TEMPORAL PATTERNS (TRAPS) IN ASR OF NOISY SPEECH

Hynek Hermansky[1,2] *and 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

In this paper we study a new approach to processing temporal information for automatic speech recognition (ASR). Specifically, we study the use of rather long-time TempoRAl Patterns (TRAPs) of spectral energies in place of the conventional spectral patterns for ASR. The proposed Neural TRAPs are found to yield significant amount of complementary information to that of the conventional spectral feature based ASR system. A combination of these two ASR systems is shown to result in improved robustness to several types of additive and convolutive environmental degradations.

1. INTRODUCTION

1.1. Spectral features

Spectrum-based techniques form the basis of most feature extraction methods in current ASR. A drawback of the spectral features is that they are quite sensitive to changes in the communication environment e.g. characteristics of different communication channels or environmental noise. Subsequently, recognizers based on spectral features exhibit rapid degradation in performance in realistic communication environments and supplementary techniques need to be applied to address this problem.

1.2. Temporal Processing

Many of the noise-robust techniques employ the temporal domain. Some of these are reviewed in [7]. This paper suggests an extreme position by challenging the accepted concept of finding acoustic correlates of phonetic categories in speech spectrum.

1.3. Phonetic Classification using TRAPs

As an alternative to spectrum-based feature vectors we have proposed in our recent work [8] the use of TempoRAl Patterns (TRAPs [1]) for phonetic classification. In this technique, we substitute a conventional spectral feature vector in ASR by a 1 sec long temporal vector of critical band logarithmic spectral energies [6] from a single frequency band (Fig 1). The phonetic class is defined with respect to the center of this 101 point (at 10ms frame rate) temporal vector similar to our earlier work on data-driven design of RASTA filters [11]. The idea here is to capture the temporal evolution of the band-limited spectral energy in a vicinity of the underlying phonetic class.

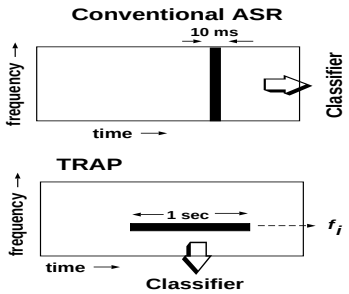

Figure 1: Temporal Paradigm for ASR

In our earlier work [8] we have mainly examined a special class of TRAPs called the Mean TRAPs. The Mean TRAPs are data-driven templates obtained by averaging 1 sec long temporal vectors for each of the phonetic classes in each frequency band independently. A simple correlation classifier is then used to perform the phonetic classification of the incoming temporal trajectories in each of the 15 critical frequency bands using the respective Mean TRAPs. We demonstrated that even with this rather simplistic approach it is possible to classify phonemes with reasonable accuracy based on rather long temporal patterns of spectral energy in a single critical band. Subsequent combination

[1]TRAP stands for the TempoRAl Pattern

of results from individual frequency bands resulted in performance close to that of the conventional spectral based systems.

Neural TRAPs provide a generalization of the Mean TRAPs and could provide means to improve the performance of the TRAPs. In this work we examine the Neural TRAPs and study their performance in several noise environments.

2. 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 phonetically-labeled speech from this corpus 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. Three independent subsets of this database of approximately 1.7 hours, 0.6 hours and 0.2 hours respectively have been used in experiments as described in the following sections. The 1.7 hours subset is the training set, the 0.2 hours subset forms the cross-validation set on which the frame-level errors for the 29 phonetic classes present in the Numbers corpus are reported, and the 0.6 hours subset (4670 words) comprises the test set on which the word-level errors are reported.

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 (ICSI), 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 estimated 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.

3. NEURAL TRAPS

Fig. 2 represents a single Neural TempoRAl Pattern classifier. As the name suggests a feed-forward multi-layer perceptron (MLP) is used to classify the central

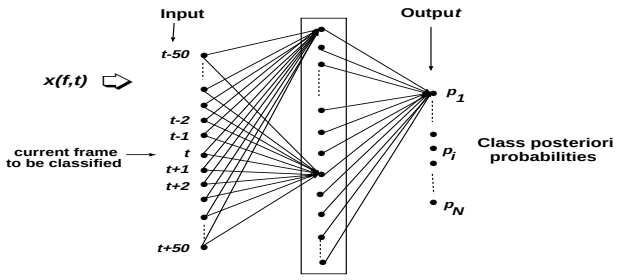

Figure 2: Neural TRAP

frame of a 1 sec long (101 points) temporal trajectory in each critical band. The temporal trajectory comprises of logarithmic energies in the particular critical band. To make the input representation robust to convolutive channel distortions the mean is subtracted from the available 1 sec trajectory. To compensate for the decrease in variance of the temporal trajectory commonly observed in presence of additive environmental noise conditions, each input 1 sec time-trajectory is normalized to have unity variance. In order to de-emphasize the contributions of the spectral energies towards the edges of the time trajectory, each input pattern is further weighted by a Hamming window as in the case of Mean TRAPs [8].

System	Frame error (%) for each critical band
Mean TRAPs	78 - 81 %
Neural TRAPs	66 - 74 %

Table 3: Frame-level performance of different TRAPs on OGI Numbers corpus

Each Neural TRAP classifier has 300 hidden units and 29 outputs. It is trained on 2 hours of the OGI Stories corpus for 29 phonetic classes. The trained TRAPs are tested on the OGI Numbers corpus. As seen from Table 3 the performance of each of the sub-band Neural TRAPs is significantly better than the performance of the corresponding Mean TRAPs. Also the performance of the individual TRAPs, though not too high, is much better than chance (96.5% error for 29 classes).

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 35% of the performance of the baseline system which uses all spectral information and around 170ms of temporal information.

System	Clean		Convolutive		Sine 1KHz.	
	Frame	Word	Frame	Word	Frame	Word
Baseline	21	6.5	22.5	7	42.57	36.9
TRAP	20	8.8	21.1	10.2	29.15	18.1

Table 1: Frame and Word errors (%)

System	Clean		White		Pink		Factory		Engine	
	Frame	Word	Frame	Word	Frame	Word	Frame	Word	Frame	Word
Baseline	21	6.5	41.59	23.4	49.39	33.5	42.15	24.3	41.56	24.9
TRAP	20	8.8	37.52	25.1	45.81	33.9	39.21	26.6	34.25	21.6
Combined	17.8	5.8	34.71	20.9	44.21	29.9	36.41	22.4	32.84	18.9

Table 2: Frame and Word errors (%). Engine refers to destroyer-engine noise.

3.1. Combination of TRAPs.

Each input speech frame is classified by 15 Neural TRAPs corresponding to the 15 critical bands [6]. To obtain a single classification result we use a MLP for combining the outputs obtained from each of the 15 TRAPs as in our previous work on multi-band ASR [10]. The input to the combining network is the concatenated vector of the class conditional log-likelihoods 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. The combiner network is trained on 1.7 hours subset of the Numbers corpus.

Table 1 compares the frame error and word error rate of the baseline system and the Neural TRAP-based recognizer. It is seen that on the frame level, the performance of the Neural TRAP-based recognizer is better than that of the baseline (spectrum-based) recognizer. However, on the word level, the baseline recognizer performs better than the Neural TRAP-based recognizer.

3.2. Combination of the Baseline and TRAP-based Recognizer

An analysis of the frame errors indicates that 40% of the errors made by the baseline system would be corrected by the TRAP-based system (i.e. the TRAP-based system makes the correct decision) while 38% of the TRAP-based system errors are not being made by the conventional system. This shows that both systems yield significant complementary information. Such a situation makes both systems good candidates for merging [9].

Based on this observation we combined the outputs of the baseline system and the TRAP-based recognizer at the frame level using a 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. This combiner is also trained on the 1.7 hours subset of the Numbers corpus. From Table 2 it is seen that the combination results in improved performance as compared to the baseline system both at the frame-level and word-level.

It should be noted that the results on the clean speech that we are reporting in Tables 1, 2 for Neural TRAPs are slightly different than the results reported in [8] since in our current system we do additional input mean and variance normalization. This normalization results in slight degradation in performance on clean data but makes the system more robust in presence of noise.

4. EXPERIMENTS IN NOISE

To test the performance of the TRAP-based system in degraded environments we tested it on speech artificially degraded by various types of noise. The recognizer was always trained only on the clean speech.

4.1. TRAPs in convolutive noise

The baseline system uses utterance-based cepstral mean subtracted features which is known to be robust to convolutive noise. TRAPs should also be robust to such distortion because of local (1 sec) input mean removal. To simulate convolutive distortion the test data was pre-processed by a pre-emphasis filter.

The performance of the baseline system without cepstral mean subtraction degrades rapidly from 21.8% frame error and 8% word error on clean test data

to 33.3% frame error and 16% word error on pre-emphasized data. On the other hand as indicated in Table 1, both the baseline system with mean subtraction and the TRAP-based system show only a slight degradation in performance to such convolutive distortion as compared to the clean test case. This demonstrates an inherent robustness of TRAPs to convolutive channel distortion.

4.2. TRAPs in additive sinusoidal noise

We tested the performance of the TRAP recognizer on additive sinusoidal noise at 1 KHz. and SNR 10dB. From Table 1 it is seen that the TRAP-based system results in half the error rate as compared to the baseline system.

4.3. TRAPs in realistic additive noise

Realistic noises (white, pink, factory and destroyer-engine) from the NOISEX-92 database were added to the data. Table 1 compares the performance of the baseline, TRAP-based and combined systems in presence of these noise conditions. It is seen that the TRAP-based system consistently gives reduced frame error as compared to the baseline system and gives quite comparable performance on the word level. The combined baseline and TRAP system results in significant improvement in both the frame and the word level performances. Specifically, the combined system results in around 15% reduction in frame error and 13% reduction in word error (average reduction on the four noise conditions) as compared to the baseline system.

5. DISCUSSION

The present work represents a continued effort in the direction of moving away from the conventional *across spectrum processing* technique towards that of *across time processing*. This effort can be considered as an extreme generalization of multi-band ASR [1, 10, 2]. We present a complete ASR system based on this concept of independent processing of temporal trajectories and show that the system is competitive with the current conventional ASR system. We demonstrate the potential robustness of this system in noisy environments. We also show that the complementary information provided by the TRAP system can be further used to improve robustness of ASR by using the TRAP system in combination with the conventional system.

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. We would also like to acknowledge ICSI for providing the special purpose hardware (SPERT board) and software for training and testing the HMM/MLP hybrid classifiers. 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? *Speech Communication, Invited paper*, 25(1-3):3–27, 1998.

[8] H. Hermansky and S. Sharma. TRAPs - classifiers of temporal patterns. *Proc. ICSLP 98, to be published*, November 1998.

[9] S. Sharma, H. Hermansky, and P. Vermuulen. Combining information from multiple classifiers for speaker verification. *Proc. Speaker Recogniton and its Commercial and Forestic Applications, Avignon, France*, 1998.

[10] S. Tibrewala and H. Hermansky. Sub-band based recognition of noisy speech. *Proc. ICASSP 97*, II:1255–1258, 1997.

[11] S. van Vuuren and H. Hermansky. Data-driven design of RASTA-like filters. *Proc. Eurospeech 97*, pages 409–412, 1997.