Temporal Envelope Subtraction for Robust Speech Recognition Using Modulation Spectrum

Sriram Ganapathy [#1], Samuel Thomas [#2] and Hynek Hermansky [#,*3]

[#] *Department of Electrical and Computer Engineering,*
[*] *Human Language Technology Center of Excellence*
Johns Hopkins University, USA
{[1] ganapathy, [2] samuel, [3] hynek}@jhu.edu

***Abstract*—In this paper, we present a new noise compensation technique for modulation frequency features derived from syllable length segments of subband temporal envelopes. The subband temporal envelopes are estimated using frequency domain linear prediction (FDLP). We propose a technique for noise compensation in FDLP where an estimate of the noise envelope is subtracted from the noisy speech envelope. The noise compensated FDLP envelopes are compressed with static (logarithmic) and dynamic (adaptive loops) compression and are transformed into modulation spectral features. Experiments are performed on a phoneme recognition task as well as a connected digit recognition task where the test data is corrupted with variety of noise types at different signal to noise ratios. In these experiments with mismatched train and test conditions, the proposed features provide considerable improvements compared to other state of the art noise robust feature extraction techniques (average relative improvement of 25 % and 35 % over the baseline PLP features for phoneme and word recognition tasks respectively).**

I. Introduction

The performance of a typical automatic speech recognition (ASR) system severely degrades when it encounters speech from noisy environments. Such performance degradation is mainly caused by mismatch in training and operating conditions. A survey of the main approaches that have been pursued in the direction of reducing this mismatch is reported in [1]. These approaches can be classified as noise robustness in features (for example [2]), enhancement of speech (for example [3], [4]) and acoustic model compensation (for example [5]). Although the problem of suppressing the uncorrelated additive noise has been widely studied in the past, single channel noisy speech recognition continues to be a challenging task.

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

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

where $x[m]$ is the discrete representation of the input signal, $s[m]$ represents the clean speech signal which is corrupted by noise $n[m]$. Assuming that the speech and noise are uncorrelated, we obtain

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

where $P_X(m, \omega_k)$, $P_S(m, \omega_k)$ and $P_N(m, \omega_k)$ 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 [6]. 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 (for example [7]).

Alternatively, features for speech recognition can be derived from trajectories of spectral energies in the individual frequency subbands (for example [8]). Spectral components of long term amplitude modulations in individual frequency subbands are called modulation spectra. The modulation spectral representations have been used in the past for predicting speech intelligibility in reverberant environments [9]. They are now widely applied in many engineering applications like audio coding [10], noise suppression [11], etc. Feature extraction techniques based on modulation spectrum have also been proposed for ASR (for example [12], [13]).

In our previous work [14], we have shown that a combination of static and dynamic modulation frequency features perform well for telephone channel speech recognition. Here, the input speech signal is decomposed into a number of critical bands. In each subband, long term envelopes are extracted using frequency domain linear prediction (FDLP) [15], [16]. FDLP envelopes are compressed using a static and a dynamic compression. The static compression stage is a logarithmic operation and dynamic compression stage uses adaptive compression loops [17]. The compressed envelopes are transformed into modulation spectral components which are used as features for a phoneme recognition system.

In this paper, we propose a noise compensation technique for these modulation frequency features based on temporal envelope subtraction. In each subband, 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. Then, the noise compensated FDLP envelopes are used to derive static and dynamic modulation frequency features. These features are used for a phoneme recognition task using the hybrid hidden Markov model - artificial neural network (HMM-ANN) phoneme recognition system [18] as well as a connected digit recognition task using the Tandem system [19]. The test data for these tasks consists of speech corrupted with variety of

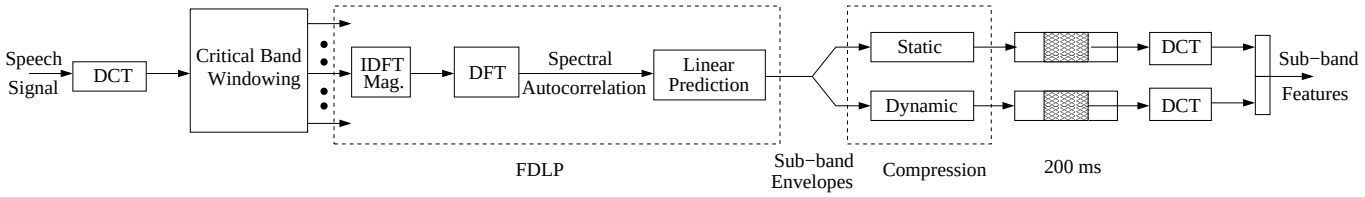

Fig. 2. Block schematic for the modulation spectrum based feature extraction technique.

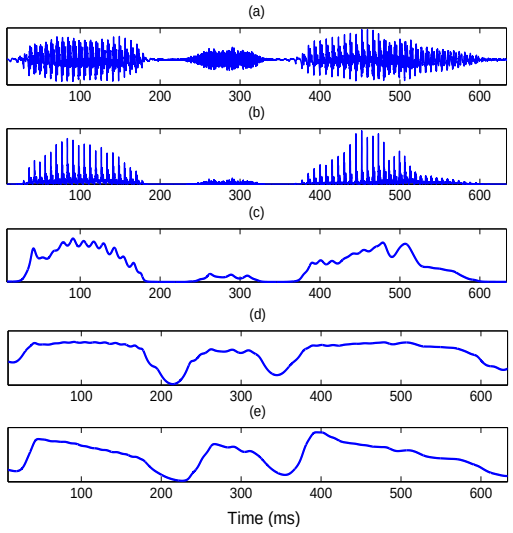

Fig. 1. Static and dynamic compression of the temporal envelopes: (a) a portion of speech signal, (b) the temporal envelope extracted using the Hilbert transform [20], (c) the FDLP envelope, which is an all pole approximation to (b) estimated using FDLP, (d) static compression of the FDLP envelope and (e) dynamic compression of the FDLP envelope.

real world noises at different signal to noise ratios (SNR). In these experiments, the proposed noise compensation technique provides considerable improvements in phoneme/word recognition accuracies over other robust feature extraction techniques.

The rest of the paper is organized as follows. In Sec. II, we describe the FDLP technique for the estimation of the temporal envelopes using linear prediction in spectral domain. The extraction of modulation frequency features from the temporal envelopes is explained in Sec. III. The proposed noise compensation technique is described in Sec. IV. Experiments performed with these modulation frequency features for phoneme and word recognition tasks are reported in Sec. V. In Sec. VI, we conclude with a discussion of the proposed features.

II. Frequency Domain Linear Prediction

The Hilbert envelope, which is the magnitude of the analytic signal, represents the instantaneous energy of a signal in the time domain. Hilbert envelopes are typically computed using the Hilbert transform operator in the time domain or by exploiting the causality of discrete Fourier transforms (DFT) [20]. Alternatively, a parametric model of the Hilbert envelopes can be extracted using linear prediction in frequency domain [15], [16].

FDLP is an efficient technique for auto regressive (AR) modelling of temporal envelopes of a signal. Typically, autoregressive (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) [21]). This 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. For the FDLP technique, the 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 [21]).

Fig. 1 shows the AR modelling property of FDLP. It shows (a) a portion of speech signal, (b) its Hilbert envelope computed using the Fourier transform technique [20] and (c) an all pole approximation for the Hilbert Envelope using FDLP.

III. Feature Extraction

The block schematic for the modulation spectrum based feature extraction technique [14] is shown in Fig. 2. Long segments of the speech signal (1000 ms) are decomposed into frequency subbands by windowing the discrete cosine transform (DCT). In our experiments, we use a critical band decomposition. 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 in the DCT domain. For deriving the spectral autocorrelations (defined as the Fourier transform of temporal envelope [16]), the subband DCT signal is converted back to the time domain using inverse discrete Fourier transform (IDFT). The IDFT is performed on the entire subband DCT for 1000 ms signal. The IDFT length corresponds to original signal length (8000). The IDFT phase containing the subband carrier signal is ignored as the FDLP operation is independent of this component. The magnitude IDFT component represents a non parametric Hilbert envelope [15]. The Hilbert envelope is transformed using DFT into spectral autocorrelations of the subband signal, which are used for linear prediction. The order of the linear prediction corresponds to 1 pole per 10 signal samples. It can be mathematically shown that the application of linear prediction

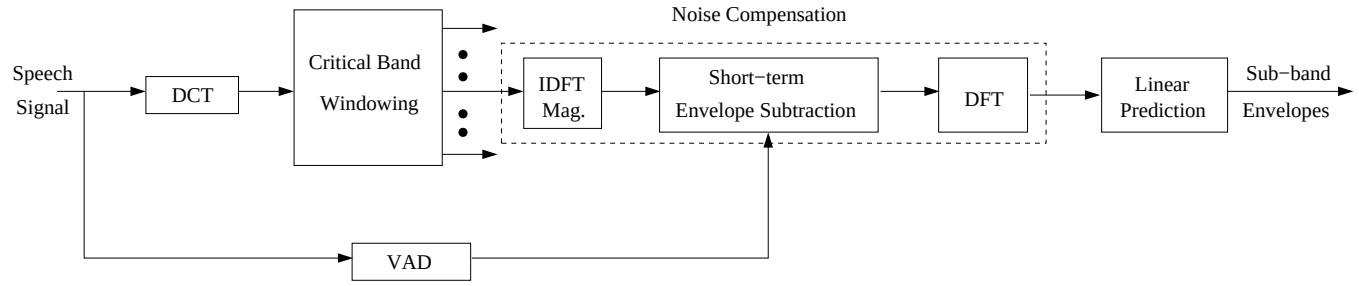

Fig. 4. Noise compensation in frequency domain linear prediction.

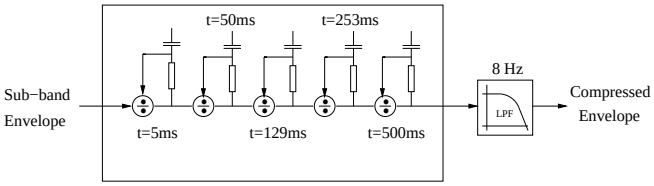

Fig. 3. Dynamic compression of the subband FDLP envelopes using adaptive compression loops [17].

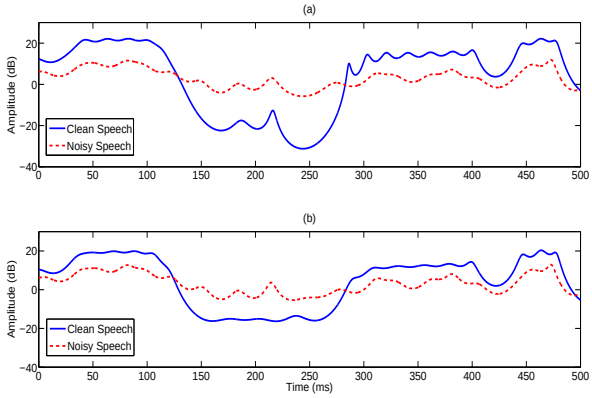

Fig. 5. Log FDLP envelopes for the fifth critical band of clean speech and speech corrupted with babble noise at 10 dB SNR. (a) without noise compensation (b) with noise compensation.

in DCT domain approximates the temporal envelope of the signal [16]. The steps involved in converting the subband DCT signal into envelope AR model parameters are referred to as FDLP. In our experiments, we use the gain normalized FDLP envelopes as these are found to be more robust to channel noise [22]. The whole set of subband temporal envelopes forms a two dimensional (time-frequency) representation of the input signal energy.

The subband temporal envelopes are then compressed using a static compression which is a logarithmic function and a dynamic compression scheme [17]. The dynamic compression is realized by an adaptation circuit consisting of five consecutive nonlinear adaptation loops as shown in Fig. 3. Each of these loops consists of a divider and a lowpass filter with time constants ranging from 5 ms to 1000 ms. The input signal is divided by the output signal of the lowpass filter in each adaptation loop. Sudden transitions in the subband envelope that are fast compared to the time constants of the adaptation loops are amplified linearly at the output, whereas the slowly changing regions of the input signal are suppressed. In this way, changes in the input signal like onsets and offsets are emphasized in the dynamic compression stage. This is also illustrated in Fig. 1, where we show the static (Fig. 1.(d)) and dynamic compression (Fig. 1.(e)) of the FDLP envelopes. The dynamic compression stage is followed by a low pass filter [17].

Since speech recognition system requires speech features sampled at 100 Hz (i.e one feature vector every 10 ms), the compressed temporal envelopes are divided into 200 ms segments with a shift of 10 ms. The temporal envelopes from the two compression streams are then converted into modulation spectral components using DCT, corresponding to the static and the dynamic modulation spectrum. We use 14 modulation frequency components from each of these streams, yielding modulation spectrum in the $0 - 35$ Hz range with a resolution of 2.5 Hz. This choice of modulation frequencies is obtained using phoneme recognition experiments on the cross validation data in TIMIT database [14].

IV. Temporal Envelope Subtraction

When speech signal is corrupted by noise, the FDLP envelopes are modified in such a way that their dynamic range is reduced. This is illustrated in Fig. 5.(a), where we plot the subband FDLP envelopes in clean and noisy conditions for the same speech utterance. The effect of noise is more pronounced in the valleys of the subband envelopes, where the mismatch between clean and noisy speech is significant. When modulation frequency features are derived from the uncompensated FDLP envelopes, the performance of the ASR system degrades significantly in noisy conditions.

The proposed noise compensation technique for FDLP is shown in Fig. 4. A voice activity detector (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 [7]. 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 before (Sec. III), long segments of the input speech signal are transformed to DCT domain where a critical

TABLE I

PHONEME RECOGNITION ACCURACIES (%) FOR DIFFERENT FEATURE EXTRACTION TECHNIQUES ON CLEAN TIMIT TEST DATA AS WELL AS THE AVERAGE PERFORMANCE FOR THE FOUR NOISE TYPES - "RESTAURANT", "BABBLE", "SUBWAY" AND "EXHIBITION HALL" WITH SNRS 0,5,10,15 AND 20 DB .

SNR (dB)	PLP-9	PLP-SS-9	MVA-9	ETSI-9	FDLP	FDLP-NC
clean	66.8	60.7	63.8	65.7	67.6	65.4
0	14.6	11.0	24.3	28.2	24.6	30.1
5	20.6	25.1	32.8	37.3	33.3	39.9
10	28.9	39.1	41.2	46.5	42.7	50.0
15	38.7	49.6	48.3	53.8	52.0	57.9
20	48.9	56.1	53.7	58.9	58.6	62.4
Avg.	30.3	36.2	40.1	44.9	42.2	48.1

band sized windowing is applied. The subband Hilbert envelopes are obtained as the magnitude IDFT of the DCT signal. We apply short term envelope subtraction on these subband Hilbert envelopes for noise compensation. This is achieved in two steps. In the first step, we window the Hilbert envelopes 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 envelope from these segments.

Since the noise component is assumed to be additive in signal domain (Eq. 1), we can write

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

where $X[k]$, $S[k]$ and $P[k]$ are the k^{th} DCT coefficient of noisy speech, clean speech and noise respectively. By virtue of the orthogonality property of the DCT matrix, the speech and noise continue to be uncorrelated in the DCT domain. Further, the application of magnitude DFT gives

$$E_X(m, b_i) = E_S(m, b_i) + E_N(m, b_i), \tag{4}$$

where $E_X(m, b_i)$, $E_S(m, b_i)$ and $E_N(m, b_i)$ are the short term non parametric Hilbert envelopes of the noisy speech, clean speech and noise respectively for the subband b_i. Eq. 4 shows that the effect of noise can be alleviated if an estimate of $E_N(m, b_i)$ is subtracted from the short term noisy speech envelope $E_X(m, b_i)$.

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 conventional spectral subtraction technique [4]. The noise compensated short term envelopes are synthesized using overlap-add to obtain the long term subband envelopes. These are converted back to subband DCT domain and used for FDLP. Static and dynamic modulation frequency features are derived from the noise compensated FDLP envelopes as described in Sec. III.

Fig. 5.(b) provides an illustration of the effect of this noise compensation technique on the subband 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. This is due to the fact when the compensated value reduces below zero, we employ the corresponding magnitude value. Although this method of compensation slightly reduces the information in valleys of clean speech signal (as illustrated by the drop in recognition performance in clean conditions), it significantly 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.

V. EXPERIMENTS AND RESULTS

A. Phoneme Recognition Task

The phoneme recognition system is based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [18]. 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 $P(x_t|q_t = i)$ is given by the Bayes rule,

$$\frac{p(x_t|q_t = i)}{p(x_t)} = \frac{P(q_t = i|x_t)}{P(q_t = i)}. \tag{5}$$

It is shown in [18] that the neural network with sufficient capacity and trained on enough data estimates the true Bayesian aposteriori probability. The scaled likelihood in an HMM state is given by Eq. 5, where we assume equal prior probability $P(q_t = i)$ for each phoneme $i = 1, 2...39$. 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 our 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 (no language model). The optimal phoneme insertion penalty that gives maximum phoneme accuracy on the cross validation data is used for the test data.

TABLE II

WORD RECOGNITION ACCURACIES (%) FOR DIFFERENT FEATURE EXTRACTION TECHNIQUES ON CLEAN OGI TEST DATA AS WELL AS THE AVERAGE PERFORMANCE FOR THE FOUR NOISE TYPES - "RESTAURANT", "BABBLE", "SUBWAY" AND "EXHIBITION HALL" WITH SNRS 0,5,10,15 AND 20 DB .

SNR (dB)	PLP-D-A	PLP-9	PLP-SS-9	MVA-9	ETSI-9	FDLP	FDLP-NC
clean	95.9	96.4	94.0	95.7	96.5	96.5	95.7
0	25.3	24.0	37.8	47.5	43.4	16.0	44.4
5	47.7	47.0	59.2	67.5	66.3	40.0	69.5
10	67.0	70.3	74.8	80.7	81.3	69.0	83.6
15	78.9	84.4	83.7	88.4	89.5	87.1	90.8
20	86.5	91.4	88.9	92.5	93.3	94.0	94.1
Avg.	61.1	63.4	68.9	75.3	74.8	61.2	76.5

Experiments are performed on TIMIT database containing speech sampled at 16 kHz. The 'sa' dialect sentences are excluded in the experiments. The 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 [23]. We do not apply any speaker based normalization on the input features.

For testing the robustness of the proposed features, a noisy version of the test data is created by adding various types of noise at different SNRs (similar to Aurora 2 database [24]). The noise types chosen are the "Restaurant", "Babble", "Subway" and "Exhibition Hall" obtained from [25]. These noises are added at various SNRs using the FaNT tool [26]. The generation of the noisy version of the test data is done using the setup described in [27].

In all the experiments, the system is trained only on the original TIMIT data, representing clean speech without the distortions introduced by the additive noise but tested on the clean TIMIT test set as well as the noisy test set (mismatched train and test conditions). The results for the proposed noise compensation technique are compared with those obtained for several other robust feature extraction techniques namely PLP features with a 9 frame context [23], Advanced ETSI (noise robust) distributed speech recognition front end [7] with a 9 frame context, Mean Variance ARMA processing [28] applied on PLP features (MVA) with a 9 frame context and spectral subtraction, proposed in [29], applied on PLP features (PLP-SS) with a 9 frame context. Among these features, the Advanced ETSI front end forms the standard feature extraction for speech recognition in noise [7]. For the modulation frequency features, we use 19 critical bands in the $300-8000$ Hz range. The FDLP based proposed modulation frequency features are tested without and with the noise compensation which are denoted as FDLP and FDLP-NC respectively.

Table I summarizes the results for the phoneme recognition experiments in TIMIT database with clean test set as well as the average performance for the four noise types with SNRs in the 0-20 dB range. Spectral subtraction [29], which is a speech enhancement technique, improves the performance of the baseline PLP features for all the noise conditions except at 0 dB. MVA processing [28], which is feature normalization method, results in good improvements over the PLP-9 features in all SNR conditions without much degradation in clean conditions. Advanced ETSI front end [7] provides the best performance among the various short term spectral features considered here.

In the case of the modulation frequency features, the application of the proposed noise compensation technique provides good robustness in all SNR conditions. For all noise types and SNR conditions, the proposed FDLP-NC features provide an average relative improvement of about 25 % over the baseline PLP features and about 6 % over the ETSI feature extraction technique.

B. Connected Digit Recognition Task

Experiments are performed with small vocabulary continuous digit recognition task (OGI-Digits database). The vocabulary consists of eleven ($0-9$ digits and "Oh") digits in 28 different pronunciations. We use the Tandem system which is based on HMM-ANN framework [19]. Features extracted from speech for every 10 ms are used to train an MLP with 1800 hidden nodes. The MLP estimates posterior probabilities of 29 English phonemes [30]. The training data consists of the whole Stories database plus the training part of the Numbers95 database. Around 10 % of the data is used for cross validation. Log and Karhunen Loeve (KL) transforms are applied on these features. This is done in order to convert the phoneme posterior probabilities into features appropriate for a conventional HMM-GMM recognition system [19]. The HMM based recognizer, trained on the training part of the OGI-Digits database, is used for classification.

The test data is corrupted with additive noise as explained in Sec. V-A. Since the Numbers data was collected over telephone channels, we applied the MIRS filter from ITU Software Tools Library [31] to the noises before adding them to Numbers data (similar to the generation of noisy Numbers data [27]). The HMM-ANN models are trained on clean condition but tested on clean as well noisy versions of the test set.

Table II summarizes the results for the connected digit recognition task using the various extraction techniques described in Sec. V-A. We also report the performance with 39 dimensional PLP features (PLP-D-A) on the HMM-GMM system (without the use of TANDEM setup). It can be seen that the HMM-ANN framework using Tandem setup generally results in increased robustness compared to the conventional

HMM-GMM system using the 39 dimensional PLP features. This validates the claims made in [32] regarding the improvements in additive noise for discriminative classifiers.

All the other feature extraction techniques are used along with the Tandem setup. Among the short term spectral features, the MVA processing provides the best robustness performance especially in the low SNR conditions. For the modulation frequency features, we use 15 critical bands in $300-3400$ Hz range. Without much degradation in clean conditions, the proposed noise compensation technique (FDLP-NC) provides an average relative improvement of 35 % over the baseline PLP-9 features and about 5 % over the MVA features. The application of the proposed noise compensation (FDLP-NC) gives significant robustness compared to the uncompensated FDLP features.

VI. Conclusions

We have proposed a noise compensation technique for modulation frequency features based on temporal envelope subtraction. Subband temporal envelopes, estimated using FDLP, are processed by both a static and a dynamic compression and are converted to modulation frequency features. For noise compensation, an estimate of the temporal envelope of noise is subtracted from the noisy speech envelope. Although the proposed technique involves simple operation of envelope subtraction in time domain (similar to the conventional spectral subtraction technique), these features provide considerable improvements over the other noise robust feature extraction techniques for phoneme and word recognition tasks in various noise and SNR conditions. In future, we wish to experiment with real world noisy speech and standard ASR systems that include model adaptation schemes.

Acknowledgment

The authors would like to thank the Medical Physics group at the Carl von Ossietzky-Universitat Oldenburg for code fragments implementing adaptive compression loops, Sivaram Garimella for the code fragments implementing the TANDEM system, spectral subtraction technique and David Gelbart for the code setup involved in the generation of noisy speech data.

References

[1] Y. Gong, "Speech Recognition in Noisy Environment", *Speech Communication*, Vol. 16 (3), Apr. 1995, pp. 261-291.

[2] B.H. Juang, L.R. Rabiner and J.G. Wilpon, "On the use of bandpass lifting in speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 35 (7), Jul. 1987, pp. 947-954.

[3] Y. Ephraim and D. Malah, "Speech enhencement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 32 (6), Dec. 1984, pp. 1109-1112.

[4] S.F. Boll,"Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 27 (2), Apr. 1979, pp. 113-120.

[5] M.J.F. Gales and S. Young, "Robust continuous speech recognition using parallel model combination", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 4 (5), Sep. 1996, pp. 352-359.

[6] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 28 (4), Aug. 1980, pp. 357-366.

[7] "ETSI ES 202 050 v1.1.1 STQ; Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", 2002.

[8] M. Athineos, H. Hermansky and D. P. W. Ellis, "LP-TRAPS: Linear predictive temporal patterns", *Proc. of INTERSPEECH*, Sept. 2004, pp. 1154-1157.

[9] T. Houtgast, H.J.M. Steeneken and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function, I. General room acoustics", *Acoustica* , Vol. 46 (1), Sept. 1980, pp. 60-72.

[10] M.S. Vinton and L.E. Atlas, "Scalable and progressive audio codec", *Proc. of ICASSP*, May 2001, pp. 3277-3280.

[11] T.H. Falk, S. Stadler, W.B. Kleijn and W.Y. Chan,"Noise Suppression Based on Extending a Speech-Dominated Modulation Band", *Proc. of INTERSPEECH*, Aug. 2007, pp. 970-973.

[12] H. Hermansky and S. Sharma, "TRAPS - Classifiers of Temporal Patterns", *Proc. of ICSLP*, Dec. 1998, pp. 1003-1006.

[13] B.E.D. Kingsbury, N. Morgan and S. Greenberg, "Robust speech recognition using the modulation spectrogram", *Speech Communication*, Vol. 25(1-3), Aug. 1998, pp. 117-132.

[14] S. Ganapathy, S. Thomas and H. Hermansky, "Modulation frequency features for phoneme recognition in noisy speech", *JASA Express Letters*, Vol. 125 (1), Jan. 2009, pp. EL8-EL12.

[15] 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.

[16] M. Athineos and D.P.W. Ellis, "Autoregressive modelling of temporal envelopes", *IEEE Trans. Speech and Audio Process.*, Vol. 55 (11), Nov. 2007, pp. 5237-5245.

[17] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition", *J. Acoust. Soc. Am.*, Vol. 106 (4), Oct. 1999, pp. 2040-2050.

[18] H. Boulard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[19] H. Hermansky, D.P.W. Ellis and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems", *Proc. of ICASSP*, Apr. 2000, pp. 1635-1638.

[20] L.S. Marple, "Computing the Discrete-Time Analytic Signal via FFT", *IEEE Trans. Signal Process.*, Vol. 47 (9), Sept. 1999, pp. 2600-2603.

[21] J. Makhoul, "Linear Prediction: A Tutorial Review",*IEEE Proceed.*, Vol 63 (4), Apr. 1975, pp. 561-580.

[22] S. Thomas, S. Ganapathy and H. Hermansky, "Hilbert Envelope Based Specto-Temporal Features for Phoneme Recognition in Telephone Speech", *Proc. of INTERSPEECH*, Sept. 2008.

[23] J. Pinto, B. Yegnanarayana, H. Hermansky and M.M. Doss, "Exploiting Contextual Information for Improved Phoneme Recognition", *Proc. of INTERSPEECH*, Aug. 2007, pp. 1817-1820.

[24] H.G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", *Automatic Speech Recognition: Challenges for the New Millenium (ASR2000)*, France, 2000.

[25] H.G. Hirsch and H. Finster, "The Simulation of Realistic Acoustic Input Scenarios for Speech Recognition Systems", *Proc. of INTERSPEECH*, Sept. 2005, pp. 2697-3000.

[26] H.G. Hirsch, "FaNT: Filtering and Noise Adding Tool", *http://dnt.kr.hsnr.de/download.html*.

[27] D. Gelbart, "Ensemble Feature Selection for Multi-Stream Automatic Speech Recognition", *Ph. D. Thesis*, University of California, Berkeley, 2008.

[28] C. Chen and J.A. Bilmes, "MVA Processing of Speech Features", *IEEE Trans. Audio, Speech and Lang. Process.*, Vol. 15 (1), Jan. 2007, pp. 257-270.

[29] S. Rangachari and P.C. Loizou, "A noise-estimation algorithm for highly non-stationary environments", *Proc. Speech Communication*, Vol.48 (2), Feb. 2006, pp. 220-231.

[30] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR", *Proc. of INTERSPEECH*, Sept. 2005, pp. 361-364.

[31] S.F.D.C. Neto, "The ITU-T Software Tool Library", *International Journal of Speech Technology*, Vol. 2 (4), May 1999, pp. 259-272.

[32] K.K. Paliwal, "Neural Net Classifiers for Robust Speech Recognition under Noisy Environments", *Proc. of ICASSP*, Apr. 1990, pp. 429-432.