TRAPS - CLASSIFIERS OF TEMPORAL PATTERNS

Hynek Hermansky[1,2] *Sangita Sharma*[1]

[1]Oregon Graduate Institute of Science and Technology,
Portland, Oregon , USA.
[2]International Computer Science Institute,
Berkeley, California, USA.
Email: hynek,sangita@ece.ogi.edu

ABSTRACT

The work proposes a radically different set of features for ASR where TempoRAl Patterns of spectral energies are used in place of the conventional spectral patterns. The approach has several inherent advantages, among them robustness to stationary or slowly varying disturbances.

1. INTRODUCTION

1.1. Spectral features

In 1665 Isaac Newton made the following observation: *'The filling of a very deepe flaggon with a constant streame of beere or water sounds yer vowells in this order w, u, ω, o, a, e, i, y'* [8]. What young Newton observed was the spectral resonance peak which enhanced the spectrum of the beer pouring sound and moved up in frequency as the "deepe flaggon" was filling up. Since then, attempts to find acoustic correlates of phonetic categories mostly followed Newton's lead and studied the spectrum of speech.

Spectrum-based techniques form the basis of most feature extraction methods in current ASR. A problem with the spectrum of sound is that it can easily be modified by variety of relatively benign means such as frequency characteristics of the communication channel or narrowband noise. Subsequently, the spectrum-based features are inherently fragile and various supplementary techniques need to be applied to combat the effects of realistic communication environments.

1.2. Temporal Processing?

Many of the noise-robust techniques employ the temporal domain. Some of these are reviewed in [7] where the extreme position is taken by challenging the early Newton view and most of the current speech recognition wisdom and proposing that *'...put in a question the whole concept of spectral analysis for deriving an internal representation of acoustic signal in human cognition. Even though there is a strong evidence that human auditory perception does some sort of spectral analysis of the incoming acoustic signal, it may be that the main reason for frequency selectivity of human auditory system is not to derive frequency content of a given segment for phonetic classification but rather to provide means for optimal choice of high signal-to-noise (SNR) regions for deriving reliable sub-band based features by temporal analysis of the high SNR sub-bands of the signal.'*

The current work examines this proposal.

2. PHONETIC CLASSIFICATION USING TRAPS

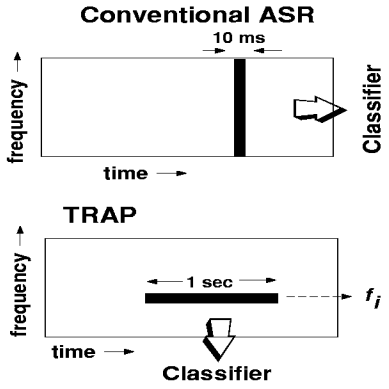

Figure 1: Temporal Paradigm for ASR

We substitute a conventional spectral feature vector in phonetic classification by a 1 sec long temporal vector of critical band logarithmic spectral energies [6] from a single frequency band (Fig 1). Similar to our earlier work on data-driven design of RASTA filters [10], the phonetic class is defined with respect to the center of this 101 point (at 10ms frame rate) temporal vector.

Since the classifier based on temporal vectors is attempting to capture the appropriate temporal pattern from the acoustic stream, we call such temporal sub-band classifier TRAP [1].

2.1. Experimental setup

We have used two databases for our work, the OGI-Stories corpus [4] and OGI Numbers corpus [5]. The OGI Stories database consists of telephone quality conversational speech. A subset of approximately 2 hours of phonemically-labeled speech was used for training the temporal classifiers. The OGI Numbers corpus consists of a set of continuous, naturally spoken utterances collected from many different speakers over the telephone. A subset of this database (approximately 0.2 hours) formed the cross-validation set for testing the trained temporal classifiers. Two more independent subsets of this database of approximately 1.7 hours and 0.6 hours were also used for experiments as described in the following sections.

[1]TRAP of course stands for the TempoRAl Pattern

2.2. TRAPs For Phonemes

To understand the nature of the information that is available in the time trajectories of spectral energy, we first examine for patterns in the temporal evolution of the basic sound units typically used in ASR, i.e phonemes.

We analyzed the OGI-Stories corpus and considered the 29 phonetic classes which also occur in the OGI-Numbers corpus. For each phoneme, a set of vectors representing its temporal evolution are extracted from a particular critical band. Each vector is an approximately 1 sec long (101 frames at 10ms frame rate) time trajectory centered around the frame which belongs to the class (phoneme) under consideration. The mean of these vectors is then used to represent the pattern corresponding to the temporal evolution of that class. It should be noted that only the center frames in all these vectors belong to the same class. Other frames may belong to any other class in whose context the center class can occur in conversational speech. Thus the mean operation averages over all the surrounding context of the phoneme under consideration. The variance of these vectors is minimum around the center since the center frames belong to the same class and increases away from the center due to the change in context.

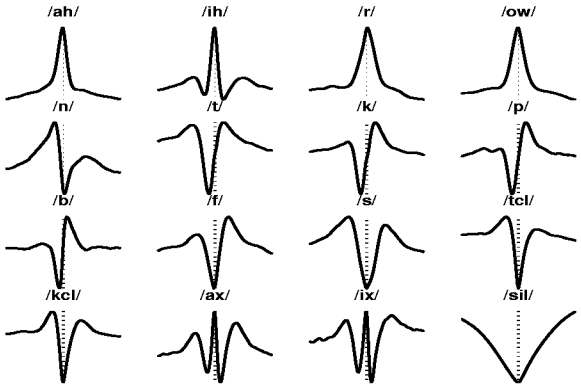

Figure 2: Mean TRAPs for 16 phonemes at the fifth critical band time trajectory. The dotted line represents the center frame

Fig. 2 represents the mean temporal patterns (which we call Mean TRAPs to emphasize their urge to capture an appropriate phoneme) of 16 phoneme classes. It can be seen from the figure that some of the classes have very distinctly different Mean TRAPs, e.g. the vowel /ah/ compared to the stop-consonant /b/ and some of the Mean TRAPs are very similar, e.g. the vowels.

2.3. Classification Using TRAPs

As a first test of utility of TRAPs, we have tried to use a simple template matching approach on the Mean TRAPs in which the correlation coefficient was used as a measure of similarity. To classify a frame in a particular critical band, a 1 sec time trajectory around that frame is matched to the 29 Mean TRAPs for that critical band. The frame is assigned to the class which gives the maximum correlation with the given TRAP. In order to de-emphasize the contributions of the spectral energies towards the edges of the time trajectory, we first removed the mean and weighted each pattern by a Hanning window.

As seen from the Table 1 the performance in the individual bands is not high but is well above the chance.

A further significant improvement was obtained from a nonlinear (Multi-Layer Perceptron based) TRAP (which we call a Neural TRAP). The Neural TRAP used in our experiments is a single hidden layer MLP with 101 dimensional input vector, a 300 unit hidden layer and 29 outputs. The input consists of 1 sec long spectral energy time trajectory centered around the frame to be classified.

The baseline system used is the standard hybrid hidden Markov model/multi-layer perceptron (HMM/MLP) speech recognizer [3] from the International Computer Science Institute, Berkeley, California, in which phonetic classification is performed by a single hidden layer MLP. The features used for the baseline system consist of 8 PLP cepstral coefficients [6] with utterance-based cepstral mean subtraction along with 9 delta and 9 acceleration coefficients. The input to the MLP consists of 9 frames of context with the current frame at the center of this context window (234 dimensional input). The hidden layer has 500 units and the output of the MLP consists of posteriori probabilities of the 29 phonetic categories occurring in the Numbers corpus. The baseline system is trained on the 1.7 hours subset of the Numbers corpus. This baseline system yields 21 % frame-level error and 6.5 % word-level error.

It is interesting to see that based only on a 1 sec time trajectory of spectral energy in a single critical band, the performance of each Neural TRAP is approximately 40% of the performance of the baseline system which uses all spectral information and around 170ms of temporal information.

System	Frame error (%) for each critical band
Mean TRAPs	78 - 81 %
Neural TRAPs	65 - 69 %

Table 1: Frame-level performance of different TRAPs on OGI Numbers corpus

Table 1 gives the range of the frame errors for the 29 TRAPS based on phonetic categories for each of the 15 critical bands when tested on the 0.2 hours subset of the Numbers database. It is encouraging to note that the performance in each critical band is approximately 80% error even for the simplest of the TRAPs - the Mean TRAPs, and goes well below 70 % error for the nonlinear Neural TRAPs. This is significantly higher than chance (96.5% error for 29 classes) in spite of the fact that none of the TRAPs have access to any information about spectral correlations between neighboring bands.

2.4. Combination of TRAPs.

Since there are about 15 critical bands available within the telephone bandwidth, we have at our disposal 15 outputs from 15 different TRAPs. The question is how the performance improves by combining their outputs.

As in our previous work on multi-band ASR [9], we use a MLP for combining the outputs obtained from each of the 15 TRAPs. The input to the combining network is the concatenated vector of the correlations (in case of Mean TRAPs) or class conditional log-likelihoods (in case of Neural TRAPs) of the 29 phonetic classes from each of the 15 TRAPs (435 dimensional input). The network has

a single hidden layer of 300 units and 29 outputs which represent the merged estimate of the class posteriori probabilities. The combination network thus has 139200 parameters which is comparable to the 131500 parameters of the baseline system.

Table 2 compares the frame error and word error rate of the baseline system, the Mean TRAP-based recognizer, and the Neural TRAP-based recognizer. The frame-level error was derived on the the 0.2 hours cross-validation subset of the OGI Numbers database, the word-level error on the 0.6 hours test subset (4670 words) of the OGI Numbers database.

It is seen that on the frame level, the performance of the baseline (spectrum-based) and the Mean TRAP combiner is comparable, the Neural TRAP-based recognizer performs better. On the word level, the baseline recognizer is the best, but the Neural TRAP-based recognizer is close behind. The word error of the simple Mean TRAP-based recognizer is about twice the error of the baseline system.

An analysis of the frame errors (not shown here) indicated that the TRAP system typically corrects about 40% of errors made by the baseline system and hence has significant complementary information.

System	Frame error	Word error
Baseline	21%	6.5%
Mean TRAP-based	22%	11.5%
Neural TRAP-based	18.7%	7.6 %
Combined Baseline and TRAP System		
Mean Trap-based	19%	6.0%
Neural Trap-based	16.9%	5.5%

Table 2: Performance of the Baseline system and the TRAP-based systems on the OGI Numbers corpus

2.5. Combination of The Baseline and TRAP-based Recognizer

As discussed above the TRAP-based recognizer has significant amount of complementary information as compared to the baseline system. In an attempt to capitalize on this observation we combined the outputs of the baseline system and the TRAP-based recognizers at the frame level using an MLP classifier. This classifier had 58 inputs (concatenation of the 29 class-conditional log-likelihoods from each of the systems), 500 hidden units and 29 outputs. From Table 2 it is seen that the combination (especially with the Neural TRAP recognizer) significantly improves the performance as compared to the baseline system.

3. PRELIMINARY EXPERIMENTS WITH NOISE

To assess possible advantages of the TRAP-based recognizer we investigated its performance in several artificially degraded situations. The recognizer was always trained only on the clean speech. For these experiments we have used only the basic template correlation-matching of Mean TRAPs.

3.1. TRAPs in additive white noise

As a preliminary experiment we added white noise from the NOISEX-92 database at signal-to-noise ratio of 10dB to the OGI Numbers database.

System	Frame error	Word error
Baseline	41.6.%	21%
TRAP-based	37.7%	27%
Combination	33.6%	19%

Table 3: Performance on white noise

Table 3 shows that the baseline-TRAP combination gives significant improvement in performance at both the frame and the word levels.

3.2. TRAPs in convolutive noise

The utterance-based cepstral mean subtraction technique is known to be robust to convolutive noise. TRAPs should also be robust to such distortion because of local (1 sec) mean removal inherent in the TRAP matching procedure. To simulate convolutive distortion the test data was pre-processed by a pre-emphasis filter.

System	Frame error		Word error	
	Clean	Noise	Clean	Noise
Baseline without CMS	21.8%	33.3%	8.0%	16%
Baseline	21%	22.5%	6.5%	7%
TRAP-based	22%	23%	11.5%	13%

Table 4: Comparison of degradation in performance from clean test condition to condition corrupted by convolutive distortion

Table 4 compares the performance of the baseline system and the the TRAP-based system to the baseline system without cepstral mean subtraction. The performance of the system without cepstral mean subtraction degrades rapidly from 21.8% frame error on clean test data to 33.3% on pre-emphasized data as shown in the the table. On the other hand both the baseline system with mean subtraction and TRAP-based system (where no explicit mean subbtraction is done on the data) show only a slight degradation in performance as compared to the clean test case. This indicates an inherent robustness of TRAPs to this often encountered source of error in ASR.

4. PHONETIC CLASSIFICATION USING BROAD TRAPS

4.1. Clustering of TRAPs

As noted in Section 2.2, some of the classes have very similar TRAPs. Based on this observation we clustered TRAPs using a hierarchical clustering algorithm with a correlation based similarity measure. The clustering resulted in 5 distinct broad-category TRAPs (which we call Broad TRAPs) as shown in Fig 3 (only 4 Broad TRAPs are shown, the fifth Broad TRAP contained only the silence class).

It is interesting to note that although no assumptions were made for the clustering algorithm, the TRAPs cluster into the five broad phonetic categories i.e 1) vowels and

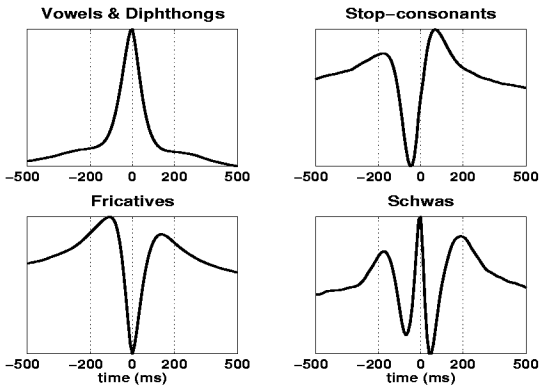

Figure 3: 4 Broad TRAPs clusters of the fifth critical band time trajectory

diphthongs 2) stop-consonants 3) fricatives 4) schwas (reduced vowels) and 5) silence. It is also interesting to note that irrespective of the phoneme duration which varies from approximately 30ms for stop-consonants to 100ms for vowels, the effect of the center phoneme lasts for a considerable time.

These Broad TRAPs have distinct and intuitive temporal patterns, e.g. the Broad TRAP corresponding to the vowel cluster has a peak in the center since vowels are characterized by high energy as compared to the other sounds. The stop-consonant Broad TRAP has a dip off-center to the left, since a stop-consonant is usually preceded by a closure characterized by low energy.

4.2. Classification Experiment

The Broad TRAPs obtained in each critical band can further be used for classification into the 5 broad categories similar to the template matching technique as described in Section 2.3. The frame-level error for such a classification in each critical band is in the range of 22% - 28%. The 5 correlations obtained in each critical band can further be used for phonetic classification. This is achieved by concatenating the correlation vectors from the Broad TRAPs in each critical band (75 dimensional vector) and using it as input to an MLP with 500 hidden units and 29 outputs.

Test condition	Frame error	Word error
Clean	24.6%	12.8%
White noise	40.8%	28.5%

Table 5: Performance with the Broad TRAPs

The performance based on Broad TRAPs is not all that different from the performance achieved by the Mean TRAPs. This result is consistent with Allen's suggestion for a partial recognition of features within each critical band [1] and suggests that the full phoneme classification on each sub-band temporal energy pattern may not be necessary.

5. DISCUSSION

As would be obvious to those readers who are familiar with Allen's interpretation of Fletcher's research [1], this work represents a further development of the Fletcher/Allen model of speech recognition. Movement away from the conventional *across spectrum processing* has recently emerged in works on multi-band ASR [9, 2]. The notion of the *across time processing* has been present in the work on RASTA processing for quite some time [7]. The current work carries both concepts to the extreme and attempts to get away with conventional spectral correlations altogether and to rely exclusively on temporal energy patterns with subsequent merging of partial recognitions in the individual frequency channels. We demonstrate that it is possible to classify phonemes with a reasonable accuracy based on rather long (much longer than a single phoneme) temporal pattern of spectral energy in a single critical band alone. We also demonstrated that by combining classification results from the individual critical bands one can achieve recognition performance quite competitive with the current state-of-art spectral-based ASR techniques.

This result opens ways for dramatically different approaches to acoustic modeling in ASR.

ACKNOWLEDGMENTS

The TRAP approach emerged from experiments with temporal spectral patterns carried out at the 1997 Summer Research Workshop at Johns Hopkins University with Juergen Luettin, Terri Kamm, and Sarel van Vuuren, and was inspired by Jont Allen's interpretation of early Fletcher's experiments in human recognition of meaningless syllables. This work was supported by NSF (IRI-9713583, IRI-9712579), DoD (MDA904-98-1-0521, MDA904-97-1-0007) and by industrial grant from Intel to Anthropic Signal Processing Group at OGI.

6. REFERENCES

1. J.B. Allen. How do humans process and recognize speech? *IEEE Trans. on Speech and Audio Processing*, 2(4):567–577, 1994.
2. H. Bourlard and S. Dupont. A new ASR approach based on independent processing and re-combination of partial frequency bands. *Proc. ICSLP 96*, 1:426–429, 1996.
3. H. Bourlard and N. Morgan. *Connectionist Speech Recognition — A Hybrid Approach,*. Kluwer Academic Publishers, Boston, 1994.
4. R. Cole, M. Noel, and T. Lander. Telephone speech corpus development at CSLU. *Proc. ICSLP 94*, September 1994.
5. R. A. Cole, M. Noel, T. Lander, and T.Durham. New telephone speech corpora at CSLU. *Proc. Eurospeech 95*, 1:821–824, September 1995.
6. H. Hermansky. Pereceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.
7. H. Hermansky. Should recognizers have ears?, invited paper. *Speech Communication*, 25(1-3):3–27, 1998.
8. P. Ladefoged. *Three Areas of Experimental Phonetics*. Oxford University Press, 1967.
9. S. Tibrewala and H. Hermansky. Sub-band based recognition of noisy speech. *Proc. ICASSP 97*, II:1255–1258, 1997.
10. S. van Vuuren and H. Hermansky. Data-driven design of RASTA-like filters. *Proc. Eurospeech 97*, pages 409–412, 1997.