Temporal envelope compensation for robust phoneme recognition using modulation spectrum

Sriram Ganapathy,[a] Samuel Thomas, and Hynek Hermansky
Department of Electrical and Computer Engineering, Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, Maryland 21218

(Received 17 December 2009; revised 17 September 2010; accepted 24 September 2010)

A robust feature extraction technique for phoneme recognition is proposed which is based on deriving modulation frequency components from the speech signal. The modulation frequency components are computed from syllable-length segments of sub-band temporal envelopes estimated using frequency domain linear prediction. Although the baseline features provide good performance in clean conditions, the performance degrades significantly in noisy conditions. In this paper, a technique for noise compensation is proposed where an estimate of the noise envelope is subtracted from the noisy speech envelope. The noise compensation technique suppresses the effect of additive noise in speech. The robustness of the proposed features is further enhanced by the gain normalization technique. The normalized temporal envelopes are compressed with static (logarithmic) and dynamic (adaptive loops) compression and are converted into modulation frequency features. These features are used in an automatic phoneme recognition task. Experiments are performed in mismatched train/test conditions where the test data are corrupted with various environmental distortions like telephone channel noise, additive noise, and room reverberation. Experiments are also performed on large amounts of real conversational telephone speech. In these experiments, the proposed features show substantial improvements in phoneme recognition rates compared to other speech analysis techniques. Furthermore, the contribution of various processing stages for robust speech signal representation is analyzed. [DOI: 10.1121/1.3504658]

PACS number(s): 43.72.Ne, 43.72.Ar [SSN] Pages: 3769–3780

I. INTRODUCTION

Conventional speech analysis techniques start with estimating the spectral content of relatively short (about 10–20 ms) segments of the signal (short-term spectrum). Each estimated vector of spectral energies represents a sample of the underlying dynamic process in production of speech at a given time-frame. Most of the information contained in these acoustic features relate to formants which provide important cues for recognition of basic speech units. Further, additional information about the dynamics of the underlying speech signal is incorporated with these feature vectors using the derivative features. Stacking such estimates of the short-term spectra in time provides a two-dimensional (time–frequency) representation of speech that forms the basis for most speech features (Hermansky, 1990).

An alternate way to describe a speech signal is a summation of a number of amplitude modulated narrow frequency bands. In this view, every frequency band can be considered to consist of a carrier signal (fine structure) and a time-varying envelope (Kumerasan and Rao, 1999). One can directly estimate trajectories of spectral energies in the individual frequency sub-bands, each estimated vector then representing the underlying dynamic process in a given sub-band. Such estimates, stacked in frequency, also form a two-dimensional representation of speech (Athineos *et al.*, 2004).

Spectral components of long-term amplitude modulations in individual frequency sub-bands are called modulation spectra. The modulation spectral representations have been used in the past for predicting speech intelligibility in reverberant environments (Houtgast *et al.*, 1980). They are now widely applied in many engineering applications [for example, audio coding (Vinton and Atlas, 2001), noise suppression (Falk *et al.*, 2007), etc.]. Feature extraction techniques that are based on modulation spectrum have also been proposed for automatic speech recognition (ASR) (Hermansky and Sharma, 1998; Kingsbury *et al.*, 1998).

The importance of various modulation frequency components in human phoneme recognition has been reported in the past (Riesz, 1928). Speech intelligibility experiments have been studied by presenting speech stimulus with varying amounts of modulation spectral information (Drullman *et al.*, 1994). In these experiments, it has been shown that important information for phoneme perception lies in the 1–16 Hz range of the modulation frequencies. The recognition of consonants, especially the stops, suffers more when the temporal modulations below 16 Hz are filtered out. In another study, the interaction between the spectral and temporal modulations has been analyzed by varying the number of sub-bands in speech analysis (Shannon *et al.*, 1995). Even when the spectral information is limited to four sub-bands, the use of temporal amplitude modulations alone provides good human phoneme recognition. However, in the presence of noise, the number of spectral channels needed for good vowel recognition increases, whereas the contribution of temporal modulations remain similar in clean and noisy conditions (Xu and Zheng, 2007).

[a]Author to whom correspondence should be addressed. Electronic mail: ganapathy@jhu.edu

For machine recognition of phonemes in noisy speech, there is considerable benefit in using larger temporal context for feature representation of a single phoneme (Morgan *et al.*, 1992; Pinto *et al.*, 2008). The techniques that are based on deriving long-term modulation frequencies do not preserve fine temporal events like onsets and offsets which are important in separating some phoneme classes. On the other hand, signal adaptive techniques, which try to represent local temporal fluctuation, cause strong attenuation of higher modulation frequencies which makes them less effective even in clean conditions (Tchorz and Kollmeier, 1999). Furthermore, the performance of most of these feature extraction techniques degrades significantly in the presence of additive or convolutive noise (mismatched train/test conditions).

In our previous work (Ganapathy *et al.*, 2009), we have proposed a combination of static and dynamic modulation frequency features for phoneme recognition. Here, the input speech signal is decomposed into a number of critical bands. In each sub-band, long-term envelopes are extracted using frequency domain linear prediction (FDLP), which is an efficient technique for auto-regressive (AR) modeling of temporal envelopes of a signal (Athineos and Ellis, 2007; Kumerasan and Rao, 1999). FDLP envelopes are then compressed using a static and a dynamic compression. The static compression stage is a logarithmic operation and dynamic compression stage uses adaptive compression loops (Tchorz and Kollmeier, 1999). The compressed envelopes are transformed into modulation spectral components which are used as features for a phoneme recognition system.

Although the features proposed in Ganapathy *et al.* (2009) perform well in clean conditions, the performance degrades in noisy environments and mismatched train/test conditions. This degradation is severe in the presence of additive noise where the low energy region of the speech signal gets masked in noise. This causes a mismatch in the representation of speech in clean and noisy condition which results in high error rates.

In this paper, we propose a noise compensation technique for modulation frequency features based on temporal envelope subtraction. In each sub-band, an estimate of the noise envelope is derived from the input noisy speech. This estimate is subtracted from the noisy speech envelope before the application of linear prediction in frequency domain. The noise compensation tries to create an invariance in representation of speech in clean and noisy conditions.

The noise compensated envelopes are further gain normalized to suppress the convolutive artifacts in speech like linear distortions due to frequency characteristics of the communication channel and reverberations. The gain normalization procedure was previously developed for deriving short-term spectral features (Thomas *et al.*, 2008). In this paper, we apply this procedure on long temporal envelopes to derive static and dynamic modulation frequency features.

Experiments are done on a phoneme recognition using the hybrid hidden Markov model-artificial neural network (HMM-ANN) phoneme recognition system (Bourlard and Morgan, 1994). The test data in these experiments consist of speech corrupted with variety of real world additive noises at different signal-to-noise ratios (SNRs), convolutive distortions introduced by different room impulse response functions and multiple telephone channel speech recordings with different frequency characteristics. In this paper, we show that the noise compensation technique provides considerable robustness in additive noise conditions. When this is used along with gain normalization technique, there is further improvement in phoneme recognition accuracy in reverberant environments and telephone channel speech recordings over the other state-of-the-art robust feature extraction techniques. We also illustrate the usefulness of the proposed features for phoneme recognition task on large amounts of conversational telephone speech (CTS) data.

In our previous work (Ganapathy *et al.*, 2009), modulation features were derived from uncompensated envelopes. The main goal of this paper is to develop a robust representation of speech in additive noise conditions. Using these techniques, considerable improvements are shown on all types of additive noise and SNR conditions. Furthermore, these techniques also enhance the robustness in the presence of other acoustic distortions like reverberation and telephone channel artifacts. Experiments are done only with distortions which involve non-speech interference. The usefulness of the proposed techniques for speech interference (like recognition of overlapped speech and speaker separation) is not addressed in this paper.

Robust phoneme recognition has a huge impact in wide range of applications like language identification, large vocabulary continuous speech recognition (LVCSR), keyword-spotting, and voice activity detection (VAD) (Schwarz, 2008). Most of these systems have a front-end phoneme recognition followed by further processing stages which use this phoneme sequence. Hence, almost all these applications benefit from improved robustness in phoneme recognition.

II. FREQUENCY DOMAIN LINEAR PREDICTION

Typically, 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) (Makhoul, 1975)]. This paper utilizes AR models for obtaining smoothed, minimum phase, and 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.

The Hilbert envelope, which is the squared magnitude of the analytic signal, represents the instantaneous energy of a signal in the time domain. A discrete time analytic signal can be obtained by forcing the causality of the discrete Fourier transform (DFT) and by ensuring the orthogonality of real and imaginary parts (Marple, 1999). Mathematically, it can be shown that the autocorrelation of discrete cosine transform (DCT) of the input signal and the discrete time Hilbert envelope are Fourier transform pairs (Athineos and Ellis, 2007). This means that the application of linear prediction on the cosine transform of the signal yields an AR model of the Hilbert envelope of the signal. Thus, a parametric model for the Hilbert envelopes can be obtained using FDLP (Athineos and Ellis, 2007; Kumerasan and Rao, 1999).

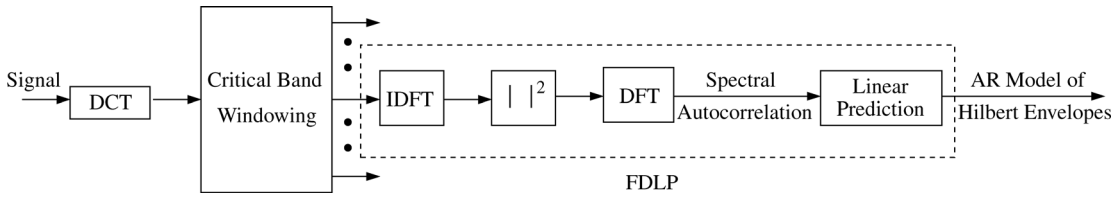

FIG. 1. Block schematic for the FDLP. The steps involved are application of DCT, estimation of spectral autocorrelations, and linear prediction to estimate the AR model of Hilbert envelope.

Figure 1 shows the block schematic for the implementation of FDLP technique. The input signal is transformed into frequency domain using DCT. The full-band DCT is windowed using bark-spaced windows to yield sub-band DCT components. In each sub-band, the inverse discrete Fourier transform (IDFT) of the DCT coefficients represents the discrete time analytic signal (Athineos and Ellis, 2007). Spectral autocorrelations are derived by the application of DFT on the squared magnitude of analytic signal. These autocorrelations are used for linear prediction [similar to the application of TDLP using time domain autocorrelations (Makhoul, 1975)]. The output of linear prediction is a set of AR model parameters which characterize the sub-band Hilbert envelopes.

For example, if the signal is sampled at 8 kHz, we get 8000 DCT coefficients for a 1000 ms window of the signal. These 8000 coefficients are windowed into 15 critical bands using bark-spaced windows (approximately 1 bark) in the DCT domain. The IDFT is performed on the sub-band DCT for 1000 ms signal. Then, spectral autocorrelations are derived using DFT operation and are used for FDLP. The order of the linear prediction is one pole per ten sub-band DCT samples.

As the conventional AR models are used effectively on signals with spectral peaks, the AR models of the temporal envelope are appropriate for signals with peaky temporal envelopes (Kumerasan and Rao, 1999). The individual poles in the resulting polynomial are directly associated with specific energy maxima in the time domain waveform. For signals that are expected to consist of a fixed number of distinct energy peaks in a given time interval, the AR model could well approximate these perceptually dominant peaks and the AR fitting procedure removes the finer-scale detail. This suppression of detail is particularly useful in speech recognition applications, where the goal is to extract the general form of the signal by means of a parametric model. An illustration of the all-pole modeling property of the FDLP technique is shown in Fig. 2, where we plot a portion of sub-band speech signal (frequency range of 500–700 Hz), its Hilbert envelope computed from the DFT (Marple, 1999) and the AR model fit to the Hilbert envelope using FDLP with a model order of 80.

A. Noise compensation in FDLP

When speech signal is corrupted by additive noise, the signal that reaches the microphone can be written as

$$x[t] = s[t] + n[t], \tag{1}$$

where $x[t]$ is the discrete representation of the input signal, $s[t]$ represents the clean speech signal which is corrupted by noise $n[t]$.

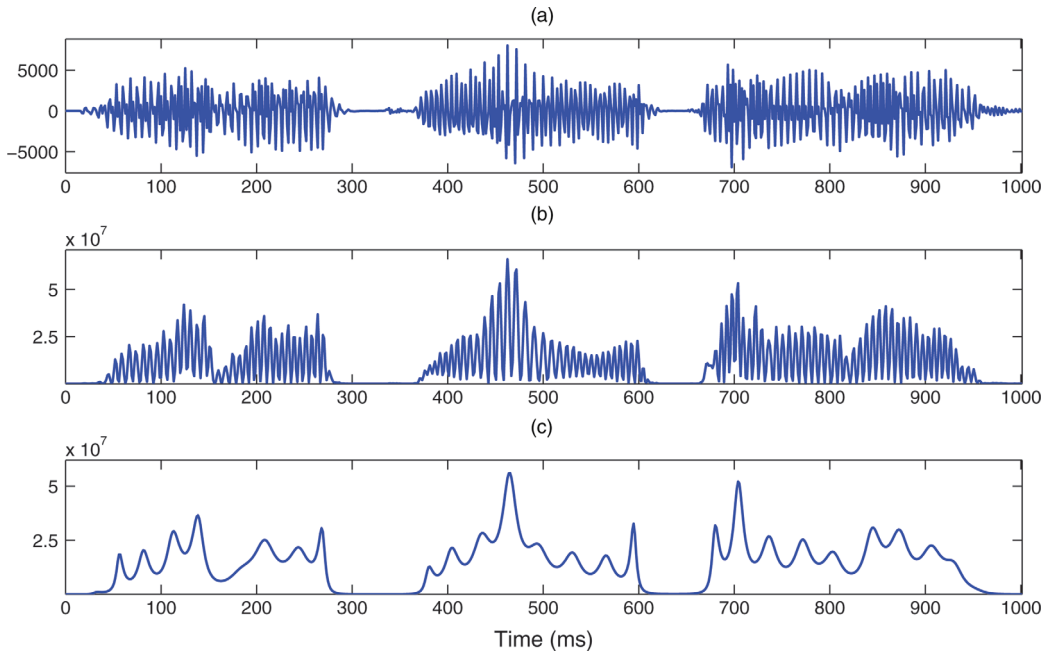

FIG. 2. (Color online) Illustration of the all-pole modeling property of FDLP. (a) A portion of the sub-band speech signal (with frequency range of 500–700 Hz), (b) its Hilbert envelope, and (c) all-pole model obtained using FDLP with model order of 80.

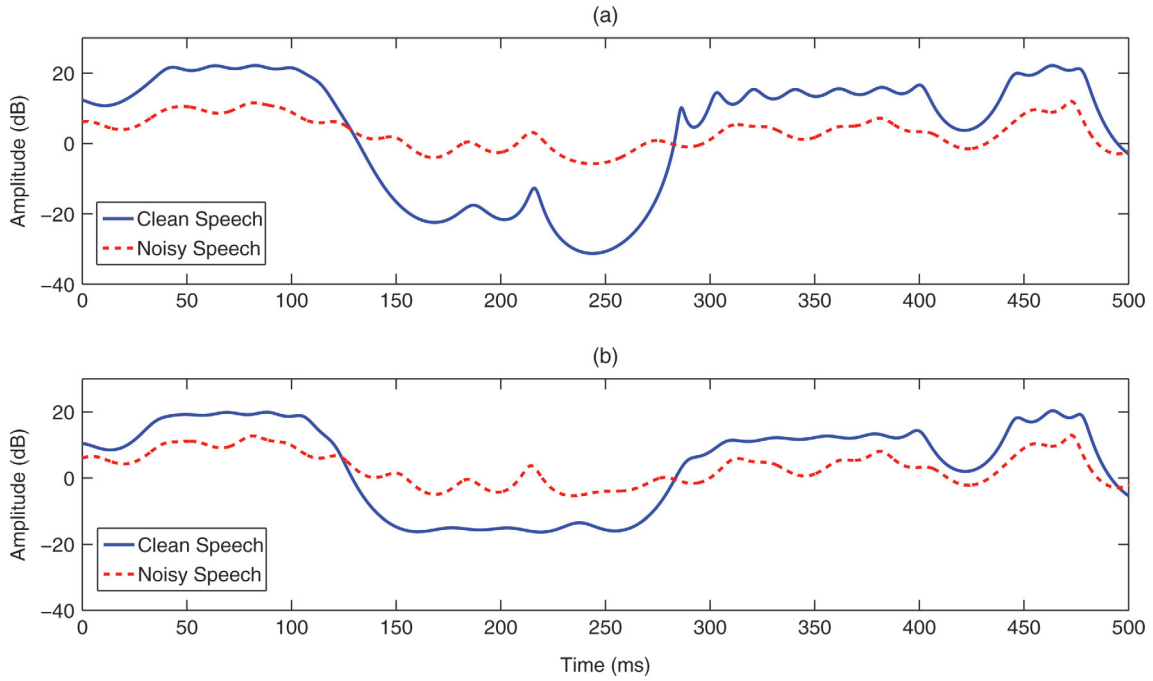

FIG. 3. (Color online) Log-FDLP envelopes for a sub-band of clean speech and speech corrupted with babble noise at 10 dB SNR. (a) Without noise compensation and (b) with noise compensation.

Assuming that the speech and noise are uncorrelated, we obtain

$$P_X(t,\omega_k) = P_S(t,\omega_k) + P_N(t,\omega_k), \tag{2}$$

where $P_X(t, \omega_k)$, $P_S(t, \omega_k)$, and $P_N(t, \omega_k)$ are the short-term power spectral densities (PSD) at frequency ω_k of the noisy speech, clean speech, and noise, respectively.

Conventional feature extraction techniques for ASR estimate the short-term (10–30 ms) PSD of speech in bark or mel scale (Hermansky, 1990). Hence, most of the recently proposed noise robust feature extraction techniques apply some kind of spectral subtraction in which an estimate of the noise PSD is subtracted from the noisy speech PSD (ETSI, 2002).

The proposed noise compensation technique for FDLP is shown in Fig. 3. A VAD operates on the input speech signal to indicate the presence of non-speech frames. The VAD is implemented using the same technique proposed in ETSI (2002). The VAD output is a flag indicating the speech/non-speech decision for every short-term frame of speech (with a length of 25 ms and a shift of 10 ms).

As mentioned in Sec. II, long segments of the input speech signal are transformed to DCT domain where they are decomposed into sub-band DCT components. The discrete time analytic signal is obtained as the squared magnitude IDFT of the DCT signal. We apply short-term noise subtraction on the analytic signal. This is achieved in two steps. In the first step, we window the analytic signal into short-term segments (of length 25 ms with a shift of 10 ms). The next step is to subtract an estimate of the short-term noise component from these segments.

Since the noise component is assumed to be additive in signal domain, we can write

$$X[k] = S[k] + P[k], \tag{3}$$

where $X[k]$, $S[k]$, and $P[k]$ are the kth DCT coefficient of noisy speech, clean speech, and noise, respectively. If speech and noise are uncorrelated, they continue to be uncorrelated in the DCT domain by virtue of the orthogonality property of the DCT matrix. Further, the application of squared magnitude IDFT gives (using the assumption of uncorrelated speech and noise)

$$A_X[t] = A_S[t] + A_N[t], \tag{4}$$

where $A_X[t]$, $A_S[t]$, and $A_N[t]$ are the short-term analytic signal representations of the noisy speech, clean speech, and noise, respectively. The previous equation shows that the effect of noise can be alleviated if an estimate of $A_N[t]$ is subtracted from the short-term noisy speech analytic signal $A_X[t]$.

An estimate of the short-term noise envelope is obtained by averaging the envelope segments in the non-speech region (from the beginning and end of speech utterance). This estimate is subtracted from the short-term envelopes of speech similar to the spectral subtraction technique (ETSI, 2002). The noise compensated short-term envelopes are synthesized using overlap-add to obtain the long-term sub-band envelopes. These are converted back to sub-band DCT domain and used for FDLP.

Figure 4 shows the effect of the proposed noise compensation technique on sub-band envelopes. The additive noise present in speech signal modifies the FDLP envelope in such a way that the dynamic range is reduced. This is illustrated in Fig. 4(a), where we plot the sub-band FDLP envelopes in clean and noisy conditions for the same speech utterance. The effect of noise is more pronounced in the valleys of the log-FDLP envelopes, where there is substantial mismatch between clean and noisy speech. When features are derived

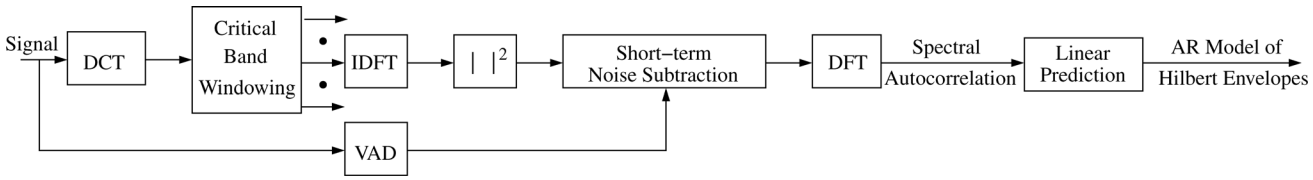

FIG. 4. Noise compensation in FDLP.

from noisy speech using the uncompensated FDLP envelopes, the performance of the phoneme recognition system degrades significantly.

Figure 4(b) provides an illustration of the effect of this noise compensation technique on the sub-band FDLP envelopes for clean and noisy speech. The noise compensation procedure modifies the clean envelopes in such a way that the valleys of trajectory are deemphasized. When the compensated value reduces below zero, the corresponding magnitude value is used. Although this method of compensation affects the information in valleys of clean speech signal, it reduces the mismatch between FDLP envelopes extracted from clean and noisy speech. In this view, the proposed approach operates like an envelope normalization procedure as opposed to a noise removal technique.

B. Gain normalization in FDLP

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.,

$$r[t] = s[t] * h[t], \tag{5}$$

where $s[t]$, $h[t]$, and $r[t]$ denote the original speech signal, the room impulse response, and the reverberant speech, respectively.

Let $s[t]$ be 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 $E_{sn}[t]$ and an instantaneous phase function $p_{sn}[t]$ (Mourjopoulos and Hammond, 1983) such that

$$s[t] = \sum_{n=1}^{N} s_n[t] = \sum_{n=1}^{N} E_{sn}[t]\cos(p_{sn}[t]). \tag{6}$$

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

$$\begin{aligned} r[t] &= \sum_{n=1}^{N} r_n[t] = \sum_{n=1}^{N} E_{rn}[t]\cos(p_{rn}[t]) \\ &\simeq \sum_{n=1}^{N} h_n[t] * s_n[t] \\ &= \sum_{n=1}^{N} E_{hn}[t]\cos(p_{hn}[t]) * E_{sn}[t]\cos(p_{sn}[t]), \end{aligned} \tag{7}$$

where E_{rn}, E_{sn}, and E_{hn} represent the envelope functions of the band passed 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 typical room impulse responses, it has been shown in Mourjopoulos and Hammond (1983) that the envelope functions are related by

$$E_{rn} \simeq \frac{1}{2} E_{hn} * E_{sn}. \tag{8}$$

If E_{rn} represents the Hilbert envelope of the nth sub-band, Eq. (8) 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.

The Hilbert envelope and the spectral autocorrelation function form Fourier transform pairs. Thus, the Hilbert envelope convolution model in Eq. (8) 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 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 (Thomas *et al.*, 2008). In our experiments, gain normalization is implemented by setting the prediction gain [gain of inverse linear prediction (LP) filter] to unity.

C. Robustness in telephone channel noise

When speech signal is passed through a telephone channel, the output signal can be modeled as a combination of back-ground additive noise and a convolutive noise in the channel

$$x[t] = s[t] * h[t] + n[t]. \tag{9}$$

In such conditions, the combination of noise compensation and gain normalization provides suppression of additive and convolutive distortions. This is illustrated in Fig. 5, where we plot the log-FDLP envelopes for clean and telephone speech in two conditions: (a) without gain normalization and noise compensation and (b) with gain normalization and noise compensation. This figure shows that the application of these techniques reduce the mismatch between the FDLP envelopes extracted from clean and noisy speech. Hence, these techniques provide significant robustness to features derived from FDLP envelopes.

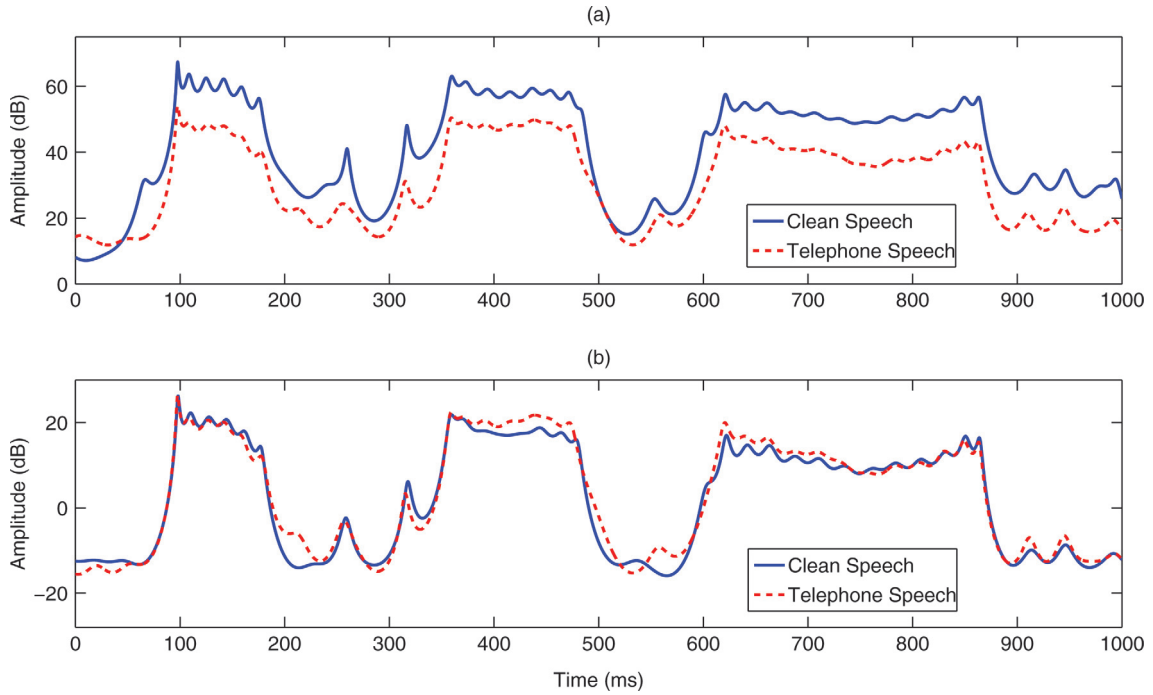

FIG. 5. (Color online) Log-FDLP envelopes for a sub-band of clean speech and telephone speech. (a) Without gain normalization and noise compensation, and (b) with gain normalization and noise compensation.

III. FEATURE EXTRACTION

The block schematic for the proposed feature extraction technique is shown in Fig. 6. Long segments of the speech signal (full speech utterances which are typically 2–3 s long for TIMIT sentences) are decomposed into frequency sub-bands by windowing the DCT. In our experiments, we use a critical band decomposition (Hermansky and Fousek, 2005). FDLP is applied on the sub-band DCT components to derive sub-band temporal envelopes of speech. The whole set of sub-band temporal envelopes forms a two-dimensional (time–frequency) representation of the input signal energy.

The sub-band temporal envelopes are then compressed using a static compression scheme which is a logarithmic function and a dynamic compression scheme (Dau *et al.*, 1996). The use of the logarithm is to model the overall non-linear compression in the auditory system which covers the huge dynamical range between the hearing threshold and the uncomfortable loudness level. The adaptive compression is realized by an adaptation circuit consisting of five consecutive nonlinear adaptation loops (Dau *et al.*, 1996). This is shown in Fig. 7. Each of these loops consists of a divider and a low-pass filter with time constants ranging from 5 to 500 ms. The input signal is divided by the output signal of the low-pass filter in each adaptation loop. Sudden transitions in the sub-band envelope that are very fast compared to the time constants of the adaptation loops are amplified linearly at the output due to the slow changes in the low-pass filter output, whereas the slowly changing regions of the input signal are compressed. The dynamic compression stage is followed by a low-pass filter with a cut-off frequency of 8 Hz (Tchorz and Kollmeier, 1999).

The static and dynamic compression of FDLP envelopes is illustrated in Fig. 8. Here we plot (a) a portion of sub-band temporal envelope derived using FDLP, (b) logarithmic compression of temporal envelope (static compression scheme), and (c) adaptive compression of the temporal envelope (using the adaptive compression loops).

The architecture of the conventional speech recognizer is typically set for speech features sampled at 100 Hz (i.e., one feature vector every 10 ms). For using our speech representation in a conventional recognizer, the compressed temporal envelopes are divided into 200 ms segments with a shift of 10 ms. DCT of both the static and the dynamic segments of temporal envelope yields the static and the dynamic modulation spectrum, respectively.

We use 14 modulation frequency components from each cosine transform, yielding modulation spectrum in the

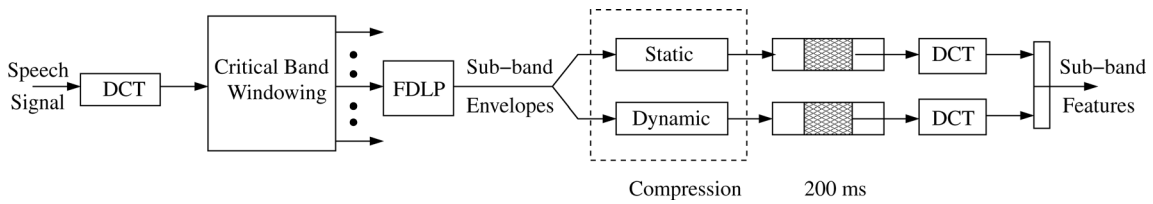

FIG. 6. Block schematic for the sub-band feature extraction. The steps involved are critical band decomposition, estimation of sub-band envelopes using FDLP, static and adaptive compression, and conversion to modulation frequency components by the application of cosine transform.

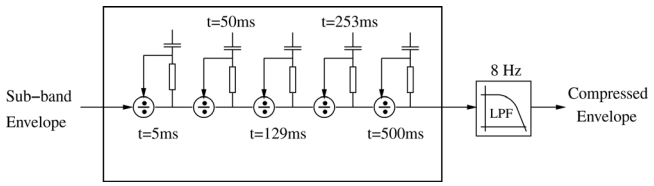

FIG. 7. Dynamic compression scheme using adaptive compression loops.

0–35 Hz region with a resolution of 2.5 Hz. This choice of modulation frequencies is found from a set of cross-validation experiments on the TIMIT database. In these cross-validation experiments, the upper cut-off frequency in the modulation spectrum is varied from 20 to 45 Hz in steps of 5 Hz and the frequency resolution is varied from 1.5 to 3 Hz. The clean cross-validation set of TIMIT database is used for phoneme recognition. In these experiments, the best phoneme recognition accuracy is obtained for an upper cut-off frequency of 35 Hz and a resolution of 2.5 Hz in the modulation spectrum. In all the subsequent experiments, we use this choice of parameters.

IV. EXPERIMENTS

A. Phoneme recognition task

The phoneme recognition system is based on the HMM-ANN paradigm (Bourlard and Morgan, 1994). The multi-layer perceptron (MLP) estimates the posterior probability of phonemes given the acoustic evidence $P(q_t = i|x_t)$, where q_t denotes the phoneme index at frame t, x_t denotes the feature vector taken with a window of certain frames. The relation between the posterior probability $P(q_t = i|x_t)$ and the likelihood $P(x_t|qt = i)$ is given by the Bayes rule,

$$\frac{p(x_t|q_t = i)}{p(x_t)} = \frac{P(q_t = i|x_t)}{p(q_t = i)}. \tag{10}$$

It is shown in Bourlard and Morgan (1994) that the neural network with sufficient capacity and trained on enough data estimates the true Bayesian *a-posteriori* probability. The scaled likelihood in an HMM state is given by Eq. (10), where we assume equal prior probability $P(q_t = i)$ for each phoneme $i = 1,2,\ldots,39$. The state transition matrix is fixed with equal probabilities for self and next state transitions. Viterbi algorithm is applied to decode the phoneme sequence.

A three layered MLP is used to estimate the phoneme posterior probabilities. The network is trained using the standard back propagation algorithm with cross entropy error criteria. The learning rate and stopping criterion are controlled by the frame classification rate on the cross-validation data. In the TIMIT phoneme recognition system, the MLP consists of 1000 hidden neurons, and 39 output neurons (with soft max nonlinearity) representing the phoneme classes. The performance of phoneme recognition is measured in terms of phoneme accuracy. In the decoding step, all phonemes are considered equally probable (i.e., there is no language model deployed). The optimal phoneme insertion penalty that gives maximum phoneme accuracy on the cross-validation data (which is a sub-set of the database excluding the train and the test set) is used for the test data. The partition of the database into train, test, and cross-validation data is described below.

B. TIMIT database

Experiments are performed on TIMIT database down-sampled to 8 kHz. In the TIMIT database, there are two "sa" dialect sentences spoken by all speakers in the corpus. The

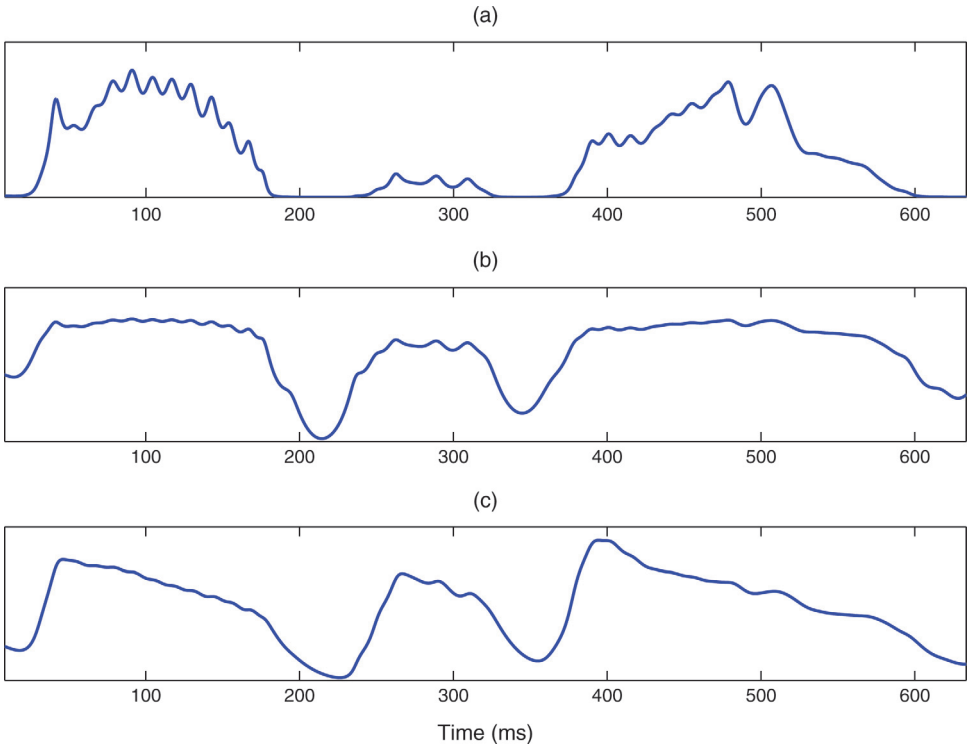

FIG. 8. (Color online) Static and dynamic compression of the FDLP envelopes. (a) A portion of sub-band FDLP envelope, (b) logarithmic compression of the FDLP envelope, and (c) adaptive compression of the FDLP envelope.

TABLE I. Recognition accuracies (%) of individual phonemes for different feature extraction techniques on clean speech, speech with additive noise (average performance of four noise types at 0, 5, 10, 15, and 20 dB SNRs), reverberant speech (average performance for nine room impulse response functions), and telephone speech (average performance for nine channel conditions). The best performance for each condition is indicated in bold.

			Clean speech				
PLP	RASTA	MRASTA	LDMN	LTLSS	MVA	ETSI	FDLP
65.4	61.2	62.8	64.8	64.8	61.9	64.0	62.1
			Speech with additive noise				
PLP	RASTA	MRASTA	LDMN	LTLSS	MVA	ETSI	FDLP
28.2	29.4	30.2	36.0	32.5	36.4	41.6	**43.9**
			Reverberant speech				
PLP	RASTA	MRASTA	LDMN	LTLSS	MVA	ETSI	FDLP
20.3	22.7	22.1	30.0	29.4	29.4	22.7	**33.6**
			Telephone speech				
PLP	RASTA	MRASTA	LDMN	LTLSS	MVA	ETSI	FDLP
34.3	45.4	48.0	50.1	37.3	49.9	47.7	**55.5**

use of these "sa" sentences in training leads to the learning of certain phoneme contexts. This may result in artificially high recognition scores (Lee, 1989) and bias the context independent phoneme recognition experiments. In order to avoid any such unfair bias for certain phonemes in certain contexts, we remove the "sa" dialect sentences from the training and test data (Lee, 1989). The remaining training data consists of 3000 utterances from 375 speakers, cross-validation data set consists of 696 utterances from 87 speakers and the test data set consists of 1344 utterances from 168 speakers. The TIMIT database, which is hand-labeled using 61 labels, is mapped to the standard set of 39 phonemes (Pinto *et al.*, 2008). We do not apply any speaker based normalization on the input features.

The robustness of the proposed features is tested on three versions of the test data corresponding to distortions introduced by additive noise, convolutive noise, and telephone channel. In the case of additive noise conditions, a noisy version of the test data is created by adding various types of noise at different SNRs [similar to Aurora 2 database (Pearce and Hirsch, 2000)]. The noise types chosen are the "Restaurant", "Babble," "Subway," and "Exhibition Hall" obtained from Hirsch and Finster (2005). These noises are added at SNRs 0, 5, 10, 15, and 20 dB using the FaNT tool (Hirsch, 2001). The generation of the noisy version of the test data is done using the set-up described in Gelbart (2008). Thus, there are four real noise types and five SNR yielding 20 versions of the test data each with 1344 utterances.

For phoneme recognition experiments with reverberant speech, the clean TIMIT test data are convolved with a set of nine different room responses collected from various sources (Gelbart and Morgan, 2001; Dhillon, 2002) with spectral coloration (defined as the ratio of the geometric mean to the arithmetic mean of the spectral magnitudes) ranging from −2.42 to −0.57 dB and reverberation time (T60) ranging from 100 to 500 ms. The use of nine different room responses results in nine reverberant test sets consisting of 1344 utterances each. For phoneme recognition experiments in telephone channel, speech data collected from nine telephone sets in the handset TIMIT (HTIMIT) database (Reynolds, 1997) are used. For each of these telephone channels, 842 test utterances, also having clean recordings in the TIMIT test set, are used.

In all the experiments, the system is trained only on the training set of TIMIT database, representing clean speech without the distortions introduced by the additive or convolutive noise but tested on the clean TIMIT test set as well as the noisy versions of the test set in additive, reverberant, and telephone channel conditions (mismatched train and test conditions).

C. Results

The baseline experiments use perceptual linear prediction (PLP) features with a context of nine frames (Morgan *et al.*, 1992; Pinto *et al.*, 2008). The results for the proposed technique are also compared with those obtained for several other robust feature extraction techniques namely:

(1) Modulation spectrum based features—Relative spectra (RASTA) (Hermansky and Morgan, 1994) features with nine frame context and multi-resolution RASTA (MRASTA) (Hermansky and Fousek, 2005),
(2) Features proposed for robustness in additive noise—advanced-ETSI (noise-robust) distributed speech recognition front-end (ETSI, 2002) and mean-variance auto regressive moving average (MVA) processing (Chen and Bilmes, 2007) with nine frame context (MVA),
(3) Robust features for reverberant speech recognition—long-term log spectral subtraction (LTLSS) (Gelbart and Morgan, 2002) and log-DFT mean normalization (LDMN) (Avendano and Hermansky, 1997) with nine frame context.

These techniques are chosen as baseline features as they are commonly deployed in ASR and phoneme recognition systems. For the proposed FDLP based modulation frequency features, we use 15 critical bands in the 300–4000 Hz with an equal band-width (in the bark frequency scale) of approximately 1 bark. Table I shows the average phoneme recognition performance for the various feature extraction techniques on clean speech, speech with additive noise, reverberant speech, and telephone channel speech. In

TABLE II. Phoneme recognition accuracies (%) for different feature extraction techniques for four noise types ("Restaurant," "Babble," "Subway," and "Exhibition Hall") at 0, 5, 10, 15, and 20 dB SNRs. The best performance for each condition is indicated in bold.

SNR (dB)	PLP	GFCC	MRASTA	LDMN	LTLSS	MVA	ETSI	FDLP
				Restaurant noise				
0	13.2	12.5	7.8	19.8	14.4	18.8	**23.2**	23.0
5	18.1	24.2	17.4	25.8	21.1	26.2	31.2	**32.0**
10	25.7	36.6	28.5	33.6	30.1	35.0	40.5	**43.4**
15	35.1	46.0	39.1	41.9	40.8	43.6	48.3	**52.0**
20	45.4	51.9	47.6	49.2	51.9	50.4	54.3	**58.1**
				Babble noise				
0	12.2	10.5	6.0	18.8	13.9	16.1	20.8	**22.4**
5	16.3	21.9	15.2	24.2	19.6	25.1	29.5	**31.3**
10	23.4	34.7	26.5	31.8	28.2	34.4	39.0	**43.2**
15	32.7	45.6	37.6	40.8	39.2	43.1	47.9	**53.0**
20	43.8	52.2	47.5	49.2	51.3	50.3	54.6	**58.7**
				Subway noise				
0	16.6	18.2	19.9	28.1	20.3	27.5	32.6	**34.5**
5	23.0	31.3	30.3	35.3	27.4	35.4	41.3	**42.6**
10	31.0	42.6	38.4	42.2	35.8	42.5	48.5	**50.6**
15	39.6	49.5	45.3	48.8	43.7	47.9	54.3	**56.2**
20	48.3	53.6	50.8	54.7	51.1	52.5	58.6	**59.9**
				Exhibition hall noise				
0	14.7	9.2	8.6	20.9	17.3	20.5	24.4	**25.4**
5	19.7	21.1	18.9	27.1	23.2	28.0	33.1	**34.7**
10	26.6	34.1	29.7	34.5	31.0	36.3	42.0	**45.0**
15	34.8	45.1	39.9	43.0	40.3	43.9	50.3	**53.5**
20	44.0	52.0	48.5	50.6	50.3	50.4	55.5	**58.7**

clean conditions, the baseline PLP feature extraction technique provides the best performance. However, the performance of the PLP based phoneme recognition system degrades significantly in all the mismatched conditions. In the case of additive noise, the ETSI features give good robustness among the short-term spectral features. For phoneme recognition in reverberant speech and telephone speech, LDMN and MVA features provide good performance among the short-term spectral features.

In all the mismatched conditions, the FDLP features provide significant robustness compared to other feature extraction techniques. On the average, the relative performance improvement over the other feature extraction techniques is about 4% for speech in additive noise, 5% for reverberant speech, and about 11% for telephone speech.

The phoneme recognition performance on the individual noise types ("Restaurant," "Babble," "Subway," and "Exhibition Hall") and SNR conditions (0–20 dB) is shown in Table II. Since the RASTA technique was mainly proposed for robustness in convolutive distortions, we replace the RASTA features with the gammatone frequency cepstral coefficients (GFCC) (Shao *et al.*, 2009) for additive noise experiments reported in this table. These features are auditory model based and the cepstral coefficients are derived directly from sub-band energies (instead of log energies). The features are 29 dimensional and are appended with first order derivatives (Shao *et al.*, 2009). We also apply a nine frame context yielding GFCC features of dimension 522.

In the experiments reported in Table II, the ETSI technique (ETSI, 2002) provides the best baseline performance in all noise conditions. For almost all noise types and SNR conditions, the proposed FDLP features provide good improvements over the best baseline features.

D. Phoneme recognition in CTS

The CTS database consists of 300 h of conversational speech recorded over a telephone channel at 8 kHz (Hain *et al.*, 2005). The training data consist of 250 h of speech from 4538 speakers, cross-validation data set consists of 40 h of speech from 726 speakers and the test data set consists of 10 h from 182 speakers. The CTS data are labeled using 45 phonemes. The phoneme labels are obtained by force aligning the word transcriptions to the previously trained hidden Markov model-Gaussian mixture model (HMM-GMM) models (Hain *et al.*, 2005).

We use the phoneme recognition system based on HMM-ANN system (described in Sec. IV A). For the CTS experiments, the MLP consists of 8 270 hidden neurons, and 45 output neurons (with soft max nonlinearity) representing the phoneme classes. Table III reports the results for the phoneme recognition experiments on CTS database. The proposed modulation features result in improved phoneme

TABLE III. Phoneme recognition accuracies (%) for different feature extraction techniques on CTS database. The best performance is indicated in bold.

PLP	RASTA	MRASTA	ETSI	FDLP
52.3	52.8	52.2	54.0	**56.6**

TABLE IV. Various modifications to the proposed feature extraction and their meanings.

Name	Meaning
V1	Short-term critical band energies
V2	Hilbert envelopes without FDLP
V3	Without gain normalization and noise compensation
V4	Only gain normalization
V5	Only noise compensation
V6	Only static compression
V7	Only adaptive compression
Prop.	Proposed technique using static and adaptive compression of gain normalized and noise compensated FDLP envelopes

recognition rate compared to other feature extraction techniques (a relative improvement of 6%).

V. RELATIVE CONTRIBUTION OF VARIOUS PROCESSING STAGES

The previous section showed that the proposed feature extraction provides promising improvements in various types of distortions. In this section, we analyze the contribution of the various processing stages of the proposed feature extraction technique for robust phoneme recognition. This is done by a set of phoneme recognition experiments on the TIMIT database with various modifications of the proposed technique. As before, the system is trained only on clean TIMIT training data, while the test data consists of clean speech, one condition of additive noise (Babble noise at 10 dB SNR), reverberant speech from one room response (with a reverberation time of 300 ms), and telephone channel speech from one set in HTIMIT database.

A. Modifications

The main processing stages in the proposed technique are the FDLP processing, gain normalization, and noise compensation and the use of two-stage compression scheme. Here, we modify these processing stages in various ways to determine their relative importance in robust phoneme recognition. The various modifications (V1–V7) with their meanings are listed in Table IV.

In the first modification (V1), the envelope estimation is done using with trajectories of short-term critical band energies instead of the FDLP processing. This is similar to the representation of speech used in MRASTA (Hermansky and Fousek, 2005). Speech signal in short analysis windows (of length 25 ms) is transformed into spectral domain and the spectral content in individual critical band is integrated. The remaining processing stages described in Sec. III are applied on these critical band energies.

In the second modification (V2), all steps described in Sec. III are performed except for the linear prediction step. This would mean that the features are derived from sub-band Hilbert envelopes directly without the use of FDLP.

In modification V3, we implement the FDLP technique without gain normalization and noise compensation. Modification V4 implements our previous work (Ganapathy *et al.*, 2009) with the gain normalization procedure (Thomas *et al.*, 2008). In V4, we omit the step of noise compensation and for V5 we omit the gain normalization step in the proposed feature extraction method. These modifications are intended to analyze the contribution of these steps in realizing robust representations of speech corrupted with additive and convolutive distortions.

In modifications V5 and V6, we analyze the use of two-stage compression mechanism. This is done by using only one type of compression (either static V5 or dynamic V6) in the proposed feature extraction technique.

TABLE V. Phoneme recognition accuracies (%) for various modifications to the proposed feature extraction in clean speech, with one condition of additive noise (Babble noise at 10 dB SNR), reverberant speech (with a reverberation time of 300 ms), and one condition of telephone channel speech. The phoneme recognition results without any modification to the proposed technique are shown at the bottom.

Feature extraction	Clean	Additive noise	Reverberant speech	Telephone channel speech
V1	56.9	38.2	37.9	50.8
V2	60.9	41.5	36.5	52.7
V3	66.5	28.6	28.3	43.0
V4	65.0	33.9	31.9	51.4
V5	62.7	38.7	30.8	46.6
V6	61.1	40.7	34.0	51.6
V7	59.0	38.0	34.2	49.7
Prop.	62.1	43.2	36.9	55.5

B. Results

The phoneme recognition accuracies obtained for the various modifications are reported in Table V. The last row of the table shows the result for the proposed feature extraction technique without any modification (Sec. III). The comparison of V1 with V2 shows that the Hilbert envelopes form an improved representation compared to short-term critical band energy trajectories. The modification V2 improves over V1 in clean and noisy conditions. The improvement in performance for the proposed feature extraction over V2 shows that the application of FDLP for deriving AR models of Hilbert envelopes improves the overall performance in clean and noisy conditions.

The performance of V3 forms the baseline for the proposed noise compensation technique. Although, V3 provides good performance in clean conditions, its performance degrades considerably in all noise conditions. The noise compensation technique provides good robustness in additive noise conditions (V5). When this is applied along with the gain normalization procedure, the resulting features (Prop.) improve significantly on all types of distortions. The application of these techniques results in a drop in performance for clean speech. The gain of the sub-band envelope can be a useful cue for phoneme recognition of clean speech (as indicated by a moderate drop in performance in clean conditions for V3 and V4). Furthermore, noise compensation technique tends to deemphasize the valleys of the envelope trajectory [Fig. 4(b)]. As the valleys of the envelope contain information in discriminating certain phoneme classes (like nasals), there is a reduction in the recognition accuracy in

clean conditions (comparison of V3 and V5). However, the improvements obtained for all types of mismatched conditions justify the employment of these normalization techniques in the proposed features.

The proposed approach (Prop.) also improves over V4 [which forms a combination of our past approaches (Ganapathy *et al.*, 2009; Thomas *et al.*, 2008)]. The improvement is consistent on all types of noise conditions with substantial improvements on additive noise. This improvement is attributed to the noise compensation procedure. Application of log compression or adaptive compression alone is worser than the joint application of these two compression schemes (V6–V7). Although this was reported in Ganapathy *et al.* (2009) for clean conditions, we find here that the joint application of static and dynamic compression schemes improved the performance in noisy conditions as well. The static compression scheme provides good robustness for fricatives and nasals (which is due to modeling property of the signal peaks in static compression), whereas the dynamic compression scheme provides good robustness for plosives and affricates (where the fine temporal fluctuations like onsets and offsets carry the important phoneme classification information). Hence, the joint application of these feature streams results in considerable improvement in performance for most of the phonetic classes.

VI. SUMMARY

We have proposed a robust feature extraction technique based on modulation spectrum of speech derived from normalized sub-band temporal envelopes. The main findings in this work can be summarized as follows:

(1) The application of linear prediction in frequency domain forms an efficient method for deriving sub-band modulations.
(2) The two-stage compression scheme of deriving static and dynamic modulation spectrum results in good phoneme recognition for all phoneme classes even in the presence of noise.
(3) The noise compensation technique provides a way to derive robust representation of speech in almost all types of noise and SNR conditions.
(4) The robustness of the proposed features is further enhanced by the application of gain normalization technique.
(5) These envelope normalization techniques provide substantial improvements in noisy conditions over the previous work (Ganapathy *et al.*, 2009).

ACKNOWLEDGMENTS

The authors would like to thank the Medical Physics group at the Carl von Ossietzky-Universitat Oldenburg for code fragments implementing adaptive compression loops, Joel Pinto, Petr Motlicek, and Fabio Valente for helpful discussions and phoneme recognition code segments. Furthermore, the authors would also like to thank Marios Athineos and Dan Ellis for PLP and FDLP feature extraction codes.

Athineos, M., and Ellis, D. P. W. (**2007**). "Autoregressive modelling of temporal envelopes," IEEE Trans. Signal Process. **55**(11), 5237–5245.
Athineos, M., Hermansky, H., and Ellis, D. P. W. (**2004**). "LP-TRAPS: Linear predictive temporal patterns," in *Proceedings of Interspeech*, pp. 1154–1157.
Avendano, C., and Hermansky, H. (**1997**). "On the effects of short-term spectrum smoothing in channel normalization," IEEE Trans. Speech Audio Process. **5**(4), 372–374.
Bourlard, H., and Morgan, N. (**1994**). *Connectionist Speech Recognition—A Hybrid Approach* (Kluwer Academic Publishers, Boston), pp. 59–115.
Chen, C., and Bilmes, J. A. (**2007**). "MVA Processing of Speech Features," IEEE Trans. Audio, Speech, Lang. Process. **15**(1), 257–270.
Dau, T., Püschel, D., and Kohlrausch, A. (**1996**). "A quantitative model of the 'effective' signal processing in the auditory system: I. Model structure," J. Acoust. Soc. Am. **99**(6), 3615–3622.
Dhillon, R., Bhagat, S., Carvey, H., and Shriberg, E. (**2002**) "The ICSI Meeting Recorder Project," http://www.icsi.berkeley.edu/Speech/mr (Last viewed August 18, 2009).
Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**(2), 1053–1064.
ETSI (**2002**). *ETSI ES 202 050 v1.1.1 STQ*; Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. http://www.etsi.org/deliver/etsi_es/202000_202099/202050/01.01.05_60/es_202050v010105p.pdf
Falk, T. H., Stadler, S., Kleijn, W. B., and Chan, W. Y. (**2007**). "Noise suppression based on extending a speech-dominated modulation band," in *Proceedings of Interspeech*, pp. 970–973.
Ganapathy, S., Thomas, S., and Hermansky, H. (**2009**). "Modulation frequency features for phoneme recognition in noisy speech," J. Acoust. Soc. Am., Express Lett. **125**(1), EL8–EL12.
Gelbart, D. (**2008**). "Ensemble feature selection for multi-stream automatic speech recognition," Ph.D. thesis, University of California, Berkeley.
Gelbart, D., and Morgan, N. (**2001**). "Evaluating long-term spectral subtraction for reverberant ASR," in *Proceedings of IEEE Automatic Speech Recognition and Understanding*, pp. 190–193.
Gelbart, D., and Morgan, N. (**2002**). "Double the trouble: Handling noise and reverberation in far-field automatic speech recognition," in *Proceedings of Interspeech*, pp. 2185–2188.
Hain, T., Burget, L., Dines, J., McCowan, I., Karafiat, M., Lincoln, M., Moore D., Garau, G., Wan, V., Ordelman, R., and Renals, S. (**2005**). "The development of AMI system for transcription of speech in meetings," in *Proceedings of Machine Learning for Multimodal Interaction*, pp. 344–356.
Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**(4), 1738–1752.
Hermansky, H., and Fousek, P. (**2005**). "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Proceedings of Interspeech*, pp. 361–364.
Hermansky, H., and Morgan, N. (**1994**). "RASTA processing of speech," IEEE Trans. Speech Audio Process. **2**, 578–589.
Hermansky, H., and Sharma, S. (**1998**). "TRAPS—Classifiers of Temporal Patterns," in *Proceedings of Interspeech*, pp. 1817–1820.
Hirsch, H. G. (**2001**) "FaNT: Filtering and noise adding tool," http://dnt.kr.hsnr.de/download.html (Last viewed September 18, 2009).
Hirsch, H. G., and Finster, H. (**2005**). "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proceedings of Interspeech*, pp. 2697–3000.
Houtgast, T., Steeneken, H. J. M., and Plomp, R. (**1980**). "Predicting speech intelligibility in rooms from the modulation transfer function, I. General room acoustics," Acoustica **46**, 60–72.
Kingsbury, B. E. D., Morgan, N., and Greenberg, S. (**1998**). "Robust speech recognition using the modulation spectrogram," Speech Commun. **25**(1–3), 117–132.
Kumerasan, R., and Rao, A. (**1999**). "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," J. Acoust. Soc. Am. **105**(3), 1912–1924.
Lee, K. F. (**1989**). "Speaker independent phone recognition using hidden Markov models," IEEE Trans. Acoust., Speech, Signal Process. **37**(11), 1641–1648.
Makhoul, J. (**1975**). "'Linear prediction: A tutorial review," Proc. IEEE **63**(4), 561–580.
Marple, L. S. (**1999**). "Computing the discrete-time analytic signal via FFT," IEEE Trans. Signal Process. **47**(9), 2600–2603.
Morgan, N., Hermansky, H., Bourlard, H., Kohn, P., and Wooters, C. (**1992**). "Continuous speech recognition using PLP analysis with multilayer perceptrons," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 49–52.

Mourjopoulos, J., and Hammond, J. K. (**1983**). "Modelling and enhancement of reverberant speech using an envelope convolution method," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1144–1147.

Pearce, D., and Hirsch, H. G. (**2000**). "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA Tutorial and Research Workshop ASR2000*, pp. 29–32.

Pinto, J., Yegnanarayana, B., Hermansky, H., and Doss, M. M. (**2008**). "Exploiting contextual information for improved phoneme recognition," in *Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing*, pp. 4449–4452.

Reynolds, D. A. (**1997**). "HTIMIT and LLHDB: Speech corpora for the study of hand set transducer effects," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1535–1538.

Riesz, R. R. (**1928**). "Differential sensitivity of the ear for pure tones," Phys. Rev. **31**, 867–875.

Schwarz, P. (**2008**). "Phoneme recognition based on long temporal context," Ph.D. thesis, BUT, Brno, CZ.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**(5234), 303–304.

Shao, Y., Jin, Z., Wang, D. L., and Srinivasan, S. (**2009**). "An auditory-based feature for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4625–4628.

Tchorz, J., and Kollmeier, B. (**1999**). "A model of auditory perception as front end for automatic speech recognition," J. Acoust. Soc. Am. **106**(4), 2040–2050.

Thomas, S., Ganapathy, S., and Hermansky, H. (**2008**). "Recognition of reverberant speech using frequency domain linear prediction," IEEE Signal Process. Lett. **15**, 681–684.

Vinton, M. S., and Atlas, L. E. (**2001**). "Scalable and progressive audio codec," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3277–3280.

Xu, L., and Zheng, Y. (**2007**). "Spectral and temporal cues for phoneme recognition in noise," J. Acoust. Soc. Am. **122**(3), 1758–1764.