

Band-independent speech-event categories for TRAP based ASR

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

^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

Band-independent categories are investigated for feature estimation in ASR. These categories represent distinct speech-events manifested in frequency-localized temporal patterns of the speech signal. A universal, single estimator is proposed for estimating speech-event posterior probabilities using temporal patterns of critical-band energies for all the bands. The estimated posteriors are used as the input features (referred to as speech-event features) to a back-end recognizer. These features are evaluated on continuous OGI-Digits task. The features are also evaluated on Aurora-2 and Aurora-3 tasks in a Distributed Speech Recognition (DSR) framework. These features are compared with earlier proposed broad-phonetic TRAPs features estimated from temporal patterns using independent estimators in each critical-band.

1. Introduction

In the original formulation of speech recognition based on TempoRAI Patterns (TRAP), frequency-localized posterior probabilities of sub-word units (phonemes) are estimated from temporal evolution of critical band spectral densities within a single critical band [2]. Such estimates are then used in another class-posterior estimator which estimates the overall phoneme probability from the probabilities in the individual critical bands. Clearly, one should not expect very accurate estimation of phoneme probability from the limited evidence within a single critical band and the error rates at individual frequencies were rather high.

However, 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 [12]. 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.

Allen [5] suggests that some speech-events are first detected early in the human speech recognition process before the phones-like speech sub-units are identified, followed by identification of larger speech units such as syllables or words. This strategy was evidently followed by Saul et.al [4] where

narrow-band phonetic cues were detected using a probabilistic network. In the TRAP framework, several alternatives were examined, eventually settling on estimation of probabilities of broad-phonetic categories [6].

2. Band-independent speech-events categories

When investigating mean temporal patterns of spectral densities of different phonemes within the individual critical bands, we observed that they often follow very similar trends. For example, for a vowel /iy/, mean patterns are very similar in both high-spectral-energy lower and higher bands (Figure 1). Further, similar patterns can be also found in another phonemes, although not necessarily at the same frequency locations. For example, in higher frequency bands of the fricative /sh/, patterns are very similar as in the high-energy bands of the /iy/ vowel (Figure 1). Similarly, the mean temporal patterns of a vowel /iy/ in the middle frequency bands and of another vowel /ax/ in the higher frequency bands are very similar (Figure 3), as well as mean temporal patterns for a schwa /ax/ in the lower bands and that of another schwa /ix/ in the middle frequency bands (Figure 4), and fricative /sh/ and a nasal /m/ in low frequency critical-bands (Figure 2).

These similar mean temporal patterns across different phones and different bands indicate that there is a finite set of distinct temporal speech activities occurring in each critical-band. Consistently with [5] we choose to call these different distinct temporal activities the speech **events**. Then, by identifying an event from this set for each band, a speech-sound (phone) can be quantified or characterized.

This lead to the investigation of new band-independent categories, which are based on distinct temporal events manifested in temporal patterns of log critical-bands spectral densities of the speech signal. We refer these categories to as speech event categories. We propose a single, universal estimator for estimating speech event class-posterior probabilities which are then used as input features to a TANDEM-based speech recognizer.

3. Events by clustering of the individual mean TRAPs

This section describes how we obtain band-independent speech-event categories for feature estimation. The mean temporal patterns are computed using 101-sample (1 s in temporal context) mean subtracted, variance normalized, and hamming windowed, temporal patterns of log critical-band energies for each phone on a labeled dataset (TIMIT) [2]. An agglomerative hierarchical clustering technique is used to obtain new categories. A correlation measure is used as a similarity measure for clus-

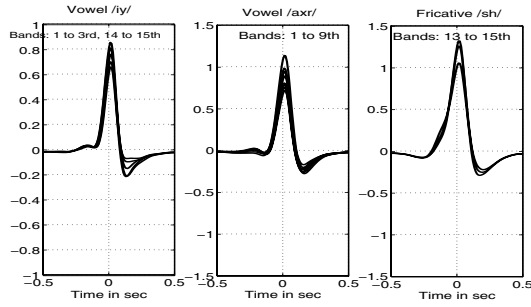


Figure 1: Similar mean temporal patterns in the different bands of same phone and in the same band of different phones.

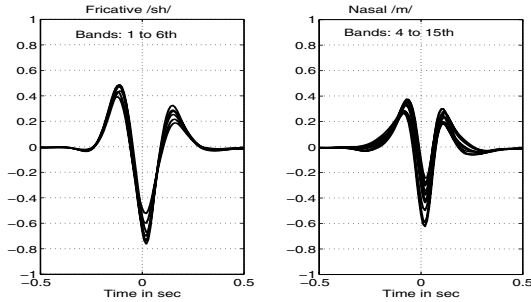


Figure 2: Similar mean temporal patterns in the same bands of different phones.

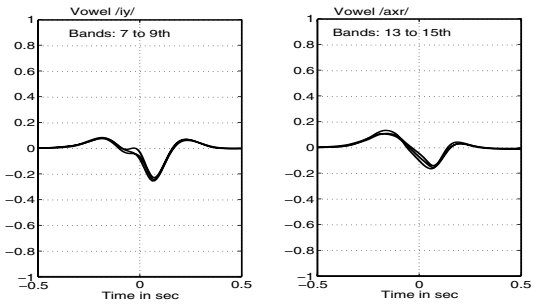


Figure 3: Similar mean temporal patterns in the different bands of different phones.

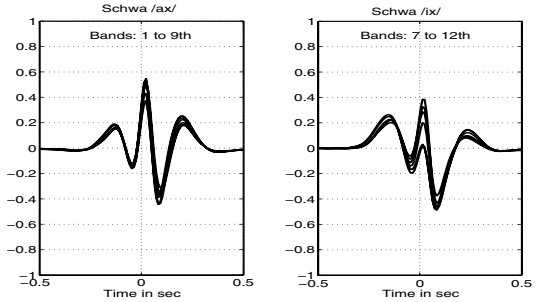


Figure 4: Similar mean temporal patterns in the same bands of different phones, in the different bands of same phone.

tering the temporal patterns. It is given by,

$$S(x, y) = \frac{\sigma_{xy}^2}{\sigma^x \sigma^y}$$

The agglomerative clustering procedure starts with $56 * 15$ (where number of TIMIT phones = 56, number of bands = 15) mean temporal patterns as the singleton clusters and at each iteration the two closest together clusters are merged with each other. This is performed iteratively until the number of clusters reach finally to 9 clusters.

This stopping point in clustering is based on some heuristics - had we continued further in the clustering, final clusters would not be able to capture the distinct 'flap' or 'schwa' mean temporal patterns, and we have chosen to keep these distinct patterns in our inventory of speech events.

The clustering techniques described above is obviously only one of many alternatives which may be employed for deriving the speech events from speech data. Overall, we feel that the particular technique for deriving speech events is also very much an open issue.

The final nine clusters are shown in the Figure 5. They represent distinct frequency-localized temporal patterns of the speech signal.

1. SILENCE-like - e.g. mean temporal patterns of the silent speech regions for all the critical-bands.
2. PLOSIVE-like - e.g. mean temporal characteristics of most of the plosives which shows a dip in energy off-center to the left as plosives are usually preceded by a stop-closure. This pattern is also found in the 8-9 bands of the glide /w/ and in the 5-6 bands of the glide /y/.

3. NASAL-like - e.g. mean temporal characteristics of a nasal /em/ in the lower 1-7 bands.
4. GLIDE-like - a peak in energy off-center to the right which represents mean temporal characteristics of glides /r/, /w/ in 1-6 bands, and of /y/ in the lower 1-4 bands and the higher 10-15 bands.
5. LOW VOCALIC ENERGY - a small burst in the energy off-center to the left followed by a small dip in energy off-center to the right. This pattern is often seen in the middle bands of vowel /iy/ and higher bands of vowel /axr/ etc.
6. SCHWA-like - mean temporal characteristics of the schwa sounds such as /ix/, /ax/ in most of the bands.
7. FLAP-like - mean temporal characteristics of the flap-like sounds such as /nx/, /dx/ in most of the bands.
8. HIGH VOCALIC ENERGY - mean temporal characteristics of most of the vowels such as /aa/, /ae/ in 1-15 critical-bands, of a vowel /iy/ in the lower 1-5 and upper 10-15 bands, and of a diphthong /oy/ in the lower 1-6 bands. This pattern is also found in the higher 12-15 bands of fricatives such as /zh/, /sh/ etc.
9. FRICATIVE-like - The ninth cluster represents mean temporal patterns of most of the fricatives such as /s/, /sh/, /f/ and of affricatives such as /ch/, /jh/ in the lower 1-9 bands. It also represents mean temporal characteristics of nasals /m/, /n/ in the higher bands

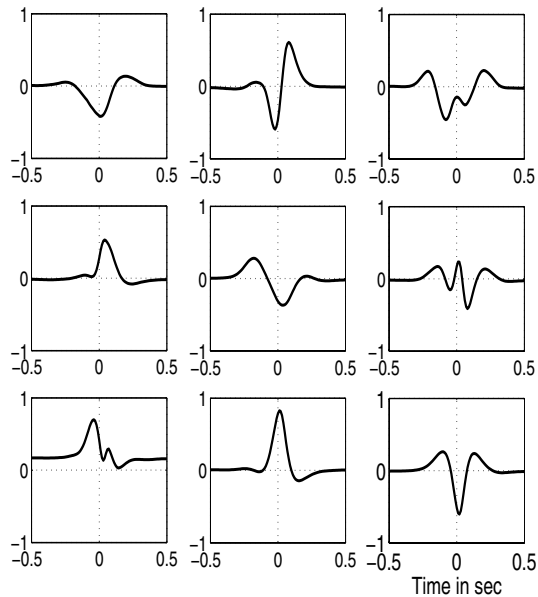


Figure 5: The 9 mean temporal patterns for the speech-event based clusters obtained by using an agglomerative clustering technique

4. Training of the Universal Temporal Pattern (UTRAP) system

A single posterior estimator is trained with 9 speech-event categories as the target classes using a Multilayer perceptron (MLP). We call this estimator Universal TRAP (UTRAP) system. The MLPs in the UTRAP technique are trained on a re-labeled dataset. The phone-labels of the training dataset were obtained by manual transcription. The new class-labels are obtained by mapping mean temporal patterns of phones to one of the closest 9 speech-event categories using the same similarity measure (correlation coefficient). The MLP is trained with back-propagation algorithm with cross-entropy as the error criterion. At every frame, speech-event posteriors are obtained by doing a forward pass for each temporal pattern from individual critical bands through the MLP. This estimated posteriors are used as the features for final recognition. We refer these features to as speech-event UTRAPs features.

The training of a UTRAP system is still an open issue. Currently training procedure starts with using the temporal patterns from the 1-band and the trained parameters from this stage are used to initialize the parameters for the second stage in of training. In the second stage, patterns from the 2-band are used for training the speech-event categories. This procedure continues up to 15 critical-band. Although not investigated yet, we feel that the temporal patterns from all critical-bands should have been randomized before the training.

5. Experimental setup and results

First, we evaluate speech-events on the continuous OGI-Digits task. For this experiment, the UTRAP MLP is trained on OGI-Stories. The 50 DCT components of mean and variance normalized, hamming windowed, 101-sample (1-s temporal context) temporal patterns of log critical-band energies are used as

input features to this MLP. It is trained with 101 hidden units and 9 target categories at the output layer. Every 10 ms, $15 * 9$ class-posteriors are estimated. They are mapped to 29 phonetic class posteriors using the TANDEM technique. The TANDEM MLP is trained with 200 hidden units and 29 target phone categories at its output layer using OGI-Numbers dataset. The 29 phone posteriors are used as input features to a HMM recognizer. The HMM models are based on 5-states and 3-mixtures context-independent mono-phone models. To determine relative effectiveness of our new technique, we compare word error rates (WERs) of this system to the baseline system with PLP cepstral features.

Table 1 shows that the speech-event UTRAP features give 8.9% WER whereas the PLP cepstral features give 5% WER. On the same task, the broad-phonetic TRAP features give 8.8% WER. This result is encouraging, given that the UTRAPs technique is in its infancy and the number of parameters of the UTRAP system (5.8 k) is more than one order-of-magnitude smaller than in the TRAPs system (86 k).

PLP cepstra	speech-event UTRAPs features	broad-phonetic TRAPs features
5.0	8.9	8.8

Table 1: Average Word error rate (%) on OGI-Digits task using PLP cepstral features, speech-event features alone, broad-phonetic TRAPs features alone

We also evaluated these features on the Aurora tasks. For this experiment, we used clean as well as noisy TIMIT corpus for training of both the UTRAP MLP and the TANDEM MLP. The TIMIT is artificially corrupted with additive subway, babble, exhibition, and car noises at SNR ranging from 5-20 dB. These are the same noises that are present in the training set of Aurora-2 (TIDIGITs) dataset. The features are evaluated on continuous digits Aurora-2 (TIDIGITs) and Aurora-3 (SpeechDat car: Spanish, Italian, and Finnish) tasks. There were three different testing conditions: Well-matched (WM), Medium-mismatched (MM), and High-mismatched (HM) conditions. The detail description of these conditions and datasets can be found in [9, 10, 11, 8]. The configuration of the HTK HMM recognizer was given by the European Telecommunications Standards Institute. In particular, each digit was modeled using a whole-word model and the models were represented by 16-states, 3-mixtures whole word HMMs. The silence model had 3-state and 6-mixture per state and one-state short pause model was used and tied to the middle state of the silence model.

We evaluate the UTRAP-based features against the robust MFCC features. These features were used in our official ETSI Aurora submission and their detail description can be found in [8].

For the UTRAPs feature computation, critical-band energies are in this case reconstructed from the transmitted modified cepstral features at the server-end. The modifications of the cepstrum consist of Wiener filtering the speech signal and LDA-RASTA filtering of temporal trajectories of the critical band energies prior to the computation of the cepstral features. The 50 DCT components of mean and variance normalized, hamming windowed, 101-sample (1-s temporal context) temporal patterns of reconstructed log critical-band energies are used as input to the UTRAP MLP. This MLP is trained with 101 hidden units and 9 target categories at the output layer. Every 10 ms, $15 * 9$ class-posteriors are estimated. They are mapped

to 7 broad-phonetic (plosives, flaps, fricatives, nasals, vowels, schwa, and silence) class posteriors using a TANDEM MLP. The Tandem MLP is trained with 200 hidden units and 7 target categories at its output layer. These 7 posteriors are derived by the TANDEM MLP without applying the softmax non-linearity and then concatenated with the modified MFCC features at the server-end. The final 52 dimensional feature vector is decorrelated using a whitening transform. The whitening transform is computed a priori using the principal component analysis applied to 52 dimensional feature vectors on the TIMIT dataset. The decorrelated feature vectors are used as the input features to the HMM based recognizer.

Table 2 and Table 3 show that in conjunction with the modified cepstral features, the both the TRAP [6] and the UTRAP features give consistent gain in word recognition performance on the Aurora-2 and Aurora-3 datasets.

	robust MFCC	with broad-phonetic TRAPs features	with speech-event UTRAPs features
WM	8.9	7.2	7.4
MM	9.4	7.1	7.2
HM	9.9	8.3	8.2

Table 2: Average Word error rates (%) on the Aurora-2, TIDIG-ITs data and by feature set (noise-robust MFCCs alone, robust MFCCs augmented with broad-phonetic TRAPs features or the UTRAPs features), with context window = 1 s. Multistyle training was used

	robust MFCC	with broad-phonetic TRAPs features	with speech-event UTRAPs features
WM	3.3	3.0	2.9
MM	8.5	7.8	7.8
HM	13.0	14.2	11.2

Table 3: Average Word error rates (%) on the Aurora-3, by testing conditions and by feature set (noise-robust MFCCs alone, robust MFCCs augmented with broad-phonetic TRAPs features or robust MFCCs augmented with the UTRAPs features), with context window = 1 s.

6. Conclusion

We propose and study a fundamental modification of the TRAP ASR technique, the Universal TRAP (UTRAP) technique. The modification consists of replacing the frequency specific sub-word unit (phonemes or broad-phonetic classes) probability estimators by a single frequency-independent estimator of probability of speech events. The speech events are currently derived by clustering all the mean temporal patterns where mean temporal patterns are computed by averaging temporal trajectories of critical band spectral densities centered within boundaries of phonemes on a phoneme-labeled speech development data. Nine specific speech events are derived in this way, roughly corresponding to meaningful phonetic events as seen in the temporal trajectories of critical-band spectral densities.

The band-independent categories allow for using a universal class posterior estimator, applied at all frequencies in all the critical bands. This results in more than one order-of-magnitude reduction in the number of system parameters.

The stand-alone performance of the speech event UTRAP features is already about equal to the performance of broad-phonetic TRAP features. A consistent gain in the recognition

performance is achieved by augmenting speech-event UTRAPs features with cepstral features. This indicates features are complementary to short-term cepstral features and are able to generalize well across wide noisy environments.

More work is needed in identifying the distinct speech-events as well as in techniques for training the UTRAP estimator.

7. Acknowledgements

This work is supported by the DARPA, EARS grant under MDA-972-02-01-0024.

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'98, Sydney, Australia, 1998.
- [3] S. Sharma, "Multistream approach to robust speech recognition", PhD thesis, 1999.
- [4] Lawrence K. Saul, Mazin G. Rahim, and Jont B. Allen, "A statistical model for robust integration of narrowband cues in speech", Computer Speech and Language, vol 15, 175-194, 2001.
- [5] J.B. Allen, "How do humans process and recognize speech?", IEEE Trans. on Speech and Audio Processing, vol 2, 567-577, 1994.
- [6] 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.
- [7] Katrin Kirchhoff, "Combining Articulatory and Acoustic Information for Speech Recognition in Noisy and Reverberant Environments", Proc. of ICSLP, 891-894, Sydney, Australia, 1998.
- [8] Andre Adami et.al, "Qualcomm-ICSI-OGI Features For ASR", Proc. of ICSLP, 21-24, Denver, USA, 2002.
- [9] Ulf Knoblich, Alcatel, "Description and Baseline Results for the Subset of the Speechdat-Car Italian Database used for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation", STQ Aurora DSR working group, au23700, 2000.
- [10] Ulf Knoblich, Alcatel, "Spanish SDC-Aurora Database for ETSI STQ Aurora WI008 Advanced DSR Front-end Evaluation : Description and Baseline Results", STQ Aurora DSR working group, auxxx00, 2000.
- [11] Nokia, "Availability of Finnish SpeechDat-Car database for ETSI STQ WI008 front-end standardisation", STQ Aurora DSR working group, au23700, 2000.
- [12] 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.