PHONEME CLASSIFICATION USING NAIVE BAYES CLASSIFIER IN RECONSTRUCTED PHASE SPACE

Jinjin Ye, Richard J. Povinelli, Michael T. Johnson

Department of Electrical and Computer Engineering
Marquette University, Milwaukee, WI

ABSTRACT

A novel method for classifying speech phonemes is presented[1]. Unlike traditional cepstral based methods, this approach uses histograms of reconstructed phase spaces. A Naïve Bayes classifier uses the probability mass estimates for classification. The approach is verified using isolated fricative, vowel, and nasal phonemes from the TIMIT corpus. The results show that a reconstructed phase space approach is a viable method for classification of phonemes, with the potential for use in a continuous speech recognition system.

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. By using the phase space reconstruction technique [1] to capture the nonlinear information not preserved by traditional speech analysis techniques, improved speech recognition accuracy may be achieved.

Various signal-processing techniques have been proposed for phoneme recognition. The most successful are Hidden Markov Models (HMM) [2, 3], often based on Gaussian Mixture Model (GMM) observation probabilities. Common features used are Linear Predictive Coding (LPC) and cepstral coefficients. Hybrid HMMs and neural networks have also been applied to phoneme classification [4]. Continuous speech recognition accuracy is typically reported using Word Error Rate (WER), or sometimes Phoneme Error Rate (PER). Isolated phoneme results using pre-segmented data are usually reported using overall classification accuracy. On the TIMIT corpus, the data set studied in this work, phoneme recognition and classification accuracies in the range 40-77% have been published [2-5].

[1]This paper is based upon work supported by NSF under Grant No. IIS-0113508

The phase space reconstruction approach is a dynamical systems method used to capture the nonlinear structure. The results discussed here show that a Naïve Bayes classifier, using features extracted from phoneme reconstructed phase spaces, can be effective in classifying phonemes. It is reasonable to expect that this classification accuracy will translate long-term to more effective continuous speech recognition systems. In general, good phoneme classifiers lead to good word classifiers, and the ability to recognize phonemes accurately provides the basis for an accurate word and continuous speech recognizer.

The proposed method is based on our previous work in classifying motor faults [6-9] and heart arrhythmias [10, 11]. In that work [6-11], reconstructed phase spaces were formed from sampled signals. For motor fault identification, the signals were torque profiles and current waveforms. For heart arrhythmia classification, the signals were leads II and V1 from a twelve lead electrocardiogram (ECG). Statistical characterizations of the reconstructed phase spaces were used as features for a neural network classifier in the case of the heart arrhythmia classification and a nearest neighbor algorithm in the case of the motor fault identification.

The work discussed here in this paper uses estimates of the probability masses (histograms) of phoneme reconstructed phase spaces as input to a Naïve Bayes classifier. The Naïve Bayes classifier is trained on six male speakers and tested on three different male speakers.

2. METHOD

2.1. Phase Space Reconstruction

Phase space reconstruction techniques are founded on underlying principles of dynamical system theory [13, 14] and have been practically applied to a variety of time series analysis and nonlinear signals processing applications [15, 16]. Given a time series

$$X = \{x_t, t = 1, \ldots, N\},$$

where t is a time index, and N is the number of observations, a reconstructed phase space is formed, according to Takens' delay method [13],

$$\overrightarrow{X_t} = \left(x_{t-(m-1)\tau}, \ldots, x_{t-2\tau}, x_{t-\tau}, x_t \right),$$

where τ is the time delay and m 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. If the phase space reconstruction is performed correctly, the result is topologically equivalent to the original system [13, 14]. Figure 1 provides an illustrative phoneme reconstructed phase space with trajectory information. Figure 2 provides an illustrative phoneme reconstructed phase space with density information.

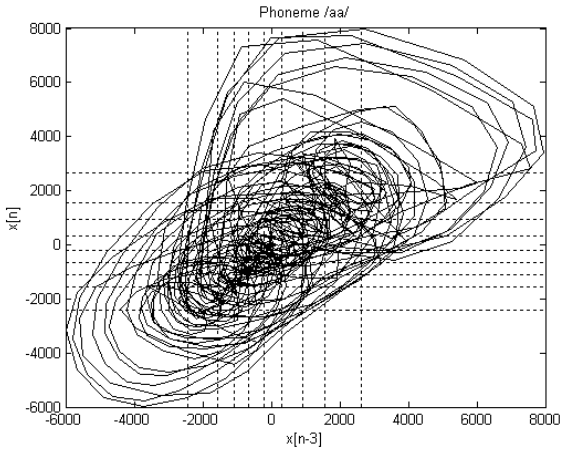

Figure 1 – Reconstructed phase space of the vowel phoneme /*aa*/ illustrating trajectory

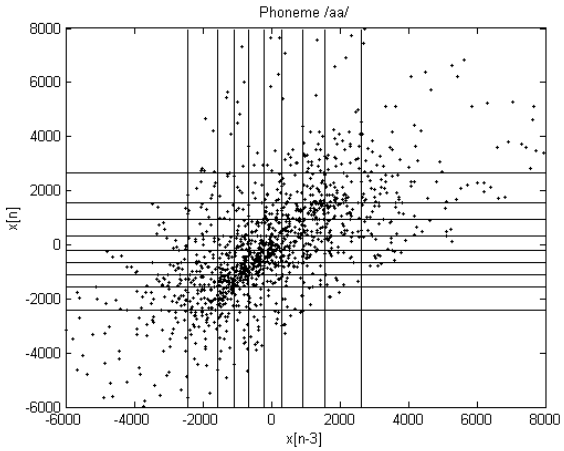

Figure 2 – Reconstructed phase space of the vowel phoneme /*aa*/ illustrating density

Although time delay and embedding dimension are important reconstructed phase space parameters, they have not been extensively studied in this preliminary work on phoneme classification. An embedding dimension of two and a time delay of three were used.

2.2. Normalization

In order to counteract the amplitude inconsistency between phoneme signals, a normalization method is used. This method uses radius of gyration, a function of the 2nd moment of the distance of the points, as a factor for phoneme signal normalization. The radius of gyration for two-dimensional reconstructed phase space is calculated as follows:

$$r = \sqrt{\frac{\sum_{k=1}^{N} d^2(k)}{N}}, \qquad (1)$$

where

$$d(k)^2 = \left(x_1(k) - \mu_1 \right)^2 + \left(x_2(k) - \mu_2 \right)^2. \qquad (2)$$

Here $x_1(k)$ denotes the k^{th} point of the signal x_t while $x_2(k)$ is the k^{th} point of the signal $x_{t-\tau}$. N is the number of samples in the phoneme signal, while the value μ_i is the mean of the signal amplitude for each reconstructed phase space dimension.

2.3. Features of Reconstructed Phase Space

A statistical characterization (estimates of the probability masses) of the reconstructed phase space [10] is formed 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 [10].

A typical phoneme reconstructed phase space is shown in Figure 1 with the corresponding intercepts, which clearly shows the structure of the embedded signal. Figure 2 gives a portrait of the reconstructed phase space based on the distribution of points. The estimates of probability mass for each phoneme class are calculated.

2.4. Naïve Bayes Classifier

The estimates of the probability masses are used as input for a Naïve Bayes classifier [17]. 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 v from

set of possible classes V according to a Maximum a Posteriori (MAP) Naive Bayes classifier is:

$$v = \underset{v_j \in V}{\arg\max}\, p\left(v_j\right) \prod_{i=1}^{n} p\left(a_i \middle| v_j\right) \tag{3}$$

The conditional probabilities in the above formula is obtained from the estimates of the probability mass function using training data. The class probability is not used in these experiments, since no prior phoneme distribution information is available, and thus we are implementing Maximum Likelihood (ML) classification. 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 speaker-independent phoneme recognizers. The TIMIT corpus consists of 630 speakers, each saying 10 sentences, including:

- 2 "sa" sentences, which are the same across all speakers;
- 5 "sx" sentences, which were read from a list of 450 phonetically balanced sentences selected by MIT;
- 3 "si" sentences, which were randomly selected by TI.

The TIMIT corpus has phoneme labeling, which makes it a useful database for phoneme classification.

In our experiments, we use training data from six male speakers and testing data from three different male speakers. Three types of phonemes are tested, which are vowels, fricatives, and nasals. A total of seven fricatives, seven vowels, and five nasals are selected for the test. A two-dimensional reconstructed phase space with a time delay of three is formed for each phoneme, and the 100 estimated probability masses are calculated.

The results in Tables 1-3 are without normalization. The result for the seven fricatives classification is shown in Table 1. The result for the seven vowels classification is shown in Table 2. The result for the five nasals classification is shown in Table 3.

Phone	Correct
'dh'	16.67%
'f'	60.00%
's'	74.47%
'sh'	77.78%
'th'	50.00%
'v'	56.25%
'z'	47.62%
Overrall	58.94%

Table 1 – Phoneme recognition results of fricatives

Phone	Correct
'aw'	26.67%
'ay'	2.63%
'ey'	21.05%
'ix'	53.15%
'iy'	34.19%
'ow'	5.26%
'oy'	14.29%
Overall	33.00%

Table 2 – Phoneme recognition results of vowels

Phone	Correct
'en'	60.00%
'm'	4.00%
'n'	9.43%
'ng'	16.66%
'nx'	85.71%
Overall	16.67%

Table 3 – Phoneme recognition results of nasals

The accuracy for vowel recognition and nasals are worse than that of fricatives. This is consistent with the results from traditional methods [2].

The result for the seven fricatives classification with normalization is shown in Table 4. The result for the seven vowels classification with normalization is shown in Table 5. The result for the five nasals classification with normalization is shown in Table 6.

Phone	Correct
'dh'	38.89%
'f'	72.00%
's'	59.57%
'sh'	94.44%
'th'	66.67%
'v'	87.50%
'z'	23.81%
Overrall	61.59%

Table 4 – Phoneme recognition results of fricatives with normalization

Phone	Correct
'aw'	13.33%
'ay'	50.00%
'ey'	23.68%
'ix'	27.97%
'iy'	44.44%
'ow'	36.84%
'oy'	21.43%
Overall	34.49%

Table 5 – Phoneme recognition results of vowels with normalization

Phone	Correct
'en'	0.00%
'm'	12.00%
'n'	39.62%
'ng'	16.67%
'nx'	57.14%
Overall	30.21%

Table 6 – Phoneme recognition results of nasals with normalization

The second set of results shows that by normalizing the phoneme reconstructed phase spaces, the classification accuracy for each class improves. In order to compare with traditional methods, different data sets will be used in the future. These preliminary results show that the proposed method is promising and has the potential to be applied to a phoneme recognizer as well as a continuous speech recognition system.

4. CONCLUSIONS

In this paper, we have presented the Naïve Bayes Classifier method for phoneme classification in the reconstructed phase space. This method is a novel approach substantially different from existing techniques. Preliminary results show that our method is a promising way for building a phoneme recognizer. Further work will focus on using Gaussian Mixture Models (GMMs) built from the reconstructed phase space as well as other dynamic features, comparison to existing methods, and integrating into a continuous recognition system. We expect to integrate features of the nonlinear systems with traditional statistical features to finally improve the overall accuracy of continuous speech recognition.

5. REFERENCES

[1] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, "Geometry from a time series," *Physical Review Letters*, vol. 45, pp. 712-716, 1980.

[2] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 1641-1648, 1989.

[3] S. Young, "The general use of tying in phoneme-based HMM speech recognition," proceedings of ICASSP, 1992, pp. 569-572.

[4] S. A. Zahorian, P. Silsbee, and X. Wang, "Phone classification with segmental features and a binary-pair partitioned neural network classifier," proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97), 1997, pp. 1011 -1014.

[5] H. C. Leung, B. Chigier, and J. R. Glass, "A comparative study of signal representations and classification techniques for speech recognition," proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93), 1993, pp. 680 -683.

[6] R. J. Povinelli, J. F. Bangura, N. A. O. Demerdash, and R. H. Brown, "Diagnostics of Faults in Induction Motor ASDs Using Time-Stepping Coupled Finite Element State-Space and Time Series Data Mining Techniques," proceedings of Third Naval Symposium on Electric Machines, 2000.

[7] J. F. Bangura, R. J. Povinelli, N. A. O. Demerdash, and R. H. Brown, "Diagnostics of Eccentricities and Bar/End-Ring Connector Breakages in Polyphase Induction Motors through a Combination of Time-Series Data Mining and Time-Stepping Coupled FE-State Space Techniques," proceedings of IEEE Industry Application Society 2001 Annual Meeting, 2001, pp. 1579-1586.

[8] R. J. Povinelli, J. F. Bangura, N. A. O. Demerdash, and R. H. Brown, "Diagnostics of Bar and End-Ring Connector Breakage Faults in Polyphase Induction Motors Through a Novel Dual Track of Time-Series Data Mining and Time-Stepping Coupled FE-State Space Modeling," *IEEE Transactions on Energy Conversion*, vol. 17, pp. 39-46, 2002.

[9] R. J. Povinelli, M. T. Johnson, J. F. Bangura, and N. A. O. Demerdash, "A Comparison of Phase Space Reconstruction and Spectral Coherence Approaches for Diagnostics of Bar and End-Ring Connector Breakage Faults in Polyphase Induction Motors using Current Waveforms," proceedings of IEEE Industry Application Society 2002 Annual Meeting, in press.

[10] F. M. Roberts, R. J. Povinelli, and K. M. Ropella, "Identification of ECG Arrhythmias using Phase Space Reconstruction," proceedings of Principles and Practice of Knowledge Discovery in Databases (PKDD'01), Freiburg, Germany, 2001, pp. 411-423.

[11] R. J. Povinelli, F. M. Roberts, K. M. Ropella, and M. T. Johnson, "Are Nonlinear Ventricular Arrhythmia Characteristics Lost, As Signal Duration Decreases?," proceedings of Computers in Cardiology, in press.

[12] C. Borgelt, "A Naive Bayes Classifier Plug-In for DataEngine," proceedings of Proceedings of the 3rd Data Analysis Symposium, 1999, pp. 87-90.

[13] F. Takens, "Detecting strange attractors in turbulence," proceedings of Dynamical Systems and Turbulence, Warwick, 1980, pp. 366-381.

[14] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65, pp. 579-616, 1991.

[15] H. D. I. Abarbanel, *Analysis of observed chaotic data*. New York: Springer, 1996.

[16] H. Kantz and T. Schreiber, *Nonlinear time series analysis*. Cambridge: Cambridge University Press, 1997.

[17] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.