Recognition Of Reverberant Speech Using Frequency Domain Linear Prediction

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

IDIAP–RR 08-41

June 2008

to appear in
IEEE Signal Processing Letters 2008

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

IDIAP Research Report 08-41

Recognition Of Reverberant Speech Using Frequency Domain Linear Prediction

Samuel Thomas Sriram Ganapathy Hynek Hermansky

June 2008

to appear in
IEEE Signal Processing Letters 2008

Abstract. Performance of a typical automatic speech recognition (ASR) system severely degrades when it encounters speech from reverberant environments. Part of the reason for this degradation is the feature extraction techniques that use analysis windows which are much shorter than typical room impulse responses. We present a feature extraction technique based on modeling temporal envelopes of the speech signal in narrow sub-bands using Frequency Domain Linear Prediction (FDLP). FDLP provides an all-pole approximation of the Hilbert envelope of the signal obtained by linear prediction on cosine transform of the signal. ASR experiments on speech data degraded with a number of room impulse responses (with varying degrees of distortion) show significant performance improvements for the proposed FDLP features when compared to other robust feature extraction techniques (average relative reduction of 24% in word error rate). Similar improvements are also obtained for far-field data which contain natural reverberation in background noise. These results are achieved without any noticeable degradation in performance for clean speech.

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 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 reverberant speech recognition continues to be a challenging task.

In reverberant environments, the speech signal that reaches the 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, that accounts for individual reflection delays, with the original speech signal, i.e.,

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

where $s(t)$, $h(t)$ and $r(t)$ denote the original speech signal, the room impulse response and the reverberant speech respectively. The effect of reverberation on the short-time Fourier transform (STFT) of the speech signal $s(t)$ can be represented as

$$R(t, \omega_k) = S(t, \omega_k)H(t, \omega_k), \tag{2}$$

where $S(t, \omega_k)$ and $R(t, \omega_k)$ are the STFT's of the clean speech signal $s(t)$ and reverberant speech $r(t)$ respectively and $H(t, \omega_k)$ denotes the STFT of the room impulse response $h(t)$. The amount of reverberation in speech is generally characterized by reverberation time (T_{60}) and the magnitude distortion in the frequency domain. For analysis windows which are longer than T_{60}, the effect of reverberation can be approximated as multiplicative in the frequency domain [3], i.e., $H(t, \omega_k)$ is not a function of time and Eq. (2) becomes

$$R(t, \omega_k) \simeq S(t, \omega_k)H(\omega_k). \tag{3}$$

In the techniques reported in [4, 5], the effect of reverberation is compensated by subtracting from $\log\big(R(t, \omega_k)\big)$, its mean.

In this letter, we propose a technique that uses gain normalized temporal trajectories of sub-band energies to compensate for the room reverberation artifacts. Hilbert envelopes of sub-band signals are estimated by applying linear prediction in the frequency domain [6] (Sec. 2). Unlike conventional approaches that use mean compensation for reverberant speech recognition [4, 5], the proposed technique alleviates the reverberation artifacts present in long temporal envelopes of narrow frequency sub-bands. For the 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 [10] (Sec. 3). When linear prediction is applied in the frequency domain, the Hilbert envelope convolution model suggests that the artifacts present in reverberant speech affect the gain of the sub-band temporal envelopes. This causes the mis-match between the features trained from clean and reverberant environments. In order to reduce the mismatch, the proposed FDLP technique normalizes the gain of the auto-regressive models in narrow frequency bands (Sec. 3.2). The application of the proposed compensation technique to the FDLP features significantly improves the recognition accuracies for reverberant speech (Sec. 4).

2 Feature Extraction based on Frequency Domain Linear Prediction

Typically, Auto-Regressive (AR) models have been used in speech/audio applications for representing the envelope of the power spectrum of the signal (Time Domain Linear Prediction (TDLP) [9]). This

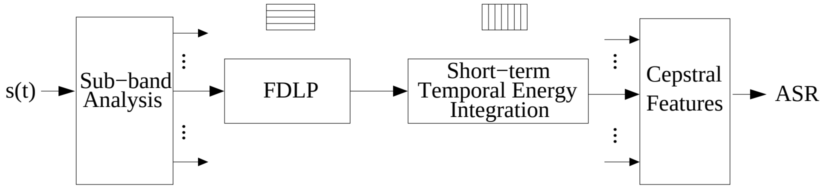

Figure 1: FDLP feature extraction for ASR

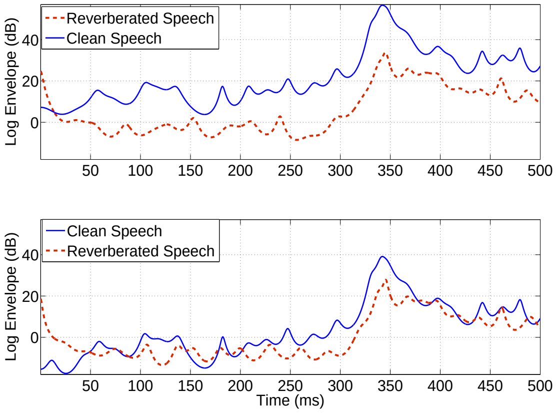

Figure 2: FDLP envelopes for clean and reverberant speech for second sub-band (a) without gain normalization (b) with gain normalization.

paper utilizes AR models for obtaining smoothed, minimum phase, parametric models for temporal rather than spectral envelopes. The duality between the time and frequency domains means that AR modeling can be applied equally well to discrete spectral representations of the signal instead of time-domain signal samples. Since we apply the LP technique to exploit the redundancies in the frequency domain, this approach is called Frequency Domain Linear Prediction (FDLP) [6, 8]. For the FDLP technique, the squared magnitude response of the all-pole filter approximates the Hilbert envelope of the signal (in a manner similar to the approximation of the power spectrum of the signal using TDLP [9]).

For the purpose of feature extraction, segments of the input speech signal (of the order of 1000 ms) are decomposed into sub-bands, where FDLP is applied to obtain a parametric model of the temporal envelope. For short utterances, the input signal is zero-padded to obtain sufficient number of samples prior to the sub-band decomposition. The whole set of sub-band temporal envelopes forms a two dimensional (time-frequency) representation of the input signal energy. This two-dimensional representation is convolved with a rectangular window of duration 25 ms and resampled at a rate of 100 Hz (10 ms intervals, similar to the estimation of short term power spectrum in conventional feature extraction techniques). These sub-sampled short-term spectral energies are converted to short-term cepstral features similar to the PLP feature extraction technique [7]. In our experiments, we use 39 dimensional cepstral features containing 13 cepstral coefficients along with the delta and double-delta features. The block schematic for the FDLP feature extraction technique is shown in Fig. 1.

3 Hilbert Envelope Convolution Model

3.1 Mathematical Framework

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 $p_{sn}(t)$ [10] such that

$$s(t) = \sum_{n=1}^{N} s_n(t) = \sum_{n=1}^{N} A_{sn}(t) \cos\big(p_{sn}(t)\big). \tag{4}$$

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

$$\begin{aligned} r(t) &= \sum_{n=1}^{N} r_n(t) = \sum_{n=1}^{N} A_{rn}(t) \cos\big(p_{rn}(t)\big) \\ &\simeq \sum_{n=1}^{N} h_n(t) * s_n(t) \\ &= \sum_{n=1}^{N} A_{hn}(t) \cos\big(p_{hn}(t)\big) * A_{sn}(t) \cos\big(p_{sn}(t)\big), \end{aligned} \tag{5}$$

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 [10] that the envelope of $r_n(t)$ can be represented as

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

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

$$A_{rn} \simeq \frac{1}{2} A_{hn} * A_{sn}. \tag{6}$$

If A_{rn} represents the Hilbert envelope of the n^{th} sub-band, Eq. (6) 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 FDLP features.

3.2 Gain Normalization for Reverberant Speech

The Hilbert envelope and the spectral autocorrelation function form Fourier transform pairs [6]. The Hilbert envelope convolution model in Eq. (6) 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 FDLP envelopes suppresses the multiplicative effect present in the spectral autocorrelation function of the reverberant speech. For example, Fig. 2 provides an illustration of the effect of gain normalization on the sub-band FDLP envelopes for clean and for reverberant speech.

Table 1: Word Accuracies (%) for FDLP features with and without the gain of the inverse filter and different window shapes

Features	Clean Speech	Revb. Speech
PLP	99.68	80.12
FDLP-MEL-G-WN	99.67	78.70
FDLP-MEL-G-GN	99.53	85.32
FDLP-MEL-R-GN	99.43	89.07
FDLP-UNF-R-GN	99.18	89.49

Table 2: Word Accuracies (%) for FDLP features for clean and reverberant speech using different number of sub-bands

Number of sub-bands	Clean Speech	Revb. Speech
24	99.18	89.49
33	99.13	91.86
67	99.09	92.93
76	99.16	93.60
96	99.07	94.79
108	99.03	94.63
120	98.91	94.55

4 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 test data was convolved with a set of 8 different room responses collected from various sources [14, 15, 16] with spectral coloration (defined as the ratio of the geometric mean to the arithmetic mean of the spectral magnitudes) ranging from -2.42 dB to -0.57 dB. The use of 8 different room responses results in 8 test sets consisting of 3003 utterances each. One of these test sets (obtained using the impulse response with a spectral coloration of -1.92 dB) is used to investigate the effect of varying the number of frequency sub-bands.

The first set of experiments compare the performance of FDLP based features with the conventional features for clean input conditions. Here, we also investigate the effect of gain normalization of the FDLP envelopes on the final recognition rate for clean and reverberant speech (for one of the impulse responses with a spectral coloration of -1.92 dB). The front-end sub-band decomposition is achieved by windowing the Discrete Cosine Transform (DCT) of relatively long segments of the input signal (1000 ms) [8].

Table 1 shows the word accuracies for PLP features (PLP) and FDLP features extracted using a Gaussian shaped mel-filter bank without and with the gain normalization on the temporal envelopes (FDLP-MEL-G-WN, FDLP-MEL-G-GN respectively) and using a rectangular shaped mel spaced filter bank with gain normalization (FDLP-MEL-R-GN). For simplicity, we also experiment with uniformly spaced DCT windows (FDLP-UNF-R-GN). Although the uniform windows cause a slight drop in performance for clean conditions, they provide an easy framework for increasing the spectral resolution of reverberant speech in further experiments.

These results show that FDLP-MEL-G-WN features perform similar to PLP features for clean speech and the gain normalized FDLP-MEL-G-GN features provide significant improvement for the reverberant speech without any noticeable degradation in performance for clean speech. Further, the improvement obtained for FDLP-MEL-R-GN over the FDLP-MEL-G-GN is due to the application of the rectangular windows, which cause the least amount of temporal smearing to the sub-band FDLP

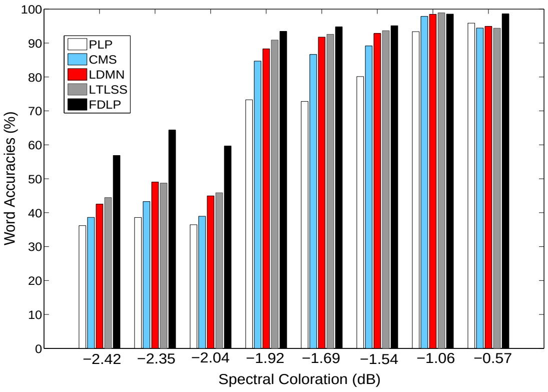

Figure 3: Comparison of Word Accuracies (%) using different techniques for ASR with reverberant speech

Table 3: Word accuracies (%) using a far-field mic

Features				
PLP	CMS	LDMN	LTLSS	FDLP
75.50	77.39	80.35	80.92	88.07

envelopes although, the Gaussian shape for DCT windows provides better spectral autocorrelation estimates for the clean signal [8]. In all further experiments, we employ the gain normalized temporal envelopes along with rectangular windows in the DCT domain (experiments using other window shapes were also performed and the rectangular shape was found to provide the best performance for reverberant speech).

In order to study the effect of finer spectral resolution for the proposed compensation technique, we increase the number of frequency sub-bands from 24 to 120. This is accomplished by increasing the duration of the temporal analysis (from 1000 ms to 2400 ms) for a constant width and overlap of the DCT windows. The test data consist of the reverberant speech using the same impulse response as before. Table 2 shows the recognition accuracies for the FDLP features when the number of sub-bands is varied.

Increasing the frequency resolution strengthens the validity of the assumptions made for the compensation technique (Eq. 6) and hence, significantly improves the recognition accuracies. In order to maintain a sufficient number of frequency samples per sub-band (which in turn decides the maximum FDLP model order), the width of the temporal analysis window needs to be increased along with the increase in number of sub-bands. However, the finite duration of the speech utterances in the database causes the need for zero padding the speech signals to obtain the desired sub-band decomposition with a fixed number of spectral components in each sub-band. For large number of sub-bands, the speech utterances get significantly zero padded resulting in a small drop in performance (for ex. decomposition into 120 sub-bands).

In Fig. 3, the results for the proposed FDLP technique are compared with those obtained for several other robust feature extraction techniques proposed for reverberant ASR namely Cepstral Mean Subtraction (CMS) [13], Long Term Log Spectral Subtraction (LTLSS) [5] and Log-DFT Mean Normalization (LDMN) [4]. This is done for the 8 different room impulse responses.

In our LTLSS experiments, we calculated the means 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. For the FDLP features, we fix the number of sub-bands to 96 and use a temporal analysis window of duration 2000 ms. For the different room responses, the proposed FDLP features, on the average, provide a relative error improvement of 24% over the other feature extraction techniques considered. The relative improvements are similar for most of the room responses, although the absolute improvements are higher for room impulse responses with higher spectral coloration (Fig. 3).

To investigate the performance of the proposed feature extraction for naturally reverberant speech in background noise, we also perform experiments on a set of connected digits recorded in an ICSI meeting room using a far-field mic (channel F in [14]). As before, we use the HMM models trained with the clean speech in the training set of modified Aurora task. The test data consist of 2790 utterances containing 9169 digits. Table 3 shows the word accuracies for the different feature extraction techniques using the far-field test data, where we obtain a relative error improvement of about 36%.

5 Conclusions

Unlike many single microphone based reverberant speech recognition approaches, the proposed technique does not normalize speech signals using long term mean subtraction in spectral domain. We show that the effect of reverberation is reduced when features are extracted from gain normalized temporal envelopes of long duration in narrow sub-bands. FDLP provides an efficient way to suppress the reverberation artifacts and hence, FDLP features extracted in reverberant environments 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.

6 Acknowledgements

In addition to the acknowledgements earlier, 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 Mathew Magimai Doss and Weifeng Li for the initial discussions in setting up the experiments and 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. 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.

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

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

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

[10] 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.

[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] 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.

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

[15] "ICSI Room Responses," http://www.icsi.berkeley.edu/speech/papers/ asru01-meansub-corr.html.

[16] "ISCA Speech Corpora," http://www.isca-students.org/corpora.