

PHONE CLASSIFICATION WITH SEGMENTAL FEATURES AND A BINARY-PAIR PARTITIONED NEURAL NETWORK CLASSIFIER

Stephen A. Zahorian, Peter Silsbee, and Xihong Wang

Department of Electrical and Computer Engineering
Old Dominion University, Norfolk, Va 23529

ABSTRACT

This paper presents methods and experimental results for phonetic classification using 39 phone classes and the NIST recommended training and test sets for NTIMIT and TIMIT. Spectral/temporal features which represent the smoothed trajectory of FFT derived speech spectra over 300 ms intervals are used for the analysis. Classification tests are made with both a binary-pair partitioned (BPP) neural network system (one neural network for each of the 741 pairs of phones) and a single large neural network. Classification accuracy is very similar for the two types of networks, but the BPP method has the advantage of much less training time. The best results obtained (77% for TIMIT and 67.4% for NTIMIT) compare favorably to the best results reported in the literature for this task.

1. INTRODUCTION

Acoustic phonetic classification with the TIMIT data base is a standard benchmark for comparing different speech signal processing front ends and classification strategies. In this study, a compact spectral/temporal feature set is used along with a binary-pair partitioned neural network classifier to obtain a high level of performance for automatic phone classification with TIMIT. The analysis consists of first computing features which represent the smoothed spectral/temporal envelope of a 300 ms speech segment for each phone. Typically about 60 parameters are computed for each segment. The analysis is very flexible, allowing tradeoffs in time and frequency resolution, variable frequency resolution versus frequency, and variable time resolution within a segment. This basic analysis method and classification strategy has been previously applied to vowel classification, also using TIMIT [1,2]. The present paper is the first application of this method to a more comprehensive phonetic classification task— all the phones in TIMIT. The best test results obtained (77.0% for TIMIT and 67.4% for NTIMIT using the 39 phone classes of the MIT/CMU standard) are very similar to the most comparable best results reported in the literature [3]. To our knowledge they are the highest obtained for 39 phones using the NIST recommended

training and test sets. Although some studies have compared a cepstral analysis front end with an auditory model front end [4], and found the auditory model to perform better (particularly in noise), our work shows that a properly refined cepstral analysis performs just as well.

2. METHOD

Spectral Analysis

Phones were represented by 300 ms segments centered at the midpoint of each labeled phone. Second order preemphasis was then applied to each segment, which was then further divided into 10 ms frames, spaced 2 ms apart. Frames were multiplied by a Hamming window and the FFT was computed. This two-dimensional time-frequency data was morphologically dilated over frequency using an 80 Hz wide window and logarithmically amplitude scaled. Features were then extracted from a portion of the 2-D spectrum (75 Hz to 6000 Hz for TIMIT and 300 Hz to 4000 Hz for NTIMIT), as follows.

The features were based on a discrete cosine transform (DCT) of the log of the DFT, incorporating a frequency warping function, $F = g(f)$:

$$DCTC(i) = \int_0^1 X(f') \cos(\pi if') df'$$

where the desired frequency range is normalized to the interval [0, 1].

By changing variables, the equation can be rewritten as

$$DCTC(i) = \int_0^1 X(f) \Phi_i(f) df$$

where the original cosine basis vectors are modified to account for the warping effect resulting in the basis vectors

$$\Phi_i(f) = \cos[\pi ig(f)] \frac{dg}{df}$$

Figure 1 depicts the first 3 basis vectors, using this approach, based on a bilinear warping using a coefficient of .45. In practice, the DCTC terms are computed using sums involving FFT samples and samples of the modified basis vectors over a specified frequency range. These DCTC terms are thus very similar to cepstral coefficients typically used in speech analysis.

In a similar manner, the DCSC terms for several frames within a speech segment are also represented with a cosine expansion over time using

$$DCSC(i, j) = \int_{-1/2}^{1/2} DCTC(i, t') \cos(\pi j t') dt'$$

The variable $t' = h(t)$ is considered to be a "warped" version of t where the function h is chosen to emphasize the center section of the segment. The time interval is normalized to $[-1/2, 1/2]$.

By again changing variables, the equation can be rewritten as

$$DCSC(i, j) = \int_{-1/2}^{1/2} DCTC(i, t) \Theta_j(t) dt$$

where the basis vectors $\Theta_j(t)$ are the modified cosines for the segment interval

$$\Theta_j(t) = \cos[\pi j h(t)] \frac{dh}{dt}$$

Figure 2 depicts the first 3 basis vectors, using a Kaiser window for dh/dt , with a Kaiser factor of 10, for a segment spanning 300 ms.

In effect, this two step procedure filters the log magnitude spectrum over frequency and then over time. It can be viewed as a partitioned 2-D filter, whose outputs are the features used for classification. The FIR filter coefficients (basis vectors over frequency and time) are designed so that the features represent the spectrum with non-uniform frequency and time resolution. In particular the frequency resolution approximates a

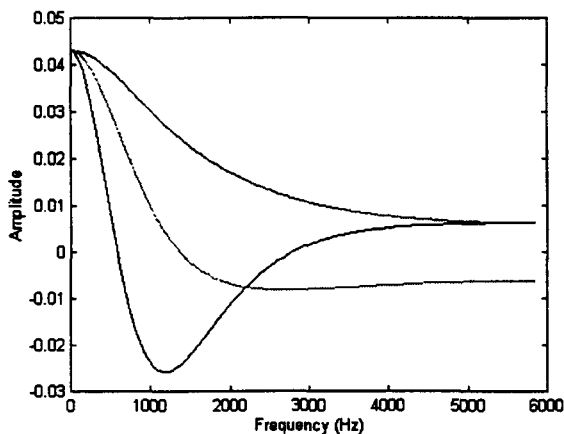


Fig.1 First three basis vectors over frequency

Bark scale and the time resolution is highest near the center of each segment and smoothly tapered to each endpoint. From another point of view, the features encode DCTC (cepstral coefficient) trajectories over each segment. As in our previous work, we refer to the terms over frequency as DCTCs and the terms over time as DCSCs. All features were scaled for a mean of 0.0 and standard deviation of .2 prior to classification.

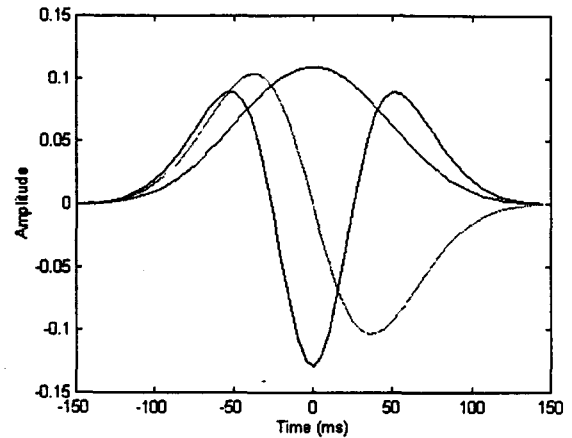


Fig. 2 First three basis vectors over time.

3. CLASSIFIER

All classifiers used in these tests were multilayer perceptrons with one hidden layer trained with backpropagation. For the case of the binary-pair partitioned (BPP) system, 741 classifiers, each with 10 hidden nodes and 1 output node, were trained with 200,000 updates. For the large neural network (LNN), a network with 500 hidden nodes was trained with 2,000,000 updates.

These network configurations were determined from experiments with developmental data, and achieved good performance with relatively fast training times. The initial learning rate was .45 for the BPP networks, and .15 for the LNN networks. The update rates were gradually reduced during the training iterations (using a multiplicative factor of .96 for every 5000 network updates for the BPP case, and a factor of .96 for every 78000 networks for the LNN case).

For the case of the BPP method, final decisions were made by combining the 2-way decisions from each of the networks. In particular, for each phone category, there were 38 classifiers used to discriminate that phone from each of the other 38 possible phones. By summing these 38 outputs, for each of the 39 phone categories, an estimate of the probability of each phone was obtained. The phone with the highest probability measure was then selected as the most probable phone.

4. EXPERIMENTAL TESTING

Data Base

All experiments were done with the TIMIT and NTIMIT data bases. Training was based on SX and SI sentences using the full training set (528 speakers, 3696 sentences, 131,890 tokens). All phones were included in the analysis. Experiments were conducted for 39 phone groups, using the CMU/MIT standard for combining phone categories [4]. For the results reported in this paper, the core test set developed by NIST (24 speakers, 168 sentences, 6790 tokens) was used. Development work was done with the 499 speaker training set, and 50 speaker test set used by [6].

Experiment 1

For each data set, 60 total features were used, with 5 terms encoding the trajectory of each of 12 cepstral coefficient terms. Except for the frequency range differences, as noted above, processing was identical for both cases. Table 1 gives test results for the TIMIT and NTIMIT data using the BPP and LNN classifiers. Test results were scored considering (a) only the top choice of the classifier, and (b) the top 3 choices of the classifier.

Test results for the TIMIT data are over 75% correct, considering only the top choice, and over 94%, using the top 3 choices of the classifier. The performance of the BPP and LNN classifiers are very similar. However the LNN classifier required about 8 times as much training time (65 hours versus 8 hours on a Pentium 133). For the NTIMIT data, the BPP classifier performance was superior to that of the LNN, with both much lower than that obtained with the TIMIT data. For example, for condition (a), the BPP results in about 67% versus 61% for the LNN. (Generally, in results with development data, the LNN and BPP results were much closer.)

Table 1. Phone classification results obtained with 60 features.

Data	Method	(a) Top 1	(b) Top 3
TIMIT	BPP	76.5%	94.5%
TIMIT	LNN	77.0%	93.8%
NTIMIT	BPP	66.6%	89.5%
NTIMIT	LNN	60.8%	82.9%

Experiment 2

Several other conditions were examined, in terms of the number of DCTCs and number of DCSCs for the both the NTIMIT and TIMIT data. In particular, the number of DCTCs was varied from 7 to 12, and the number of DCSC terms varied from 4 to 8 (but with total features varying from 48 to 96). We hypothesized that for the NTIMIT data, better performance might be achieved by increasing time resolution (i.e., more DCSC terms) and reducing frequency resolution (fewer DCTC terms). These cases were investigated only with the BPP classifier, since this classifier required much less training time than the LNN classifier. For all

cases tried, classification accuracy for NTIMIT ranged from 64.4% to 67.0%, and from 74.0% to 76.6% for TIMIT. Thus the classification accuracy was quite insensitive to the number of total terms used and to tradeoffs between time and frequency resolution.

Nevertheless, two attempts were made at feature "optimization." In the first of these, 35 features from a set of 96 (12 DCTC trajectories, each encoded with 8 DCS terms) were selected for each pair of phones, to optimally discriminate that pair of phones using a Gaussian Bayes classifier for feature evaluation. A modified add-on procedure was used [7,8] to determine feature sets. The classification results for TIMIT data, using the 35 selected features, are given in the first row of Table 2. In the second attempt, 60 features were chosen as indicated in Table 3. These particular features, which were used for all phone pairs, were chosen based on the features typically selected from the automatic feature ranking experiment for the 741 phone pairs. Results using these "best" 60 features are listed in the remaining rows of Table 2, for the TIMIT and NTIMIT data respectively. The results based on the 35 features are slightly inferior to those listed in Table 1. The results based on the 60 "best" features are in fact only very slightly different (but generally higher) than for the 60 features used in experiment 1. Thus the main conclusion of experiment 2 is that phonetic classification accuracy, using the features analysis methods described, and the BPP classifier is insensitive to the number of features used. Also many combinations of DCTC and DCSC terms yield very similar performance.

Table 2. Phone classification results for 39 phones for four cases.

Case		Top 1	Top 3
TIMIT/BPP	Best 35	75.7%	94.6%
TIMIT/BPP	Best 60	76.4%	94.7%
TIMIT/LNN *	Best 60	76.3%	93.4%
NTIMIT/BPP	Best 60	66.8%	90.1%
NTIMIT/LNN	Best 60	67.4%	88.1%

(* This case only trained for 1.1 million updates)

Table 3. List of 60 "best" features.

DCTC	0	1	2	3	4	5	6
#DCSC	5	6	9	9	7	6	5

DCTC	7	8	9	10	11	12	
#DCSC	4	3	2	2	1	1	

5. CONCLUSIONS

The results obtained with the signal analysis and partitioned neural network classifier used in this study are among the best reported for TIMIT and NTIMIT. Unfortunately, despite the use of a "standard" data base, it is still difficult to compare results because of differences in selection of training and test data, and also because of differences in selection of phone groups. The results in this paper are based on the complete training and core test sets, which we believe are the most reasonable choices for comparisons of phonetic classification methods with TIMIT. We also advocate the use of 39 phones as per MIT/CMU (or else the full set of phones). To our knowledge the results reported in this paper are the highest obtained for 39 phones using the NIST recommended training and test sets. The most directly comparable results appear to be those of [3] (about 78% classification for 39 phones from TIMIT, and 69.2% classification for 39 phones from NTIMIT, both using 610 training speakers and 20 test speakers) and [9] (about 72% classification for 39 TIMIT phones using articulatory based modeling).

The signal analysis used to obtain these results is quite conventional (i.e., FFT derived cepstral coefficients), except for several refinements as reported previously in studies of vowel classification [2, 10], and as summarized in this paper. Additionally, a second step of analysis is used to compute a compact set of segment features from a collection of frame based features. The final features thus represent spectral/temporal information in an integrated form. The analysis also performs well for telephone speech, provided small changes are made in the computations such as a reduced frequency range. Thus our analysis appears to be as robust to noise as are the auditory model based front ends advocated in some studies [4].

The binary-pair partitioned classifier, which we have previously shown to be effective for automatic speaker identification [11] and more limited phone classifications tasks, scales up well to large phone classification tasks. The system of neural networks used for the present work was found to perform comparably to a large neural network classifier, but requires much less training time. The rapid training time enables more thorough examination of different configurations of signal processing front ends. The partitioned classifier also allows for convenient individual optimization of features and classifier for every pair of phones. Using feature optimization techniques for each pair of phones, classification obtained using 35 features for each phone pair was obtained nearly equivalent to that obtained using 60 features.

ACKNOWLEDGMENTS

Portions of this work were supported by NSF grants IRI-9217436 and BES-9411607.

6. REFERENCES

- [1]. Zahorian, S. A., Nossair, Z. B., and Norton, C. A. III, "A Partitioned Neural Network Approach for Vowel Classification Using Smoothed Time/Frequency Features," EUROSPEECH-93, pp II:1225-1228, 1993.
- [2]. Nossair, Z. B., Silsbee, P. L., and Zahorian, S. A., "Signal Modeling Enhancements for Automatic Speech Recognition," ICASSP-95, pp. 824-827, 1995.
- [3]. Leung, H. C., Chigier, B., and Glass, J. R., "A Comparative Study of Signal Representations and Classification Techniques for Speech Recognition," ICASSP-93, pp II-680 - II-683, 1993.
- [4]. Meng, H. and Zue, V., "A Comparative Study of Acoustic Representations of Speech for Vowel Classification Using Multi-Layer Perceptrons, ICSLP-90, 1053-1056, 1990.
- [5]. Lee, K. F., and Hon, H. W., "Speaker Independent Phone Recognition Using Hidden Markov Models," IEEE Trans. ASSP, Vol. 37, No. 11, 1641-1648, 1989.
- [6]. Goldenthal, W. D., and Glass, J. R., "Modeling Spectral Dynamics for Vowel Classification," EUROSPEECH-93, pp 289-292, 1993
- [7]. Parsons, T., *Voice and Speech Processing*, McGraw-Hill Book Company, 1987
- [8]. Zahorian, S. A. and Jagharghi, A. J., "Spectral-Shape Features versus Formants as Acoustic Correlates for Vowels" J. Acoust. Soc. Am. 94-4, pp 1966-1982, (1993).
- [9]. Deng L. And Sun, D., "Phonetic Classification and Recognition Using HMM Representation of Overlapping Articulatory Features for all Classes of English Sounds" , ICASSP-94, 1994
- [10]. Silsbee, P. L., Zahorian, S. A., and Nossair, Z. B., "A Warped Time-Frequency Expansion for Speech Signal Representation, " IEEE-SP Intl. Symp. on Time-Frequency and Time-Scale Analysis, 1994.
- [11]. Rudasi, L. and Zahorian, S. A., "Text Independent Talker Identification with Neural Networks, ICASSP-91, pp. 389-392, 1991