STUDY OF ATTRACTOR VARIATION IN THE RECONSTRUCTED PHASE SPACE OF SPEECH SIGNALS

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

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

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

ABSTRACT

This paper presents a study of the attractor variation in the reconstructed phase spaces of isolated phonemes. The approach is based on recent work in time-domain signal classification using dynamical signal models, whereby a statistical distribution model is obtained from the phase space and used for maximum likelihood classification. Two sets of experiments are presented in this paper. The first uses a variable time lag phase space to examine the effect of fundamental frequency on attractor patterns. The second focuses on speaker variability through an investigation of speaker-dependent phoneme classification across speaker sets of increasing size. The results are discussed at the end of the paper.

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 (Deller et al., 2000). However, recent work in phase space reconstruction techniques (Abarbanel, 1996; Kantz and Schrieber, 2000) for nonlinear modeling of time-series signals has motivated investigation into the efficacy of using dynamical systems models in the time-domain for speech recognition (Ye et al., 2002). In theory, reconstructed phase spaces capture the full dynamics of the underlying system, including nonlinear information not preserved by traditional spectral techniques, leading to possibilities for improved recognition accuracy.

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

The classical technique for phoneme classification is 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).

In contrast, the reconstructed phase space is a plot of the time-lagged vectors of a signal. Such phase spaces have been shown to be topologically equivalent to the original system, if the embedding dimension is large enough (Sauer et al., 1991). Structural patterns occur in this processing space, commonly referred to as trajectories or attractors, which can be quantified through invariant metrics such as correlation dimension or Lyapunov exponents or through direct models of the phase space distribution. Previous results on phoneme classification (Ye et al., 2002) have shown that a Bayes classifier over statistical models of the reconstructed phase spaces are effective in classifying phonemes.

Phase space reconstructions are not specific to any particular production model of the underlying system, assuming only that the dimension of the system is finite. We would like to be able to take advantage of our *a priori* knowledge about speech production mechanisms to improve usefulness of phase space models for speech recognition in particular.

In pursuit of this, we have implemented two sets of experiments to study attractor variation, the first to look at fundamental frequency effects in vowels and the second to look at speaker variability issues. Fundamental frequency, as a parameter that varies significantly but does not contain information about the generating phoneme, should clearly affect the phase space in an adverse way for classification. This hypothesis is examined through a compensation technique using variable lag reconstructions. Speaker variability is

an as of yet unknown factor with regard to the amount of variance caused in underlying attractor characteristics, and is an important issue in the question of how well this technique will work for speaker-independent tasks. Initial experiments have shown some significant discriminability in such tasks, but performed at a measurably lower accuracy than that for speaker-dependent tests.

Each of these experiments use a nonparametric distribution model based on bin counts with a maximum likelihood phoneme classifier, as described in more detail in the next section. The TIMIT database is used for both tasks.

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 signals processing applications [refs xx] (Abarbanel, 1996; Kantz and Schrieber, 2000). Given a time series

$$x = x_n, \qquad n = 1 \ldots N \tag{1}$$

where n is a time index and N is the number of observations, the vectors in a reconstructed phase space are formed as

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

where τ is the time lag and d is the embedding dimension. Taken as a whole, the signal forms a trajectory matrix compiled from the time-lagged signal vectors:

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

Figure 1 shows an illustrative two-dimensional reconstructed phase space trajectory, while Figure 2 shows the same space with individual points only, as modeled by the statistical approach used here.

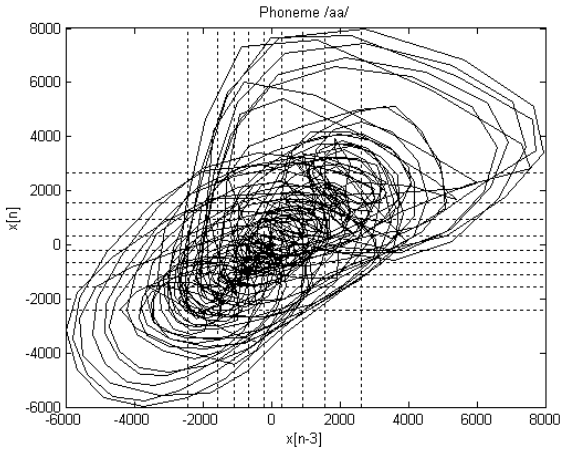

Figure 1 – Reconstructed phase space of the vowel phoneme /*aa*/ illustrating trajectory

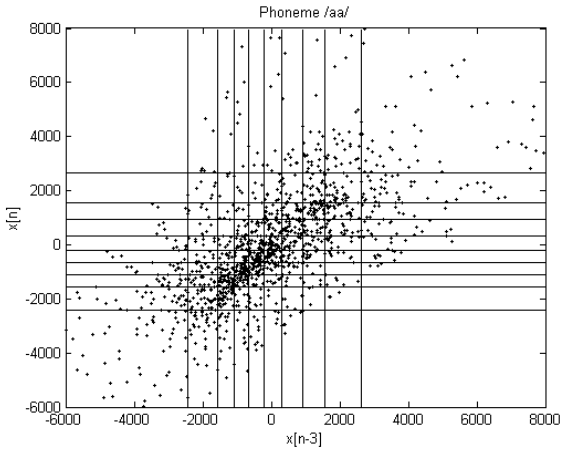

Figure 2 – Reconstructed phase space of the vowel phoneme /*aa*/ illustrating density

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). Using such measures, a time lag of six is appropriate for isolated phoneme recognition using TIMIT. For the variable lag experiments baseline time lags of 6 and 12 are used, as discussed in details later.

In practice, the phase-space reconstruction is zero-meaned and the amplitude variation is radially normalized via:

$$\mathbf{x}'_n = \frac{\mathbf{x}_n - \boldsymbol{\mu}_{\mathbf{X}}}{\sigma_r} \tag{4}$$

$$\sigma_r \triangleq \sqrt{\frac{1}{N-(d-1)\tau} \sum_{n=1+(d-1)\tau}^{N} \left\| \mathbf{x}_n - \boldsymbol{\mu}_{\mathbf{x}} \right\|^2} \tag{5}$$

where $\mathbf{x}_n$ is an original point in the phase space, $\boldsymbol{\mu}_{\mathbf{X}}$ is the sample mean of the columns of $\mathbf{X}$.

2.2. Nonparametric Distribution Model of Reconstructed Phase Space

A statistical characterization of the reconstructed phase space , related to the natural measure or natural distribution of the underlying attractor (Abarbanel, 1996; Kantz and Schrieber, 2000), 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 the data points. The intercepts of the bins are determined using all the training data.

2.3. The Bayes Classifier

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

$$\hat{\omega} = \arg\max_{i=1\ldots C} \left\{ \hat{p}_i(\mathbf{X}) \right\} = \arg\max_{i=1\ldots C} \left\{ \prod_{n=1}^{N} \hat{p}_i(\mathbf{x}_n) \right\} \tag{6}$$

where $\hat{p}_i(\mathbf{x}_n)$ is the bin-based likelihood of a point in the phase space, C is the number of phoneme classes, and $\hat{\omega}$ is the resulting maximum likelihood hypothesis.

3. EXPERIMENT DESIGN

3.1. Variable Lag Model for Vowels

The basic idea of this experiment is to use variable time lags instead of a fixed time lag for embedding vowel phonemes, as a function of the underlying fundamental frequency of the vowel. An estimate of the fundamental frequency is used to determine the appropriate embedding lag.

The fundamental frequency estimate algorithm for vowels used here is based on the computation of autocorrelation in the time domain as implemented by the Entropic ESPS package (Entropic, 1993).

The typical vowel fundamental frequency range for male speakers is 100~150Hz, with an average of about 125Hz, while the typical range for female speakers is 175~256Hz, with an average of about 200Hz. For this experiment only male speakers were used. In the reconstructed phase space, a lower fundamental frequency has a longer period, corresponding to a larger time lag. With a baseline time lag and mean fundamental frequency given as τ and f_0 respectively, we perform fundamental frequency compensation via the equations

$$\tau f_0 = \tau' f'_0 \tag{7}$$

and

$$\tau' = \frac{\tau f_0}{f'_0} \tag{8}$$

where τ' is the new time lag and f'_0 is the fundamental frequency estimate of the phoneme example. This time lag is rounded and used for phase space reconstruction, for both estimation of the phoneme distributions across the training set and maximum likelihood classification of the test set examples.

Two different baseline time lags are used in the experiments. A time lag of 6 corresponds to that chosen through examination of the mutual information and autocorrelation heuristics; however, rounding effects lead to quite a low resolution on the lags in the experiment, which vary primarily between 5, 6, and 7. To achieve a slightly higher resolution, a second set of experiments at a time lag of 12 is implemented for comparison. Since the final time lags used for reconstruction are given by a fundamental frequency ratio, the value of the baseline frequency is not of great importance, but should be chosen to be near the mean fundamental frequency. Baselines of 120Hz and 130Hz were investigated. The final time lag is given in accordance with equation (8) above.

The data set used here includes 6 male speakers for training and 3 different male speakers for testing, all within the same dialect region.

3.2. Speaker Variability

Using the phase space reconstruction technique for speaker-independent tasks clearly requires that the attractor pattern across different speakers is consistent. Inconsistency of attractor structures across different speakers would be expected to lead to smoothed and imprecise phoneme models with resulting poor classification accuracy. The experiments presented here are designed to investigate the inter-speaker variation of attractor patterns. Although a number of different attractor distance metrics could be used for this purpose, the best such choice is not readily apparent and we have instead focused on classification accuracy as a function of the number of speakers in a closed-set speaker dependent recognition task. The higher the consistency of attractors across speakers, the less accuracy degradation should be expected as the number of speakers in the set is increased.

All speakers are male speakers selected from the same dialect region within the TIMIT corpus. The bin-based models and maximum likelihood classification methods discussed previously are used in all cases. The only variable is the number of speakers for isolated phoneme classification tasks.

To examine speaker variability effects across different classes of phonemes, vowels, fricatives and nasals are tested separately. The overall data set is a group of 22 male speakers, from which subsets of 22, 17, 11, 6, 3, 2 or 1 speaker(s) have been randomly selected. Classification experiments are performed on sets of 7 fricatives, 7 vowels, and 5 nasals.

4. RESULTS

4.1. Variable Lag for Vowels

The 7 vowel set { /aw/ /ay/ /ey/ /ix/ /iy/ /ow/ /oy/ } is used for these experiments. As described previously, data are selected from 6 male speakers for training and 3 different male speakers for testing, all within the same dialect region.

There are a total of six experiments, three with a baseline lag of 6 and three with a baseline lag of 12. In each case the tests are run with a fixed lag as well as with variable lags based on both 120Hz and 130Hz. The six experiments are summarized as follows:

Exp 1: $d = 2, \tau = 6, \tau' = \tau$,

Exp 2: $d = 2, \tau = 6, f_0 = 120Hz, \tau' = \frac{\tau f_0}{f_0'}$,

Exp 3: $d = 2, \tau = 6, f_0 = 130Hz, \tau' = \frac{\tau f_0}{f_0'}$,

Exp 4: $d = 2, \tau = 12, \tau' = \tau$,

Exp 5: $d = 2, \tau = 12, f_0 = 120Hz, \tau' = \frac{\tau f_0}{f_0'}$,

Exp 6: $d = 2, \tau = 12, f_0 = 130Hz, \tau' = \frac{\tau f_0}{f_0'}$,

where d is the embedding dimension, τ is the default time lag, f_0 is the default fundamental frequency, f_0' is the estimated fundamental frequency, and τ' is the actual embedding time lag.

Table 1 shows the resulting ranges for τ' given the parameters, while Table 2 and 3 show the classification results.

τ	f_0	τ'
6	120Hz	5~7
6	130Hz	5~8
12	120Hz	10~14
12	130Hz	10~16

Table 1 – Range of τ' given τ and f_0

Exp 1	Exp 2	Exp 3
27.70%	33.78%	**35.14%**

Table 2 – Vowel phoneme classification results for lag of 6

Exp 4	Exp 5	Exp 6
39.19%	35.14%	**39.86%**

Table 3 – Vowel phoneme classification results for lag of 12

As can be seen from the above results, the best classification accuracy is obtained by using a variable lag model with default fundamental frequency of 130Hz.

4.2. Speaker Variability

The evaluation of speaker variability was carried out using the leave-one-out cross validation. The overall classification accuracies for the three types of phonemes are shown in Table 4, Table 5 and Table 6.

Spkr#	1	2	6	22
Acc(%)	58.00	51.06	49.26	48.58

Table 4 – Phoneme classification results of fricatives with different number of speakers

Spkr#	1	2	6	11	17	22
Acc(%)	61.90	49.09	49.58	46.92	45.46	46.03

Table 5 – Phoneme classification results of vowels with different number of speakers

Spkr#	1	2	3	6	22
Acc(%)	51.79	40.00	31.91	29.95	26.76

Table 6 – Phoneme classification results of nasals with different number of speakers

Figure 3 is a visual interpretation of the results presented above. The classification results are plotted against the number of speakers for vowels, fricatives and nasals respectively.

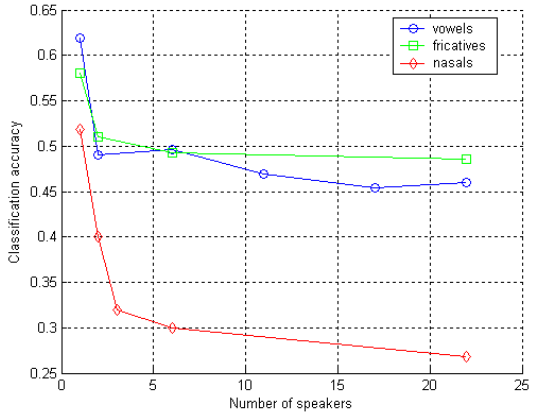

Figure 3 – The classification accuracy vs. number of speakers

As can be seen from Figure 3, the degree of attractor variation across speakers is different for these three types of phonemes. Nasals appear to have the largest variability while the fricatives have the least, which is consistent with the results reported in the previous paper (Ye et al., 2002), for a speaker-independent task over vowels, nasals, and fricatives. In all three phoneme types, the accuracy was relatively unchanged after 2 or 3 speakers.

5. WORK IN PROGRESS

Because of the large difference between the baseline numbers in the variable lag experiments, we will verify and investigate them by running other experiments. The experiments presented are performed on a speaker-independent dataset, which could affect the results due to speaker variability. The additional investigation on the effect of fundamental frequency on attractor patterns will be included in the final paper.

6. DISCUSSION AND CONCLUSIONS

We have examined attractor variation in the reconstructed phase space of isolated phonemes. The variable lag experiments show an improvement when the phase space incorporates fundamental frequency compensation. The speaker variability experiments indicate that there is a significant amount of consistency across attractor patterns between speakers, which shows that the reconstructed phase space representation of speech signal has the discriminative power for the speaker-independent tasks. Future work would include discovering the method to compensate speaker variability for the phase space reconstruction technique on speech recognition applications.

REFERENCES

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

Deller, J.R., Hansen, J.H.L. and Proakis, J.G., 2000. Discrete-Time Processing of Speech Signals. IEEE Press, New York, 908 pp.

Entropic, 1993. ESPS Programs Manual. Entropic Research laboratory.

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

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.