In *International Joint Conference on Neural Networks*, Baltimore, 1992.

Feature Selection for Improved Classification

Fred E. Shaudys and Todd K. Leen

Dept. of Computer Science and Engineering
Oregon Graduate Institute of Science and Technology
19600 N.W. von Neumann Drive, Beaverton OR 97006-1999

Abstract

We apply the feature-selection technique of Fukunaga and Koontz, an extension of the Karhunen-Loève transformation, to spoken letter recognition. Feedforward networks trained for letter-pair discrimination with the new features show up to 37% reduction in classifier error rate relative to networks trained with spectral coefficients. This performance increase is accompanyed by a 91% reduction in feature dimension. For three-letter discrimination, the new features perform comparably to spectral coefficients, with a 90% reduction in feature dimension.

1 Introduction

It is generally recognized that compact feature sets ameliorate the computational burden for classifiers. Researchers have used algebraic and neural network implementations of principal component analysis to compress high-dimensional data for speech [1], acoustic emission [2], and vision [3, 4] problems.

Beyond compression, statistical feature-selection techniques should provide features with good discriminatory power. Since principal component analysis, or the Karhunen-Loève (KL) transformation, is based on a reconstruction criteria, there is no guarantee that discriminatory information is retained in the compression. Fukunaga and Koontz [5] (FK) have developed a feature selection technique aimed at extracting features that are important for classification [6, 7, for additional detail]. The technique is a supervised extension of the KL transformation.

In this paper, we apply the FK transformation to spectral data from spoken letters. We compare the performance of classifiers trained on the raw spectral data, on principal components, and on the FK features.

2 Feature Selection

The FK transformation proceeds as follows. We assume that there are m classes of objects. From the training data, we construct the correlation matrix

$$R \;=\; E[x\,x^T] \;=\; \sum_{I=1}^{m} P_I\, R_I \tag{1}$$

where E is the average over the training data, x is the N-dimensional input vector, P_I is the a priori probability for the I^{th} class and R_I is the correlation matrix for the I^{th} class. The unit norm eigenvectors of the R are denoted e_i, $1 \leq i \leq N$ with corresponding eigenvalues λ_i. Next construct the matrix T whose i^{th} row is $e_i/\sqrt{\lambda_i}$, and transform the original vectors under T,

$$y \;=\; T\,x. \tag{2}$$

It is straightforward to verify that $E[y\, y^T] = I$. Next construct class correlation matrices from the vectors y,

$$C_I = P_I\, E_I[y\, y^T], \tag{3}$$

where E_I is the average over the vectors in class I. Clearly

$$\sum_{I=1}^{m} C_I = E[y\, y^T] = I. \tag{4}$$

Next the eigensystem for C_I is found. Let f^I_j and λ^I_j denote, respectively, the j^{th} eigenvector and eigenvalue of C_I. From (4) it follows that

$$\bar{C}_I\, f^I_j \equiv (\sum_{K \neq I} C_K)\, f^I_j = (I - C_I)\, f^I_j = (1 - \lambda^I_j)\, f^I_j \equiv \bar{\lambda}^I_j\, f^I_j. \tag{5}$$

Hence the f^I_j are eigenvectors of both C_I and the matrix $\bar{C}_I$ formed by summing all the class correlation matrices except C_I. The eigenvalues of these two matrices are related by

$$\bar{\lambda}^I = 1 - \lambda^I. \tag{6}$$

Since C_I and $\bar{C}_I$ are positive semidefinite, (6) requires that the eigenvalues λ^I and $\bar{\lambda}^I$ all lie between 0 and 1. Owing to (6), eigenvectors corresponding to large eigenvalues for C_I correspond to small eigenvalues for $\bar{C}_I$. Hence, directions for large variance of class I are directions of small variance for the other classes.

This last result forms the basis for selecting discriminatory features. To simplify the discussion, assume for now that there are two classes of objects, $I = 1, 2$. Select those eigenvectors f^1_j corresponding to eigenvalues λ^1_j close to unity and close to zero. Project the y vectors onto this set of eigenvectors to recover the final feature vectors z. The set of z-vectors drawn from class 1 will have large variance along the directions corresponding to large λ^1, and small variance along the directions correspond to small λ^1. The reverse is true for vectors from class 2. Thus the $z-$vectors from the two classes lie near orthogonal subspaces.

The number of eigenvectors chosen is a parameter in the design space, but will typically be far less than the dimension N of the original data. Hence, like dimension reduction by principal component analysis, the FK features form a more compact set than the original data. However these features should have more discriminatory power than the principal components.

3 Application to Spoken Letter Recognition

3.1 Data Description

The aim of this study was to improve the performance of a neural network spoken letter recognition system. The data for our experiments was drawn from the spoken letter database described in [8]. The database consists of 7800 examples of English letters spoken in isolation. There are two utterances of each letter (A-Z) by each of one hundred fifty speakers. Each utterance in the database is digitized at 16 kHz sampling rate and lowpass filtered at 7.6 kHz. For our experiments the data was broken into subsets consisting of utterances from thirty speakers, fifteen male and fifteen female.

Discrete Fourier transform (DFT) coefficients serve as the raw features for this study. First each utterance was segmented into a sequence of broad phonetic categories using a rule-based segmenter. The segmenter identifies closure, sonorant, fricative and stop segments of the utterance [9]. Then, from the digitized speech, we obtained a 128-point DFT computed over a 10 msec Hanning window, every 3 msec.

We compare the performance of three-layer backpropagation networks trained for letter recognition using the DFT data, principal components of the DFT data, and features selected by the FK algorithm. The backpropagation networks were trained using conjugate gradient optimization [10]. The next two subsections describe the details of the feature selection, and the classifier results.

3.2 Two-letter Discrimination

We chose three letter pairs for study: (F,H), (A,E) and (M,N). Here we report only the results for (M,N), the most challenging pair. Only the sonorant segment of the data is required to discriminate between these letters. The sonorant was divided into four equal length portions. The lowest 32 DFT coefficients, spanning the range 0-4kHz, were time-averaged over each of the four portions of the sonorant. The remaining 32 spectral coefficients, spanning the range from 4-8kHz, were averaged over 8 frequency bands for each 3msec time slice. These were then time-averaged over each of the four portions of the sonorant. This left a total of 160 features for the entire sonorant.

3.2.1 Feature Generation

Four-fifths of the data (120 speakers) served as training data and one-fifth was retained as test data. The training data was used to generate the correlation and class-correlation matrices required for the FK algorithm. The eigenvectors f_j^I recovered from these correlation matrices were used to generate feature sets for both the training and test data. Thus, test set features were generated based only on the statistical properties of the training data.

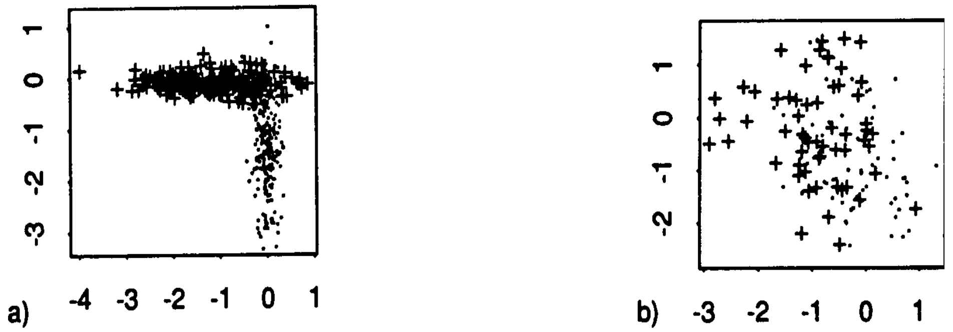

Figure 1: Scatter plots of the first two features for the letters M (+) and N (·).

Figure 1 shows scatter plots of the first two FK features, corresponding to the highest and lowest λ^1, for the letters M and N. The training data is plotted in figure 1a and the test data in figure 1b. For the training data, the two classes are nearly contained in orthogonal subspaces. For the test data, on the other hand, there is no clear separation. We found that classifiers trained on a set of FK features generated by the algorithm as described in section 2 performed poorly on the test data, relative to classifiers trained on the raw spectral coefficients.

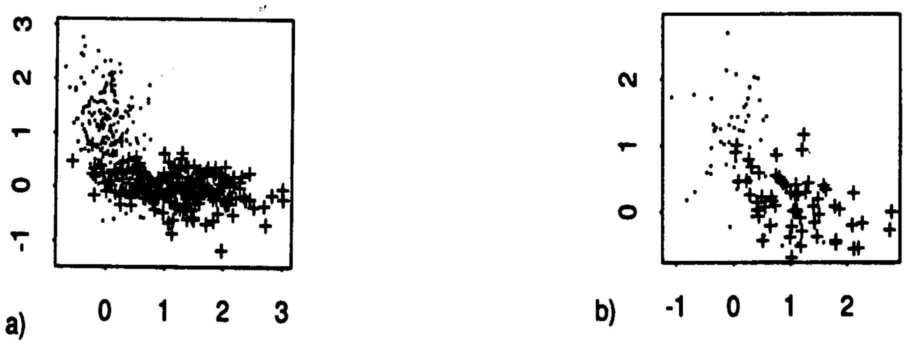

Figure 2: Scatter plots of the first two features derived from 60 eigenvectors of R.

The gross difference between the scatter plots for the training and test data is caused by overfitting to the training data during the feature selection process. To alleviate this overfitting, we reduced the number of free parameters in the feature-extraction algorithm. We accomplished this by reducing the dimension of the intermediate feature vector y in equation (2). We used only the leading m eigenvectors of the original correlation matrix to construct the transformation matrix T in (2). This leaves m-dimensional vectors y for the rest of the transformation process. The value of m was chosen to give good classifier performance.

Figure 2 shows scatter plots of the first two FK features derived using the leading 60 of the 160 total eigenvectors of the original correlation matrix. The training data is plotted in figure 2a, and the test data in figure 2b. The training data has been smeared out relative to figure 1a; there is higher variance orthogonal to the class subspaces. However, the test data is better separated than in figure 1b.

3.2.2 Classifier Performance

To assess the discrimination ability of the FK features, we first developed a performance profile for classifiers trained on the raw spectral coefficients. The results serve as the experimental control. Using the original 160-dimensional data, we trained classifier networks with hidden layers ranging in size from 5 to 45 nodes. The peak test set performance was obtained for networks with 10 and 15 hidden nodes.

Table 1 summarizes the best performance of classifiers trained on the spectral coefficients. Speaker differences between the five data subsets produced significant variation in training and test set performance depending on which data subset was reserved as the test set. Accordingly, we report results for each of five configurations. The first column in the table lists the data subsets used to train the classifier, followed by the data subset used for test. The second column gives the number of input, hidden, and output nodes in the networks. The third column gives the number of training epochs at which the test set performance peaked. The fourth column gives the percentage of correct classification for the test set. The table shows an average peak test set performance of 91.98% correct. On average, the training required 142 epochs.

Table 1: M/N Raw Data

Data	Net Config.	Epochs	%
1234-5	160 10 2	185	89.92
1235-4	160 15 2	115	94.17
1245-3	160 10 2	190	90.0
1345-2	160 10 2	85	95.0
2345-1	160 15 2	135	90.83
Ave.		142	91.98

Table 2: M/N FK Features

Data	Net Config.	Epochs	%
1234-5	12 15 2	165	91.6
1235-4	6 15 2	205	96.67
1245-3	18 15 2	85	94.12
1345-2	12 15 2	70	95.83
2345-1	18 15 2	155	96.67
Ave.	14	136	94.98

Table 2 presents the best performance of classifiers trained on the FK features. We found that using 60 correlation eigenvectors e_i to compute the FK features provided good results. We varied the number of input features between 6 and 54, in increments of 6. We used networks with 15 hidden nodes for all experiments.

Networks trained on the FK features outperformed the control. The average peak test performance is 94.98% with an average training time of 136 epochs. The average reduction in classification error, relative to the control is a little over 37%. Most of the information needed for good class separation seems to be captured in the first fourteen features, on average. This represents an order of magnitude decrease in the feature vector dimension relative to the spectral coefficients.

Of course, the basic Karhunen-Loève transformation can also be used to reduce the dimension of the original features, and one might question whether the FK transformation offers an advantage over the former. To address this, we trained classifier networks on the principal components. As above, we varied the number of network inputs between 6 and 54, in increments of 6. All networks had 15 hidden nodes. The results are shown in Table 3.

Table 3 M/N Principal Components

Data	Net Config.	Epochs	test
1234-5	54 15 2	45	90.0
1235-4	30 15 2	110	95.0
1245-3	16 15 2	50	90.76
1345-2	12 15 2	215	93.33
2345-1	48 15 2	100	90.76
Ave.	32	104	91.97

Clearly, the FK features provided higher classification performance with greater compression than the principal components. Averaged over the five dataset configurations, networks trained on the principal components performed about the same as the control. The number of principal components for optimal performance averaged 32, an 80% reduction in the dimension of the feature vectors.

3.3 Three-letter Discrimination

In this section we present results for classification of the letters (B,D,V). We chose this triple based on preliminary experiments on the triples: (B,D,V), (B,D,E), (P,T,G), and (V,Z,C). The triple (B,D,V) is most challenging, and we limit our discussion to those results.

Because the sonorant portion of each of these letters is identical, more information is required than for the (M,N) experiments above. Therefore, the portion of each utterance before the sonorant (combinations of stop and fricative) was used along with the first part of the sonorant itself. The portion preceding the sonorant was divided into three equal length parts. The sonorant was divided into seven equal length parts, of which only the first two were used. From each part, 32 time-averaged coefficients from 0-4 KHz and 8 time and frequency averaged coefficients from 4 to 8 KHz were generated. The resulting feature vectors have 200 components. The data was divided into four subsets for training and test.

We trained nets with 200 input nodes, hidden layers ranging from 5 to 40 nodes and 3 output nodes. The peak test set performance was obtained for networks with 10 hidden nodes. Averaged over the four dataset configurations, the peak test set performance was 92.5% correct. Training required an average of 178 epochs.

For the experiments on the FK features and the principal components, all nets contained ten hidden nodes and three output nodes. The number of input nodes ranged from 8 to 64 nodes in increments of 8. We generated the FK features using 80 correlation eigenvectors.

Classification performance for networks trained on the FK features was comparable to the control, though with considerable compression and reduction in the number of training epochs required. Averaged over the dataset configurations, the peak test set score was 92.86%. For two of the dataset configurations the FK features outperformed the control, one configuration showed comparable performance, and one configuration performed worse than the control. Training required an average of 63 epochs. The best performing networks averaged 20 input nodes, a 90% compression relative to the spectral coefficients.

Networks trained on principal components performed comparably to those trained on the raw spectral coefficients and the FK features. Averaged over the dataset configurations, the peak test set score was 92.5%. Training required an average of 207 epochs. The best performing networks averaged 46 inputs, a 72% compression relative to the spectral coefficients.

4 Summary

Our results show that the Fukunaga-Koontz transformation provides considerable data compression, and can improve the classification performance of feed-forward networks. For the (M,N) experiments, networks trained on spectral coefficients achieve peak performance of about 92%, while networks trained on the FK features achieve peak performance of about 95%. This performance increase is accompanied by a 91% reduction in the dimension of the feature vector used as input to the classifier. For our (B,D,V) experiments, the FK features achieved peak performance comparable to the raw spectral coefficients, though with compression of 90%.

The major drawback to the procedure is the need to tune several parameters to achieve these results. The number of correlation eigenvectors used to compute the FK features, the number of FK features to retain as input to the classifier, and the number of hidden nodes in the classifier network all affect performance.

Acknowledgements - We thank Mark Fanty for help with data selection and Ron Cole for continued interest. This work was supported in part by the Office of Naval Research.

References

[1] Todd K. Leen, M. Rudnick, and D. Hammerstrom. Hebbian feature discovery improves classifier efficiency. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, pages I–51 to I–56, June 1990.

[2] Jian Yang and Guy A. Dumont. Classification of acoustic emission signals via Hebbian feature extraction. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, pages I–113 to I–118, JULY 1991.

[3] Garrison W. Cottrell and Janet Metcalfe. EMPATH: Face, emotion, and gender recognition using holons. In R. Lippmann, John Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 564–571. Morgan Kauffmann, 1991.

[4] B.A. Golomb, D.T. Lawrence, and T.J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In R. Lippmann, John Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 572–577. Morgan Kauffmann, 1991.

[5] Keinosuke Fukunaga and Warren Koontz. Application of the Karhunen-Loève expansion to feature selection and ordering. *IEEE Transactions on Computers*, C-19:311–318, 1970.

[6] Josef Kittler. Feature selection methods based on the Karhunen-Loeve expansion. In K.S. Fu and A.B. Whinston, editors, *Pattern Recognition – Theory and Applications, NATO Advanced Study Institute Series E, No. 22*. Noordhoof, Leyden, 1977.

[7] Josef Kittler. The subspace approach to pattern recognition. In R. Trappl, G.J. Klir, and L. Ricciardi, editors, *Progress in Cybernetics and Systems Research*, page 92. Hamisphere Publ. Co., Washington, 1978.

[8] Ron Cole, Yeshwant Muthusamy, and Mark Fanty. The ISOLET spoken letter database. Technical Report CSE 90-004, Oregon Graduate Institute of Science & Technology, March 1990.

[9] Ronald Cole, Mark Fanty, Yeshwant Muthusamy, and Murali Gopalakrishnan. Speaker-independent recognition of spoken english letters. In *Proceedings of the International Joint Conference on Neural Networks*, pages II–45, June 1990.

[10] Etienne Barnard and Ronald A. Cole. A neural-net training program based on conjugate-gradient optimization. Technical Report CSE 89-014, Oregon Graduate Institute of Science and Technology, July 1989.