# Modulation spectrum analysis for recognition of reverberant speech

*Sri Harish Mallidi*[1], *Sriram Ganapathy*[1], *Hynek Hermansky*[1,2]

[1]Department of Electrical and Computer Engineering
[2]Human Language Technology Center of Excellence
Johns Hopkins University, USA
{smallid1,ganapathy,hynek}@jhu.edu

## Abstract

Recognition of reverberant speech constitutes a challenging problem for typical speech recognition systems. This is mainly due to the conventional short-term analysis/compensation techniques. In this paper, we present a feature extraction technique based on modeling long segments of 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 by linear prediction on cosine transform of the signal. We show that the FDLP modulation spectrum plays an important role in the robustness of the proposed feature extraction. Automatic speech recognition (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 $40\%$ in word error rate). Similar improvements are also obtained for far-field data which contain natural reverberation in background noise.
**Index Terms**: Frequency Domain Linear Prediction, Reverberant Speech, Automatic Speech Recognition.

## 1. Introduction

When speech is corrupted by room reverberation, the short-term spectral estimates are smeared. This causes a mismatch in the features extracted from clean speech and results in a degradation in the ASR performance. 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 can be modelled 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 reverberant speech respectively. The effect of reverberation on the short-time Fourier transform (STFT) of the speech signal $s(t)$ can be represented as

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

where $S(n, \omega_k)$ and $R(n, \omega_k)$ are the STFTs of the clean speech signal $s(t)$ and reverberant speech $r(t)$ respectively. Here, $H(n, \omega_k)$ denotes the STFT of the room impulse response $h(t)$, $n$ denotes the frame index and $w_k$ denotes the $k$th frequency bin.

The amount of reverberation in speech is generally characterized by reverberation time ($T_{60}$) (time required for reflections of a direct sound to decay by $60$ dB below the level of the direct sound, typically in the range of $200$-$700$ms). The main assumption in conventional short-term channel compensation techniques is $H(n, \omega_k) = H(\omega_k) \forall n$. While this assumption is reasonable for distortions like linear telephone channel noises, it is not valid for long-term artifacts like room reverberations. Thus, by using conventional approaches like cepstral mean subtraction [3], feature warping [4], (where analysis windows for deriving cepstral features are much shorter than $T_{60}$), the effect of reverberation cannot be suppressed.

The use of long-term mean subtraction has also been studied in the past for the suppression of room reverberation [5, 6]. This approach involves the subtraction of a mean estimate of the log spectrum using a long-term (2s) analysis window, followed by an overlap-add re-synthesis. In our past work, the application of gain normalization of 1s long sub-band temporal envelopes has also shown to be useful for speech recognition in room reverberations [7]. Sub-band temporal envelopes of speech are derived using FDLP [8, 9]. The sub-band envelopes in long-term analysis windows and narrow sub-bands, are gain normalized to provide robustness in reverberant environments [7]. These long-term sub-band envelopes are integrated in short-term windows ($25$ ms with a shift of $10$ ms) and are converted to cepstral features similar to conventional feature extraction techniques [10].

In this paper, we propose to analyze the robustness of the FDLP feature extraction techniques using the concept of modulation spectrum. Spectral representation of amplitude modulation in sub-bands are called "Modulation Spectra" [11]. It has been shown that important information for speech perception lies in the $1 - 16$ Hz range of the modulation frequencies [12]. In the FDLP framework, the modulation spectrum is defined as the spectrum of the log FDLP envelope in sub-band. The modulation spectrum in FDLP analysis is determined by a number of parameters like the model order in FDLP, band-width of the sub-band and the expansion factor used in the estimation. We analyze the effects of each of these components in terms of the average modulation spectrum as well as in terms of the robustness of the final ASR system for reverberant speech recognition.

The proposed features are used for a connected digit recognition task in TIDIGTS database. For reverberant speech recognition experiments, the test data was convolved with a set of $8$ different room responses collected from various sources [13, 14]. The ASR models are trained on clean TIDIGITS

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government.

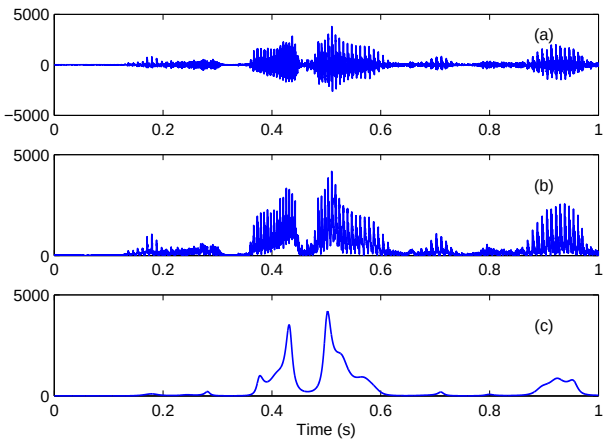


Figure 1: Illustration of the all-pole modelling with FDLP. (a) a portion of speech signal, (b) its Hilbert envelope and (c) all-pole model obtained using FDLP.

data and are tested using clean as well as reverberant speech data. In these experiments, the proposed features provide significant improvement over the baseline features (average improvement of 40 %). Further, we also show consistent improvements with experiments on naturally reverberant connected digits data recorded using a far-field microphone.

The rest of the paper is organized as follows. In Sec. 2, the FDLP technique for feature extraction is explained. The speech recognition setup using connected digits is described in Sec. 3. The modulation spectrum analysis using FDLP envelopes is detailed in Sec. 4. Speech recognition experiments with the proposed features are reported in a 5. In Sec. 6, we conclude with a discussion of the proposed features.

## 2. Frequency domain linear prediction

Linear prediction (LP) analysis exploits a simple form of redundancy in a signal by modelling the current sample as a linear combination of a fixed number of past samples. By extracting the linear dependence, the original signal is described as a result of passing a temporally uncorrelated (white) excitation sequence passed through a fixed all-pole digital filter. When LP analysis is applied in time domain, the filter comprises a parametric approximation of its power spectrum. The duality of time and frequency domain means LP can be applied to discrete spectral representation of a signal. This process is called as frequency domain linear prediction (FDLP). In a manner similar to parametric representation of power spectrum by time domain linear prediction, FDLP provide a parametric representation of Hilbert envelope of the signal [9]. Fig. 1 plots the FDLP envelope which approximates the Hilbert envelope of the signal.

### 2.1. Estimating robust temporal envelopes

For long segments of the signal in narrow sub-bands, Hilbert envelope of the reverberant speech can be approximated as the convolution of Hilbert envelope of clean speech and Hilbert envelope of room impulse response. Hilbert envelope and the spectral autocorrelation form Fourier transform pairs. Therefore, spectral autocorrelation of reverberant speech is product of spectral autocorrelation of clean speech and spectral autocorrelation of room impulse response. The spectral autocorrelation

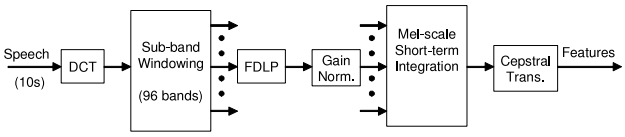


Figure 2: Block schematic of FDLP feature extraction.

of room impulse response can be assumed to be slowly varying compared to clean speech. In a first approximation, normalizing gain in sub-band FDLP envelopes suppresses the multiplicative effect present in spectral autocorrelation function of the reverberant speech [7]. This technique is called as gain normalization and it helps in supressing the effect of reverberation.

### 2.2. Cepstral features

For the purpose of feature extraction, the input speech signal is decomposed into sub-bands, where FDLP is applied in each sub-band to obtain a parametric model of the temporal envelope. The whole set of sub-band temporal envelopes forms a two dimensional time-frequency representation (similar to conventional short-term spectrogram) 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 conventional PLP feature extraction technique [10]. 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.

## 3. Experimental setup

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 [16]. We use the complex version of the back end proposed in [17]. 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] and natural farfield data [13].

## 4. Modulation spectrum and robustness

The relation between modulation spectrum and parameters of FDLP are analyzed in this section. Modulation spectrum is the spectral representation of amplitude modulation component of sub-band signal. Average modulation spectrum (AMS) is the Monte-Carlo average estimate of modulation spectrum from a large number of sub-bands of various speech utterances. In order to compute AMS, we choose 120 utterances from Aurora speech database (60 male and 60 female) and average the modulation spectral estimate over all the sub-bands of all the utterances. For each utterance, the signal is decomposed into a number of sub-bands and the FDLP envelope is computed in each sub-band as described in section 2. Modulation spectrum is the Fourier transform of the log FDLP envelope.

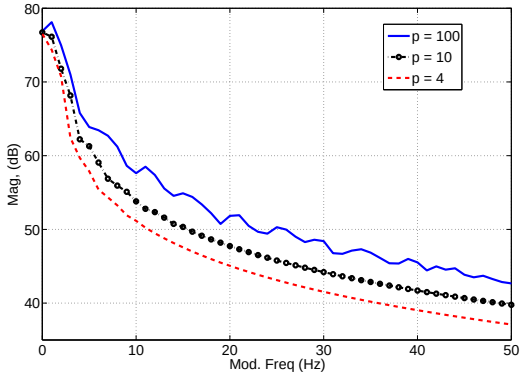


Figure 3: AMS for various model orders $p$ per second per sub-band.

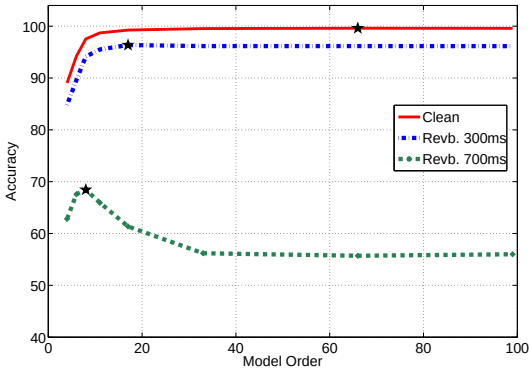


Figure 4: Word recognition accuracy as function of the model order for clean and two types of reverberant data. The best performance in each condition is highlighted using the star sign.

### 4.1. Model order

In a linear prediction scenario, the model order corresponds to number of previous samples used in prediction. In the FDLP analysis, the model order controls the number of distinct temporal peaks in the sub-band envelope. Model order has a direct influence on the AMS, which is illustrated in Fig. 3. As seen here, the higher the model order, the lower the roll-off of AMS in the modulation frequency domain.

When speech is corrupted by room reverberation, the sub-band envelopes are smeared in time. The degree of smearing is determined by the reverberation time (T60). In this case, higher order FDLP results in the estimation of large number of signal peaks which are not robust. On the other hand, a lower model order fails to capture enough information needed for good ASR performance in clean conditions (or when there is a lower degree of reverberation). This tradeoff is illustrated in Fig. 4, where we plot the ASR accuracy for clean conditions and on two types of reverberant data (which has reverberation time of 300 and 700 ms) as a function of the FDLP model order. The best performance in each condition is also highlighted. It can be seen that a lower model order is good when there is significant amount of reverberation, while a higher model order is preferred for clean conditions.

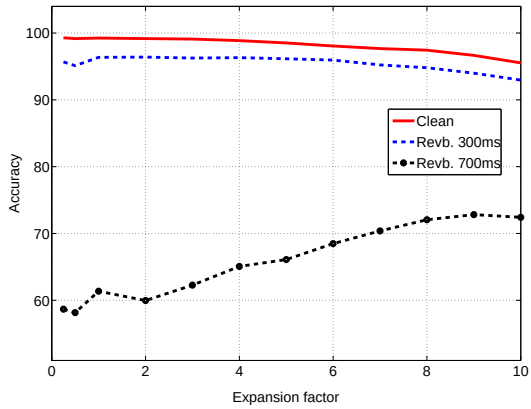


Figure 5: Word recognition accuracy as function of the expansion factor for clean and two types of reverberant data.

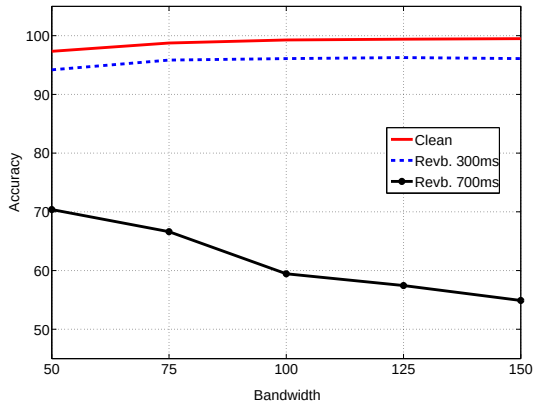


Figure 6: Word recognition accuracy as function of the bandwidth of the sub-band for clean and two types of reverberant data.

### 4.2. Envelope expansion

In the past, it has been shown that the time domain linear prediction can be modified to estimate a transformed spectral envelope instead of the original spectrum [18]. The autocorrelations derived from the modified power spectrum are used for linear prediction. In FDLP framework, spectral autocorrelations can be derived from transformed Hilbert envelopes where the transformation here corresponds to raising the original Hilbert envelope to a power $r$. When the Hilbert envelope is compressed ($r < 1$), the resulting model tends to approximate the valleys of the envelope better [18]. However, expansion of the envelopes ($r > 1$) results in enhanced modelling of the peaks of the envelope.

We apply the transform linear prediction in FDLP and derive features for ASR. When speech is corrupted by room reverberation, the high energy peaks (where the signal to reverberant component ratio is high) can be more robustly estimated as compared to the valleys of the envelope. Thus, FDLP features derived using expanded envelopes ($r > 1$) are more robust in reverberant environments. This is illustrated in Fig. 5, where we plot the ASR accuracy for clean conditions as well as the two reverberant conditions as function of the the expansion factor $r$.

Table 1: Word accuracies (%) using a clean test data, average word accuracy for 8 conditions of artificial reverberant data and average word accuracy for 4 conditions of natural far-field microphone data.

| | PLP | CMS | LDMN | LTLSS | FDLP |
|---|---|---|---|---|---|
| Clean | 99.7 | 99.7 | 99.6 | 99.6 | 98.9 |
| Art. Revb. | 65.6 | 71.9 | 75.7 | 76.6 | 86.7 |
| Far fiel | 69.1 | 73.6 | 76.3 | 76.8 | 86.4 |

### 4.3. Bandwidth

In the past work [7], a decomposition of 96 bands was found to be robust in reverberant environments. However, for a fixed number of sub-bands, the bandwidth of the sub-bands can be varied keeping the band overlap constant. As mentioned before, the use of narrow sub-band increased the validity of the assumptions made in the gain normalization. But, narrow sub-band also means that the modulation extent of the corresponding AMS reduces (given by half of bandwidth of the sub-band). As seen in Fig. 6, as the bandwidth reduces the robustness in reverberant environment improves significantly while the performance in clean conditions degrades moderately.

## 5. Results

In this section, we use the proposed features for recognition of reverberant speech from 8 different artificial room responses collected from various sources [13, 14, 15] with reverberation time ranging from 200 to 800ms. The use of 8 different room responses results in 8 test sets consisting of 3003 utterances each. 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 [13]). The test data consist of four parallel channels with 2790 utterances each . As before, we use the HMM models trained with the clean speech in the training set of modified Aurora task.

For the proposed FDLP features, we use a expansion factor $r = 4$, with 15 poles per second per sub-band and sub-band bandwidth of 100 Hz. 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) [3], Long Term Log Spectral Subtraction (LTLSS) [6] and Log-DFT Mean Normalization (LDMN) [5]. The average performance in clean conditions as well as in 8 conditions of artificial reverberation and 4 conditions of natural far-field data is shown in Table 1. For the different artificial room responses, the proposed FDLP features, on the average, provide a relative error improvement of 43% over the other feature extraction techniques considered. Further, on the 4 conditions of far-field test data, we obtain a relative error improvement of about 41%. The performance improvement is achieved with a moderate degradation in clean conditions.

## 6. Conclusions

In this paper, we have studied the effect of various parameters in deriving FDLP modulation spectrum for robust representation of speech features. These parameters include the model order, the bandwidth of the sub-band and the expansion factor used in transform linear prediction and provide substantial robustness in reverberant environments. Once the FDLP envelopes are estimated, these are converted to short-term features and are used for ASR similar to conventional short-term spectral features. In reverberant environments, the proposed features provide significant improvements compared to other feature extraction techniques. The application of the proposed techniques for larger vocabulary tasks and speaker recognition tasks as well as for signals distorted by additive and convolutive noise are currently pursued.

## 7. 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] 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.

[4] J. Pelecanos, and S. Sridharan, "Feature warping for robust speaker verification", *Proc. Speaker Odyssey 2001 Speaker Recognition Workshop*, Greece, pp. 213-218, 2001.

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

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

[7] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Proc. Letters*, Vol. 15, pp. 681-684, 2008.

[8] R. Kumerasan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications", *Journal of Acoustical Society of America*, Vol. 105 (3), Mar. 1999, pp. 1912-1924.

[9] M. Athineos, and D. Ellis, "Autoregressive modelling of temporal envelopes," *IEEE Tran. Signal Proc.*, Vol. 55, pp. 5237-5245, 2007.

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

[11] T. Houtgast, H. J. M. Steeneken and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function, I. General room acoustics," *Acoustica 46*, pp. 60-72, 1980.

[12] R. Drullman, J.M. Festen and R. Plomp,"Effect of Reducing Slow Temporal Modulations on Speech Reception", *J. Acoust. Soc. Am.*, Vol. 95(5), pp. 2670-2680, 1994.

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

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

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

[16] 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.

[17] 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.

[18] H. Hermansky, H. Fujisaki, Y. Sato, "Analysis and synthesis of speech based on spectral transform linear predictive method", em Proc. ICASSP, April, 1983.