# SMOOTHED TIME/FREQUENCY FEATURES FOR VOWEL CLASSIFICATION

**Zaki B. Nossair and Stephen A. Zahorian**
Department of Electrical and Computer Engineering
Old Dominion University
Norfolk, VA, 23529

## ABSTRACT

A novel signal modeling technique is described to compute smoothed time-frequency features for encoding speech information. These time-frequency features compactly and accurately model phonetic information, while accounting for the main effects of contextual variations. These segment-level features are computed such that more emphasis is given to the center of the segment and less to the end regions. For phonetic classification, the features are relatively insensitive to both time and frequency resolution, as least insofar as changes in window length and frame spacing are concerned. A 60-dimensional feature space based on this modeling technique resulted in 70.9 % accuracy for classification of 16 vowels extracted from the TIMIT data base in speaker-independent experiments. These results are higher than any other results reported in the literature for the same task.

## Introduction

One of the fundamental issues in feature selection for automatic speech recognition is the representation of spectral/temporal information which best captures the phonetic content of the speech signal. Since time and frequency resolution are inversely related, in practice a tradeoff must be made between these two resolutions. Ideally these resolutions should also depend on frequency. Another important consideration, at least for any statistically-based recognizer, is the desirability of using as few features as possible. In this paper signal processing strategies are described to compute smoothed time-frequency features for encoding speech information of speech segments in a compact form.

There are at least two primary techniques which appear in the literature for modeling phonemes extracted from continuous speech. In one of these, each phoneme is represented by a sequence of feature vectors that are extracted from equally spaced frames of speech. Typically an HMM is used for modeling this sequence of feature vectors [1]. In the second method each phoneme is represented by three feature vectors extracted at the beginning, the middle, and the end of the labeled acoustic signal [2,3,4]. These feature vectors are then concatenated to form one longer fixed-length vector and then used as the input to a classifier such as a neural network. Neither of these methods is particularly effective at capturing the temporal history of the underlying features. Although the basic features are often augmented with some type of delta terms, the temporal modeling deficiency is only partially overcome.

In the present technique each acoustic segment is initially represented with multiple feature vectors which are extracted from equally spaced frames of speech. Then each feature trajectory across these multiple frames is represented by a low-order time-warped cosine basis vector expansion. The coefficients of these cosine expansions are then used as a representation for the underlying phonetic information. The use of this low-order cosine expansion over time enables modeling the temporal or dynamic information as well as the contextual information in a compact and integrated form. Using a weighted basis vector expansion over a sufficiently long acoustic segment, we are able to emphasize the center of the segment, which contains most of the information about the underlying phone, but also include contextual information from the neighboring phonemes. The use of only a few low-order terms in the expansion over time also smoothes out noise due to signal processing artifacts such as window length and frame spacing.

In the remaining sections of this paper we provide a detailed explanation for the feature computation procedures, the classification technique, experimental procedures and experimental results.

## COMPUTATION OF TIME/FREQUENCY FEATURES

The features are computed in a multi-stage process as follows. The first step is to high-frequency preemphasize the speech signal using a second order FIR filter given by

$$y[n] = 0.3426\, x[n] + 0.4945\, x[n-1] - 0.64\, x[n-2],$$

where the coefficients are for a 16 kHz sampling rate. This pre-emphasis, which has a broad peak near 3 kHz, and which approximates the inverse of the equal-loudness contour, results in slightly better performance (about 1%) than does a first order pre-emphasis, ($y[n] = x[n] - .95\, x[n-1]$ ). The next step of processing is to compute a 1024 point FFT from each Kaiser-windowed (coefficient of 5.33) frame of speech data. The magnitude spectrum is then determined, logarithmically amplitude scaled, and frequency warped with a bilinear transformation with a warping coefficient of .45. The scaled FFT spectrum is then reduced (or smoothed) using a cosine transform, computed over a frequency range of 75 Hz to 6000 Hz. These coefficients which we call discrete cosine transform coefficients (DCTCs), are essentially cepstral coefficients. Each DCTC trajectory is then represented by the coefficients in a modified cosine expansion over the segment interval as follows:

$$DCTC_i(n) = \sum_{k=0}^{k=M-1} DCS_{ik}\, BV_k(n)$$

where $DCTC_i(n)$ is the ith cosine coefficient of the magnitude spectrum of the nth frame,
$BV_k(n)$ is the nth value of the kth basis vector,
M is the number of basis vectors used,
and $DCS_{ik}$ is kth coefficient of the modified cosine expansion of $DCTC_i$.
The basis vectors are computed as "time-warped" cosine basis vectors, using a Kaiser-window weighting function, such that the data are more accurately represented in the center of the interval than near the endpoints, using

$$BV_k(n) = KW(n) \cos(Wk/L).$$

where KW(n) is the Kaiser-window weighting function,
L is the length of each trajectory (number of frames),

$$W = (0.5\, \pi /L) + \sum_{j=1}^{j=n-1} DW(j),$$

and $$DW(j) = (KW(j)+KW(j+1))\pi(L-1)/L(\sum_{m=1}^{m=L-1}(KW(m)+KW(m+1))).$$

Figure 1 depicts the first three basis vectors, using a coefficient of 5 for the Kaiser warping function.

The methodology described above allows considerable flexibility for examining tradeoffs between time and frequency resolution. For example to increase frequency resolution, the frame length should be increased and the number of DCTC terms should be increased, whereas the number of DCS terms can be reduced. To increase time resolution, the frame length and frame spacing and number of DCTC terms should be reduced, but the number of DCS terms should be increased. The tradeoff between the resolution of the representation at the center of the segment relative to the endpoints can be examined by varying the coefficient in the time-warping function. Also very importantly, the procedure results in considerable data reduction relative to the original features. For example, 15 features sampled at 30 frames (450 total features) can be reduced to 75 features if 5 basis vectors are used for each expansion. We conducted several experiments designed to evaluate the usefulness of these features for automatic vowel classification and to investigate the tradeoffs mentioned above.

**Figure 1** First three basis vectors used to encode trajectories of spectral features

## CLASSIFIER

The pattern classification approach used in this study is called a binary paired partitioning (BPP) neural network [5,6]. This classification approach partitions an N-way classification task into N*(N-1)/2 two-way classification tasks. Each two-way classification task is performed using a neural network classifier which is trained to discriminate one pair of categories. The two-way classification decisions are then combined to form the N-way decisions. For all experiments reported in this paper each pair-wise network was a memoryless, feed-forward, multi-layer perceptron and was configured to have one hidden layer of 5 nodes, unless otherwise stated, and one output node. Back-propagation was used for training these networks with 160,000 network updates using an initial learning rate of .45 and momentum term of .6. The learning rate was reduced by a factor of .96 every 5000 network updates.

## Experiments

The experiments were performed with vowel data from the DARPA/TIMIT data base (October 1990 version). The vowels used were /iy,ih,eh,ey,ae,aa,aw,ay,ah,ao,oy,ow,uh,ux,er,uw/. In this paper we give only test results based on the SX sentences using 499 speakers (356 male and 143 females) for training and 50 speakers for testing (33 male 17 females). The vowels and speakers used in these tests were the same as those used in previously reported tests with the TIMIT vowels ([2], [3]), in order to facilitate comparisons of our experimental results with previously published results.

### Experiment 1--Control

This experiment was conducted to determine the extent to which temporal information for vowels can be captured using only 1, 3, or 5 frames extracted at uniformly spaced time points with respect to the labeled section of each vowel. This method is very similar to that used in previous studies of vowel classification using TIMIT data. Ten DCTCs were computed for each frame (since, as described in the following experiment, this number was found to be sufficient). The vowel classification results for these three cases were 53.9%, 63.8%, and 65.4%. These results are similar to those obtained in previous studies for similar conditions.

### Experiment 2--Time/Frequency Resolution.

This experiment was designed to examine tradeoffs in time/frequency resolution. That is, we wanted to examine classification performance as a function of the number of DCTCs used and the number of DCSs used to encode each DCTC. For this experiment we used a fixed time interval of 300 msec centered at the labeled midpoint of each vowel and a time warping of 10. Figure 2 shows the evaluation results of this experiment in bargraph form. The results show that to some extent time and frequency resolution can be traded off in the sense that as more DCTCs are used (better frequency resolution), fewer DCSs are needed (poorer time resolution). For optimum results, 10 or more DCTCs are required. The overall best result of 70.9% s was obtained with 12 DCTCs and 5 DCSs (60 features), corresponding to a relatively smooth resolution in both time and frequency.

**ContainsDatafor**

**PostscriptOnly.**

**Figure 2** Vowel classification rate as a function of the number of DCTCs used and the number of DCSs used to encode each DCTC.

**Experiment 3 -- A More Detailed Examination of Temporal Resolution**

In this experiment we examined the effects of segment length and temporal resolution within the segment (midpoint versus endpoints) for vowel classification. We varied the segment duration from 50 msec up to 300 msec and the amount of time warping from 0 to 12. Except for the 50 msec segment, we also used a fixed number of features (10 DCTCs * 5 DCSs, or 50 features). For the case of the 50 msec time window we used 30 features (10 DCTCs * 3 DCSs), since there is very little temporal variation over this short interval. Table 1 shows performance as a function of amount of time warping for different window lengths. Note that for each duration, as the warping factor increases, the basis vectors increasingly emphasize the center of the interval, thus reducing the effective time duration of the window.

Table 1. Classification rates for 16 vowels as a function of the amount of time warping for different time window lengths.

| warping | Time Duration(ms) | | | |
|---|---|---|---|---|
| Factor | 50 | 100 | 200 | 300 |
| 0 | 63.7 | 67.1 | 69.6 | 66.2 |
| 2 | 64.4 | 65.8 | 70.3 | 69.0 |
| 4 | 62.9 | 66.9 | 69.3 | 69.3 |
| 6 | 61.5 | 65.5 | 68.7 | 69.8 |
| 8 | 61.0 | 64.4 | 67.9 | 70.5 |
| 10 | 61.1 | 63.6 | 68.5 | 70.4 |
| 12 | 61.4 | 63.8 | 67.8 | 69.3 |

The results in Table 1 clearly show that performance improves as the segment interval increases from 50 ms to

100 ms to 200 ms. There is only a slight improvement resulting from increasing the interval to 300 ms from 200 ms (70.5% versus 70.3%). However, as expected, as the segment becomes longer, the best results are obtained with a larger warping factor. The absolute best results were obtained using a warping factor of 8 and segment length of 300ms.

## Summary

A spectral/temporal feature set has been described for speech analysis and has been evaluated with vowel classification experiments. The spectral/temporal features result in substantially higher classification rates for vowels than can be obtained by simply concatenating multiple frames of static features. This new feature set has been used to obtain vowel classification results of 70.9% for 16 vowels of the DARPA/TIMIT data base, higher than any other previously reported results ([1], [4], [5]).

## References

[1] K.-F. Lee and H.-W Hon (1989), "Speaker-Independent Phone Recognition Using Hidden Markov Models," IEEE Trans. Acoust. Speech, Sig. Process. 37, 1641-1648.

[2] H. Leung and V. Zue (1988), "Some Phonetic Recognition Experiments Using Artificial Neural Nets," ICASSP-88, pp. I: 422-425.

[3] H. Leung and V. Zue (1990), "Phonetic Classification Using Multi-Layer Perceptrons," ICASSP-90, pp. I: 525-528.

[4] Z. B. Nossair and S. A. Zahorian (1991). "Dynamic Spectral Shape Features as Acoustic correlates for Initial Stop Consonants," J. Acoust. Soc. Am.- 89-6, pp. 2978-2991.

[5] L. Rudasi and S. A. Zahorian (1991), "Text-Independent Talker Identification with Neural Networks," ICASSP-91, pp. 389-392.

[6] L. Rudasi and S. A. Zahorian (1992), "Text-Independent Speaker Identification using Binary-pair Partitioned Neural Networks," IJCNN-92, pp. IV: 679-684.

Figure 1. First three basis vectors used to encode trajectories of spectral features.

Figure 2. Vowel classification rate as a function of the number of DCTCs used and the number of DCSs used to encode each DCTC.