Comparison of Different Approaches for Speech Recognition in Hands-free Mode

Hans-Günter Hirsch [*1], *Sriram Ganapathy* [#2], *Hynek Hermansky* [#3]

[*] Institute for Pattern Recognition, Niederrhein University of Applied Sciences, Krefeld, Germany
[#] Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA
Email: [1] hans-guenter.hirsch@hs-niederrhein.de [2,3] {ganapathy, hynek}@jhu.edu
Web:: [1] dnt.kr.hs-niederrhein.de [2,3] www.clsp.jhu.edu

Abstract

To improve speech recognition in case of a hands-free speech input we apply the signal processing technique known as frequency domain linear prediction (FDLP). By analyzing the effects of reverberation we prove that this method is well suited to create robust acoustic features for this mode of speech input. Furthermore, we compare the efficiency of this robust feature extraction scheme with the alternative approach of adapting the Hidden Markov Models (HMMs). We especially investigate the influence of a varying distance between the speaker and the microphone in a reverberant room.

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 reverberation. Thus, a lot of approaches have been developed in the field of signal processing to compensate the influence of reverberation. Most of these approaches are based on the usage of two or more microphones. Some investigations have been carried out to compensate the effects of reverberation in case only one microphone is available, e.g. [1],[2],[3]. In most applications of speech recognition only one microphone is applied. Another approach is the adaptation of the HMMs that can be seen as alternative technique in comparison to extracting robust features.

In this paper, we present the signal processing approach known as FDLP [4], [5]. By analyzing the effects of reverberation we prove that this technique allows the extraction of robust features. Its efficiency to improve the recognition performance is compared with the alternative method of adapting the HMMs. We apply an approach where the cepstral coefficients and the energy coefficient as means of the Gaussian distributions in each HMM state are adapted based on an estimation of the reverberation time [6].

In this work, we put a special focus on the distance between the speaker and the microphone. Most investigations so far assume that the speaker is fairly far away from the microphone. But, this is not true in a lot of applications, e.g., thinking at the control of a telephone or a PC on an office desk when sitting at the desk. Therefore, we measured a set of impulse responses at different distances in different rooms so that we can investigate the influence of a varying distance [7].

The comparison against other approaches like the adaptation of HMMs and the investigation of the recognition performance dependent on the distance between speaker and microphone make the difference of this work and the application of FDLP on reverberant data in [8].

2 Frequency Domain Linear Prediction (FDLP)

FDLP is based on the idea of treating the short-term energy contour in a subband as the spectrum 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. This analysis of short segments and the transmission of the filter coefficients is known as linear predictive coding (LPC).

The spectral characteristics of a DPCM filter can be seen as a kind of spectral envelope of the corresponding DFT spectrum. It describes and contains the frequency characteristics of the vocal tract. This 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. 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 IDFT on the contour. This signal is taken as input to estimate the parameters of the DPCM filter. The order of the filter is chosen dependent on the length of the energy contour reflecting the length of the whole speech signal. We define the filter complexity with a value describing the filter order per second.

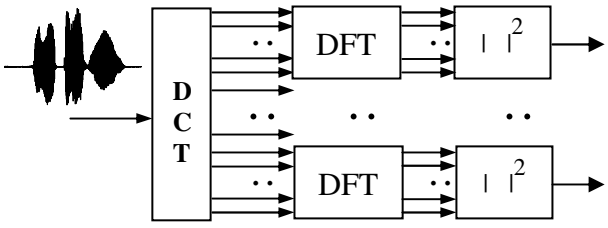

Figure 1: Processing scheme to estimate a set of Hilbert envelopes.

The implementation of FDLP is based on the analysis of the Hilbert envelopes in a number of subbands [8]. As shown in Figure 1 all samples of an utterance are transformed by means of a DCT. The DCT spectrum contains the same number of components as the number of samples taken as input to the DCT. Several subsets of the

DCT coefficients are extracted according to the desired bandwidths and characteristics of the subband filters. Each subset of extracted DCT components is transformed by applying a DFT. The result is an estimate of the Hilbert envelope in this subband. A setup with 96 filters of equal bandwidth gave best recognition results for most conditions.

The Hilbert envelope is transformed by means of an IDFT. Thus, a kind of corresponding "time" signal is determined. The well known LPC analysis is applied on this signal where the order of the filter is defined dependent on the length of the envelope. 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 Hilbert envelope as the spectral characteristics of an all-pole filter:

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

It turned out in earlier investigations that it can be of advantage to neglect the gain factor g [8]. This has the effect of gain normalization when comparing the envelopes in the different subbands. The effect will be similar to applying the well known cepstral mean normalization. In case of a spectral weighting due to e.g. the microphone, the degradation of the error rates should be reduced by such a type of frequency dependent gain normalization.

To apply FDLP for speech recognition the set of smoothed Hilbert envelopes are taken to calculate the Mel cepstral coefficients for short speech segments of about 25 ms duration. The envelopes are still sampled at the high sampling rate of the speech signal. A spectrum is estimated every 10 ms by weighting the values of each Hilbert envelope with a window function of 25 ms duration. The energy is calculated in each subband for this segment from the weighted values of the envelope. Thus, a sequence of spectra is created at a rate of 100 Hz. The Mel cepstral coefficients can be calculated for each spectrum by summing up the spectral subband energies according to a Mel filterbank and transforming the logarithmic Mel spectrum to the cepstral domain.

3 Applying FDLP to Reverberant Signals

Looking at the condition of a hands-free speech input to a speech recognition system 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. The transmission in a room can be described as a convolution of the speech signal and the room impulse response (RIR) in the field of signal processing.

The influence of a hands-free speech 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 modelled as the multiplication of the speech spectrum and the corresponding transfer function. But, looking at the analysis of short segments as it is done in the field of speech recognition this consideration 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. Considering the influence of a hands-free speech input as a "spectral coloration" might only be approximately correct when either the RIR is fairly short or the speaker is close to the microphone so that the energy of the late reflections is relatively low in comparison to the energy of the direct sound. Thus, all approaches that try to compensate the influence of a nearly stationary transfer function, like e.g. cepstral mean subtraction, will not be able to handle the effects of reverberation in general.

The second aspect is the examination of the contours of the short-term energy in small subbands to cover the effects over the fairly long time period of the RIR. The transmission in a reverberant environment can be approximately described as a low-pass filtering of these energy contours [9]. In Figure 2 two versions of an energy contour are shown for a speech signal that contains the recording of three spoken digits.

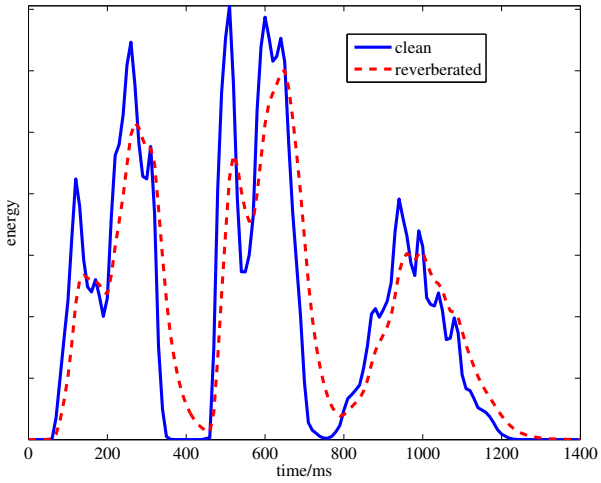

Figure 2: Energy contours of a clean signal and after recording in hands-free mode.

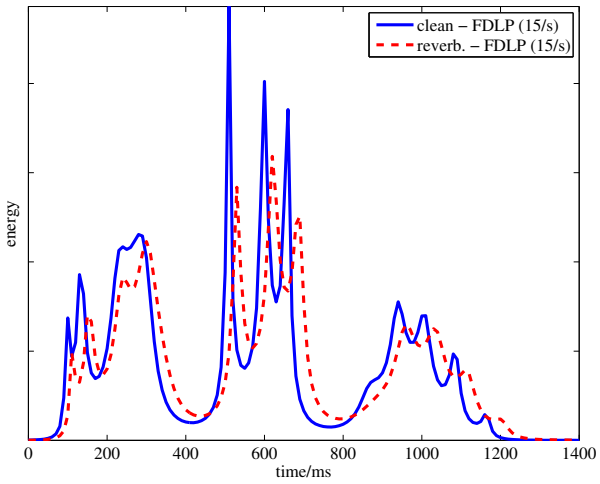

Figure 3: Energy contours after processing the clean signal and the reverberant signal by FDLP with a filter order of 15 per second.

The envelope of the clean signal as well as the contour after the transmission in a room are shown assuming an ideally 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 e.g. the pauses between the words are no longer characterized by energy close to zero.

To analyze the influence of FDLP processing two further contours are visualized in Figure 3. Both contours shown in Figure 2 have been processed by FDLP with a filter order of 15 per second. Besides a small shift in time due to the reverberation the contours 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. This might allow a good usability for achieving a high recognition performance of reverberant signals recorded in hands-free mode. The creation of similar features is exactly the goal of a "robust" feature extraction scheme even though the former characteristics of clean signals with low energy in short speech pauses is not present anymore.

4 Recognition Experiments

We ran a first set of recognition experiments to investigate the usability and efficiency of FDLP for the recognition of speech signals that have been recorded in hands-free mode. The training set of the clean TIDigits speech data has been used to create two gender dependent whole-word HMMs for each English digit (zero – nine, oh) applying HTK. Each HMM consists of 16 states with a mixture of 2 Gaussians to model the occurrence of each acoustic parameter in each state.

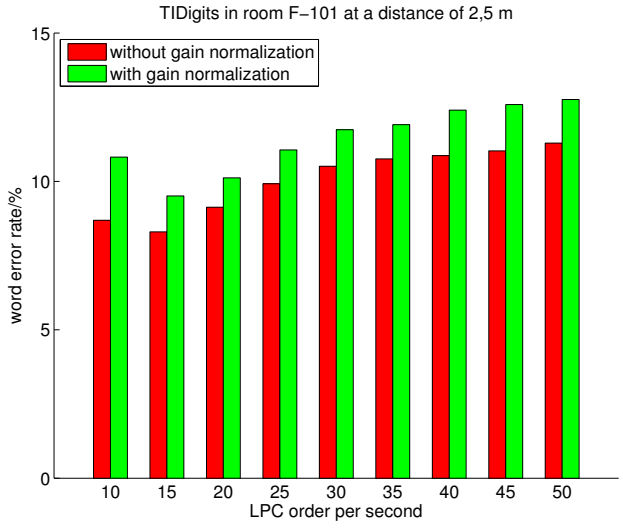

Figure 4: Word error rates for the recognition in hands-free mode.

The word error rates are presented in Figure 4 for the recognition of the speech signals inside the test set of the TIDigits. All speech signals have been convolved with the RIR of a meeting room (F-101) at a distance of 250 cm between speaker and microphone. With these experiments, we wanted to investigate the dependency on the order of the DPCM filter. Training data have been processed in the same way so that a separate set of HMMs has been determined for each filter order. It turns out that the lowest error rate is achieved for the low filter order of 15 per second. In Figure 3, it was visualized that FDLP processing with a low filter order causes the creation of energy contours that include similar smearing or low-pass effects as the corresponding contours of the reverberant signal. Thus, the reduction of the error rate for a decreasing filter order could be expected to some extent. But with this first experiment, we looked at reverberation as the only effect for the modification of the speech signal. In practice, there will occur also additive noise from the recording equipment and from the acoustic environment. We could see in our further experiments that the fairly low filter order of 15 per second is not the optimal choice for other acoustic conditions where also additive noise is included. Comparing the performance with and without gain normalization, it seems to be of disadvantage to apply the gain normalization for this condition of a high distance between speaker and microphone.

To take into account also other distorting effects, we looked at a subset of the so called "meeting recorder digits" [10]. This subset consists of 2715 utterances with a total of about 9000 digits that have been recorded with a setup of 4 microphones in a conference room and that were spoken by several participants during different meetings. As HMMs still the ones trained on clean TI-Digits are used. Word error rates are shown in Table 1 where the recordings of just one microphone (mic. E) have been used as input for the recognition. The results for the other microphones look similar. It turns out that the lowest error rates are achieved for a filter order of about 20 to 30.

filter order per second	Word error rate (%)	
	WITHOUT gain norm.	WITH gain norm.
10	19,03	15,34
15	15,34	11,51
20	14,38	10,38
25	14,43	10,12
30	14,83	10,22
35	15,42	10,48
40	15,48	10,55
45	15,56	10,67
50	15,78	10,63

Table 1: Word error rates for meeting recorder digits.

Furthermore, we looked at the performance for recognizing the clean TIDigits data without any effects of a hands-free speech input. Word error rates are presented in table 2. The lowest error rates are achieved for a filter order of about 30 to 40. We achieve an error rate of 0,55 % when applying an usual Mel cepstral analysis. The modification of the energy contours due to the FDLP processing causes a deterioration of the recognition performance on clean data.

filter order per second	Word error rate (%)	
	WITHOUT gain norm.	WITH gain norm.
10	2,23	2,11
15	1,35	1,22
20	1,11	0,98
25	1,06	0,88
30	0,98	0,85
35	1,02	0,84
40	1,02	0,78
45	1,03	0,81
50	1,00	0,83

Table 2: Word error rates for the clean TIDigits.

We choose a filter order of 30 per second for all further experiments. We could verify this choice as being almost optimal by running some of the later experiments also with a different filter order without presenting the results in this paper. Analyzing the effect of gain normalization, we observed a reduction of the error rates besides in the experiment on the artificially created, reverberant TIDigits. Especially, we find a gain for the real recordings in the meeting room where the sound level of most utterances is fairly low and different from the average level of the TIDigits.

We put our further focus on looking at the conditions of a varying distance between speaker and microphone. We created several sets of the TIDigits by convolving the clean signals of the TIDigits test set with a set of RIRs measured in the meeting room (F-101) at different distances between speaker and microphone [7]. The word error rates are shown in Figure 5 for 7 different positions of the microphone in the meeting room.

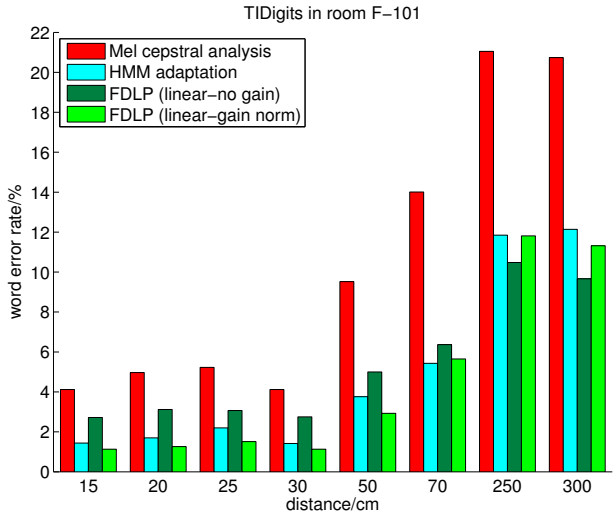

Figure 5: Word error rates for the recognition in hands-free mode at different distances.

We compared the results applying 4 different analysis and recognition schemes. The highest error rates are achieved for a Mel cepstral analysis without any additional processing or adaptation to compensate the effects of reverberation. 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. Based on the Mel cepstral analysis we applied an adaptation of the HMMs. The adaptation scheme was developed to compensate the influence of background noise, unknown frequency characteristics and reverberation [6].

The adaptation as well as the FDLP processing lead to a considerable improvement of the recognition performance. The application of gain normalization is of advantage for small distances up to 70 cm. This can be explained by the influence of spectral coloration that will be more involved at low distances. This spectral weighting will be compensated by normalizing the gain in the different subbands. At higher distances the dominant effect is the smearing of the energy contours due to the reverberation. The gain normalization causes a small deterioration for this condition.

5 Conclusions

FDLP is a processing technique that can be applied in speech recognition to create robust acoustic features for the condition of a hands-free speech input in reverberant environments, especially in case the speech is recorded with only one microphone. The technique as it was realized within this work comes along with the small disadvantage that it causes a certain processing delay because the energy contours can not be processed until the whole utterance is available.

References

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

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

[3] 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.

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

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

[6] H.G. Hirsch, H. Finster, “A New Approach for the Adaptation of HMMs to Reverberation and Background Noise”, *Speech Communication*, Vol.50, 2008.

[7] H.G. Hirsch, A. Kitzig, K. Linhard, “Simulation of the Hands-free Speech Input to Speech Recognition Systems by Measuring Room Impulse Responses”, *9. ITG-Fachtagung Sprachkommunikation,* Bochum, 2010.

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

[9] 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.

[10] A. Janin et al., “The ICSI meeting corpus”, *ICASSP*, 2003.