IDIAP RESEARCH REPORT

Front-end for Far-field Speech Recognition based on Frequency Domain Linear Prediction

Sriram Ganapathy [a b] Samuel Thomas [a b]
Hynek Hermansky [a b]

IDIAP–RR 08-17

July 2008

[a] IDIAP Research Institute, Martigny, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-17

Front-end for Far-field Speech Recognition based on Frequency Domain Linear Prediction

Sriram Ganapathy Samuel Thomas Hynek Hermansky

July 2008

Abstract. Automatic Speech Recognition (ASR) systems usually fail when they encounter speech from far-field microphone in reverberant environments. This is due to the application of short-term feature extraction techniques which do not compensate for the artifacts introduced by long room impulse responses. In this paper, we propose a front-end, based on Frequency Domain Linear Prediction (FDLP), that tries to remove reverberation artifacts present in far-field speech. Long temporal segments of far-field speech are analyzed in narrow frequency sub-bands to extract FDLP envelopes and residual signals. Filtering the residual signals with gain normalized inverse FDLP filters result in a set of sub-band signals which are synthesized to reconstruct the signal back. ASR experiments on far-field speech data processed by the proposed front-end show significant improvements (relative reduction of 30% in word error rate) compared to other robust feature extraction techniques.

1 Introduction

Even a small amount of reverberation causes significant degradation in ASR performance. This is primarily due to the temporal smearing of the short-term spectra (which are used for deriving conventional features for speech recognition). Since reverberation is a long term phenomenon, techniques based on short term spectra generally result in increased word error rates for far-field speech as the models trained in clean environments fail to match the test conditions. Although several approaches have been proposed for recognition of multi-channel reverberant speech (for example [1, 2]), single channel far-field speech recognition continues to be a challenging task.

In reverberant environments, the speech signal that reaches the far-field microphone is superimposed with multiple reflected versions of the original speech signal. These superpositions can be modeled by the convolution of the room impulse response with the original speech signal. This can be written as

$$r(t) = s(t) * h(t), \tag{1}$$

where $s(t)$, $h(t)$ and $r(t)$ denote the original speech signal, the room impulse response and the far-field reverberant speech respectively.

The amount of reverberation in speech is generally characterized by the reverberation time ($RT60$) and the magnitude distortion in the frequency domain. $RT60$ for a room is defined as the time required, in seconds, for the sound in a room to decrease by 60 decibels after the sound source is removed. The magnitude distortion is represented by the spectral coloration in the room impulse response (defined as the ratio of the geometric mean to the arithmetic mean of the spectral magnitudes).

For analysis windows which are longer than $RT60$, the effect of reverberation can be approximated as multiplicative in the frequency domain [3]. This fact has been exploited in [4, 5], where the effect of reverberation is compensated by mean subtraction in long-term log-spectral domain.

In this paper, we propose a technique that uses gain normalized temporal trajectories of sub-band energies to compensate for the room reverberation artifacts. For reverberant speech, the sub-band Hilbert envelopes can be assumed to be a convolution of the sub-band Hilbert envelope of the clean speech with the sub-band Hilbert envelope of the room impulse response [6]. This Hilbert envelope convolution model is valid for long temporal analysis (relative to $RT60$) in narrow frequency sub-bands.

For the proposed front-end, Hilbert envelopes of sub-band signals are estimated by FDLP [7]. This results in an auto-regressive model of the Hilbert envelope (inverse FDLP filter) and a prediction residual (FDLP residual) [8]. The sub-band signals in the frequency domain can be reconstructed by filtering the FDLP residual using the inverse FDLP filter.

When linear prediction is applied in the frequency domain, the Hilbert envelope convolution model [6] suggests that the artifacts present in reverberant speech alter the gain of the sub-band temporal envelopes. In order to avoid the mis-match in gain of sub-band Hilbert envelopes for the clean and reverberant speech, the FDLP residual signals are filtered by the gain normalized inverse FDLP filter. This is followed by sub-band synthesis to yield a signal which is input to the conventional ASR.

The rest of the paper is organized as follows. Sec. 2 describes the FDLP technique for extracting the sub-band Hilbert envelopes. Sec. 3 details the underlying mathematical details for the Hilbert envelope convolution model. The proposed front-end for far-field reverberant speech is explained in Sec. 4. The ASR experiments with the proposed front-end are reported in Sec. 5 followed by the conclusions in Sec. 6.

2 Frequency Domain Linear Prediction

Linear prediction is a mathematical operation where present value of a discrete-time signal is estimated as a linear function of previous samples. When linear prediction is done in the time domain,

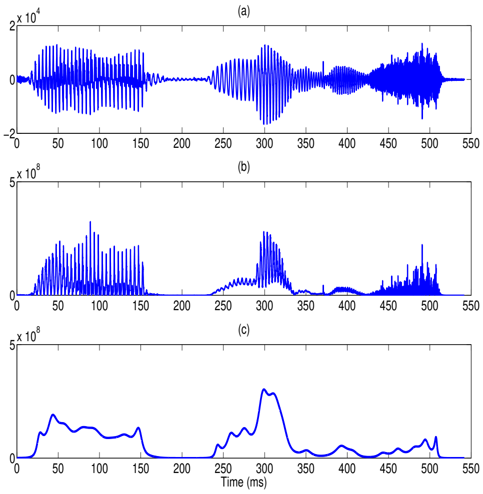

Figure 1: *Illustration of the all-pole modelling with FDLP. (a) a portion of the speech signal, (b) its Hilbert envelope (c) all pole model obtained using FDLP*

Auto-Regressive (AR) models are obtained that represent the envelope of the power spectrum of the signal [9]). The duality between the time and frequency domains means that AR modeling can be applied equally well to spectral representations of the signal instead of time-domain signal samples. This paper utilizes linear prediction in the frequency domain for obtaining smoothed, minimum phase, parametric models for temporal rather than spectral envelopes.

For the FDLP technique, the squared magnitude response of the inverse filter approximates the Hilbert envelope of the signal (in a manner similar to the approximation of the power spectrum of the signal using Time Domain Linear Prediction (TDLP) [9]). For signals that are expected to consist of a number of distinct transients, fitting an AR model can constrain the modeled envelope to be a sequence of maxima, and the AR fitting procedure can remove finer-scale detail. This suppression of detail is particularly useful in classification applications, where the goal is to extract the general form of the signal regardless of minor variations.

In our case, the FDLP technique is implemented in two parts - first, the discrete cosine transform (DCT) is applied on long segments of speech to obtain a real valued spectral representation of the signal. Then, linear prediction is performed on the DCT coefficients to obtain a parametric model of the temporal envelope. Fig. 1 shows an illustration of the AR modelling property of FDLP It shows (a) a portion of speech signal of 500 ms duration, (b) its Hilbert envelope computed using the Fourier transform technique [10] and (c) an all pole approximation (of order 50) for the Hilbert Envelope using FDLP.

3 Hilbert Envelope Convolution Model

Let $s(t)$ denote a long term speech signal, which is decomposed into contiguous frequency bands denoted as band limited signals $s_n(t)$. Each of these sub-band signals can be modeled in terms of product of a slowly varying, positive, envelope function $A_{sn}(t)$ and an instantaneous phase function $p_{sn}(t)$ [6] such that

$$s(t) = \sum_{n=1}^{N} s_n(t) = \sum_{n=1}^{N} A_{sn}(t) \cos\left(p_{sn}(t)\right). \tag{2}$$

Reverberant speech $r(t)$, can similarly be expressed as sum of band limited signals $r_n(t)$ in sub-bands as

$$\begin{aligned} r(t) &= \sum_{n=1}^{N} r_n(t) = \sum_{n=1}^{N} A_{rn}(t) \cos\big(p_{rn}(t)\big) \\ &\simeq \sum_{n=1}^{N} h_n(t) * s_n(t) \\ &= \sum_{n=1}^{N} A_{hn}(t) \cos\big(p_{hn}(t)\big) * A_{sn}(t) \cos\big(p_{sn}(t)\big), \end{aligned}$$

where A_{rn}, A_{sn} and A_{hn} represent the envelope functions of the bandpassed reverberant speech, the original speech and the room impulse response; their corresponding phase functions are given by $p_{rn}(t)$, $p_{sn}(t)$ and $p_{hn}(t)$. For room impulse responses, it has been shown in [6] that the envelope of $r_n(t)$ can be represented as

$$A_{rn}(t)e^{jp_{rn}(t)} \simeq \frac{e^{j(\omega_n t+\phi_n(t))}}{2} \int_{-\infty}^{t} A_{hn}(t-t_1)A_{sn}(t_1)dt_1,$$

where ω_n is the center frequency of each band and $\phi_n(t)$ is the phase difference between the original speech and the room response. In narrow sub-bands, the envelope functions are related by

$$A_{rn} \simeq \frac{1}{2} A_{hn} * A_{sn}. \tag{3}$$

If A_{rn} represents the Hilbert envelope of the n^{th} sub-band, Eq. (3) shows that the Hilbert envelope of the sub-band signal for the reverberant speech can be approximated as the convolution of the Hilbert envelope of the clean speech signal in that sub-band with that of the room impulse response. All these results assume analysis windows longer than the duration of the room impulse response. Since FDLP is performed on long temporal segments, the Hilbert envelope convolution model can be exploited for compensating reverberation artifacts in an FDLP based front-end for far-field speech.

The Hilbert envelope and the spectral autocorrelation function form Fourier transform pairs [8]. The Hilbert envelope convolution model in Eq. (3) shows that the spectral autocorrelation function of the reverberant speech is the multiplication of spectral autocorrelation function of the clean speech with that of the room impulse response. For the room impulse response, the spectral autocorrelation function in narrow frequency sub-bands can be assumed to be slowly varying compared to that of the speech signal. Thus, normalizing the gain of the sub-band Hilbert envelopes suppresses the multiplicative effect present in the spectral autocorrelation function of the reverberant speech. This forms the basis for the proposed far-field speech recognition front-end.

4 Front-end for Far-field Speech Recognition

The input speech signal is analyzed into narrow frequency sub-bands by windowing the DCT of the signal using rectangular windows. Linear Prediction is applied on the windowed DCT components to obtain the FDLP envelope $A_{rn}(z)$ which is a parametric model for the sub-band Hilbert envelope and a prediction residual. Gain normalization of the Hilbert envelopes is achieved by setting the gain of the inverse FDLP filter to unity. The FDLP prediction residual is filtered with the gain normalized sub-band Hilbert envelope $\hat{A}_{rn}(z)$ to obtain the de-reverberated sub-band signal. This is followed by a sub-band synthesis to reconstruct the signal to be used by the ASR. The block schematic of the proposed front-end for far-field speech recognition is shown in Fig. 2. An illustration of the effect of gain normalization on the sub-band FDLP envelopes for clean and reverberant speech is provided in Fig. 3.

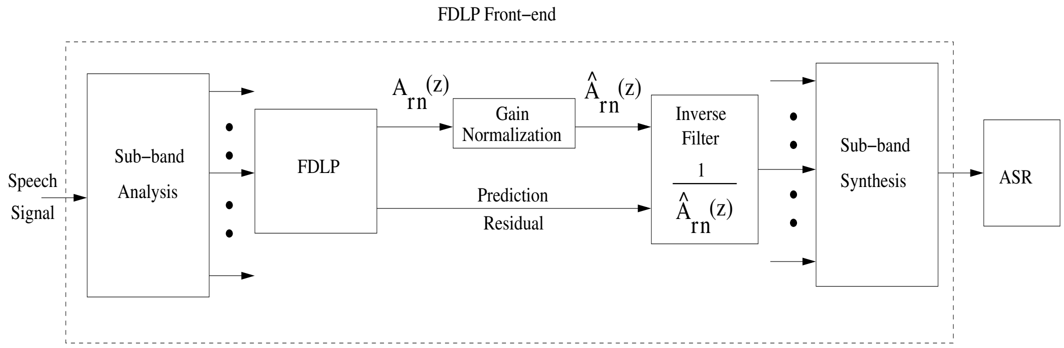

Figure 2: *Front-end for far-field speech based on FDLP*

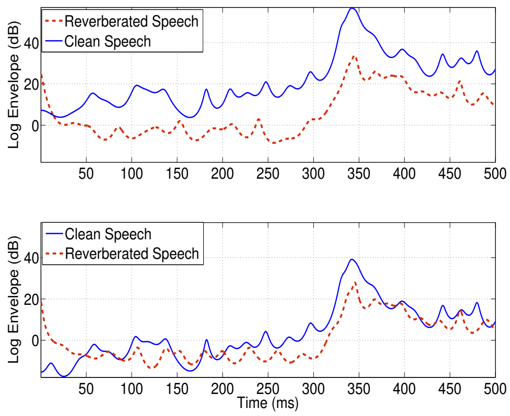

Figure 3: *FDLP envelopes for clean and reverberant speech for second sub-band (a) without gain normalization (b) with gain normalization.*

5 Experiments and Results

We apply the proposed features and techniques in a connected word recognition task with a modified version of the Aurora speech database using the Aurora evaluation system [11]. We use the "complex" version of the back end proposed in [12]. The training dataset contains 8400 clean speech utterances, consisting of 4200 male and 4200 female utterances downsampled to 8 kHz and the test set consist of 3003 utterances [5].

For reverberant speech recognition experiments, the input speech signal is processed by the proposed front-end and then used to extract 13 dimensional PLP features [13]. Delta and double delta features are appended to obtain 39 dimensional features which are input to an ASR.

The first set of experiments are conducted on artificially reverberated data obtained by convolving the clean test data with a room impulse response having a spectral coloration of -1.92 dB and a RT60 of 0.5 seconds [14]. This test data is used to determine the optimum FDLP model order. Table 1 shows the word accuracies as a function of the number of poles used in the FDLP model.

For all further experiments, the number of poles for the FDLP is fixed at 30. Table 2 compares the word accuracies for clean and artificially reverberated data using the proposed front-end as well as other robust feature extraction techniques proposed for reverberant speech namely Cepstral Mean Subtraction (CMS) [15], Long Term Log Spectral Subtraction (LTLSS) [5] and Log-DFT Mean Nor-

Number of poles	Word Acc.
16	93.10
24	93.21
30	93.49
40	92.98
50	92.82

Table 1: *Word Accuracies (%) for artificially reverberated speech as a function of the number of FDLP poles.*

Features	Clean	Revb.
PLP	99.68	72.80
CMS	99.66	86.64
LDMN	99.64	91.75
LTLSS	99.62	92.58
FDLP Front-end	99.60	93.49

Table 2: *Word accuracies (%) for clean and artificially reverberated speech with the proposed front-end as well as various other feature extraction techniques.*

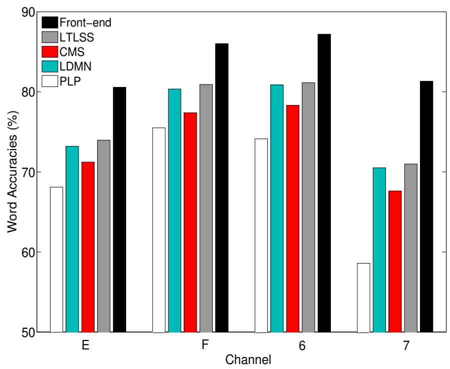

Figure 4: *Comparison of Word Accuracies (%) using different techniques for ASR for far-field reverberant*

malization (LDMN) [4]. In our LTLSS experiments, the long-term log-spectral means are computed independently for each individual utterance (which differs from the approach of grouping multiple utterances for the same speaker described in [5]) using a shorter analysis window of 32 ms, with a shift of 8 ms. Similar to the proposed front-end, PLP features are used along with all these robust feature extraction techniques to obtain 39 dimensional input vectors for the ASR. This table shows that the application of the proposed front-end significantly improves the ASR performance for the PLP features (about 76% reduction in the word error rate). These results are achieved without any noticeable degradation in performance for clean speech. The improvement provided over the other feature extraction techniques is about 12%.

The next set of experiments are performed on the digits corpus recorded using far-field microphones for the ICSI Meeting task [14]. This corpus also forms part of the Aurora-5 speech database [16]. The far-field data set consists of four sets of 2790 utterances each. These sets correspond to speech recorded simultaneously using four different far-field microphones (channels E, F, 6 and 7 in [14]). Each channel contains 9169 digits similar to those found in TIDIGITS corpus. As before, we use the HMM models trained with the clean speech in the training set of modified Aurora task. The results for the proposed front-end as well as other feature extraction techniques are shown in Fig. 4. For the different far-field microphone channels, the proposed FDLP based front-end, on the average, provide a relative error improvement of 30% over the other feature extraction techniques considered.

6 Conclusions

Unlike many single microphone based reverberant speech recognition approaches, the proposed front-end does not normalize speech signals using long term mean subtraction in spectral domain. We show that the effect of reverberation is reduced when the speech signals are reconstructed using gain normalized temporal envelopes of long duration in narrow sub-bands. FDLP provides an efficient way to suppress the reverberation artifacts and hence, PLP features extracted from far-field speech signals processed using the proposed front-end provide significant improvements over other robust feature extraction techniques. The application of the proposed techniques for larger vocabulary tasks and for signals distorted by additive and convolutive noise are currently pursued.

7 Acknowledgements

The authors would like to thank David Gelbart for helpful discussions, room impulse responses, speech databases and code fragments for LTLSS experiments. Furthermore, we would also like to thank Marios Athineos and Dan Ellis for PLP and FDLP feature extraction codes.

References

[1] J.L. Flanagan, J.D. Johnston, R. Zahn and G.W. Elko, "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms", *J. Acoust. Soc. Am.*, vol. 78, no. 11, pp. 1508-1518, Nov. 1985.

[2] H. Wang and F. Itakura, "An Approach to Dereverberation using Multi-Microphone Sub-band Envelope Estimation", in *Proc. ICA*, Toronto, Canada, 1991, pp. 953-956.

[3] C. Avendano, *Temporal Processing of Speech in a Time-Feature Space*, Ph.D. thesis, Oregon Graduate Institute, 1997.

[4] C. Avendano and H. Hermansky, "On the Effects of Short-Term Spectrum Smoothing in Channel Normalization", IEEE Trans. Speech and Audio Proc., vol. 5, issue. 4, pp. 372-374, Jul 1997.

[5] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition," in *Proc. ICSLP*, Colorado, USA, 2002, pp. 2185-2188.

[6] J. Mourjopoulos and J.K. Hammond, "Modelling and Enhancement of Reverberant Speech using an Envelope Convolution Method", in *Proc. ICA*, Boston, USA, 1983, pp. 1144-1147.

[7] M. Athineos, H. Hermansky and D.P.W Ellis, "LP-TRAPS: Linear Predictive Temporal Patterns," in *Proc. of Interspeech*, Jeju Island, Korea, pp. 1154-1157, 2004.

[8] J. Herre and J.D Johnston, "Enhancing the Performance of Perceptual Audio Coders by using Temporal Noise Shaping (TNS)", in *Proc. 101st AES Conv.*, Los. Angeles, USA, 1996, pp. 1-24.

[9] J. Makhoul, "Linear Prediction: A Tutorial Review", in *Proc. of the IEEE*, Vol 63(4), pp. 561-580, 1975.

[10] L.S. Marple, "Computing the Discrete-Time Analytic Signal via FFT", in *IEEE Trans. on Acoustics, Speech and Signal Proc.*, Vol. 47, pp. 2600-2603, 1999.

[11] H.G. Hirsch and D. Pearce,"The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", in *Proc. ISCA ITRW ASR 2000*, Paris, France, 2000, pp. 18-20.

[12] D. Pierce and A. Gunawardana, "Aurora 2.0 speech recognition in noise: Update 2", in *Proc. ICSLP Session on Noise Robust Rec.*, Colorado, USA, 2002.

[13] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, Apr. 1990.

[14] "The ICSI Meeting Recorder Project", http://www.icsi.berkeley.edu/Speech/mr.

[15] A.E. Rosenberg, C. Lee and F.K. Soong,"Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification", in *Proc. ICSLP*, Yokohama, Japan, 1994, pp. 1835-1838.

[16] "Aurora-5", http://aurora.hsnr.de/aurora-5.html.