SPEECH RECOGNITION FROM TEMPORAL PATTERNS

Sangita Sharma[1] and Hynek Hermansky[1,2]
[1]*Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA*
[2]*International Computer Science Institute, Berkeley, California, USA*
sangita,hynek@ece.ogi.edu

ABSTRACT

This paper explores the nature of linguistic information present in the temporal structure of speech. We derive TempoRAl Patterns (TRAPs) which describe the temporal evolution of different phonemes in conversational speech. The temporal spread of the TRAPs in the region preceding as well as following the occurrence of a phoneme suggests that the information regarding the phoneme is available in rather long time durations around it. We propose an approach using TRAPs for phonetic feature extraction in speech recognition instead of the conventional spectral-based features. The resulting temporal-based automatic speech recognition (ASR) system yields recognition performance which is comparable to that of the conventional spectral-based ASR system and it is inherently robust to certain types of noise.

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, w , o, a, e, i, y'* [11]. 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 systems. The speech signal is processed as a series of independent short-time (typically 10ms) frames to account for the non-stationary characteristic of the signal. Features such as filter-bank energies, linear predictive coefficients, mel-frequency cepstral coefficients (MFCC) [15] and perceptual linear predictive (PLP) coefficients [7] characterize the spectral envelope in each short-time frame

The problem with the spectrum of speech is that it can easily be modified by variety of relatively benign means such as the frequency characteristics of the communication channel or narrow-band noise. As a result, recognizers based on spectral features exhibit rapid degradation in performance in realistic communication environments.

1.2. Temporal processing

Mechanical properties of a live human cochlea suggest evidence that the human auditory system does some sort of spectral analysis [14]. Subsequently the first step in most speech processing systems is the short-term spectral analysis. At the same time psychoacoustic evidence in the form of the phenomenon of forward masking (exhibited by mammals) as well as the linguistic concept of syllable suggest that the peripheral auditory system in humans can integrate information from much larger time-spans than the temporal duration of the frame used in speech analysis (refer to [8] for review). This time span is hypothesized to be of the order of around 200ms. Many of the noise-robust techniques (such as dynamic cepstra [6], cepstral mean subtraction and RelAtive SpecTral processing (RASTA) [9]) also employ the temporal domain.

Experiments on articulation index conducted by Fletcher [5],[1] suggest that linguistic message gets decoded independently in several frequency sub-bands and the final decoding is based on merging the information from the reliable sub-bands. Further, the perceptual studies reported in [13],[10],[18],[12] which study the effect on intelligibility of filtered speech, point to the spectral redundancy in speech. These studies put into question the relevance of the spectral envelope to human speech perception.

The extreme position is taken in [8] which proposes that '*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. OUR APPROACH

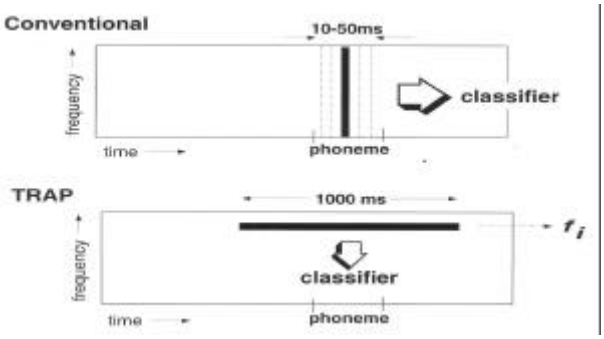

Figure 1: Temporal paradigm of ASR

Fig. 1 represents the proposed approach and also contrasts it with the conventional paradigm of speech recognition. Specifically, in contrast to doing phonetic classification based on spec-

tral features extracted from the entire frequency spectrum (and possibly augmented with dynamic features [6]) as in conventional recognition systems, we propose phonetic classification based on features extracted from rather long (as much as 1 sec) temporal trajectory of logarithmic energies from a single frequency (critical) band. The phonetic class is defined with respect to the center of this temporal trajectory.

Since this classifier based on temporal vectors is attempting to capture the appropriate temporal pattern from the acoustic stream, we call such a temporal sub-band classifier as TRAP (TempoRAl Pattern).

3. PATTERNS OF TEMPORAL EVOLUTION (TRAPS) OF PHONEMES

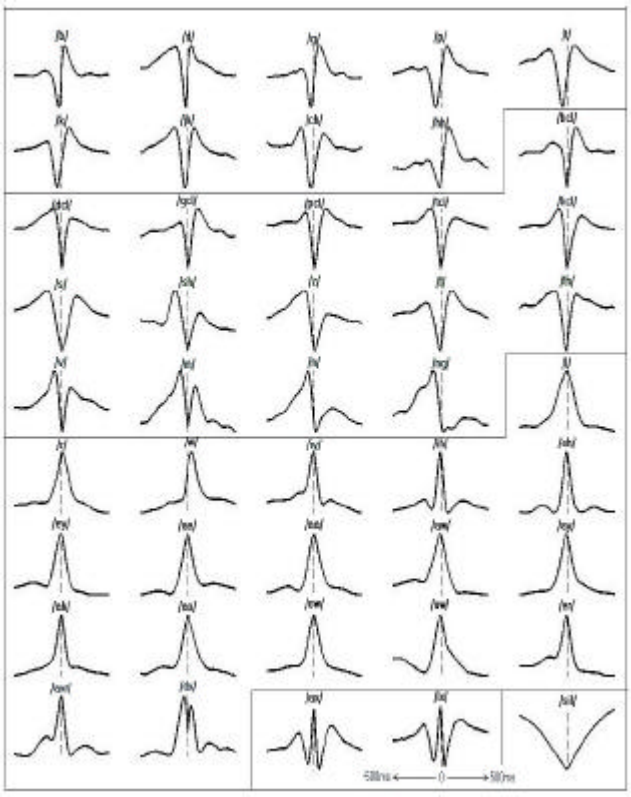

Figure 2: Mean TRAPs for 45 phonemes at the fifth critical band (~500 Hz). The dotted line in each of the TRAPs represents the center frame. The TRAPs within regions separated by solid lines are similar.

To understand the nature of linguistic information that is available in the temporal structure of speech, we analyze the time trajectories of spectral energy of the phonemic sound units typically used in ASR systems. We analyzed approximately 3 hours of phonetically hand-labeled conversational speech from the OGI-Stories corpus [4]. We considered the 45 most frequently occurring phonetic classes. 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 on the frame, which belongs to the phoneme under consideration. The mean of these vectors is then used to represent the pattern corresponding to the average temporal evolution of that class. It should be noted that only the center frames in all these vectors belong to the same phoneme. Other frames may belong to any other phonemes, which could surround the given phoneme.

Thus, the TRAP averages all occurrences of the given phoneme and of its phonetic environment in the fluent speech database used in the design. Then, both end-points of any TRAP are approaching the average spectrum of speech at the given frequency, and the center of the TRAP represents the average spectral value of the target phoneme at the given frequency. Values in between indicate the co-articulation pattern for the given phoneme, averaged over all possible phonetic environment of the given phoneme. As such, the TRAP can be also viewed as an average temporal pattern by which the given phoneme interacts with the neighboring phonemes.

Fig. 2 represents the mean temporal patterns (which we call Mean TRAPs) of the 45 phoneme classes. It can be seen from the figure that some of the classes have distinctly different Mean TRAPs, e.g. the vowel /ah/ compared to the stop-consonant /b/ and some of the Mean TRAPs are very similar.

3.1. Clustering of TRAPs

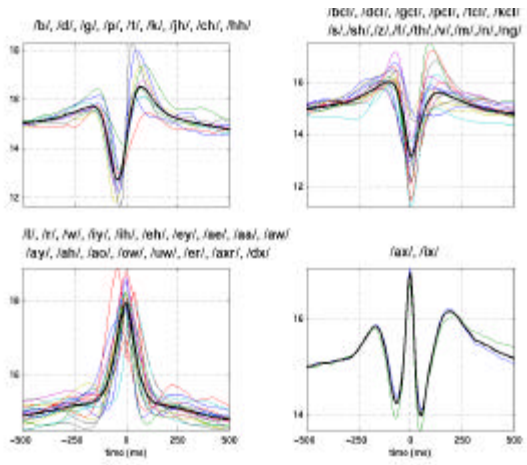

Figure 3: Broad TRAP clusters of the fifth critical band TRAPs.

Based on the observation of similarity of the TRAPs, we clustered them using a correlation-based similarity measure. The clustering results in 5 distinct clusters, indicated in Fig. 2 by solid lines. Fig. 3 represents four of the clusters, the fifth cluster consisting of the silence (/sil/) 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 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 80ms for vowels, the effect of the center phoneme lasts for approximately the same time, i.e. for about 200 ms in both time directions. This tempo-

ral spread in the evolution of the phoneme suggests that the information about the phoneme is not localized only in the duration of its occurrence, but is spread in rather long time durations around it.

These clusters have distinct and intuitive temporal patterns, e.g. the cluster 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 cluster has a dip off-center to the left, since a stop-consonant is usually preceded by a closure characterized by low energy.

4. CLASSIFICATION USING TRAPS

A simple way to use the TRAPs for phonetic classification is based on the template matching approach with correlation based similarity measure. Such a classifier is a simple linear correlation classifier. To classify a frame in a particular critical band, a 1 sec time trajectory around that frame is matched to each of the 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 remove the mean from each 1 sec time trajectory and weigh each trajectory by a Hanning window.

System	Frame Error (%) for each critical band
Baseline	21
Mean TRAPs	78 - 83
Neural TRAPs	66 - 74

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

We used the OGI-Numbers [3] database to test the classification performance of the TRAPs. This database consists of continuous strings of numbers which can be described by 29 phonetic classes. The TRAPs corresponding to these 29 phonemes derived from the OGI-Stories database are used in classification. Note that the database for deriving the TRAPS (OGI Stories) is different from the target database (OGI Numbers). The baseline system is a HMM/NN hybrid system from the International Computer Science Institute in Berkeley, CA, and it is a spectral-based system which uses PLP cepstral features along with delta and delta-delta features. This is currently our best system for use on the OGI Numbers database.

Table 1 gives the range of frame errors obtained in each of the 15 (Bark-spaced) critical frequency bands using the 29 Mean TRAPs in the template matching classifier (referred to as Mean TRAP-based classifier). It is encouraging to note that each critical band performance is approximately 80% error even for a simple TRAP classifier. This is significantly higher than chance (96.5% error for 29 classes) inspite of the fact that none of the TRAP classifiers have access to any information about spectral correlations with neighboring frequency bands.

A further significant improvement in each TRAP-classifier performance is obtained if we use a nonlinear (Multi-Layer Perceptron (MLP) based) TRAP (which we call a Neural TRAP) instead of the template-matching based correlation classifier. In addition to doing mean removal and windowing, we also normalize each 1 sec time trajectory to unity variance in order to improve robustness to noise. The use of Neural TRAPs improves the TRAP classifier performance from 80% error to around 70% error as seen in Table 1.

4.1. Combination of TRAPs

System	Frame Error	Word Error
Baseline	21 %	6.5 %
Mean TRAP-based	22 %	11.5 %
Neural TRAP-based	20 %	8.8 %
Combined Baseline and TRAP system		
Mean TRAP-based	19 %	6.0 %
Neural TRAP-based	17.8 %	5.8 %

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

There are about 15 critical bands in the range of telephone speech. Thus, at any given time we have at our disposal 15 outputs from 15 independent TRAP classifiers. Hence, for final classification, we can combine the outputs from all the TRAP classifiers. We use a non-linear classifier (MLP) for performing the combination. Table 2 compares the frame error and word error rate of the baseline system, the Mean TRAP-based and the Neural TRAP-based recognizers. It is seen that on the frame level, the performance of the baseline (spectrum-based) system and the Mean TRAP combiner is comparable; the Neural TRAP-based recognizer performs better. On the word level, the baseline recognizer yields the best performance.

A comparison of the frame errors of the baseline and the TRAP-based system indicates a trend in errors such that the TRAP-based system reduces the frame errors for fricatives and silence frames but yields errors higher than the baseline on vowels. This trend in errors is consistent with the notion that the vowels are better characterized by their spectrum and hence a spectral based system should classify them better.

4.2. Combination of spectral-based baseline system with the TRAP-based system

The analysis of frame errors also shows that the spectral-based baseline and the TRAP based system make significant amount (around 40%) of complementary frame errors. This indicates that a combination of the two systems can yield an improvement in performance. From Table 2 it is seen that the combination of the baseline system with the TRAP-based system (especially with the Neural TRAP recognizer) yields improvement in performance as compared to the baseline system.

5. EXPERIMENTS IN NOISY CONDITIONS

To assess possible advantages of the TRAP-based recognizer we investigated its performance in several artificially degraded situations. For speech corrupted with simulated convolutive noise, the TRAP-based system exhibits only a slight degradation in performance as compared to its performance on clean speech. This is achieved based on only 1 sec mean removal per frame without the need for utterance-based mean removal (which is

done in our baseline system) and shows that the TRAP-based system is inherently robust to convolutive channel distortions. For several additive noises from the NOISEX-92 database [17] (such as white, pink, destroyer-engine and factory noises) it is seen that the combined baseline and TRAP-based system yields on average a 13% reduction in word error rate as compared to the baseline system.

6. 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 [16], [2]. 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 demonstrate that by combining classification results from the individual critical bands one can achieve high recognition performance. It is further interesting to note that the conventional spectral-based system and the proposed TRAP-based system yield significant amount of complementary information. Hence, a combination of these two systems yields improvement in recognition performance especially in noisy conditions.

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.

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] R. Cole, M. Noel, and T.Lander. Telephone speech corpus development at CSLU. *Proc. ICSLP 94*, pages 1815–1818, Sep. 1994.

[4] R.A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at CSLU. *Proc. Eurospeech 95*, 1:821–824, Sep. 1995.

[5] H. Fletcher. *Speech and Hearing in Communication*, Krieger, New York, 1953.

[6] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on Acoustic, Speech and Signal Processing*, 29:254–272, 1981.

[7] H. Hermansky. Pereceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[8] H. Hermansky. Should recognizers have ears?, invited paper. *Speech Communication*, 25(1-3):3–27, 1998.

[9] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing* , 2(4):578–589, 1994.

[10] K.D. Kryter. Speech bandwidth compression through spectrum selection. *Journal of Acoustical Society of America*, 32(5):547–556, May 1960.

[11] P. Ladefoged. *Three Areas of Experimental Phonetics*, Oxford University Press, 1967.

[12] R.P. Lippmann. Accurate consonant perception without mid-frequency speech energy. *IEEE Transactions on Speech and Audio Processing*, 4(1):66–69, 1996.

[13] G.A.Miller and P.E. Nicely. An analysis of perceptual confusions among some English consonants. *Journal of Acoustical Society of America*, 27(2):338–352, 1995.

[14] Brian C.J. Moore. *An introduction to the psychology of hearing*, Academic Press, San Deigo, California, 1989.

[15] S.B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE ASSP*, 28:357–366, April 1980.

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

[17] Andrew Varga and Herman J.M. Steeneken. Assessment for automatic speech recognition II, NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–251, 1993.

[18] R.M. Warren, K.R. Riener, J.A. Bashford Jr., and B.S. Brubaker. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Perception & Psychophysics*, 57(2):175–182, 1995.