SPECTRO-TEMPORAL FEATURES FOR AUTOMATIC SPEECH RECOGNITION USING LINEAR PREDICTION IN SPECTRAL DOMAIN

Samuel Thomas, Sriram Ganapathy and Hynek Hermansky

IDIAP Research Institute, Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{tsamuel,ganapathy,hynek}@idiap.ch

ABSTRACT

Frequency Domain Linear Prediction (FDLP) provides an efficient way to represent temporal envelopes of a signal using auto-regressive models. For the input speech signal, we use FDLP to estimate temporal trajectories of sub-band energy by applying linear prediction on the cosine transform of sub-band signals. The sub-band FDLP envelopes are used to extract spectral and temporal features for speech recognition. The spectral features are derived by integrating the temporal envelopes in short-term frames and the temporal features are formed by converting these envelopes into modulation frequency components. These features are then combined in the phoneme posterior level and used as the input features for a hybrid HMM-ANN based phoneme recognizer. The proposed spectro-temporal features provide a phoneme recognition accuracy of 69.1% (an improvement of 4.8% over the Perceptual Linear Prediction (PLP) base-line) for the TIMIT database.

1. INTRODUCTION

Traditionally, acoustic features for Automatic Speech Recognition (ASR) systems are extracted by applying Bark or Mel scale integrators on power spectral estimates in short analysis windows ($10-30$ ms) of the speech signal. Typical examples of such features are the Mel Frequency Cepstral Coefficients (MFCC) [1] and Perceptual Linear Prediction (PLP) [2]. Most of the information contained in these acoustic features relate to formants which provide important cues for recognition of basic speech units. The signal dynamics are represented by a sequence of short-term feature vectors with each vector forming a sample of the underlying process. Further, additional information about the dynamics of the underlying speech signal is incorporated with these feature vectors using the derivative features. But, the problems of time-frequency resolution and efficient sampling of the short-term representation are addressed in an ad-hoc manner.

It has been shown that important information for speech perception lies in the $1-16$ Hz range of the modulation frequencies [3]. In order to exploit the information at these modulation frequencies, speech signals of relatively long temporal segments need to analyzed. An explicit incorporation of the information about the speech dynamics have been proposed for feature extraction [4, 5, 6]. Here, the long temporal trajectories (typically 1000ms) of spectral energy in critical bands are used for feature extraction.

Recently, the technique of linear prediction (LP) in spectral domain (originally proposed for temporal noise shaping in audio coding [7]) was used for ASR feature extraction [8], where a representation of the temporal envelope in different frequency sub-bands is obtained by using the dual of the conventional linear prediction in time domain. In FDLP, the poles of the auto regressive (AR) model represent the temporal peaks rather than the spectral peaks. By using analysis windows of the order of hundreds of milliseconds, the technique automatically decides the distribution of the poles to best model the temporal envelope. Further, the model has some important advantages:

- Fine time-dependent resolution provides information about transient events in time like stop bursts.
- Long-term summarization of power in spectral bands presents the ability to capture complete description of the linguistic units lasting more than 10 ms.

In this paper, we propose to exploit the above mentioned properties of FDLP by extracting spectro-temporal features for ASR. Specifically, the FDLP envelopes are used to obtain spectral and temporal features which are combined to form a joint spectro-temporal feature set. These features are input to a hybrid Hidden Markov Model - Artificial Neural Network (HMM-ANN) phoneme recognition system [9]. The HMM-ANN system has a Multi-Layer Perceptron (MLP) for estimating the phoneme posterior probabilities and a Viterbi decoder for finding the best phoneme sequence.

The rest of the paper is organized as follows. In Sec. 2, we describe the FDLP technique which approximates temporal envelopes of a signal using linear prediction in spectral domain. The extraction of spectro-temporal features from the temporal envelopes is given in Sec. 3. Experiments with the proposed features for a phoneme recognition task in TIMIT database is reported in Sec. 4 along with a comparison of the other feature extraction techniques in the literature. In Sec. 5, we conclude with a discussion of the proposed features.

2. MODELLING SUB-BAND TEMPORAL ENVELOPES USING FDLP

The Hilbert envelope, which is the squared magnitude of the analytic signal, represents the instantaneous energy of a signal in the time domain. Hilbert envelopes are typically computed either by using the Hilbert transform operator in the time domain or by exploiting the causality of Discrete Fourier Transforms (DFT) [10]. However, in order to use the

This work is partially supported by the European IST Programme Project FP6-0027787 and the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM2); managed by the IDIAP Research Institute on behalf of the Swiss Federal Authorities. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

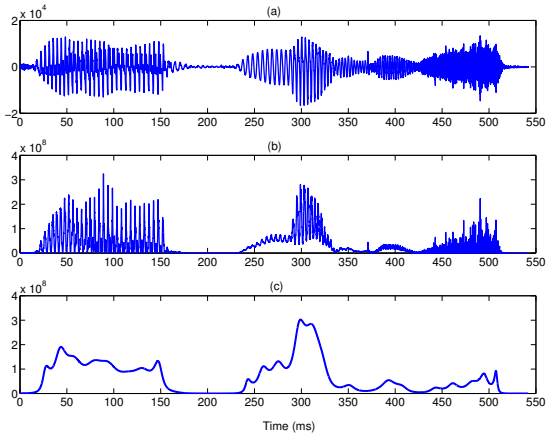

Figure 1: *Illustration of the all-pole modelling property of FDLP. (a) a portion of the speech signal, (b) its Hilbert envelope (c) all pole model obtained using FDLP*

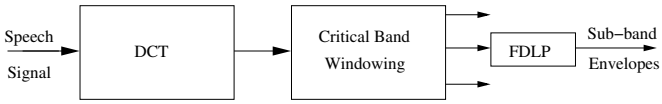

Figure 2: *Deriving sub-band temporal envelopes from speech signal using FDLP*

Hilbert envelope as a dual of the power spectrum for speech recognition tasks, we require a parametric model. FDLP is an efficient technique for auto regressive (AR) modelling of temporal envelopes of a signal [7]. It represents a dual technique to the conventional Time Domain Linear Prediction (TDLP). In the case of TDLP, the AR model approximates the power spectrum of the input signal, whereas FDLP fits an all pole model to the Hilbert envelope (squared magnitude of the analytic signal). In our case, the FDLP technique is implemented in two parts - first, the discrete cosine transform (DCT) is applied on long segments of speech to obtain a real valued spectral representation of the signal. Then, linear prediction is performed on the DCT coefficients to obtain a parametric model of the temporal envelope. Fig. 1 shows an illustration of the AR modelling property of FDLP. It shows (a) a portion of speech signal of 500 ms duration, (b) its Hilbert envelope computed using the Fourier transform technique [10] and (c) an all pole approximation (of order 50) for the Hilbert Envelope using FDLP.

For ASR tasks, the speech signal in long segments (hundreds of milliseconds) is decomposed into frequency subbands by windowing the DCT. Using FDLP, an all-pole minimum phase estimate of the temporal dynamics of each subband signal is obtained. The block schematic for the extraction of sub-band temporal envelopes from speech signal is shown in Fig. 2. The whole set of sub-band temporal envelopes forms a two dimensional (time-frequency) representation of the input signal energy.

3. SPECTRO-TEMPORAL FEATURES USING FDLP

The sub-band temporal envelopes obtained from the FDLP are used to derive spectral and temporal features. These features are then combined to obtain joint spectro-temporal features which are used for posterior based speech recognition system. The joint spectro-temporal features adaptively capture fine temporal nuances with high temporal resolution while at the same time summarize the spectral evolution in time scales of hundreds of milliseconds.

3.1 Deriving short-term spectral features from sub-band temporal envelopes

The conventional feature extraction methods obtain short-term spectral features by integrating the estimate of power spectrum of the signal in sub-bands (example PLP [2]). Similar to the representation of energy in the spectral domain in the form of power spectrum, the distribution of energy in the time domain is expressed in the form of Hilbert envelope. Since integration of signal energy is identical in time and frequency domain, the Hilbert envelope can equivalently be utilized for obtaining the sub-band energy based short term spectral features. The sub-band temporal envelopes are obtained using the FDLP technique applied on relatively long temporal segments (1000 ms) of the input signal. These sub-band envelopes are integrated in short term frames (of the order of 25 ms with a shift of 10 ms). These short term sub-band energies are converted to short-term cepstral features similar to the PLP feature extraction technique [2]. As the features are derived from short temporal segments, they capture spectral details in the order of 10 ms.

3.2 Deriving long-term temporal features from sub-band temporal envelopes

The long-term sub-band envelopes from the FDLP form a compact representation of the temporal dynamics over long regions of the speech signal. We use cepstral recursion to convert our all-pole models of the temporal trajectories into modulation spectral components [8]. Since the speech recognizer requires features at a frame rate of 10 ms, the modulation frequency components for the current frame in each sub-band are obtained along with contextual information of neighboring frames (in a manner similar to the TRAP-TANDEM setup [5]). The recognition performance depends on the number of neighboring frames forming the context for the current frame (varied from $10-40$). The temporal features for each sub-band are stacked together and fed to the posterior probability estimator.

3.3 Combining Spectro-Temporal Features

The posterior probability estimator is an ANN trained with features from the training data to estimate phoneme posterior probabilities [5]. The number of parameters to be trained in the Multi Layer Perceptron (MLP) are dependent on the input feature dimension and are limited by the amount of the training data. Since, the dimension of the temporal features is high, we use two posterior estimators for the temporal features corresponding to the even and odd bands respectively. Spectral features with a context of 9 frames are used in another posterior probability estimator and the output of these three neural net classifiers are combined using the Dempster Shafer (DS) theory of evidence [11].

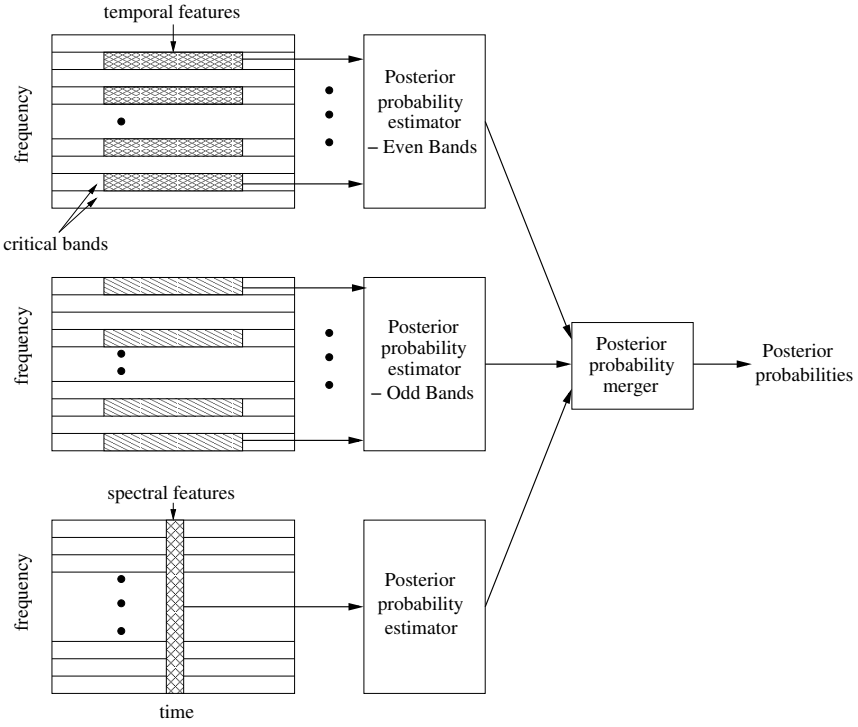

Figure 3: *Schematic of the joint spectro-temporal features for posterior based ASR*

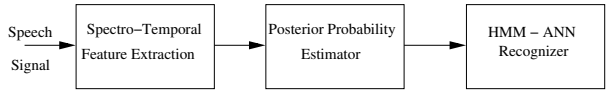

Figure 4: *HMM - ANN speech recognizer*

Fig. 3 shows the schematic of the proposed combination of spectral and temporal features. These phoneme posterior probabilities are used in a phoneme posterior based speech recognition system.

4. EXPERIMENTS AND RESULTS

The phoneme recognition system is based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [9] shown in Fig. 4. The 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,

$$\frac{p(x_t|q_t = i)}{p(x_t)} = \frac{P(q_t = i|x_t)}{P(q_t = i)}. \quad (1)$$

It is shown in [9] 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. 1, 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. Experiments were performed on TIMIT database, excluding 'sa' dialect sentences. All speech files are sampled at 16 kHz. 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 [12].

As explained in Sec. 3, the speech signal is processed to extract spectro-temporal features for every frame. These features are mean/variance normalized (across the training data set) to obtain feature vectors for every 10 ms of speech. 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.

For deriving the temporal envelopes from the speech signal, the current frame of 10 ms is appended with a number of neighboring frames in a manner similar to the TRAP-TANDEM setup [5]. The FDLP model order is fixed at an average rate of 100 poles per second for each sub-band. The sub-band decomposition for the spectral feature extraction is done on a Mel scale and 13 cepstral features are derived

Table 1: Phoneme Recognition Accuracies (%) for FDLP based Spectro-Temporal features.

# Contextual Frames	Acc.
10	68.3
20	69.1
30	69.0
40	68.4

Table 2: Phoneme Recognition Accuracies (%) for different feature extraction techniques.

	PLP 9 frame	LPTRAP	MRASTA	FDLP
Acc.	67.6	61.9	64.4	69.1

along with their first and second derivatives (similar to 39 dimensional PLP features). For deriving the temporal features, we use a critical band decomposition and obtain 21 modulation frequency components for each sub-band. These features are appended along with their first frequency derivatives (similar to M-RASTA features [6]) to obtain 42 dimensional temporal features for each sub-band. Due to the limitations in the amount of training data, we train separate MLPs for even and odd bands. The length of contextual information is varied and phoneme recognition is performed with spectro-temporal features. Table 1 summarizes the results for the experiments with FDLP based spectro temporal features.

In the base-line experiments, PLP features with a 9 frame context [12], LP-TRAP features [8] and M-RASTA features [6] are used for the phoneme recognition task with the same hybrid HMM-ANN system. These results are shown in Table 2. The best base-line phoneme recognition accuracy is 67.6% for the PLP 9 frame context.

The best phoneme recognition accuracies are obtained for spectro-temporal features derived using a context of 20 frames (i.e., with a FDLP frame length of 225 ms). The improvement over the PLP baseline is around 4.8% which is statistically significant.

5. CONCLUSIONS

We have proposed a novel method of extracting spectro-temporal features for ASR. For this purpose, temporal envelopes of critical band sized sub-bands are modelled using Frequency Domain Linear Prediction. The spectral features are derived by integrating the FDLP envelopes in short-term frames and the temporal features are obtained by converting the temporal envelopes into modulation frequency components. These features are combined in the phoneme posterior level and used as the input features to a hybrid HMM-ANN recognizer. The proposed features provide noticeable improvements over the PLP feature base-line for phoneme recognition tasks. The results are promising and encourage us to experiment on other tasks with different test and noisy conditions.

6. ACKNOWLEDGEMENTS

In addition to the acknowledgements earlier, the authors would like to thank Joel Pinto, Petr Motlicek and Fabio Valente for helpful discussions and code fragments. Furthermore, we would also like to thank Marios Athineos and Dan Ellis for PLP and FDLP feature extraction codes.

REFERENCES

[1] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", in *IEEE Trans. on Acoustics, Speech and Signal Proc.* Vol. 28, pp. 357-366, 1980.

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

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

[4] H. Hermansky and S. Sharma, "TRAPS - Classifiers of Temporal Patterns.", in *Proc. of ICSLP*, Sydney, Australia, Vol. 3, pp. 1003-1006, 1998.

[5] H. Hermansky, "TRAP-TANDEM: Data-driven Extraction of Temporal Features from Speech", in *Proc. of IEEE ASRU*, St. Thomas, US Virgin Islands, pp. 255-260, 2003.

[6] H. Hermansky and P. Fousek, "Multi-Resolution RASTA Filtering for TANDEM-Based ASR", in *Proc. of Interspeech*, Lisbon, Portugal, pp. 361-364, 2005.

[7] J. Herre and J.D Johnston, "Enhancing the Performance of Perceptual Audio Coders by using Temporal Noise Shaping (TNS)," in *Proc. of 101st AES Conv.*, Los Angeles, USA, pp. 1-24, 1996.

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

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

[10] L.S. Marple, "Computing the Discrete-Time Analytic Signal via FFT", in *IEEE Trans. on Acoustics, Speech and Signal Proc.*, Vol. 47, pp. 2600-2603, 1999.

[11] F. Valente and H. Hermansky, "Combination of Acoustic Classifiers based on Dempster-Shafer Theory of Evidence", in *Proc. of ICASSP*, Hawaii, U.S.A, pp. 1129-1132, 2007.

[12] J. Pinto, B. Yegnanarayana, H. Hermansky and M. M. Doss, "Exploiting Contextual Information for Improved Phoneme Recognition", in *Proc. of Interspeech*, Antwerp, Belgium, pp. 1817-1820, 2007.