A PARTITIONED NEURAL NETWORK APPROACH FOR VOWEL CLASSIFICATION USING SMOOTHED TIME/FREQUENCY FEATURES

Stephen A. Zahorian, Zaki B. Nossair, and Claude A. Norton III

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

(This is close to version of text submitted in paper to Eurospee, but there are a few editorial changes from this to the final version)

1. Introduction

Accurate speech recognition requires both features that are highly discriminative with respect to the categories of interest and a classifier which can form arbitrary boundaries in the feature space. Performance depends on both the features and the classifier, with some classifiers better compensating for scaling, interdependence, etc. of the features. However, for a given feature set there is an upper bound on possible classification accuracy, dependent on separability of the categories with respect to the features. Another important consideration is that statistical models that contain many parameters which must be estimated require extremely large amounts of training data. Since the number of model parameters generally increases with the number of features, this consideration often manifests itself in the so-called "curse of dimensionality" [2]. That is, for a fixed set of training data, classifier performance improves on the training data as additional features or dimensions are added, but degrades on test data. Therefore a compact set of highly discriminative features is preferred to a large set of features which (potentially) contain the same information. Some classifiers are more powerful than others, not only in the ability to form more complex decision regions in a feature space, but also in terms of the ability to utilize higher-dimensionality feature spaces.

In this paper we describe two features sets for speech recognition, both based on a cosine transform of the magnitude power spectrum. The first set encodes only static spectral information whereas the second set is obtained by combining features computed over several frames. These new features, which encode the trajectories of the initial features, provide more time resolution at the center of the segment and less at the endpoints. We show that this incorporation of temporal information substantially improves classifier performance.

We also discuss a classification method which is developed around the elementary yet fundamental idea that a successful classifier must be able to resolve differences between every pair of categories. In particular this classification approach uses a system of pair-wise classifiers, with one elementary classifier for each pair of categories that are to be distinguished. With this technique, both the classifier and features can be individually optimized for each pair-wise discrimination task. We describe the structure of this classification method in more detail and experimentally demonstrate the benefits of this approach using vowel classification experiments.

2. Features

To illustrate some of the interactions between classification performance and the features used, experiments were performed with two feature sets and with varying number of features from each set. The first step of processing was always to compute a 1024 point FFT from each 20 ms Kaiser-windowed (coefficient of 5.33) frame of speech data. The next step in processing was to compute the coefficients in a cosine transform of the scaled magnitude spectrum. The coefficients were computed over the frequency range of 75 Hz to 6000 Hz.

Feature Set 1 For this case a single frame of data, positioned at the labeled center of each vowel token, was used. Twenty-five coefficients in a cosine transform of the magnitude spectrum were computed as the features for this set. These cosine coefficients were obtained using log amplitude scaling and bilinear frequency warping (coefficient of .45) of the spectrum. Thus these coefficients are essentially cepstral coefficients.

Feature Set 2. The motivation for computing these features was to compactly represent both spectral and temporal information useful for vowel classification. These features encode the trajectory of the smoothed short-time spectra, but with the central region more heavily weighted than the endpoints. Using the processing as described for the first feature set, 15 cepstral coefficients were computed for each of thirty equally-spaced frames of data spanning 300 ms centered on the middle of each token. These 450 features were reduced in several steps to 58 total features as follows. First the time trajectory of each of the 15 cepstral coefficients was represented by an eight-term cosine time expansion. The time expansion cosine basis vectors were first warped using a Kaiser-window weighting function (coefficient = 3), such that the cepstral coefficient data are more accurately represented in the center of the interval than near the endpoints. Figure 1 depicts the first three "warped" basis vectors. Feature ranking experiments, as discussed in the experimental sections were used to reduce these 120 features (15 cepstral coefficients * 8 time expansion terms) to 58 features for use in additional experiments.

3. Classification Methods

The primary classification method used in this study, which we call binary-pair partitioning (BPP), partitions an N-way classification task with N * (N-1)/2 "elemental" classifiers, each of which discriminates a particular pair of categories. With this approach, there are two distinct steps in the classification process: (1) "Elemental" classifiers are trained to discriminate between two categories (i.e., vowels for the results presented in this paper). (2) The binary decisions from step one must be combined to form the final "top-level" N-way decision. We discuss these points in the next few paragraphs and then illustrate the method for a vowel classification task. A similar method has previously been used for automatic speaker identification ([7], [8]).

In the first step of the procedure, elemental classifiers are trained using all available training data from each category. Although any type of classifier can be used, in the experiments reported in this paper, all elemental classifiers consisted of feedforward neural networks with one hidden layer and one output node. These networks were trained with back-propagation to discriminate between a pair of vowels, using training data selected only from those vowels.

The results from elemental classifiers must be combined to form the final decision. We first note that the output of each classifier is an estimate of the conditional probability of a particular category given one of two possibilities [3]. Thus, the probability of the jth category, conditioned on the outputs of all binary classifiers, is obtained by simply summing the outputs of all binary classifiers trained to discriminate that category. The final classification rule is to decode the category with the largest sum, which corresponds to the maximum posteriori probability, or minimum probability of error.

To illustrate both the potential and limitations of the BPP classifier, two control classifiers were also used for comparison. One of these was a "large" feedforward fully interconnected neural network with one hidden layer and one output node for each category. This type of network, trained with back propagation, is typically used for pattern recognition with neural networks. The other control classifier was a Gaussian full-covariance matrix maximum likelihood classifier (MXL), which is theoretically optimum if the features are multi-variate Gaussian [2]. For this classifier, each category is "modeled" by its mean vector and covariance matrix.

4. Experiments

Several experiments were conducted to investigate the features described above and to illustrate the BPP classifier for vowel recognition. We examined performance as a function of the number of input features, number of hidden nodes, learning rates, spectral versus spectral/temporal features, amount of neural network training, etc. For each feature set, experiments were also conducted with the two control classifiers mentioned above. In this section we report results for two of these experiments. Note that for the results reported here, the large neural network contained 250 hidden nodes whereas each binary neural network contained 35 hidden nodes. These values, as well as the number of training iterations, were obtained from other experiments, not reported in this paper.

The experiments were performed with the vowel data of all speakers 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,ax/. All vowel tokens from all the training sentences (3260 male sentences and 1360 female sentences) were used for training and all tokens from all the testing sentences (1120 and 560 for male and female speakers respectively) were used for testing.

Approximately 43000 vowel tokens were used for training and 10000 tokens (from different speakers) were used for testing.

Feature evaluation techniques, using the methods described in Nossair and Zahorian [6], were used to find subsets of features with high discrimination power. This feature ranking procedure was used to determine a subset of 58 features from the 120 features computed over the 300 ms section of speech data. In particular, the feature ranking algorithm was used to find the 30 best features for each pair of vowels (120 pairs total). These 120 feature sets were pooled to form a set of 58 features, which contained all of the features found in any individual set. Within each group of 30 features, the features were ranked in terms of importance. This was done so that in experiments with the BPP classifier, a comparison could be made of results based on features "tuned" for each pair, versus a common set of features for all subclassifies. The feature ranking algorithm was also used to find a set of 30 common features for discriminating the 16 vowels. These features were also rank ordered and used in experiments. Feature ranking was based strictly on the training speakers.

Figure 2 depicts the test classification results for the three classifiers, as a function of the number of features used, for the first feature set. Figure 3 depicts test results for feature set 2. In Fig. 3, for each of the three classifiers, the features were arranged in the order selected by the feature ranking algorithm for classifying the 16 vowels. Thus, for example, the result based on six features, was obtained using the "best" six features from the feature ranking algorithm. In addition, for the BPP classifier, another set of classification results was obtained using highest ranking subsets of features, individually selected for each of the 120 binary pair classifiers (BPP2).

Several observations regaring the features and the BPP classifier are as follows:

1. For both feature sets and for every number of features, the BPP test results are higher than for either of the other two classifiers. The large neural network outperforms the Gaussian classifier for large numbers of features, but is poorer than the Gaussian classifier if only a few features are used (less than 20).

2. The classification results obtained with the second feature set are uniformly higher, for all classifiers and every number of features, than were obtained with the first feature set. The overall best rate of 72.4% is substantially higher than the best rate of 53.5% obtained with feature set one. Clearly, the temporal information included in feature set two benefits vowel classification.

3. The use of features individually optimized for each binary pair classifier enhances performance over that obtained with a common set of features. The difference in performance is largest for two features (57.4% classification rate versus 44.1%), and decreases to a very small level (71.5% versus 70.5%) as the number of features is increased to 30. This decrease in performance level as the number of features is increased is not surprising since the percentage of features which differ from classifier to classifier steadily decreases as the size of the feature set increases.

4. The classification rate of the BPP classifier increases as the number of features increases, for the second feature set. For the first set, performance "saturates" at about 15 features, thus indicating additional spectral detail is of no benefit. The overall best rate of 72.4% was obtained

using 58 spectral/temporal features. Thus the BPP classifier is able to use a large number of features effectively. Presumably even better results could be obtained if additional features were used.

5. Conclusions

A spectral/temporal feature set and a neural-network based classifier structure, which we call binary-pair partitioning, have been described and evaluated with vowel classification experiments. The spectral/temporal features result in substantially higher classification rates for vowels for all classifiers tested. The BPP classifier, which is comprised of separate classifiers for every pair of discriminations which must be made, achieved higher vowel classification rates than either of the two control classifiers used. This new classifier has been used to obtain vowel classification results of 72.4% for 16 vowels of the DARPA/TIMIT data base, higher than any other previously reported results ([1], [4], [5]). One advantage of the BPP classifier structure is that this structure provides a convenient framework for selecting and optimizing features separately for each pair of categories. The BPP classifier appears able to utilize a larger number of features than either of the other two classifiers. These properties make the BPP classifier very attractive for other applications, including a recognition system for all phones.

References

[1] D. J. Burr (1992), "Comparison of Gaussian and Neural Network Classifiers on Vowel Recognition using the Discrete Cosine Transform," ICASSP-92, pp. II: 365-368.

[2] R. O. Duda and P. E. Hart (1973). Pattern Analysis and Scene Classification (John Wiley & Sons, New York, 1973)

[3] H. Gish (1990). "A Probabalistic Approach to the Understanding and Training of Neural Network Classifiers," ICASSP-90, 1361-1364.

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

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

[6] 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.

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

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

Figure captions:

Figures should be inserted in text at first convenient place after they are mentioned. They should be sized to be square and to be one column wide.

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

Figure 2. Automatic vowel classification results for 16 vowels using varying number of spectral features for three classifiers.

Figure 3. Automatic vowel classification results for 16 vowels using varying number of spectral/temporal features for three classifiers. The curve labeled BPP1 using features jointly optimized for all elemental classifiers whereas the curve labeled BPP2 is for features individually optimized for each elemental classifier.

Table 1. Confusion matrix for test vowels. Rows are actual vowels and columns are vowels identified by classifier.

	iy	ih	eh	ey	ae	aa	aw	ay	ah	ao	oy	ow	uh	ux	er	ax
iy	88.6	5.2	0.2	2.9	0.1			0.1		0.2				1.8	0.3	0.5
ih	7.5	66.5	9.2	3.9	1.1			0.5	1.9	0.2		0.6	1.8	1.9	0.9	4.0
eh	0.7	11.5	61.1	2.9	9.6	0.6	0.5	1.2	5.6	0.8	0.1	0.9	0.3	0.2	2.1	1.8
ey	8.4	7.2	3.7	72.3	2.3			3.1	0.3	0.7	0.1	0.5	0.1	0.7	0.3	0.4
ae	0.5	1.5	11.4	1.1	76.5	1.5	1.4	1.8	2.4	0.1		0.2	0.1	0.5	0.5	0.7
aa		0.1	0.6		0.8	67.7	2.2	4.5	4.8	16.8	0.2	0.8		0.1	0.7	0.7
aw	0.5	0.5	3.8		10.0	17.1	52.9	1.4	5.7	1.4		6.2				0.5
ay	0.5	0.1	1.4	1.9	2.6	7.9	1.0	78.1	3.1	1.2	0.6	0.1		0.1	0.7	0.6
ah	0.4	2.1	11.2		3.6	8.2	1.9	3.3	46.5	3.3	0.1	6.5	1.1	0.4	0.8	10.8
ao	0.6	0.3	0.3	0.1	0.7	14.6	0.5	0.7	1.7	71.0	1.4	5.3	0.7	0.2	0.6	1.3
oy	0.4	0.8	0.4	1.2		1.6		4.1	0.8	12.2	75.2	1.6			0.4	1.2
ow	0.3	0.6	2.1	0.3	0.7	1.7	2.3	0.3	4.4	6.7	0.6	75.2	0.7	0.6	0.7	3.0
uh		19.1	4.4		1.0	1.0			11.8	2.9	0.5	3.9	31.4	3.4	2.5	18.1
ux	10.1	9.7	0.7	0.4	0.4				0.4	0.2		0.9	0.4	72.6	2.7	1.6
er	0.1	2.9	2.3	0.5	1.4	1.2		1.0	1.4	1.0	0.1	0.7	0.3	0.8	84.2	2.1
ax	0.8	6.0	3.3		0.5	0.8	0.2	0.5	7.2	1.2	0.3	2.3	1.1	0.9	0.6	74.3