INTERSPEECH 2009 BRIGHTON

Tandem Representations of Spectral Envelope and Modulation Frequency Features for ASR

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

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

10.21437/Interspeech.2009-748

Abstract

We present a feature extraction technique for automatic speech recognition that uses Tandem representation of short-term spectral envelope and modulation frequency features. These features, derived from sub-band temporal envelopes of speech estimated using frequency domain linear prediction, are combined at the phoneme posterior level. Tandem representations derived from these phoneme posteriors are used along with HMM based ASR systems for both small and large vocabulary continuous speech recognition (LVCSR) tasks. For a small vocabulary continuous digit task on the OGI Digits database, the proposed features reduce the word error rate (WER) by 13 % relative to other feature extraction techniques. We obtain a relative reduction of about 14 % in WER for an LVCSR task using the NIST RT05 evaluation data. For phoneme recognition tasks on the TIMIT database these features provide a relative improvement of 13% compared to other techniques.

Index Terms: Frequency Domain Linear Prediction (FDLP), Spectral Envelope Features, Modulation Frequency Features, Tandem based ASR systems.

1. Introduction

Feature extraction techniques for automatic speech recognition (ASR) are designed to suppress irrelevant redundancies contained in the speech signal while preserving relevant information about sound classes. These features should also be invariant across speakers, additive noise and channel distortions. In conventional feature extraction techniques like Mel Frequency Cepstral Coefficients (MFCC) [1] and Perceptual Linear Prediction (PLP) [2], acoustic features are derived by analyzing the spectrum of speech in short analysis windows (10-30 ms). Information about the dynamics of the underlying speech signal is added to these features by augmenting them with derivatives of the spectral trajectory at each instant. In more recent approaches for feature extraction [3, 4], long analysis windows (several hundred milliseconds) have been used to capture important acoustic information in the 1-16 Hz modulation frequency range [5].

Although these acoustic features represent the dynamics of the speech signal, further improvements can be achieved by reducing their dimensionality and enhancing their ability to discriminate between various sound classes [6]. For these features to be useful with statistical models like Hidden Markov Models (HMMs), they need to be transformed such that their distributions can be effectively modeled using a mixtures of Gaussians. Using the Tandem technique proposed in [7], acoustic features are trained discriminatively with a multi-layer perceptron (MLP) and transformed into features that are modeled by HMMs. It is observed that these features have less speaker variability and perform better than the conventional features for various ASR tasks [6, 8].

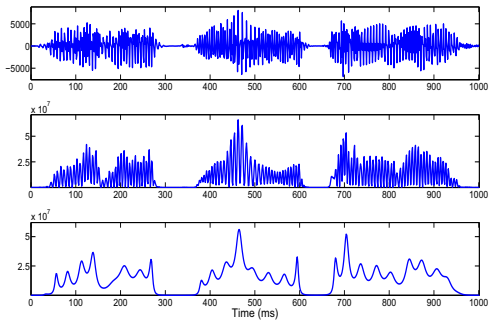

Figure 1: Illustration of the all-pole modeling property of FDLP. (a) a portion of the sub-band speech signal, (b) its Hilbert envelope (c) all pole model obtained using FDLP

In this paper, we extend our feature extraction technique [9] for Tandem processing in ASR. These features combine short-term spectral information along with long-term amplitude modulations. Unlike conventional feature extraction techniques, we analyze speech signals in frequency sub-bands over long temporal segments (several hundred milliseconds). We estimate temporal envelopes in frequency sub-bands using the dual of the conventional time domain linear prediction (TDLP). This is done by applying linear prediction on the cosine transform of sub-band signals [10]. In the same way as the TDLP fits an all pole model to the power spectrum of the signal, frequency domain linear prediction (FDLP) fits an all pole model to the Hilbert envelope which represents the instantaneous energy of the time domain signal [11]. Fig. 1 illustrates the AR modeling of FDLP. It shows (a) a portion of the sub-band speech signal, (b) its Hilbert envelope computed using the Fourier transform technique [12] and (c) an all pole approximation to the Hilbert Envelope using FDLP. These representations of the speech signal are able to capture fine temporal events associated with transient events like stop bursts while at the same time summarize the temporal evolution of the signal energy [10].

This work was partially supported by grants from European IST Programme DIRAC Project FP6-0027787; the Swiss National Center of Competence in Research (NCCR) on “Interactive Multi-modal Information Management (IM)2”

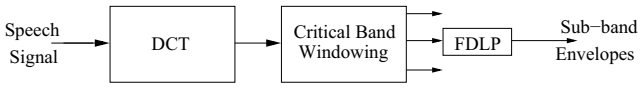

Figure 2: Deriving sub-band temporal envelopes from speech signal using FDLP

As described in [9], short-term spectral envelopes are derived from sub-band envelopes by integrating the sub-band envelopes in short analysis windows. Modulation frequency features are derived by compressing these envelopes using static and adaptive compression techniques. We apply the cosine transform in long analysis windows (200 ms) to yield modulation frequency components of speech. The spectral envelope and modulation frequency features are used to train MLPs and are combined at the phoneme posterior level. These phoneme posterior probabilities are used as input features for Tandem based ASR systems.

The rest of the paper is organized as follows. In Sec. 2, the FDLP technique for deriving sub-band envelopes is discussed. We describe the conversion of these sub-band envelopes into Tandem representations in Sec. 3. Experiments performed with the proposed features for a variety of ASR tasks are reported in Sec. 4. In Sec. 5, we conclude with a discussion of the proposed features.

2. Frequency Domain Linear Prediction

FDLP is an efficient technique for auto regressive (AR) modeling of temporal envelopes of a signal [10]. In this technique, we first apply the discrete cosine transform (DCT) on long segments of speech to obtain a real valued spectral representation of the signal. The DCT transform of the signal is decomposed using critical-band-sized windows. Linear prediction is performed on each sub-band DCT signal to obtain a parametric model of the its temporal envelope. The block schematic for extraction of sub-band temporal envelopes from speech signal is shown in Fig. 2.

3. Tandem Representations of Features from Sub-band Temporal Envelopes

We use the sub-band temporal envelopes estimated using FDLP to derive spectral envelope and modulation frequency features. Tandem representations of these features are used for ASR experiments. Fig. 3 shows the schematic of the proposed feature extraction technique.

3.1. Spectral envelope features

In conventional feature extraction techniques like PLP, short-term features are extracted by integrating the power spectral estimates on Mel or Bark scale [1, 2]. Since integration of signal energy is identical in time and frequency domain, sub-band Hilbert envelopes can equivalently be used for obtaining the short-term energy estimates in the time domain. In our technique, we derive short-term features from sub-band temporal envelopes, which are modelled using FDLP. This is done by integrating the envelopes in short term frames (of the order of 25 ms with a shift of 10 ms). These short term sub-band energies are converted into 13 cepstral features along with their first and second derivatives [9], similar to the PLP features [2]. Each frame of these spectral envelope features is used with a context of 9 frames for training an MLP network.

3.2. Modulation frequency features

In techniques like TRAPS [3] and MRASTA [4], modulation frequency features are derived by analyzing temporal trajectories of spectral energy estimates in individual sub-bands using long analysis windows. As described earlier, since FDLP estimates the temporal envelope in sub-bands, modulation features can be derived from these envelopes as well. We compress the sub-band temporal envelopes statically and dynamically. The envelopes are compressed statically using the logarithmic function. Dynamic compression of the envelopes is achieved using an adaptation circuit which consists of five consecutive non-linear adaptation loops proposed in [18]. These loops are designed so that sudden transitions in the sub-band envelope that are fast compared to the time constants of the adaptation loops are amplified linearly at the output, while the steady state regions of the input signal are compressed logarithmically. The compressed temporal envelopes are then transformed using the Discrete Cosine Transform (DCT) in long term windows (200 ms long, with a shift of 10 ms) We use 14 modulation frequency components from each cosine transform, yielding modulation spectrum in the $0 - 35$ Hz range with a resolution of 2.5 Hz [9]. The static and dynamic modulation frequency features of each critical band are stacked together and used to train an MLP network.

3.3. Tandem representations of features

We combine the spectral envelope and modulation frequency features at the phoneme posterior level using the Dempster Shafer (DS) theory of evidence [19]. These phoneme posteriors are first gaussianized by using the log function and then decorrelated using the Karhunen-Loeve Transform (KLT) [7]. This reduces the dimensionality of the feature vectors by retaining only the feature components which contribute most to the variance of the data. We use 25 dimensional features in our Tandem representations similar to [6].

4. Experiments

We perform a set of experiments using Tandem representations of the proposed spectral envelope and modulation frequency features along with other state-of-the-art features for ASR. These include a phoneme recognition task, a small vocabulary continuous digit recognition task and a large vocabulary continuous speech recognition (LVCSR) task. For each of these experiments, we train three layered MLPs to estimate phoneme posterior probabilities using these features. The proposed features are compared with three other feature extraction techniques - PLP features with a 9 frame context [14] which are similar to spectral envelope features derived using FDLP (FDLP-S), M-RASTA features [4] and Modulation SpectroGram (MSG) features [13] with a 9 frame context, which are both similar to modulation frequency features (FDLP-M).

We combine FDLP-S features with FDLP-M features using the DS theory of evidence to obtain a joint spectro-temporal feature set (FDLP-S+FDLP-M). Similarly, we derive two more feature sets by combining PLP features with M-RASTA features (PLP+M-RASTA) and MSG features (PLP+MSG). 25 dimensional Tandem representations of these features are used for our experiments. We also experiment with 39 dimensional PLP features without any Tandem processing (PLP-D).

Our first experiment is to validate the usefulness of Tandem representation of our features for a phoneme recognition task using HMMs. We perform experiments on the TIMIT database,

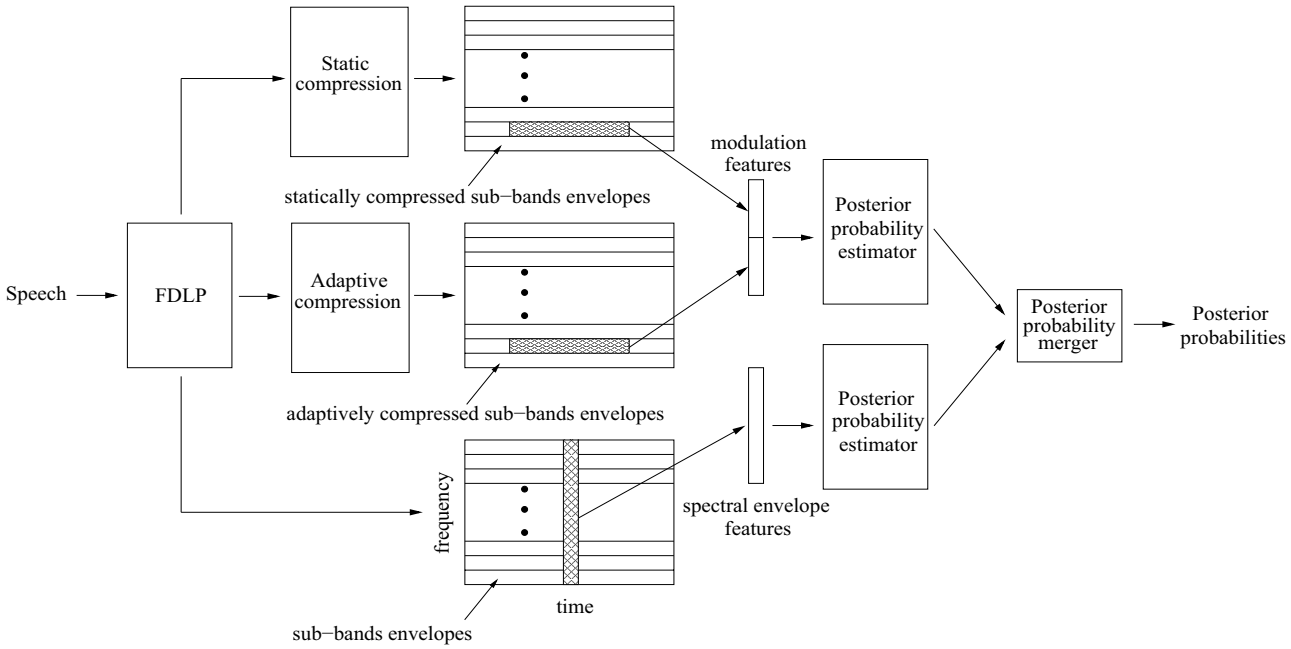

Figure 3: Schematic of the joint spectral envelope-modulation features for posterior based ASR

Table 1: Phoneme Error Rates (%) for different feature extraction techniques on the TIMIT database

Features	PER (%)
PLP-D	31.7
PLP	29.9
FDLP-S	29.9
M-RASTA	33.2
MSG	34.9
FDLP-M	29.4
PLP+M-RASTA	28.8
PLP+MSG	28.6
FDLP-S+FDLP-M	27.5

Table 2: Word Error Rates (%) on the OGI Digits database for different feature extraction techniques

Features	WER (%)
PLP-D	4.1
PLP	3.8
FDLP-S	3.4
M-RASTA	3.7
MSG	4.0
FDLP-M	3.2
PLP+M-RASTA	2.9
PLP+MSG	3.0
FDLP-S+FDLP-M	2.9

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 [14]. A three layered MLP is used to estimate the phoneme posterior probabilities. The network consisting of 1000 hidden neurons, and 39 output neurons (with soft max nonlinearity) representing the phoneme classes is trained using the standard back propagation algorithm with cross entropy error criteria. The learning rate and stopping criterion are controlled by the error in the frame-based phoneme classification on the cross validation data.

The Tandem representation of each feature set is used along with a decision tree clustered triphone HMM with 3 states per triphone, trained using standard HTK maximum likelihood training procedures. The emission probability density in each HMM state is modeled with 11 diagonal covariance Gaussians. We use a simple word-loop grammar model using the same standard set of 39 phonemes. Table 1 shows the results for phoneme error rates (PER) across all individual phoneme classes for these techniques. The proposed features (FDLP-S+FDLP-M) reduced the PER by 13% compared to PLP-D baseline feature set.

In our second experiment, we use these features on a small-vocabulary continuous digit recognition (OGI Digits database) to recognize eleven (0-9 and "zero") digits with 28 pronunciation variants [4]. MLPs are trained using these features to estimate posterior probabilities of 29 English phonemes using the whole Stories database plus the training part of Numbers95 database with approximately 10% of data for cross-validation. Tandem representation of the features are used along with a phoneme-based HMM system with 22 context-independent three-state phoneme HMMs, each model distribution represented by 32 Gaussian mixture components [4]. Table 2 shows the results for word recognition accuracies. For this task, the proposed spectral envelope features (FDLP-S) and modulation frequency features (FDLP-M) reduce the WER by 10% and 13 % compared to PLP and MRASTA features respectively.

In our third experiment, we use these features on an LVCSR task using the AMI LVCSR system for meeting transcription [15]. The training data for this system uses individual headset microphone (IHM) data from four meeting corpora; NIST (13 hours), ISL (10 hours), ICSI (73 hours) and a preliminary part

Table 3: Word Error Rates (%) on RT05 Meeting data, for different feature extraction techniques. TOT - total WER(%) for all test sets, AMI, CMU, ICSI, NIST, VT - WER (%) on individual test sets [15]

Features	TOT	AMI	CMU	ICSI	NIST	VT
PLP-D	41.9	42.4	39.4	31.3	50.9	46.4
PLP	46.4	40.9	43.7	30.0	54.7	65.1
FDLP-S	42.5	41.6	41.5	33.1	51.6	45.5
M-RASTA	45.4	46.7	41.6	36.8	53.4	49.0
MSG	44.4	43.9	40.7	34.5	52.1	52.3
FDLP-M	39.5	37.7	33.7	39.4	45.4	41.7
PLP+M-RASTA	40.5	40.5	37.8	28.5	48.9	47.9
PLP+MSG	39.6	38.8	39.3	27.3	46.6	47.6
FDLP-S+FDLP-M	35.9	36.2	34.2	27.8	42.9	39.0

of the AMI corpus (16 hours). MLPs are trained on the whole training set in order to obtain estimates of phoneme posteriors for each of the feature sets. Acoustic models are phonetically state tied triphone models trained using standard HTK maximum likelihood training procedures. The recognition experiments are conducted on the NIST RT05 [16] evaluation data. Juicer large vocabulary decoder is used for recognition with a pruned trigram language model [17]. This is used along with reference speech segments provided by NIST for decoding and the pronunciation dictionary used in AMI NIST RT05s system [15]. Table 3 shows the results for word recognition accuracies for these techniques on the RT05 meeting corpus. The proposed features (FDLP-S+FDLP-M) obtain a significant relative reduction of about 14 % in WER for the LCVSR task (compared to a relative reduction of 5% for PLP+M-RASTA and PLP+MSG features).

In all our experiments, Tandem representations of the proposed features improve ASR accuracies over other features. FDLP-S features provide similar results as PLP features. Similarly, the modulation frequency features (FDLP-M) improve ASR performances over other techniques that derive features from the modulation spectrum. Combining these feature streams results in significant improvements for all the three tasks.

5. Conclusions

Acoustic features derived from sub-band trajectories of speech estimated using FDLP provide good representations of the speech signal. Tandem processing of these features further reduces their irrelevant variabilities while increasing the discriminability between speech classes. Combining the spectral envelope and modulation features provides significant improvements over the base-line systems for ASR tasks.

6. Acknowledgements

Authors would like to thank Fabio Valente, John Dines, Philip Garner, Joel Pinto and Sivaram Garimella for their help in setting up the experiments. The authors are grateful to the AMI Consortium for providing the AMI LVCSR system.

7. References

[1] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.

[2] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech”, The Journal of the Acoustical Society of America, vol. 87, pp. 1738-1752, 1990.

[3] H. Hermansky and S. Sharma, “TRAPS - classifiers of temporal patterns”, in ISCA ICSLP, pp. 1003-1006, 1998.

[4] H. Hermansky and P. Fousek, “Multi-resolution RASTA filtering for TANDEM-based ASR”, in ISCA INTERSPEECH, pp. 361-364, 2005.

[5] R. Drullman, J.M. Festen, and R. Plomp, “Effect of reducing slow temporal modulations on speech reception”, The Journal of the Acoustical Society of America, vol. 95, pp. 2670-2680, 1994.

[6] Q. Zhu, B. Chen, N. Morgan and A. Stolcke, “On using MLP features in LVCSR”, in ISCA INTERSPEECH, pp. 921-924, 2004.

[7] H. Hermansky, D.P.W. Ellis and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems”, in IEEE ICASSP, pp. 1635-1638, 2000.

[8] A. Stolcke et.al., “Recent innovations in speech-to-text transcription at SRI-ICSI-UW”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 5, pp. 1729-1744, 2006.

[9] S. Thomas, S. Ganapathy and H. Hermansky, “Phoneme recognition using spectral envelope and modulation frequency features”, in IEEE ICASSP 2009.

[10] M. Athineos and D.P.W. Ellis, “Autoregressive Modeling of Temporal Envelopes”, IEEE Transactions on Signal Processing, vol. 55, no. 11, pp. 5237-5245, 2007.

[11] J. Herre and J.D. Johnston, “Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS)”, Proc. 101st Conv. Aud. Eng. Soc, 1996.

[12] L.S. Marple, “Computing the discrete-time ‘analytic’ via FFT”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 47, no. 9, pp. 2600-2603, 1999.

[13] B. Kingsbury, Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments, Ph.D. thesis, University of California Berkeley, 1998.

[14] J. Pinto, B. Yegnanarayana, H. Hermansky, and M.M. Doss, “Exploiting contextual information for improved phoneme recognition”, in ISCA INTERSPEECH, pp. 1817-1820, 2007.

[15] T. Hain et.al., “The 2005 AMI system for the transcription of speech in meetings”, NIST RT05 Workshop, 2005.

[16] Rich Transcription Spring 2005 Evaluation, Online: http://www.itl.nist.gov/iad/mig/tests/rt/2005-spring/index.html.

[17] D. Moore et.al., “Juicer: A weighted finite state transducer speech coder”, Lecture Notes in Computer Science, vol. 4299, pp. 285-296, 2006.

[18] T. Dau, D. Pueschel, and A. Kohlrausch, “A quantitative model of the “effective” signal processing in the auditory system. I. Model structure”, The Journal of the Acoustical Society of America, vol. 99, pp. 3615-3622, 1996.

[19] F. Valente and H. Hermansky, “Combination of Acoustic Classifiers based on Dempster-Shafer Theory of Evidence”, in IEEE ICASSP, pp. 1129-1132, 2007.