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 of Applied Sciences, Krefeld, Germany

ABSTRACT

A processing scheme is presented to create robust acoustic features for the recognition of noisy speech signals at a hands-free speech input in reverberant environments. The robustness is achieved by extending a Mel cepstral analysis scheme by the additional processing steps of a noise reduction and a frequency domain linear prediction (FDLP). FDLP is a signal processing technique to modify the energy contours of subband signals. We analyze the effects of FDLP and show its ability to create robust features by looking at the energy contours of clean and reverberant speech signals. This technique is combined with a noise reduction approach that is based on an adaptive filtering in the spectral domain. We analyze the effectiveness of our approach by running several different recognition experiments and evaluating the achieved recognition results. The experiments contain recognition tasks with different complexity and with a different modeling of speech by applying whole-word or phoneme based reference models. Besides the recognition of reverberant versions of the TIDigits we focus on the experiments as defined by the Reverb Challenge task where the focus is put on the effects of a hands-free speech input without a lot of noise in the background. Furthermore, we created two further versions of the development test data to include the effects of typical noise scenarios at a realistic signal-to-noise ratio (SNR) in room environments.

Index Terms— robust speech recognition, robust acoustic features, frequency domain linear prediction, hands-free speech input

1. INTRODUCTION

The application of speech recognition is especially useful when people do not have their hands available for controlling devices like in the situation of driving a car. Such scenarios come along with the need of a hands-free speech input. Unfortunately, the acoustic environment modifies the speech signal due to the effects of noise and reverberation. Thus, a lot of approaches have been developed in the field of signal processing to compensate these distortion effects. Many approaches are based on the usage of two or more microphones with respect to humans who perceive sound with their two ears. But so far, a single microphone is used in most applications especially in the field of speech recognition. Therefore, we focus on the processing of a single channel signal within the investigations presented in this paper. Regarding the effect of an additive noise signal in the background, most approaches are based on an adaptive filtering, e.g. [1], [2]. Based on a technique for smoothing the adaptive filter characteristics [3], we developed an analysis scheme that includes this type of filtering [4]. Investigations have been carried out to compensate the effects of reverberation in case only one microphone is available, e.g. [5], [6], [7]. FDLP can be seen as a further technique for the determination of robust features [8], [9]. We combine the adaptive filtering with the FDLP processing in our approach to compensate the effects of noise and reverberation at the same time.

An alternative approach for improving the robustness is based on the adaptation of the Hidden-Markov models (HMMs) that are used as reference models in most recognition systems nowadays. Instead of extracting features that are independent of the acoustic input conditions, a set of parameters is estimated that contains and defines the distortion effects. These parameters are taken to adapt the HMMs to the acoustic conditions of each individual utterance. We have developed such an adaptation scheme for compensating the influence of background noise, an unknown frequency characteristics and reverberation [10]. So far, the focus of our work in the field of adaptation was on the usage of whole-word HMMs. Therefore, we were not able to apply this adaptation scheme on the triphone based modeling in the context of the Reverb Challenge task within the limited time frame.

We present our analysis scheme for the extraction of robust features in the next section. Thereby, we focus on the explanation of the FDLP processing. Its ability is shown to create robust features with respect to the modifications of the signals recorded at hands-free mode in a reverberant environment. The recognition results are presented in a further section for the different recognition experiments.

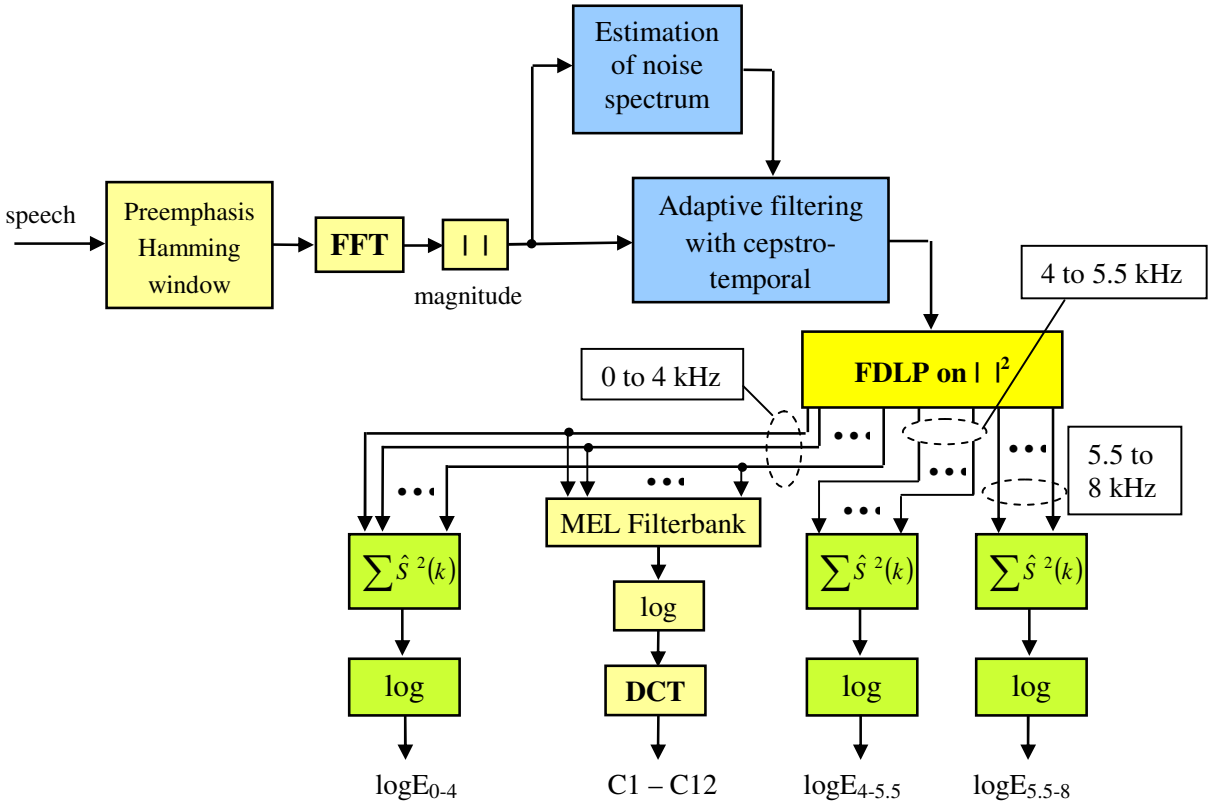

Figure 1: *Robust feature extraction scheme*

2. EXTRACTION OF ROBUST FEATURES

We present the signal processing scheme in this section as it has been used within these investigations. The two processing blocks for increasing the robustness against additive noise and against reverberation are described in separate subsections. Furthermore, we motivate the use of FDLP as a technique to create robust features in case of a hands-free speech input in a reverberant environment.

2.1. Mel cepstral based analysis scheme

The complete signal processing scheme is shown in figure 1 as it has been applied to improve speech recognition in the presence of noise at a hands-free speech input in a reverberant environment. The scheme is based on the well known Mel frequency cepstral analysis that has been extended by a few processing blocks. The additional blocks have been included with the intention to extract acoustic features that are fairly robust against the distortion effects of noise and reverberation. The main component of the cepstral analysis is the DFT (Discrete Fourier Transform) to perform a short-term spectral analysis. Within the Reverb Challenge framework [11] speech data are processed that have been sampled at a rate of 16 kHz. Frames of duration 25 ms containing 400 samples are transformed with a DFT of length 512. We apply a preemphasis filtering and the weighting of the 400 samples with a Hamming window before transforming the samples. The Hamming window is shifted by 10 ms to estimate consecutive short-term spectra. The DFT magnitude spectrum is filtered with an adaptive filter scheme to reduce the influence of a stationary noise floor. Furthermore, we apply the FDLP processing on the filtered DFT spectrum. Details of the adaptive filtering and the FDLP processing are presented in the following sections.

The filtered and modified DFT components are separately processed in the three frequency regions from 0 to 4 kHz, from 4 to 5.5 kHz and from 5.5 to 8 kHz. We use a Mel filterbank to reduce the 129 DFT components in the range from 0 to 4 kHz to 24 Mel spectral components by splitting the frequency range into the corresponding number of nonlinearly spaced frequency bands and performing a weighted adding of the DFT magnitude components in each band. Applying a DCT (Discrete Cosine Transformation) on the logarithmic Mel spectrum we determine 12 cepstral coefficients C1 to C12. These cepstral coefficients represent features that are statistically independent to a large extent. Three energy coefficients are determined as further coefficients by summing up the spectral magnitudes in the three mentioned frequency regions. Thus, we get 15 acoustic features that are taken as part of each feature vector and that are estimated every 10 ms. 15 Delta and 15 Delta-delta

coefficients are added to each feature vector by filtering the temporal contour of each feature coefficient with a set of filter coefficients that has been proposed as part of the robust front-end standardized by ETSI [2].

The motivation for calculating energy coefficients separately in each of the three frequency regions has been the result of earlier investigations [12]. At that time we designed a front-end to allow the processing of speech signals sampled at rates of 8 kHz or 11 kHz or 16 kHz. The front-end can be used as part of a distributed speech recognition system where the feature extraction is done in any type of terminal, e.g. a mobile phone or a PC, and the recognition is performed at a system located at a central position inside a network. The idea was the generation of exactly the same features in the frequency region from 0 to 4 kHz independent of the sampling rate. Thus, it would be sufficient to train the central recognizer only on speech signals sampled at 16 kHz. In case of getting the features from a terminal that works at a sampling frequency of 8 or 11 kHz the distribution functions of the features in the higher frequency regions are ignored for the calculation of the emission probabilities. Special care has been taken of the signal processing so that we create features in the lower frequency region that take the same value for signals sampled at different rates. For example, the preemphasis filtering is realized in the usual way with a FIR filter of first order for data sampled at 8 kHz. But special filters have been designed for 11 and 16 kHz so that the filter characteristics correspond with the filter shape of the 8 kHz version in the frequency range up to 4 kHz.

2.2. Adaptive Filtering

To reduce the influence of a stationary noise in the background an adaptive filtering is applied in the spectral domain. We process the magnitudes of the 257 DFT components that are calculated for each frame of the input signal and that describe the spectrum in the range from 0 to 8 kHz. To define the characteristics of the filter an estimation of the noise spectrum is needed. An own approach [13] is applied for detecting speech pauses and estimating the noise spectrum from the spectra in the pause segments. We look at the energy contours in the DFT subbands. In case we notice the exceeding of an adaptive energy threshold in a certain number of subbands we take this as indication for the beginning of speech. In a similar way the ending of speech is detected when the subband energy falls below this threshold in a predefined number of subbands. The noise spectrum is estimated by calculating the moving average of the DFT spectra during speech pauses. The estimated noise spectrum is taken to realize an adaptive filtering [1]. The adaptive filtering contains a so called cepstro-temporal smoothing of the filter characteristics [3]. The smoothing is based on transforming a first estimate of the clean speech spectrum into the cepstral domain. The contours of most cepstral coefficients are smoothed along time. We transform the modified cepstra back into the spectral domain to estimate the smoothed characteristics of the adaptive filter. This type of smoothing has been introduced for the purpose of speech enhancement to reduce the amount of musical tones and to improve the subjective quality of the noise reduced signal [3].

2.3. Frequency Domain Linear Prediction (FDLP)

The goal of FDLP is the modification of the energy contour in each subband so that this modification covers the effects of reverberation to a large extent. FDLP is based on the idea of treating the short-term energy contour in a single subband as the filter characteristic of a DPCM (Differential Pulse Code Modulation) filter. DPCM is usually applied in the field of speech coding to encode the spectral characteristics of short speech segments with a length of about 20 ms. The filter coefficients of a FIR filter are estimated for the encoding to create a differential signal with minimum energy. Taking the zeros of the encoding filter as poles of an inverse filter on the decoding side we can describe the envelope of the short-term speech spectrum with an all-pole filter characteristic. The locations of the poles correspond to the formant frequencies of the vocal tract in case of articulating a voiced sound. The analysis of short segments and the transmission of the filter coefficients are known as linear predictive coding (LPC).

The effect of analyzing the envelope or a smoothed version of a spectrum has been taken as motivation to apply the idea of DPCM filtering in a different domain [8], [9]. Creating a smoothed version of the short-term energy contour for a whole speech utterance is the idea of the method referred to as FDLP. Therefore, the contour of the short-term energy in a subband is treated as a spectrum. The corresponding “time signal” is calculated by applying an inverse DFT (IDFT) on the contour as it shown in figure 2. In our case, the sequence of the DFT coefficients in each DFT bin is considered as the energy contour in the corresponding subband. So, the processing shown in figure 2 is individually applied to each of the 257 DFT bins. The output signal of the IDFT is taken as input to estimate the parameters of the LPC filter. The order of the filter is chosen dependent on the length of the energy contour respectively the length of the whole speech signal. We define the filter complexity with a value describing the filter order per second. Typically, an order of about 30 per second is chosen.

The predictor coefficients a_i can be taken to describe a smoothed version of the energy contour as the spectral characteristics of an all-pole filter:

$$H(z) = \frac{g}{1 - \sum a_i \cdot z^{-i}}$$

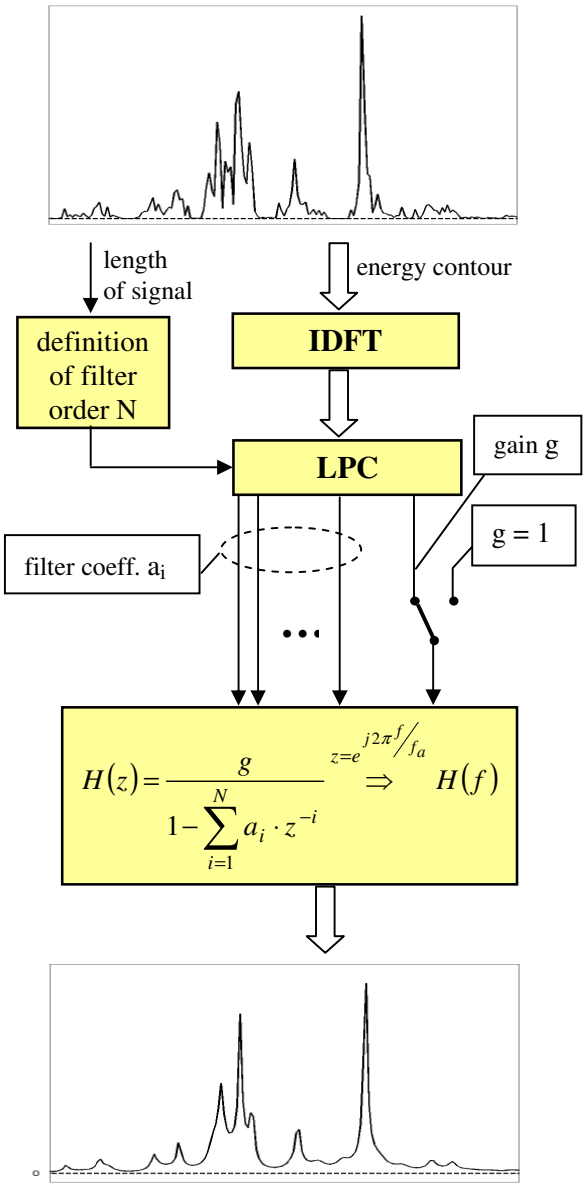

Figure 2: *FDLP processing scheme*

The poles of a DPCM filter usually define the frequencies where the spectrum has its maxima. In our case, the poles define the position of the peaks in the energy contour. Choosing an order of 30 poles per second corresponds to the definition of up to 15 peaks per second.

It turned out in earlier investigations that it can be of advantage to neglect the gain factor g [14]. This is visualized in figure 2 by moving the switch to the position where g is equal 1. The gain normalization leads to an accumulated subband energy that takes the same value in each subband. The effect will be similar as applying the well known cepstral mean normalization. In case of a spectral weighting, e.g., due to the microphone, the degradation of the error rates will be usually reduced by such a type of frequency dependent gain normalization.

2.4. The effect of FDLP in case of reverberation

Looking at the condition of a hands-free speech input inside a room the modification of the speech signal by the acoustic environment has to be considered. The sound propagation can be modeled as an additive superposition of the sound on the direct path from the speaker to the microphone and a huge number of single and multiple reflections at the walls and the interior. In terms of signal processing the transmission in a room can be described as a convolution of the speech signal and the room impulse response (RIR). The example of a RIR is shown in figure 3 as it has been estimated for the transmission in a small conference room at a distance of 2.5 m between speaker and microphone.

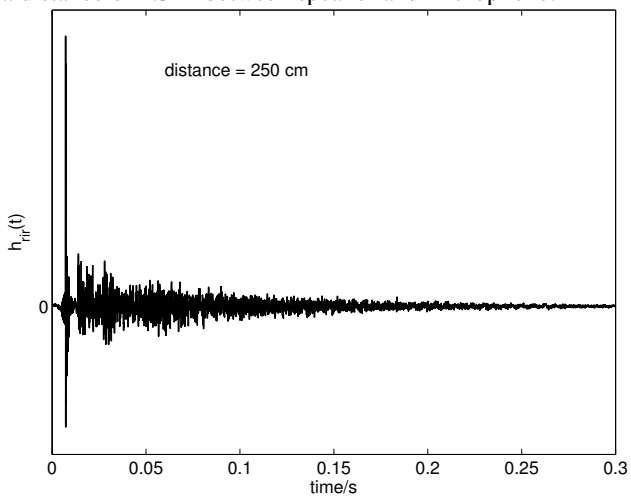

Figure 3: *RIR measured in a meeting room*

It becomes obvious that the room impulse response has a fairly long duration of several hundred milliseconds. The peak value at the beginning of the RIR corresponds to the sound on the direct path between speaker and microphone. The so called reverberation time T_{60} is the parameter that can be taken to define the decrease of the RIR amplitudes along time. The reverberation time corresponds to the time until the sound level decreases by 60 dB when switching off a stationary sound excitation inside a room. T_{60} takes a value of about 0.65 s for the RIR shown in figure 3.

In general, the influence of a hands-free input on the speech signal can be described by 2 aspects. The first one is the effect called “spectral coloration”. Transforming the RIR to the spectral domain the acoustic transmission in the room can be modeled as the multiplication of the speech spectrum and the corresponding transfer function. But, considering the analysis of short segments as it is done in the field of speech recognition the description as a multiplication of spectra does not hold due to the fact that the RIR is much longer

than the analysis window of the short-time spectral analysis. Looking at the influence of a hands-free speech input as a "spectral coloration" only might be approximately correct when either the RIR is fairly short or the speaker is close to the microphone. In case of a close speaker the energy of the late reflections is relatively low in comparison to the energy of the direct sound. As a consequence of a long RIR all approaches that try to compensate the influence of a nearly stationary transfer function, like for example cepstral mean normalization, will not be able to handle the effects of reverberation overall.

The second aspect that describes the effect of reverberation becomes obvious when examining the contours of the short-term energy in subbands. The transmission in a reverberant environment can be approximately described as a low-pass filtering of these energy contours [15]. In figure 4 two versions are shown for the energy contour of a speech signal containing the utterance of three digits.

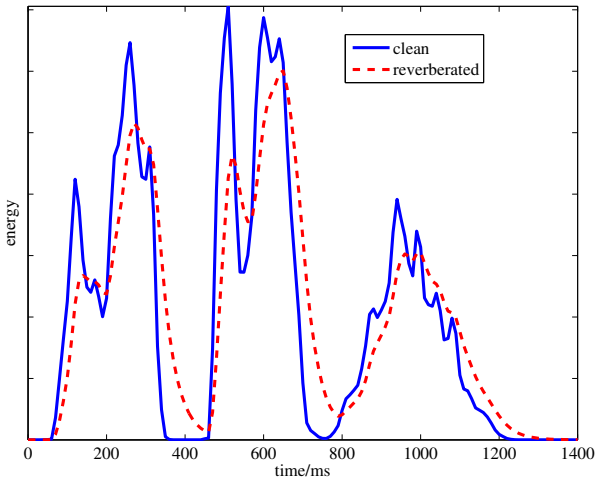

Figure 4: *Energy contours of a clean signal and after recording in hands-free mode.*

The envelope of the clean signal as well as the contour after the transmission in a room are shown assuming an exponentially decaying RIR with a reverberation time of about 0.5 s. We can see the so called reverberation "tails" due to the reflections of the sound. This leads to a smearing of the energy so that for example the pauses between the words are no longer characterized by energy close to zero. Furthermore, the spectral characteristics of a sound with low energy will be partly covered by a preceding sound with high energy.

To analyze the influence of FDLP processing in this context two further contours are visualized in figure 5. Both contours shown in figure 4 have been processed by FDLP with a filter order of 15 per second. Besides a small shift in time due to the reverberation the two contours in figure 5 look very similar. Especially during speech pauses both curves have a similar characteristic. This can be taken as an indication that the acoustic features will also be fairly similar after processing the energy contours in subbands from either clean or reverberant signals with FDLP. The creation of similar features is exactly the goal of a "robust" feature extraction scheme. This might indicate a good usability for achieving a high recognition performance of reverberant signals recorded in hands-free mode.

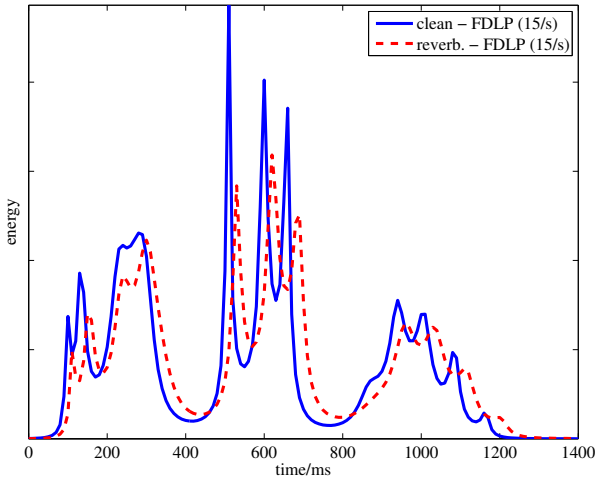

Figure 5: *Energy contours after processing the clean signal and the reverberant signal by FDLP with a filter order of 15 per second.*

3. RECOGNITION EXPERIMENTS

We present the results of several recognition experiments in this section. The first experiment is based on the TIDigits database [16] and the usage of whole word HMMs for the recognition. We focus on the effect of reverberation only by creating reverberant versions of the TIDigits with a set of RIRs that have been estimated inside a conference room at a varying distance between speaker and microphone. The Reverb Challenge task [11] is considered as the second experiment where the focus is also on the effect of reverberation. But the task contains the recognition of a large vocabulary of about 5000 words. The experiments are based on the usage of triphone HMMs. To investigate also the effects of a stronger background noise as it will occur in a lot of practical applications we created two further versions of the WSJCAM0 data [17] that have been chosen as development test data within the Reverb Challenge task. We present the results on these data in a third subsection.

3.1. Reverberant digits at a varying distance between speaker and microphone

A lot of investigations in the field of robust recognition in hands-free mode examined a speech input at a large distance between speaker and microphone. We have been interested to analyze the recognition performance dependent on the

distance between speaker and microphone. An increasing distance comes along with a decreasing ratio between the energy of the direct sound and the energy of the later reflections. This leads to a higher distortion of the speech signal for an increasing distance. We measured a set of RIRs in a small meeting room (F-101) at 7 different positions of the microphone. These impulse responses are available for download [18]. A hardware and software setup developed in the SpeeCon project [19] was applied to measure the RIRs. Several versions of the TIDigits have been created by convolving the clean signals of the TIDigits test set with each RIR. In contrast to the Reverb Challenge data we used a version of the TIDigits that have been sampled at a rate of 8 kHz. With respect to the processing scheme shown in figure 1 the energy values in the frequency regions above 4 kHz were not available as acoustic parameters.

The word error rates are shown in figure 6 for the 7 different positions of the microphone in the meeting room. We compared the results applying 4 different analysis and recognition schemes. In all cases, two gender dependent whole-word HMMs have been trained for each English digit (zero – nine, oh) based on the analysis of the complete training set of the clean TIDigits. Each HMM consists of 16 states with a mixture of 2 Gaussians to model the occurrence of each acoustic parameter in each state. A set of 39 acoustic parameters has been used in all configurations. The set consists of the 12 Mel cepstral coefficients and the frame energy as well as the corresponding Delta and Delta-delta coefficients.

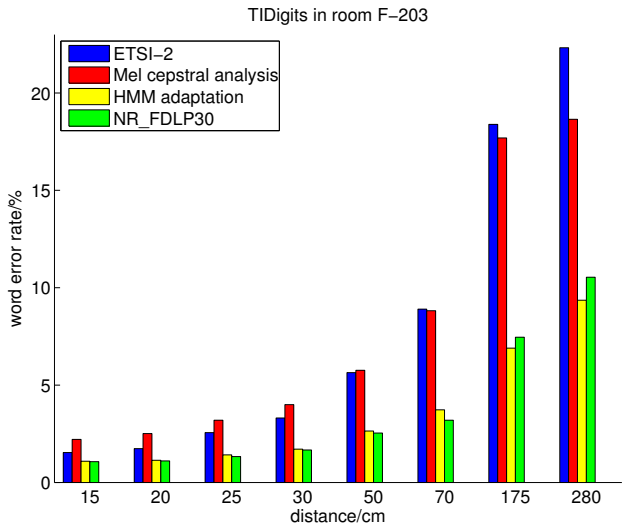

Figure 6: *Word error rates for the recognition in hands-free mode at different distances.*

The highest error rates are achieved for the robust front-end standardized by ETSI [2] as well as for an own Mel cepstral analysis without any additional processing or adaptation to compensate the effects of reverberation. The own processing corresponds to the scheme in figure 1 without the adaptive filtering and the FDLP processing. As expected, the performance decreases when moving to higher distances where the energy of the direct sound is less in relation to the energy of the reflected sound. Surprisingly, the performance of the ETSI scheme is even less than the standard cepstral analysis. But the effects of reverberation were not taken into account and were not evaluated at the time when this robust feature extraction scheme was developed.

Based on the Mel cepstral analysis we applied our HMM adaptation scheme [10]. We adapt the cepstral coefficients and the energy coefficient as well as the corresponding Delta and Delta-Delta coefficients as they occur as means of the Gaussian distributions in each HMM state. The reverberation time and the spectrum of a stationary background noise are estimated as adaptation parameter. We perform the adaptation once during each speech input when the beginning of speech is detected. The recognition performance considerably increases when applying the adaptation technique. This is also true when applying the FDLP processing. The reduction of error rates is almost the same for both techniques. FDLP is done with a filter order of 30 per second and including the gain normalization.

Looking at the word error rates in table 1 for the recognition of the clean TIDigits, FDLP causes a degradation of the recognition performance in comparison to the Mel cepstral analysis without or with HMM adaptation. This might be caused by the smearing effect on the energy contours.

Table 1: *Word error rates for clean TIDigits.*

Mel cepstral analysis	HMM adaptation	FDLP with gain normalization
0.56 %	0.55 %	0.85 %

3.2. The Reverb Challenge task

We applied our processing scheme on the data of the Reverb Challenge task. Besides substituting the feature extraction, we have been using the provided framework [11] for training triphone HMMs with the clean WSJCAM0 data [17]. Furthermore, we applied the provided scripts for performing the recognition experiments with HTK [21]. Two recognition experiments have been defined as part of the framework.

The first one is based on a set of 742 sentences from the WSJCAM0 data. These data have been used as development test data. The set has been split into 3 subsets (r1, r2 and r3). The word error rates are listed in table 2 for the recognition of the clean data. Results are presented for the feature extraction as proposed in the framework and as done with HTK in comparison to applying our feature extraction scheme. As already shown in the preceding section FDLP leads to a deterioration of the recognition performance when

applied on clean data. It looks like the smoothing of the subband energy contours causes a certain loss of information.

Table 2: *Word error rates for the clean development test set.*

analysis	devset – clean		
	set r1	set r2	set r3
MFCC_0_D_A_Z	10.50%	11.51%	10.81%
NR_FDLP30	17.65%	18.12%	18.25%

The data of the three subsets have been taken to create several reverberant versions by convolving the speech signals with a set of RIRs to simulate the recording in hands-free mode. Furthermore, recorded background noise has been added at a SNR of 20 dB. The RIRs have been measured in 3 different rooms at a smaller and at a larger distance between speaker and microphone. Each RIR of one of the 3 rooms is separately applied to one of the 3 subsets. Thus, the subsets get the indication room1, room2 and room3 in table 3 showing the word error rates for the distorted versions of the 3 subsets.

Table 3: *Error rates for the development test set.*

analysis	devset – **near**		
	room1	room2	room3
MFCC_0_D_A_Z	15.29%	43.90%	51.95%
NR_FDLP30	20.06%	26.57%	29.90%
	devset – **far**		
MFCC_0_D_A_Z	25.29%	85.80%	88.90%
NR_FDLP30	27.51%	68.33%	70.72%

The case of a smaller distance is indicated by the term near. The large distance is indicated by the term far. It turns out that the FDLP processing leads to a considerable reduction of the error rates in the presence of a stronger reverberation effect. The improvement becomes especially obvious in the case of a higher reverberation time and a larger distance between speaker and microphone.

Later on, a second subset of another 742 WSJCAM0 utterances has been released as evaluation set. The splitting in 3 subsets and the distortion of the data has been done in a similar way than for the development set. The corresponding error rates are listed in table 4. The term "SimData" is introduced in table 4 indicating the creation of these data by a simulation of the hands-free recording in a slightly noisy environment. Almost the same relative improvement or degradation occurs as for the development set when comparing the results for the two feature extraction schemes.

The second recognition experiment is based on a set of speech signals that have been recorded in hands-free mode [20]. The term "RealData" is used in table 4 as reference to these experiments. In total this set contains 186 sentences where each utterance has been recorded at a smaller distance (near) and at a larger distance (far). The word error rates are listed in table 4 for the evaluation set only. As in case of the first experiment with "SimData" there is almost no difference between the results achieved with the development set and with the evaluation set. We observe a considerable improvement of the recognition performance with our feature extraction scheme.

3.3. Noisy WSJCAM0 test sets

The Reverb Challenge task considers the presence of background noise only to a small extent. A stationary noise signal is added at the high SNR of 20 dB. In a lot of practical applications the influence of noise will be stronger than it is taken into account by the Reverb Challenge task. Thus, we created two further versions of the development test set consisting of 742 utterances. We added car noise at a SNR of 10 dB for one version and noise signals recorded inside rooms for the other version. In case of the car noise we applied also a RIR that we have measured inside a car. We randomly selected noise segments from a number of recordings inside different cars and inside different rooms. Word error rates are listed in table 5. We can proof with this experiment that our analysis technique shows a higher robustness against additive noise.

Table 4: *Word error rates (%) for the reverberant **evaluation** test sets.*

analysis	SimData							RealData		
	Room1		Room2		Room3		Average	Room1		Average
	Near	Far	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 %	89.72	87.34	88.53 %
NR_FDLP30	21.08	25.50	27.72	58.76	31.72	69.73	39.07 %	75.53	72.48	74.00 %

Table 5: *Error rates for the noisy development test set*

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

4. CONCLUSIONS

We have presented a feature extraction scheme to cope with the effects of additive noise and a hands-free recording in a reverberant environment. Recognition results have been presented for different experiments. We could proof that we can improve the recognition performance in comparison to a typical Mel cepstral analysis scheme for both distortion effects.

The FDLP processing shows the tendency to degrade the recognition of clean data. We intend to estimate the presence of reverberation in future work so that we could disable the FDLP processing or apply FDLP with a higher filter order in case of clean data to reduce the loss in performance.

5. ACKNOWLEDGEMENTS

The author was able to 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.

6. REFERENCES

[1] R.C. Hendriks, T. Gerkmann, J. Jensen, “DFT Domain based single-microphone noise reduction for speech enhancement”, Morgan & Claypool Publishers, 2013

[2] 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.

[3] C. Breithaupt, T. Gerkmann, R. Martin, “Cepstral smoothing of spectral filter gains for speech enhancement without musical noise”, IEEE Signal processing letters, 2007.

[4] H.G. Hirsch, A. Kitzig, “Robust speech recognition by combining a robust feature extraction with an adaptation of HMMs”, 9. ITG symposium on speech communication, Bochum, Germany, 2010.

[5] A. Sehr, R. Maas, W. Kellermann, “Model-based dereverberation in the logmelspec domain for robust distant-talking speech recognition”, ICASSP conference, 2010.

[6] A. Krüger, R. Haeb-Umbach, “Model based feature enhancement for automatic speech recognition in reverberant environments“, InterSpeech conference, 2009.

[7] M. Unoki, S. Morita, M. Akagi, “A Study on the IMTF-Based Filtering on the Modulation Spectrum of Reverberant Signal”, Signal Processing, Vol. 14, 2010.

[8] M. Athineos, D.P. Ellis, “Frequency domain linear prediction for temporal features”, ASRU workshop, 2003.

[9] M. Athineos, H. Hermansky, D.P. Ellis, “LP-TRAPS: Linear predictive temporal patterns”, InterSpeech conference, 2004.

[10] H.G. Hirsch, H. Finster, “A new approach for the adaptation of HMMs to reverberation and background noise”, Speech Communication, Vol.50, pp. 244-263, 2008.

[11] Kinoshita, K.; Delcroix, M.; Yoshioka, T.; Nakatani, T.; Habets, E.; Haeb-Umbach, R.; Leutnant, V.; Sehr, A.; Kellermann, W.; Maas, R.; Gannot, S.; Raj, B., "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13), 2013.

[12] H.G. Hirsch, K. Hellwig, S. Dobler, “Speech Recognition at multiple sampling rates”, Eurospeech conference, 2001.

[13] Hirsch, H.G., Ehrlicher, C., “Noise estimation techniques for robust speech recognition”, ICASSP, Detroit, pp. 153-156, 1995.

[14] S. Thomas, S. Ganapathy, H. Hermansky, "Recognition of Reverberant Speech Using Frequency Domain Linear Prediction", IEEE Signal Processing Letters, 2008.

[15] T. Houtgast, H. J. M. Steeneken, “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria”, JASA, 1985.

[16] Leonard, R.G. (1984), A Database for Speaker-Independent Digit Recognition, ICASSP, San Diego, Vol. 3, p. 42.11, 1984.

[17] Robinson, T.; Fransen, J.; Pye, D.; Foote, J.; Renals, S., "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95), vol.1, pp. 81-84, 1995.

[18] http://dnt.kr.hsnr.de/ (➔ download section)

[19] K. Linhard. Measurement of room acoustics and noise characteristics, SpeeCon project report, available at http://www.speechdat.org, 2002.

[20] Lincoln, M.; McCowan, I.; Vepa, J.; Maganti, H.K., "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," Proceedings of the 2005 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-05), pp. 357-362, 2005.

[21] Young, S. J.; Evermann, G.; Gales, M. J. F.; Hain, T.; Kershaw, D.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; Valtchev, V.; Woodland, P. C., "The HTK Book, version 3.4," Cambridge University Engineering Department, 2006.