PHONEME CLASSIFICATION OVER THE RECONSTRUCTED PHASE SPACE USING PRINCIPAL COMPONENT ANALYSIS

Jinjin Ye
Department of Electrical and Computer Engineering
Marquette University
Milwaukee, WI USA
jinjin.ye@mu.edu

Michael T. Johnson
Department of Electrical and Computer Engineering
Marquette University
Milwaukee, WI USA
mike.johnson@mu.edu

Richard J. Povinelli
Department of Electrical and Computer Engineering
Marquette University
Milwaukee, WI USA
richard.povinelli@mu.edu

ABSTRACT

Although isolated phoneme classification using features from time-domain phase space reconstruction has been investigated recently, the best representation of feature vectors for the discriminability over phoneme classes is still an open question. This paper applies Principal Component Analysis (PCA) to feature vectors from the reconstructed phase space. By using PCA projection, the basis of the feature space is orthogonalized. A Bayes classifier uses the transformed feature vectors to classify phoneme exemplars. The results show that the classification accuracy with PCA method surpasses the accuracy using only original features in most cases. PCA projection was implemented in three ways over the reconstructed phase space on both speaker-dependent and speaker-independent data. Models are trained and tested using data drawn from the TIMIT database.

1. INTRODUCTION

State of the art speech recognition systems typically use cepstral coefficient features, obtained via a frame-based spectral analysis of the speech signal. Such frequency domain approaches do not necessarily preserve the nonlinear information present in speech. Reconstructed phase space (Abarbanel, 1996; Kantz and Schrieber, 2000) could capture the nonlinear information not preserved by traditional speech analysis techniques, which could result in the improved speech recognition.

This paper is based on work supported by NSF under Grant No. IIS-0113508

The classical techniques used for phoneme classification task are Hidden Markov Models (HMM) (Lee and Hon, 1989; Young, 1992), often based on Gaussian Mixture Model (GMM) observation probabilities. The most common features are Mel Frequency Cepstral Coefficients (MFCCs). As an alternative to the traditional techniques, a nonlinear dynamical systems method called the phase space reconstruction can be applied to studying speech. The reconstructed phase space is simply a plot of the time-lagged vectors of signal, which is used to represent the nonlinear structure. Geometric structures occur in this processing space that are called trajectories or attractors. Reconstructed phase spaces are topologically equivalent to the original system, if the embedding dimension is large enough (Sauer et al., 1991). The full dynamics of the system can be recovered using phase space reconstruction. The previous results (Ye et al., 2002) showed that a Bayes classifier, using features extracted from phoneme reconstructed phase spaces, can be effective in classifying phonemes.

In order to truly represent the underlying dynamic systems that produce the speech signals, usually a high dimensional phase space reconstruction is required. Considering the computational cost associated with the phase space method and training data requirement, a lower dimensional phase space reconstruction is usually desired in practice. Principal Component Analysis (PCA) is a transformation that can be used to reduce the feature dimension. The original feature space is transformed to another feature space on a different set of orthogonal base. By doing PCA transformation over the phase space, the eigenspaces that retain the most significant amount of information are kept. Previous work on transformation over frequency domain features

(Kwon et al., 2002) and phase space features (Broomhead and King, 1986) can also be found in literature.

This paper shows that PCA projection helped improve the classification accuracy of isolated phoneme classification in general using features from the reconstructed phase space. The work uses nonparametric distribution model of phoneme reconstructed phase spaces as input to a Bayes classifier. The Bayes classifier is trained and tested on both speaker-dependent and speaker-independent corpus from the TIMIT database.

2. METHOD

2.1. Phase Space Reconstruction

Phase space reconstruction techniques are founded on underlying principles of dynamical system theory (Sauer et al., 1991; Takens, 1980) and have been applied to a variety of time series analysis and nonlinear signal processing applications (Abarbanel, 1996; Kantz and Schrieber, 2000). Given a time series

$$x = x_n, \quad n = 1...N, \tag{1}$$

where n is a time index, and N is the number of observations, the vectors in a reconstructed phase space is formed, according to Takens' delay method (Takens, 1980),

$$\mathbf{x}_n = [x_{n-(d-1)\tau} \quad \dots \quad x_{n-\tau} \quad x_n], \tag{2}$$

where τ is the time delay and d is the embedding dimension. This reconstructed phase space is in essence no more than a multi-dimensional plot of the signal against delayed versions of itself. Figure 1 provides an illustrative phoneme reconstructed phase space with trajectory information. Figure 2 provides an illustrative phoneme reconstructed phase space with density information. In practice, the attractor is zero-meaned in the phase space and the amplitude variation is normalized from phoneme to phoneme using the standard deviation of the radius.

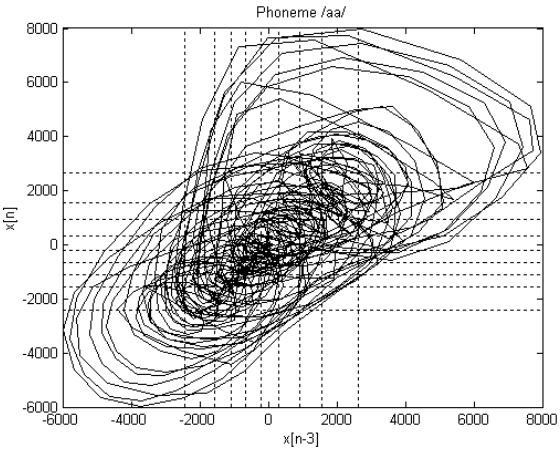

Figure 1 – Reconstructed phase space of the vowel phoneme /*aa*/ illustrating trajectory

The time lag used in the reconstructed phase space is empirical but guided by some key measures such as mutual information and autocorrelation (Abarbanel, 1996; Kantz and Schrieber, 2000). Based on these measures, a time lag of six is selected for all the experiments. The embedding dimensions before and after PCA projection are 15 and 3 respectively.

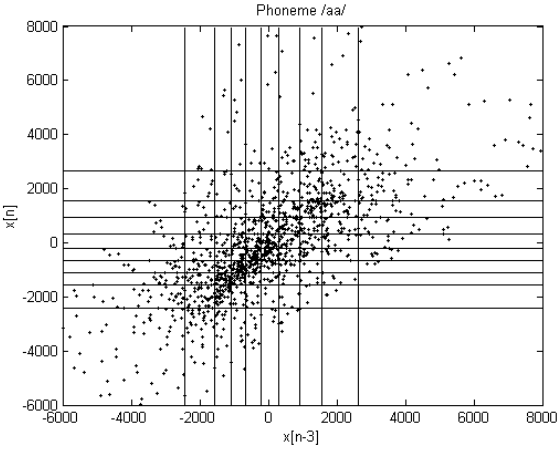

Figure 2 – Reconstructed phase space of the vowel phoneme /*aa*/ illustrating density

2.2. Principal Component Analysis

In order to perform PCA over the phase space features, a trajectory matrix is compiled from the vectors that are created by the time delay embedding method,

$$\mathbf{X} = \begin{bmatrix} x_1 & x_{1+\tau} & \cdots & x_{1+(d-1)\tau} \\ x_2 & x_{2+\tau} & \cdots & x_{2+(d-1)\tau} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N-(d-1)\tau} & x_{N-(d-2)\tau} & \cdots & x_N \end{bmatrix} \quad (3)$$

A scatter matrix is formed,

$$\mathbf{S} = \mathbf{X}^{\mathbf{T}}\mathbf{X} \quad (4)$$

and an eigendecomposition is performed such that

$$\mathbf{S} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^{\mathbf{T}} \quad (5)$$

where the eigenvalues of $\mathbf{\Lambda}$ are reordered in non-increasing order along the diagonal. Select the largest eigenvalues, and let $\mathbf{\Phi}'$ be a matrix containing corresponding columns of $\mathbf{\Phi}$. Then

$$\mathbf{Y} = \mathbf{X}\mathbf{\Phi}' \quad (6)$$

is the new PCA projected trajectory matrix.

Three types of projection were implemented and applied to each set of the experiments, which we denote PCA projection, individual projection and class-based projection. The difference between each implementation mainly depends on the various ways to compute and apply transformations over the data set.

The PCA projection learns one scatter matrix from all the training data and applies the PCA transformation to the trajectory matrix from each phoneme. The individual projection learns and applies the transformation to the trajectory matrix from each phoneme in example-by-example basis. The class-based projection involves two steps in implementation. In the training phase, it learns a scatter matrix and applies transformation over each phoneme class. Assuming there are C phoneme classes, then in test phase, it applies C different transformations on the trajectory matrix from each phoneme exemplar and these projected trajectory matrices are used to compute probabilities under the corresponding class models for the Bayes classifier.

2.3. Nonparametric Distribution Model of Reconstructed Phase Space

A statistical characterization, related to the natural measure or natural distribution of the attractor (Abarbanel, 1996; Kantz and Schrieber, 2000), of the reconstructed phase space is estimated by dividing the reconstructed phase space into 100 histogram bins as is illustrated in Figure 2. This is done by dividing each dimension into ten partitions such that each partition contains approximately 10% of all training data points. The intercepts of the bins are determined using all the training data.

A typical phoneme reconstructed phase space is shown in Figure 1 with the corresponding intercepts, demonstrating the structure of the embedded signal. Figure 2 gives a portrait of the reconstructed phase space based on the natural distribution.

2.4. The Bayes Classifier

The estimates of the natural distribution are used as input for a Bayes classifier. This classifier simply computes the conditional probabilities of the different classes given the values of attributes and then selects the class with the highest conditional probability.

If an instance is described with n attributes a_i $(i=1...n)$, then the class that instance is classified to a class c from set of possible classes C according to a Maximum Likelihood (ML) classifier is:

$$c = \arg\max_{c_j \in C} p(c_j) \prod_{i=1}^{n} p(a_i \mid c_j) \quad (7)$$

The conditional probabilities in the above formula are obtained from the estimates of the natural distribution using training data. This Bayes classifier minimizes the probability of classification error under the assumption that the sequence of points is independent.

3. EXPERIMENTS AND RESULTS

The TIMIT corpus was used to train and evaluate the phoneme classification task. The three types of projection were implemented as described before:

- PCA projection
- Individual projection
- Class-based projection

The embedding dimensions before and after PCA projection are 15 and 3 respectively for all the experiments.

The speaker-dependent experiment used data from one male speaker with 417 phoneme exemplars over standard 48 phonemes (Lee and Hon, 1989). Classification results with three types of projection were obtained from the speaker-dependent experiments, which can give a comparison between different implementations as mentioned above.

The speaker-independent test used training data from six male speakers and testing data from three different male speakers with experiments run on three types of phonemes respectively. A total of 7 fricatives, 7 vowels, and 5 nasals are selected for these experiments. Also, classification results with three types of projection were obtained from the speaker-independent experiments, which can give an idea of how the projection over the phase space affects the classification accuracy on different types of phonemes.

Table 1 shows the results of speaker-dependent experiments on a total of 48 phonemes with and without projection. Table 2 shows the results of speaker-independent experiments on a total of 7 fricative phonemes with and without projection. Table 3 shows the results of speaker-independent experiments on a total of 7 vowel phonemes with and without projection. Table 4 shows the results of speaker-independent experiments on a total of 5 nasal phonemes with and without projection.

Without Proj.	PCA Proj.	Individual Proj.	Class-based Proj.
24.33%	**28.47%**	25.30%	11.19%

Table 1 – Phoneme classification results of speaker-dependent experiments on a total of 48 phonemes

Without Proj.	PCA Proj.	Individual Proj.	Class-based Proj.
39.07%	**42.38%**	33.77%	29.14%

Table 2 – Phoneme classification results of fricatives

Without Proj.	PCA Proj.	Individual Proj.	Class-based Proj.
40.54%	**43.24%**	29.68%	8.78%

Table 3 – Phoneme classification results of vowels

Without Proj.	PCA Proj.	Individual Proj.	Class-based Proj.
55.21%	48.96%	47.92%	48.96%

Table 4 – Phoneme classification results of nasals

Table 1 can give a comparison between different projection implementations on a total of 48 phonemes. In this case, the PCA projection method works best and the class-based projection method works worst. The PCA projection method also works best for the fricative and vowel phoneme classification tasks while the class-based projection method gives the lowest classification accuracies for these two tasks. It can be observed that some phonemes tend to be classified as one particular phoneme for both fricative and vowel experiments using class-based projection method. The confusion of these phonemes in the reconstructed phase space using distribution model can be observed by investigating the confusion matrices for each case.

4. CONCLUSIONS

A novel approach for isolated phoneme classification task using the features from the reconstructed phase space is presented in this paper. In order to find the feature transformations over the reconstructed phase space that have the better discriminability in terms of classification

accuracy, the principal component analysis over the reconstructed phase space was investigated. The PCA has the potential benefits to reduce the feature dimensionality and preserves the most significant amount of information corresponding to the largest eigenvalues. Three projection implementations were tested. The experiment results showed that the PCA projection method yielded best classification accuracy and the class-based projection method yielded worst classification accuracy in overall. Future work would include investigating the feature transformations such as multiple discriminant analysis and nonlinear component analysis that can extract features possibly more useful for classification purposes.

REFERENCES

Abarbanel, H.D.I., 1996. Analysis of observed chaotic data. Springer, New York, xiv, 272 pp.

Broomhead, D.S. and King, G., 1986. Extracting Qualitative Dynamics from Experimental Data. Physica D: 217-236.

Kantz, H. and Schrieber, T., 2000. Nonlinear Time Series Analysis. Cambridge Nonlinear Science Series 7. Cambridge University Press, New York, NY, 304 pp.

Kwon, O.-W., Lee, T.-W. and Chan, K., 2002. Application of variational Bayesian PCA for speech feature extraction. ICASSP, 1: 825-828.

Lee, K.-F. and Hon, H.-W., 1989. Speaker-independent phone recognition using hidden Markov models. IEEE Transactions on Acoustics, Speech and Signal Processing, 37(11): 1641-1648.

Sauer, T., Yorke, J.A. and Casdagli, M., 1991. Embedology. Journal of Statistical Physics, 65(3/4): 579-616.

Takens, F., 1980. Detecting strange attractors in turbulence. Dynamical Systems and Turbulence, 898: 366-381.

Ye, J., Povinelli, R.J. and Johnson, M.T., 2002. Phoneme Classification Using Naive Bayes Classifier In Reconstructed Phase Space. 10th Digital Signal Processing Workshop.

Young, S., 1992. The general use of tying in phoneme-based HMM speech recognition. ICASSP, 1: 569-572.