IDIAP RESEARCH REPORT

Hilbert Envelope Based Features for Far-Field Speech Recognition

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

IDIAP–RR 08-42

June 2008

to appear in
MLMI 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 08-42

Hilbert Envelope Based Features for Far-Field Speech Recognition

Samuel Thomas Sriram Ganapathy Hynek Hermansky

June 2008

to appear in
MLMI 2008

Abstract. Automatic speech recognition (ASR) systems, trained on speech signals from close-talking microphones, generally fail in recognizing far-field speech. In this paper, we present a Hilbert Envelope based feature extraction technique to alleviate the artifacts introduced by room reverberations. The proposed technique is based on modeling temporal envelopes of the speech signal in narrow sub-bands using Frequency Domain Linear Prediction (FDLP). ASR experiments on far-field speech using the proposed FDLP features show significant performance improvements when compared to other robust feature extraction techniques (average relative improvement of 43% in word error rate).

1 Introduction

When speech is recorded in rooms using far-field microphones, 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 $h(t)$. For long analysis windows, this effect of reverberation can be approximated as multiplicative in the frequency domain [1], 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 [2, 3], the effect of reverberation is compensated by subtracting from $\log\big(R(t, \omega_k)\big)$, its mean.

In this paper, 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 [4] (Sec. 2). Unlike conventional approaches that use mean compensation for reverberant speech recognition [2, 3], the proposed technique alleviates the reverberation artifacts present in long temporal envelopes of narrow frequency sub-bands(Sec. 3). The application of the proposed compensation technique to the FDLP features significantly improves the recognition accuracies for reverberant speech recorded using far-field microphones (Sec. 4).

2 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) [5]). This paper utilizes AR models for obtaining smoothed, minimum phase, parametric models for temporal rather than spectral envelopes (Fig. 1). Since we apply the LP technique to exploit the redundancies in the frequency domain, this approach is called Frequency Domain Linear Prediction (FDLP) [4], [6]. 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 [5]).

When speech is analyzed in narrow sub-bands using such long analysis windows, each sub-band signal can be modeled in terms of the product of a slowly varying, positive, envelope function and an instantaneous phase function [7]. In the case of far-field speech, each of these sub-band signals gets modified by the room impulse response and 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 function [7]. Since the Hilbert envelope and the spectral autocorrelation function form Fourier transform pairs [4], normalizing the gain of the sub-band FDLP envelopes suppresses the multiplicative effect present in the spectral autocorrelation function of the reverberant speech.

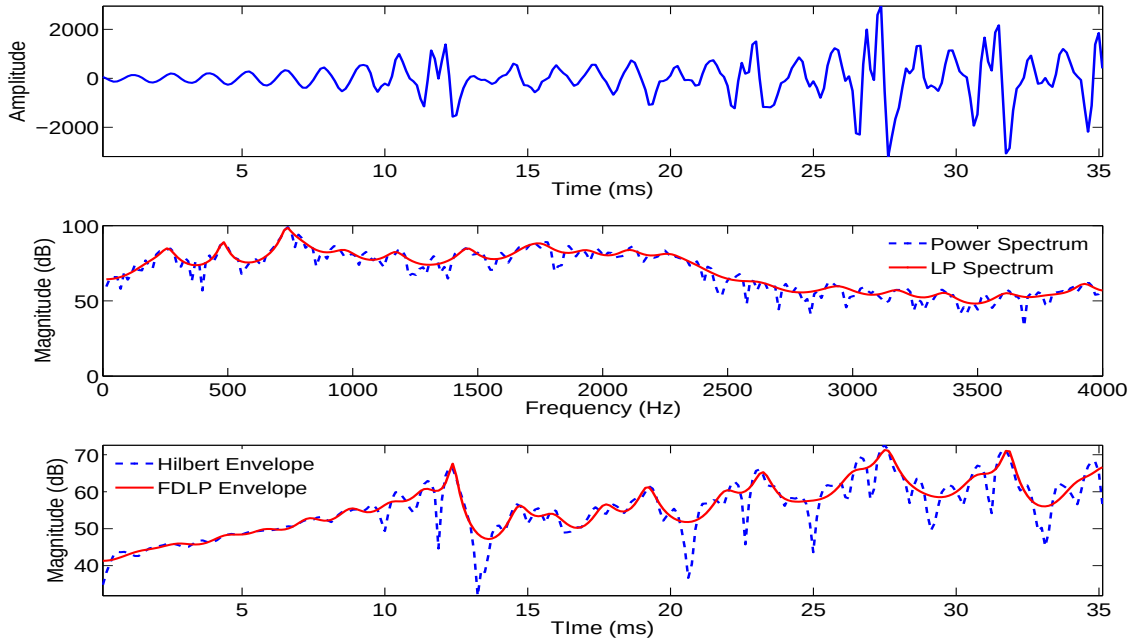

Figure 1: *Linear Prediction in time and frequency domains for a portion of speech signal*

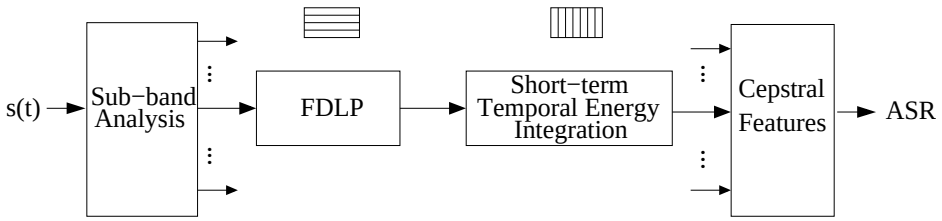

Figure 2: *FDLP feature extraction for ASR*

3 Features based on Frequency Domain Linear Prediction

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. The whole set of sub-band temporal envelopes forms a two dimensional (time-frequency) representation of the input signal energy. Each of these temporal envelopes is gain normalized to suppress the reverberation artifacts. 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 [8]. 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. 2.

4 Experiments and Results

We apply the proposed features and techniques to a connected word recognition task on a digits corpus using the Aurora evaluation system [9] along with the "complex" version of the back end proposed in [10]. We train models using a training dataset used in [3] which contains of 8400 clean speech utterances, consisting of 4200 male and 4200 female utterances downsampled to 8 kHz. In order to study the effect of finer spectral resolution for the proposed compensation technique, we first perform experiments using a test set of 3003 clean utterances also used in [3]. We also create a test set for artificially reverberated speech by convolving the clean test set with a room impulse response (with RT60 of 0.5 seconds and a direct-to-reverberant energy ratio of 0 dB [12]).

The first set of experiments compare the performance of FDLP based features with the conventional features for clean and artificially reverberated speech. We also study the effect of finer spectral resolution for the proposed compensation technique by increasing the number of frequency sub-bands. Table 1 shows the word accuracies for PLP features (PLP) and FDLP features when the number of sub-bands is varied from 24 (FDLP-24) to 120 (FDLP-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. For all these experiments we employ gain normalized temporal envelopes along with rectangular windows in the DCT domain.

Table 1: Word Accuracies (%) for PLP and FDLP features for clean and reverberant speech

Feature Set	Clean Speech	Revb. Speech
PLP	99.68	80.12
FDLP-24	99.18	89.49
FDLP-33	99.13	91.86
FDLP-67	99.09	92.93
FDLP-76	99.16	93.60
FDLP-96	99.07	94.79
FDLP-108	99.03	94.63
FDLP-120	98.91	94.55

Table 2: Word Accuracies (%) using different feature extraction techniques on far-field microphone speech

Channel	PLP	CMS	LDMN	LTLSS	FDLP
Channel E	68.1	71.2	73.2	74.0	85.2
Channel F	75.5	77.4	80.4	81.0	88.1
Channel 6	74.1	78.3	80.9	81.1	89.6
Channel 7	58.6	67.6	70.5	71.0	84.9

The next set of experiments are performed on the digits corpus recorded using far-field microphones as part of the ICSI Meeting task [11]. The corpus consists of four sets of 2790 utterances each. Each of these sets correspond to speech recorded simultaneously using four different far-field microphones [11]. Each of these sets contain 9169 digits similar to those found in TIDIGITS corpus. The number of sub-bands for the FDLP features is fixed at 96 along with a temporal analysis window of duration 2000 ms. We use the HMM models trained with the clean speech from earlier experiments. 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) [3] and Log-DFT Mean Normalization (LDMN) [2]. 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 [3]) using

a shorter analysis window of 32 ms, with a shift of 8 ms. Table 2 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 43% over the best other feature extraction technique.

5 Conclusions

Unlike many single microphone based far-field 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.

References

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

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

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

[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.