

Beyond a single critical-band in TRAP based ASR

Pratibha Jain^a, Hynek Hermansky^{a,b}

^aOGI School of Science & Engineering, Oregon Health and Science University, Portland, Oregon, USA.

^bInternational Computer Science Institute, Berkeley, California, USA.

pratibha,hynek,@ece.ogi.edu

Abstract

TRAP based ASR attempts to extract information from rather long (as long as 1 s) and narrow (one critical-band) patches (temporal patterns) from time-frequency plane. We investigate the effect of combining temporal patterns of logarithmic critical-band energies from several adjacent bands. The frequency context is gradually increased from one critical-band to several critical-bands by using temporal patterns jointly from adjacent bands as input to the class-posterior estimators. We show that up to three critical-bands of frequency context is required for achieving higher recognition performance. This work also indicates that local bands interaction is important for improved speech recognition performance.

1. Introduction

In the temporal pattern (TRAP) based ASR [2, 1], frequency-localized posterior probabilities of sub-word units (phonemes) are estimated from temporal evolution of critical band spectral densities within a single critical band. Such estimates are then used in another class-posterior estimator which estimates the overall phoneme probability from the probabilities in the individual critical bands. The frequency-localized estimates in the TRAP scheme only serve as an intermediate features for the final phoneme probability estimation and therefore the targeted frequency-localized classes do not necessarily need to be phonemes. Most often, the frequency-localized posterior probability estimates form an input to the TANDEM ASR system [10]. Briefly, the TANDEM system first derives a vector of posterior probabilities of sub-word speech classes for every speech analysis frame from some evidence presented to the input of its trained Multilayer Perceptron (TANDEM MLP). In the case of TRAP-TANDEM, this evidence itself consists of concatenated vectors of posterior probabilities of some sub-word classes (which may be but do not need to be the same at the classes utilized in the TANDEM), each estimated at the particular individual frequency. The TANDEM estimates are gaussianized and whitened. They form the feature vector for the subsequent HMM recognizer.

In the first stage of processing, TempoRAL Pattern (TRAP) system estimates features from rather long (around 1 s) temporal patterns of critical-band energies in a single critical-band [1, 2, 7]. The information about spectral-shape across the critical-bands is completely ignored. This can be inconsistent with properties of human hearing, where the phenomenon of Comodulation Masking Release (CMR) [4] indicates some interactions among the individual critical bands. This lead to the work presented in this paper where temporal patterns from several neighboring bands are used in the TRAP-like fashion and some techniques for pre-processing of data from several adjoining critical bands are investigated.

2. Task and system description

The task is recognition of eleven words (American English digits). The test set was derived from the subset of CSLU Speech Corpus [8], containing utterances of connected digits. There are 2169 utterances with total length about 1.7 hours. There are 12437 words in this set.

Recognizer is a HMM system (HTK). Each word is modeled by sequence of context independent five states, three mixture per state, phoneme models. The initial feature extraction for deriving the critical band spectrum is based on a short-term FFT spectrum, computed from 25 ms analysis frames with 10 ms analysis steps, integrated into $M = 15$, Bark-scaled trapezoidal filters [9].

2.1. Training of the feature extraction

A separate subset of CSLU Speech Corpus was used for training the TANDEM probability estimator. The TANDEM system first derives a vector of posterior probabilities of sub-word speech classes for every speech analysis frame from some evidence presented to the input of its trained Multilayer Perceptron (TANDEM MLP). In the case of TRAP-TANDEM, this evidence itself consists of concatenated vectors of posterior probabilities of some sub-word classes (which may be but do not need to be the same at the classes utilized in the TANDEM), each estimated at the particular individual frequency [2, 7]. The TANDEM estimates are gaussianized and whitened and form the feature vector for the subsequent HMM recognizer. This set contain 3590 utterances with total length about 1.8 hours. No restrictions applied to this set. So the TANDEM probability estimator sees also digits uttered in isolation and other natural numbers (the OGI Numbers contain besides digits also other natural numbers).

The subset of OGI Stories database [8] was used for training the band probability estimators (TRAP MLPs) [2]. This set contains 208 utterances with total length about 2.7 hours. So the frequency-localized probability estimators are trained on quite different speech material, containing free fluent speech from similar environment and possibly uttered by some of the same speakers as the speakers in the test set (the OGI Numbers were selected from the same recordings as the OGI Stories).

The number of target phonemes for training the probability estimators is $N = 29$. Target phonemes are these which occur in digits utterances. Others phonemes are not used for training but they create context in the TRAP vectors.

2.2. Training of the recognizer

Training set contains 2547 utterances with total length about 1.2 hours. This set is also derived from the CSLU Speech Corpus and utterances containing only connected digits are used. So the training of the recognizer is done in a usual manner, where the training set is as similar as possible to the anticipated test set.

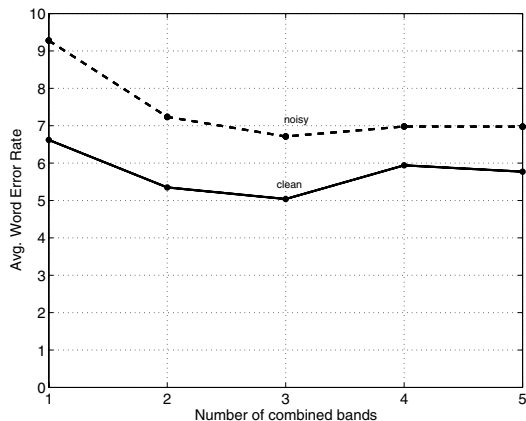


Figure 1: *Effect of the number of combined temporal patterns from adjacent critical-bands on ASR*

3. Optimal number of adjoining critical bands

The first experiment aimed at finding the optimal number of adjoining critical bands to be used in the TRAP probability estimation. Temporal patterns from several adjoining bands are concatenated. The temporal context is kept constant at 1 s (101-samples from 8 kHz speech). The temporal patterns are mean subtracted and variance normalized prior to estimating joint PCAs on the concatenated temporal vectors. These temporal vectors are formed by concatenating temporal patterns from several adjacent bands. For example for a 3-band case, a 303-sample temporal vector was formed using three 101-sample temporal patterns from adjacent bands. The joint Principal Component Analysis (joint PCA) on the OGI-Stories is used to derive bases for the projection of the concatenated vector space to a smaller 75 dimensional vector, used as an input to the subsequent TRAP probability estimators, so that the number of parameters of the estimators are kept the same in each system configuration. The block diagram of the system for a 3-band case is shown in the Figure 4. For this case, for each 3-band adjacent triplet, an independent estimator is trained for estimating the phone-posterior probabilities. The number of joint PCAs are kept at 75 which covers 96% of the total variability of 303-sample feature space. The Multilayer perceptrons (MLPs) are trained using backpropagation algorithm with cross-entropy criterion. The MLPs are trained with 300 hidden units and 29 nodes at the output layer which target context-independent phone categories. The estimated phone-posteriors from each channel is combined and used as the input to a Tandem MLP for final phone posterior estimation. Here each channel comprises group of 3 adjacent critical-bands. If temporal patterns from N adjacent bands were combined, we used $N-1$ overlap between two adjacent channels. For example, for a 3-band case, we had 1-2-3, 2-3-4, \dots , 13-14-15 groupings of three adjacent bands that define each channel.

Figure 1 shows that recognition error significantly decreases with increasing frequency context from 1 critical-band to 3 critical-bands. After that it either increases or remains the same. This trend holds good for the clean as well as for noisy environments.

This finding could be of more than engineering significance. Supporting findings were reported in the recent work in human

speech recognition by Healy et al. [6]. Unlike the earlier experiments in perception of severely band-limited speech (e.g. Warren et al. [5]) they used sinusoidal or narrow-band noise carriers modulated by the spectral envelope derived from the given critical band and reported only chance human recognition of meaningful sentences from such temporal patterns of the spectral envelope information from a single critical-band. When the envelope information from more than one critical band was used, the human recognition improved significantly. Further, they found that the simultaneous presentation of narrowband temporal patterns of the speech signal with carriers within 1 octave (about 3 critical-bands), provide highest speech intelligibility (increases from 0% when a single carrier was used to around 80% for the three carriers within the 1 octave frequency span).

4. Main directions of variability in three-band TRAP

Analysis of the main directions of variability (the main eigenvectors from the joint PCA analysis) in the localized time-frequency patterns formed from 1 s long temporal trajectories of three adjacent critical bands reveals interesting patterns. The joint PCAs are shown in Figures 2 and 3.

The most dominant eigenvectors from the joint PCA analysis shown in the Figure 2 indicate the need for averaging the time trajectories of the three adjoining critical bands prior to projecting the averaged trajectories on the cosine-like transform bases. These "averaged" bases cover around 92.7% of the total variability of a 303-sample space.

However, there is about 15 PCA bases among the preserved 75 bases, which are different. These are illustrated in the Figure 3 and as seen, they attempt to capture the local spectral slope prior to the cosine-like transform projection, essentially subtracting the time-aligned components from the individual temporal patterns from the adjoining bands broadens the studied two-dimensional vector space. In that way, they are incorporating information about local spectral-slope into the projected feature components. These "differentiating" bases cover just 3.3% of the total variability - however, we found that they are very important for achieving higher recognition performance.

It can be seen from these figures that joint PCAs can be approximated by two simple operations: 1) averaging individual 101-sample temporal patterns from individual adjacent bands, 2) subtracting individual 101-sample temporal patterns from the first and the last band of each group of adjacent bands, prior to the subsequent DCT transformation. Due to linearity of the operations, the other interpretation could be averaging the DCT components of individual 101-sample temporal patterns from individual adjacent bands and subtracting the DCT components of individual 101-sample temporal patterns from the first and the last band.

5. Results with the DCT bases

We compare the direct projection on the PCA-derived bases with the method where we would first average and differentiate the temporal patterns from the three adjacent critical bands prior to projecting the averaged and differentiated vectors on the cosine-transform bases (Figure 5). Table 1 shows the performance of 3-band joint DCT, 3-band joint PCA, and 1-band DCT system. It can be seen from the results that 3-band DCT/PCA system outperform 1-band system. Whereas the 3-band joint DCT and the 3-band joint PCA results are comparable in most of the cases.

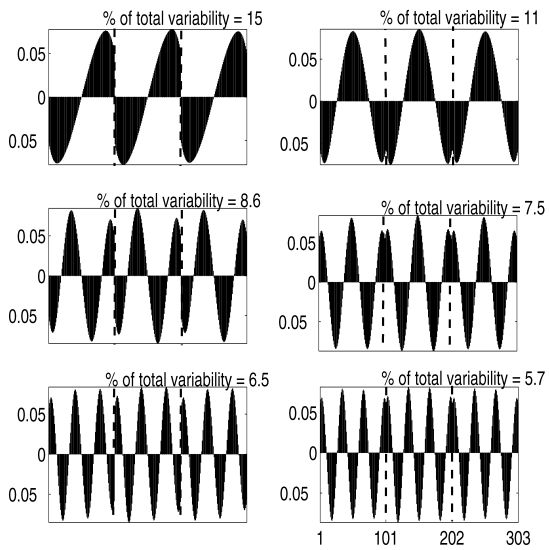


Figure 2: First six joint PCAs from 303-sample, joint temporal vector, obtained from three adjacent bands

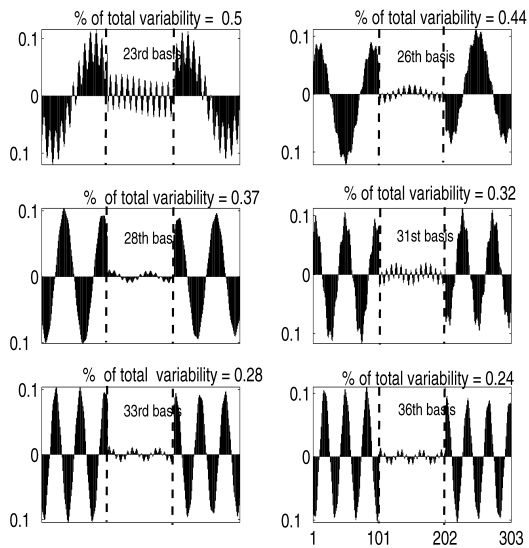


Figure 3: Some of the six joint PCAs that do spectral subtraction

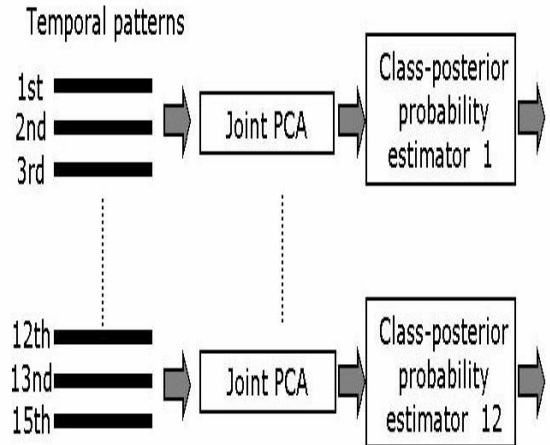


Figure 4: Block diagram of the system for combining several temporal patterns for phone posterior estimation using joint PCA components

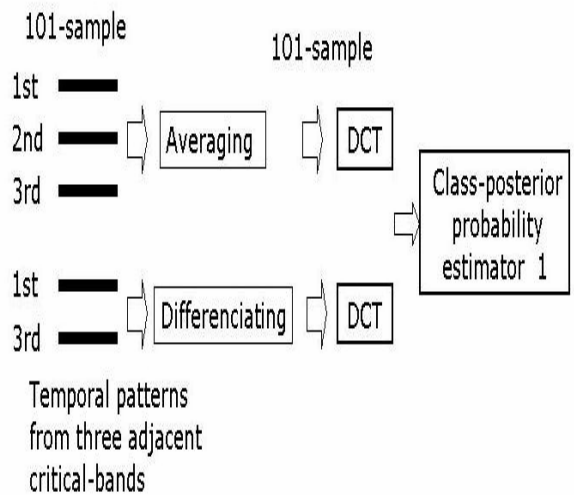


Figure 5: Block diagram of the system for combining several temporal patterns for phone posterior estimation using joint DCT components

However for the clean and white-noise case, the 3-band DCT bases perform worse than 3-band PCA bases. This could be due the fact that we had the same number of averaging and differentiating DCT bases for the projection whereas in the case of joint PCAs, in the first 75 bases, there are only around 15 differentiating bases (Figure 3).

No of Bands	clean (WER)	babble (WER)	pink (WER)	factory2 (WER)	white (WER)
3-band DCT	5.7	8.6	5.8	7.2	5.8
3-band PCAs	5.0	8.7	5.8	7.1	5.3
1-band DCT	6.2	11.1	8.0	10.0	7.0

Table 1: % Word error rate on using 76 joint DCTs on the temporal patterns of three adjacent critical-bands, 75 joint 3-bands PCAs, 75 1-band DCTs

6. Conclusion

We found the extent of the critical-band interaction and its effect on the recognition performance of the TRAP-TANDEM system. We also showed the joint PCA can be approximated by addition and subtraction of temporal patterns from individual bands, prior to the computation of DCT transform. Our results indicate that about 3-bands are required for capturing important cross spectral details for achieving high recognition performance.

7. Acknowledgements

This work is supported by the DARPA, EARS grant under MDA-972-02-01-0024. This work is also supported by the NASA, grant under NCC2/1218.

8. References

- [1] H. Hermansky, S. Sharma, "Temporal Patterns (TRAPs) in DSR of Noisy Speech", Proc. of ICASSP, Phoenix, USA, 1999.
- [2] H. Hermansky, S. Sharma, "TRAPs Classifiers of Temporal Patterns", Proc. of ICSLP, Boston, USA, 1998.
- [3] S. Sharma, "Multi-stream approach to robust speech recognition", PhD thesis, 1999.
- [4] J. W. Hall, III, John H. Grose, "Comodulation masking release and auditory grouping", JASA, vol 88, No 1, 119-125, 1990.
- [5] Warren et al., "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits", Percept. Psychophys., vol 57, 175-182, 1995.
- [6] Eric W. Healy, Richard M. Warren, "The role of contrasting temporal amplitude patterns in the perception of speech", JASA, vol 113, No 3, 2003.
- [7] Pratibha Jain, Hynek Hermansky, Brian Kingsbury, "Distributed speech recognition using noise-robust MFCC AND TRAPs-estimated manner features", Proc. of ICSLP, 473-476, Denver, USA, 2002.
- [8] R. Cole, M. Noel and T. Lander, "Telephone speech corpus development at CSLU", Proc. of ICSLP, 1815-1818, Yokohama, Japan.
- [9] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech", JASA, vol 87, no 4, 1990.

- [10] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems", Proc. Int. Conf. Acoustics, Speech and Signal Processing, Istanbul, Turkey, June 2000.