal that reaches the microphone on the direct path from the speaker. The recording contains also a huge number of delayed and

ISBN 978-3-8007-3300-2

attenuated versions of the signal due to the reflections of the sound at the walls and the furniture. This effect can be described by a system with an impulse response whose envelope ideally has an exponentially decaying characteristic. Thus, the acoustic features of a speech segment will occur over a longer time interval when recording a speech signal in hands-free mode inside a room. The modification of a speech signal due to reverberation can not be compensated by a noise reduction scheme that is based on an adaptive filtering of short-term spectra as used for the robust extraction of acoustic features. To compensate this modification to some extent we have developed a HMM adaptation scheme to adapt the means of the logarithmic energy and the cepstral coefficients in each HMM state to the acoustic environment of the room [4]. Based on an estimation of the reverberation time as single parameter we apply a model to consider the influence of the reflections in a room as a low-pass filtering of the energy and the subband energy contours along time. The mean cepstral coefficients as contained in the HMMs are transformed back to the Mel spectrum where the low-pass filtering of the subband energy contours is applied. The modified Mel spectra are transformed to the cepstral domain again to be used as adapted energy and cepstral mean parameters.

We applied this adaptation approach in combination with the robust feature extraction scheme to improve the recognition performance for the acoustic condition of a hands-free speech input. The Aurora-5 data base contains versions of the TIDigits where the speech input in hands-free mode is simulated in two different rooms. Looking at the condition without noise in the background the word error rates are achieved that are presented in table 3.

	office	living room
without adaptation	4,34 %	9,38 %
with adaptation	3,30 %	5,95 %

Table 3: Word error rates at a hands-free speech input

The additional HMM adaptation is helpful to considerably improve the error rates.

Besides adapting the HMMs to the effect of reverberation the adaptation scheme has been developed in its original version to compensate the influence of noise in the background. Instead of “subtracting” the estimated noise spectrum as the adaptive filtering approach can be described for realising a noise reduction, the estimated noise spectrum is added in the linear Mel spectral domain to adapt the mean parameters of all HMM states. This HMM adaptation scheme has been applied so far only in case of feature extraction scheme without noise reduction. Due to the fact that the adaptive filtering is applied with a limitation to a maximum attenuation we investigated the HMM adaptation to additive noise also in case of using the feature extraction including the cepstro-temporal smoothing. The limited attenuation causes a remaining amount of noise after the filtering.

Word error rates are presented in figure 4 for the hands-free speech input in a noisy office environment like in figure 3.

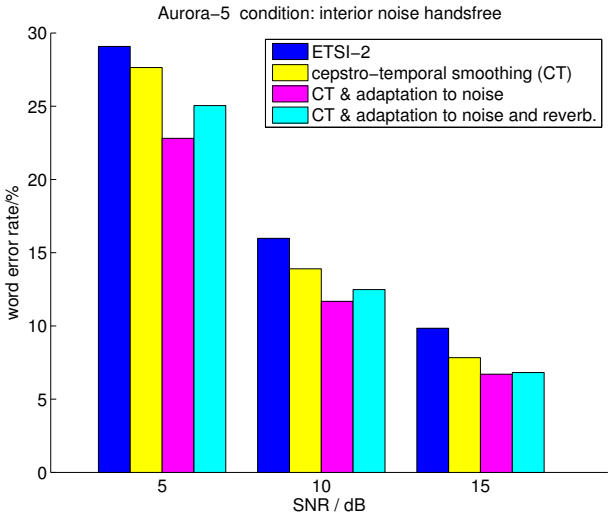

Figure 4: Word error rates for the hands-free speech input inside an noisy office

Besides the error rates for the two cases of applying the ETSI front-end and the noise reduction based on the cepstro-temporal smoothing two further results are presented for each SNR condition. These include the additional HMM adaptation to additive noise only and the combined adaptation to additive noise and to the effect of reverberation. It turns out that the adaptation to additive noise helps to improve the recognition performance. The combined adaptation to noise and reverberation allows an improved recognition in comparison to not applying an adaptation at all but it is not as efficient as adapting to noise only. The additive noise seems to be the dominant effect especially in case of a low SNR. As a conclusion the adaptation to reverberation should be disabled for a SNR below a certain threshold.

References

[1] C. Breithaupt. Noise reduction algorithms for speech communications – Statistical analysis and improved estimation procedures, dissertation at the Ruhr-University Bochum, 2008

[2] C. Breithaupt, R. Martin. DFT based speech enhancement for robust automatic speech recognition, ITG Fachtagung Sprachkommunikation, Aachen, 2008

[3] ETSI standard document. Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm. ETSI document ES 202 050 v1.1.3 (2003-11), Nov. 2003.

[4] H.G. Hirsch. Automatic speech recognition in adverse acoustic conditions, in *Advances in Digital Speech Transmission*, John Wiley and sons, 2008

[5] Aurora project. http://aurora.hsnr.de, data available at http://www.elda.org, 2007.

[6] H.-G. Hirsch, C. Ehrlicher. Noise estimation techniques für robust speech recognition, ICASSP, 2005

ISBN 978-3-8007-3300-2