7-1-2005

Time-Domain Isolated Phoneme Classification Using Reconstructed Phase Spaces

Michael T. Johnson
Marquette University, michael.johnson@marquette.edu

Richard J. Povinelli
Marquette University, richard.povinelli@marquette.edu

Andrew C. Lindgren
Marquette University

Jinjin Ye
Marquette University

Xiaolin Liu
Marquette University

See next page for additional authors

Accepted version. *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 4 (July 2005): 458-466. DOI.

Time-Domain Isolated Phoneme Classification Using Reconstructed Phase Spaces

M.T. Johnson
Department of Electrical & Computer Engineering, Marquette University, Milwaukee, WI
R.J. Povinelli
Department of Electrical & Computer Engineering, Marquette University, Milwaukee, WI
A.C. Lindgren
Department of Electrical & Computer Engineering, Marquette University, Milwaukee, WI
Jinjin Ye
Department of Electrical & Computer Engineering, Marquette University, Milwaukee, WI
Xiaolin Liu
Department of Biomedical Engineering, Marquette University, Milwaukee, WI
K.M. Indrebo
Department of Electrical & Computer Engineering, Marquette University, Milwaukee, WI

Abstract: This paper introduces a novel time-domain approach to modeling and classifying speech phoneme waveforms. The approach is based on statistical models of reconstructed phase spaces, which offer significant theoretical benefits as representations that are known to be topologically equivalent to the state dynamics of the underlying production system. The lag and dimension parameters of the reconstruction process for speech are examined in detail, comparing common estimation heuristics for these parameters with corresponding maximum likelihood recognition accuracy over the TIMIT data set. Overall accuracies are compared with a Mel-frequency cepstral baseline system across five different phonetic classes within TIMIT, and a composite classifier using both cepstral and phase space features is developed. Results indicate that although the accuracy of the phase space approach by itself is still

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

currently below that of baseline cepstral methods, a combined approach is capable of increasing speaker independent phoneme accuracy.

SECTION I.

Introduction

Current state-of-the-art speech recognition systems use frequency-domain features, such as Mel-frequency cepstral coefficients (MFCCs), which are based upon a switched linear model of the human speech production mechanism. This familiar model is a reasonable, albeit somewhat rough, approximation of the true physiological process, and has led to successful coding, synthesis, and recognition algorithms for many years.

One limitation of this frequency-domain approach is the inability of such a representation to capture the nonlinear and higher-order characteristics of the speech production process. Research in this area has suggested that there is evidence of nonlinear behavior in both voiced and unvoiced excitation patterns, and that such nonlinearity is not insignificant.[1–2,3] To capture this nonlinear information, a number of other analytical methods have been investigated as an alternative to traditional linear approaches, including the use of time-frequency and time-scale transforms, higher-order statistics, and dynamical systems and chaos theory.

The basis for the dynamical systems approach, which is the focus of the work presented here, lies in theorems showing that by embedding a signal into a sufficiently high dimensional space, a structure is formed that is topologically equivalent to the original phase space, i.e., state space, of the system generating the signal. This embedding, called a reconstructed phase space (RPS), is typically constructed by mapping time-lagged copies of the original signal onto axes of the new high dimensional space. The time evolution of the signal within the RPS traces out a trajectory pattern referred to as its attractor, a term adopted (somewhat loosely) from dynamical systems theory, which is a representation of the dynamics of the underlying system. Each point in the space, as a vector of time-lagged signal points, captures short-time dynamics, and the overall attractor structure is a full representation of those dynamics. Since the attractor of an RPS captures all information about the underlying system, it is an appealing choice for signal analysis, processing, and classification. There has been some other work in time-domain representations of speech signals, such as through autoregressive modeling,[4] but the RPS approach introduced here has the advantage of capturing both linear and nonlinear aspects of the underlying system.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

The use of RPSs is well known within the dynamical systems field, and measures taken from that field have been utilized in a number of application areas, including the tasks of speech synthesis and recognition. Examples include the use of dynamical invariants such as Lyapunov exponents and fractal dimensions[5–6,7,8,9,10] as features for recognition, as well as work in functional modeling of attractors using orthogonal polynomial bases.[11,12] Our prior work in the area of attractor modeling has focused on using statistical representations for Bayesian signal classification, with applications to heart arrhythmia identification[13,14] and motor diagnostics,[15] as well as the speech representation and recognition tasks[16–17,18,19,20,21,22,23] presented here. Advantages of such an approach over those based on invariant metric features include that it captures more aspects of the attractor and that it generalizes well to arbitrary systems, regardless of the degree of nonlinearity present.

The goal of the work presented here is to directly model reconstructed phase spaces for application to speech recognition. The current effort focuses on isolated phoneme recognition, with the goal of identifying its capability for capturing phonetic differences in a speaker independent environment.

Section II gives a detailed overview of the dynamical systems theory and terminology and examines attractor patterns for different phoneme classes. Section III introduces the statistical model used to capture these patterns, as well as the frequency domain baseline model. Issues of lag and dimension are discussed in Section IV, followed by supporting experimental results and discussion in Section V. Accuracy as a function of phoneme class is examined in Sections VI and VII discusses results of the composite RPS/MFCC classifier. We conclude in Section VIII with a discussion of the initial success of this new approach and of its potential for more complex speech recognition tasks.

SECTION II.

Basis in Dynamical Systems Theory

As introduced above, the underlying principle of this work lies in the idea that the state space of a system can be reconstructed through an embedding of a single state variable or observation sequence from that system. We denote a time series as x_n, with a time delay RPS of dimension d and time lag τ defined by the trajectory matrix

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

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

where each row vector in the matrix represents a single point in the space

$$\begin{aligned} \mathbf{x}_n &= [x_n x_{n-\tau} \cdots x_{n-(d-1)\tau}], \\ n &= (1+(d-1)\tau) \cdots N. \end{aligned} \quad (2)$$

Each point in the space captures local short-time signal dynamics, and together the entire RPS is a representation of the dynamics of the underlying system. The concept of time delay embedding was first introduced by Packard,[24] based on early theorems by Whitney[25] relating to topological embeddings in Cartesian spaces. Working from this idea, Takens[26] proved that delay coordinate maps of dimension greater than twice that of the original system are embeddings, providing an important theoretical justification for the practical use of time delay reconstructions. Sauer, Yorke, and Casdagli[27] have extended Takens' work, establishing that, except for a set of degenerate cases with measure zero, the topological equivalence property is guaranteed for time-lag reconstructed phase spaces. In addition, they tightened the bound on the required dimension to $d > 2d_0$, where d_0 is the boxcounting dimension of the attractor of the underlying system. Together, the above theorems guarantee that for almost every time delay embedding, the reconstructed dynamics of the map, including dynamical invariants such as fractal dimensions and Lyapunov exponents, are topologically identical to the true dynamics of the system.

The concept of dimension d and lag τ play a significant role in both the theoretical and practical aspects of working with reconstructed phase spaces. The topological equivalence property of the space is only guaranteed for $d > 2d_0$; however, this is a sufficient condition not a necessary one, so that often dimensions of much less than $2d_0$ are enough to fully represent the structure of the attractor. To identify the minimum possible

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

dimension, heuristic procedures such as false nearest neighbor thresholds [28] are typically used. The time lag τ has little impact from a theoretical viewpoint, and in fact there are no limitations or assumptions placed upon it with respect to the underlying time-lag reconstruction theorems for discrete-time signals.[27] However, since topological invariance of systems does not equate to identical phase spaces or attractors, from a practical viewpoint the lag must be selected with respect to some relevant criteria. Both dimension and lag, including methods for selecting them as well as their impact on classification accuracy in the speech task, will be discussed in more detail in Section 4.

Many types of signals and systems can be characterized through phase space analysis, including linear, nonlinear, chaotic, and stochastic systems. Linear systems have a fixed point or periodic attractor structure, while nonlinear systems may be aperiodic with complex attractor structure. Attractors of chaotic systems (a subset of general nonlinear systems) have several unusual characteristics such as snap back repellers, sensitivity to initial conditions, positive Lyapunov exponents, and topological transitivity. Additive noise processes add a random component to each point in the underlying phase space, obscuring the attractor and increasing the required dimension for adequate representation.

Examples of reconstructed phase spaces with dimension 3 and lag 6, taken from the TIMIT data set[29,30] for five different phonetic classes are shown in Fig. 1. The classes include vowels, semi-vowels, stops, nasals, and fricatives. The plots demonstrate that vowels, as quasiperiodic waveforms, exhibit the most distinct structure, with semi-vowels, and nasals having similar but less defined characteristics. Fricatives, generated by turbulent air flow, exhibit much less structure (and would be expected to require higher dimensions for adequate modeling), while stops and affricates have a defined nonperiodic structure.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

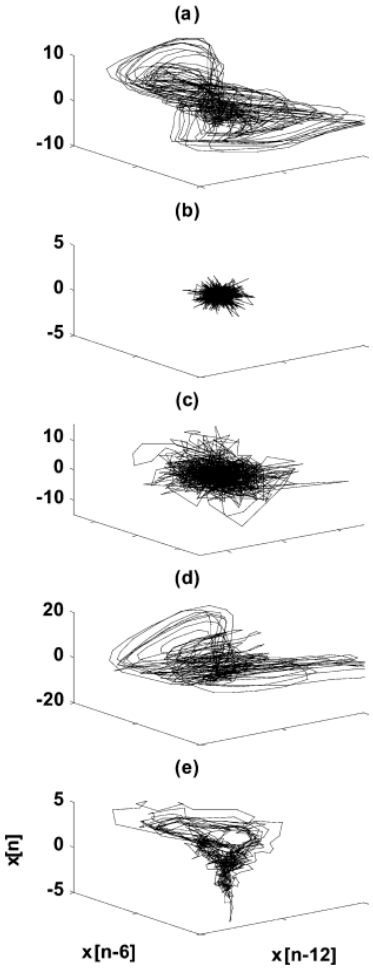

Fig. 1. Examples of reconstructed phase spaces. (a) VOWEL /ow/, (b) FRICATIVE /f/, (c) STOP /t/, (d) SEMIVOWEL /r/, and (e) NASAL /ng/.

SECTION III.

Phoneme Attractor Model

Isolated phoneme waveforms are embedded into RPSs using a pre-specified dimension d and lag τ. To address amplitude variation across phoneme instances, the reconstructed phase spaces are amplitude normalized. This is done through a radial normalization given by

$$\mathbf{x}_n = \frac{\mathbf{x}_n - \boldsymbol{\mu}_{\mathbf{x}}}{\sigma_r} \tag{3}$$

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

where

$$\sigma_r \triangleq \sqrt{\frac{1}{N-(d-1)\tau} \sum_{n=1+(d-1)r}^{N} \| \mathbf{x}_N - \boldsymbol{\mu}_{\mathbf{x}} \|_2^2 .}$$

(4)

A d-dimensional Gaussian mixture model (GMM) probability distribution is estimated over the RPS $\mathbf{X}$ for each phoneme class

$$\hat{p}(\mathbf{x}_n) = \sum_{m=1}^{M} w_m \hat{p}_m(\mathbf{x}_n) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

(5)

where M is the number of mixtures, w_m is a mixture weight and $\mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is a Gaussian distribution over $\mathbf{x}_n$ with mean $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{m}}$. These parameters are learned using the Baum Welch algorithm, beginning with a single mixture and increasing using a binary split permutation across all mixtures after each parameter estimation, until the desired number of mixtures is reached. Maximum likelihood (ML) classification is accomplished via

$$\hat{c} = \arg \max_{i=1\ldots C} \left\{ \hat{p}_i(\mathbf{X}) \right\} = \arg \max_{i=1\ldots C} \left\{ \sum_{n=1}^{N} \log \hat{p}_i(\mathbf{x}_n) \right\}$$

(6)

where C is the number of phonemes. In summary, the training process consists of learning a GMM across all the trajectory matrices data for a given phoneme, and testing consists of computing a point-by-point likelihood from those GMMs for each phoneme. The features being modeled are the time-lagged observation vectors from the original time domain signal. The statistical distribution of these observation vectors captures the attractor geometry and short-term signal dynamics, including spectral characteristics as well as nonlinear system characteristics. Long-term dynamics due to nonstationarity must be

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

captured in other ways, such as through state sequences in a Hidden Markov model or through global trajectory models[31–32,33] just as with spectral features.

The baseline method selected for comparison uses a 39-element feature vector, comprised of 12 mel-frequency cepstral coefficients (MFCCs) plus energy, augmented with delta and delta-delta (first- and second-order linear regression) coefficients. Frequency domain processing is done with the HTK toolkit,[34] using a pre-emphasis filter with frequency response of $H(z) = 1/(1 - 0.97z^{-1})$, a 25 ms hamming window and 10 ms step size, and a 24-band triangular mel-frequency filter bank with discrete cosine transformation to 12 MFCCs.

GMM implementation for both the RPS and cepstral approaches is done through a 1-state Hidden Markov Model in HTK, with a 16-mixture state distribution for the cepstral coefficients and a 128-mixture state distribution for the RPS features.

Note that since the MFCC features are frame based and the RPS features are sample point based, there are substantially more observations available for training in the RPS case, by a multiplicative factor equal to L, the frame step size. With 16 kHz signals, the 10 ms step size used here corresponds to a factor of $L = 160$. The change in observation rate also affects computation time by approximately the same linear factor.

The data set used for these experiments is TIMIT,[29,30] a speaker independent corpus that contains expertly-labeled phonetic boundary information. The original 64 phoneme TIMIT set is reduced to a 48 phoneme set for building models, and results are folded to create a 39-phoneme confusion matrix, using the approach given in.[35] For within-class recognition experiments, the five phonetic classes are given by

Vowels {tt ih ix}{tt ax ah}{tt ao aa}
tt iy eh ey ae aw ay ox ow uh uw er
Semivowels {tt el l} tt r w y hh
Stops tt b d g p t k dx
Nasals {tt n en} tt m ng
Fricatives {tt sh zh} tt jh ch s z f th v dh.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

Brackets indicate those models that are trained separately and then folded for generating accuracy results. The silence models {tt cl vcl epi sil} were not used for the within-class recognition experiments, but are included in overall accuracy numbers.

SECTION IV.

Analysis of Lag and Dimension in Phoneme Attractors

As mentioned previously, the dimension d and lag τ, the fundamental parameters of a time delay RPS, are both important and difficult to determine exactly. The dimension is perhaps of greater significance, since a sufficient dimension is a theoretical requirement for valid modeling, but lag has also been shown to have significant impact, altering the structure of the resulting RPS attractor as well as in some cases affecting the required dimension.[36,37] Methods for estimating dimension and lag are typically heuristic and sensitive to algorithm parameters, and in addition the criteria on which they are based may not be entirely generalizable to the larger goal of maximizing classification accuracy. In this section, we review the most common approaches for identifying dimension and lag, apply them across the TIMIT corpus and generate histograms of the results as a function of phonetic class. The results of these experiments are then compared with recognition accuracy results as a function of dimension and lag, with the goals of examining the impact of these parameters on accuracy and identifying whether heuristically determined values for them are adequate.

At low dimensions, there are many points along an RPS trajectory that are near each other due to projection rather than dynamics. As the dimension is increased these points, called false neighbors, "unfold" from each other into distinct neighborhoods. Once the dimension is high enough so that the attractor structure is fully unfolded, there is no benefit to any further increase, as the dimension of the attractor will be unchanged even if the dimension of the embedding space is increased. Heuristic procedures such as the false nearest neighbor method[28] take advantage of this concept to estimate the lowest dimension in which there are no false nearest neighbors. The implementation used here is taken from Abarbanel et al.[38,39] We denote $\mathbf{x}_n(d)$ as a point in an RPS of dimension d and lag τ, and define $\mathbf{x}_n^{NN}(d)$ as its nearest neighbor, the nearest point to $\mathbf{x}_n(d)$ with respect to Euclidean distance. The squared distance between these two points is

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

$$D_n(d)^2 = \| \mathbf{x}_n(d) - \mathbf{x}_n^{NN}(d) \|^2 = \sum_{i=0}^{d-1} [x_{n-i\tau}(d) - x_{n-i\tau}^{NN}(d)]^2 .$$

(7)

The difference in squared distance between dimension d and $d+1$, which indicates how far the two neighboring points have moved from each other, is then

$$D_n(d+1)^2 - D_n(d)^2 = \sum_{i=0}^{d} \left[x_{n-i\tau}(d) - x_{n-i\tau}^{NN}(d)\right]^2 - \sum_{i=0}^{d-1} \left[x_{n-i\tau}(d) - x_{n-i\tau}^{NN}(d)\right]^2 = [x_{n-d\tau}(d) - x_{n-d\tau}^{NN}(d)]^2.$$

(8)

Normalizing the square root of this difference with respect to the original distance at the lower dimension results in a ratio of how far apart two originally close points have moved, which can be compared to a threshold to identify a "false neighbor"

$$\frac{\left|x_{n-d\tau}(d) - x_{n-d\tau}^{NN}(d)\right|}{D_n(d)} > \text{Threshold} \triangleq r_T$$

(9)

and the percentage of false nearest neighbors is

$$\left(\frac{1}{N-(d-1)\tau}\right) \times \sum_{n=1+(d-1)\tau}^{N} \text{sgn}\left\{\frac{\left|x_{n-d\tau}(d) - x_{n-d\tau}^{NN}(d)\right|}{D_n(d)} - r_T\right\} \quad (10)$$

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

where $sgn(\cdot)$ is the sign function. The percentage of false nearest neighbors can then be compared to a second threshold on the order of 0.001–0.01 to select an appropriate dimension. Each of these two thresholds can have significant effect on the results of the algorithm.

There are several common techniques used for identifying the preferred time lag for an RPS, including using the first minimum of the auto-mutual information function or the first zero-crossing of the auto-correlation function.[28] Each of these functions can be poorly behaved, especially on noisy signals, occasionally giving artificially low or absurdly high values. The automutual information approach is used here, as it is slightly more common in practice. To implement this, a two-dimensional (2-D) histogram of $\{x_n, x_{n-\tau}\}$ is used to calculate the auto-mutual information function

$$I(\tau) = \sum_{i,j} p_{ij}(\tau) \ln \frac{p_{ij}(\tau)}{p_i(\tau) p_j(\tau)} \tag{11}$$

where i and j are the histogram bin indices. The first local minimum of the function $I(\tau)$ is taken as the desired lag. This process essentially finds the lag giving the least overlap of information between axes in a 2-D phase space.

Since the automutual information function is independent of RPS dimension, whereas the false nearest neighbor method requires a lag selection for embedding, the automutual information method is implemented first, and the results are used to set the reconstruction lag for the false nearest neighbor technique.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

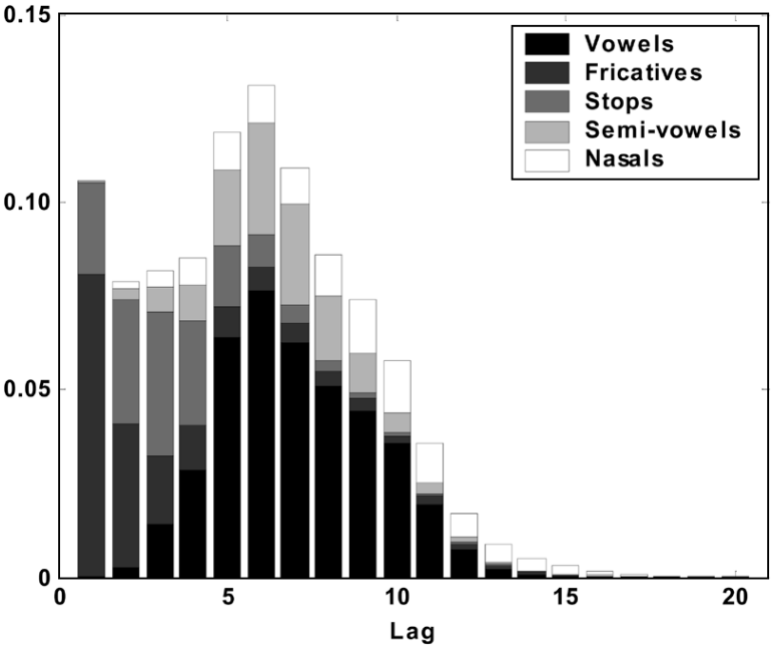

Fig. 2. Histograms of first minimum of automutual information.

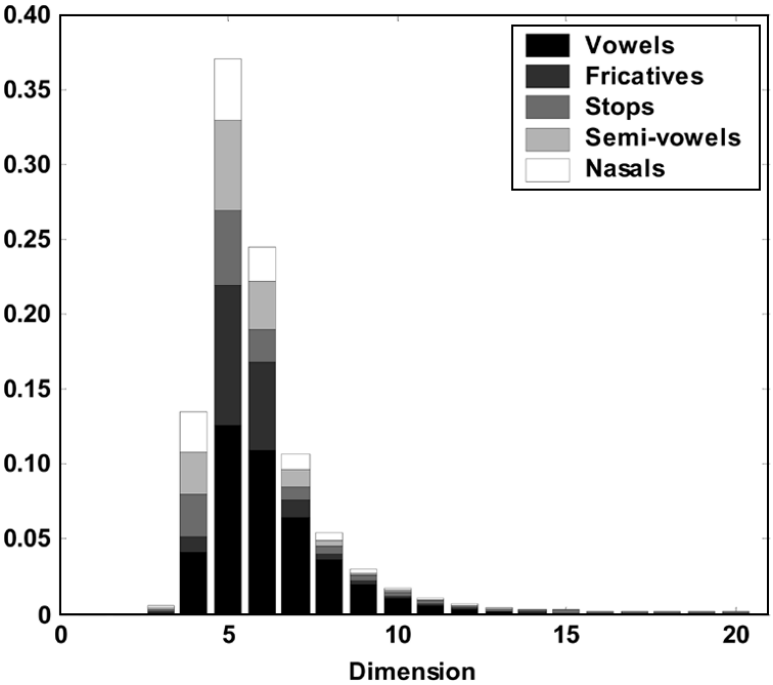

Fig. 3. Histograms of false nearest neighbor threshold crossings.

Histograms of the lag determined by the first minimum of the automutual information function across phonemes within TIMIT are shown in Fig. 2. The overall height of the bar chart represents the distribution of lags across the entire TIMIT set, while the

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

individual stacked elements within each bar indicate the breakdown across phoneme classes. There are several immediately apparent observations regarding these results, including that the distribution is quite spread out, ranging from one up to 20 or more. In addition, the breakdown of the distribution is inconsistent across the classes, indicating for example that using this criteria the selected lag for fricatives would be one whereas that for nasals would be nine.

Overall, the distribution outlined by these histograms suggests that the best lag is probably five or six based on this criterion, with six representing the peak value by a small margin. Using a lag of six as the baseline, the dimension is varied and histograms of the minimum dimension as determined by the false nearest neighbor algorithm outlined above, with $r_T = 15$, are plotted across phonemes within TIMIT. The resulting false nearest neighbor histograms are shown in Fig. 3. Again, the overall height of the chart represents the distribution of chosen dimensions across the entire TIMIT set, while the individual stacked elements within each bar indicate the breakdown across phoneme classes.

The results shown here initially seem more consistent than those used for determining lag, indicating an optimal dimension of five across all phonetic classes. This is somewhat surprising, since expectations would be that the chosen dimension for periodic signals such as vowels should be much lower than that for sounds such as fricatives. In addition, since the thresholds used in the method place a significant bias on the results, it is of interest to measure the impact of this factor as well. To

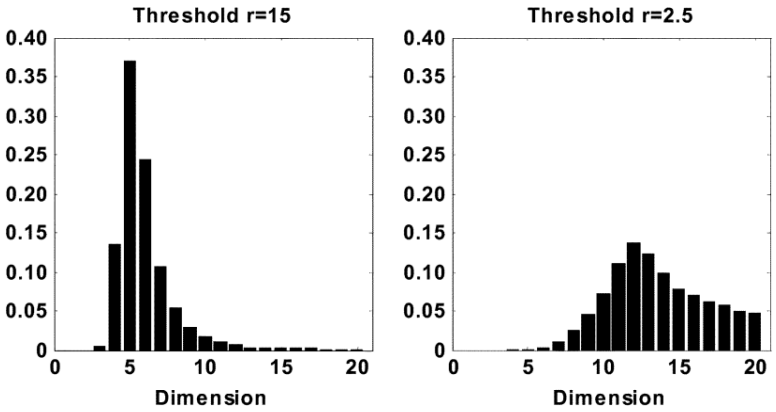

Fig. 4. Histograms of false nearest neighbor thresholds 15 and 2.5.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

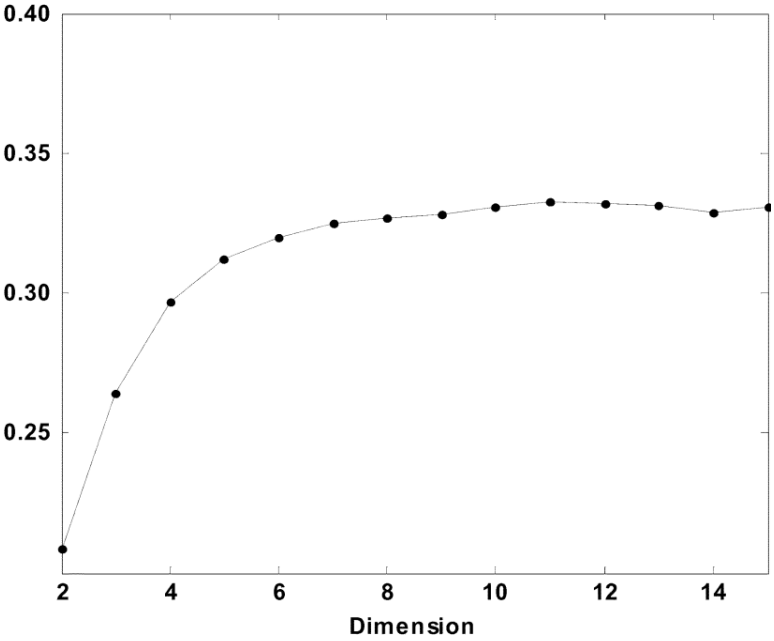

Fig. 5. TIMIT accuracy versus dimension, at lag 6.

visualize this latter effect, in Fig. 4 we compare the overall false nearest neighbor histogram from Fig. 3 with a second histogram computed using a different threshold, $r_T = 15$, on the false nearest neighbor distance ratio of (9). The resulting effect is to shift the histogram significantly to the right, indicating a much higher dimension than in the first case.

There is thus no clear interpretation regarding the best dimension to use. As a threshold of 15 is considered to be a standard value and as this value generally gives stable results in the range $10 < r_T < 50$,[39] we will use the results of the first plot of Fig. 4, which suggests that the benefits of continuing to increase dimension seem to drop off after a dimension of about 5. This indicates that a baseline choice using the standard tools might be a dimension of 5 and lag of 6.

SECTION V.

TIMIT Accuracy Results Across Lag and Dimension

To examine how well the automutual information and false nearest neighbor heuristics correlate with respect to the underlying classification task, the GMM RPS classifier described in Section 6 is tested across a wide range of lags and dimensions.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

In the first set of classification experiments the lag is held constant at 6 and the dimension is varied. Resulting accuracies across the TIMIT corpus are shown in Fig. 5.

The accuracy shown in Fig. 5 starts to asymptote around a dimension of 6, but continues increasing slowly until a dimension of about 11, at which time it plateaus and appears to begin a very gradual drop. The asymptote of 6 is consistent with the dimension chosen according to the false nearest neighbor method with a threshold of 15.

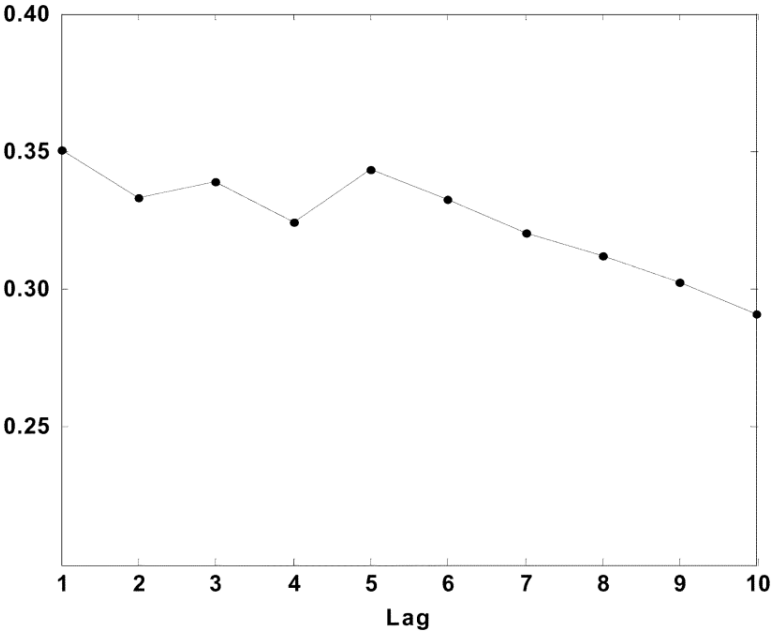

Fig. 6. TIMIT accuracy versus lag, at dimension 11.

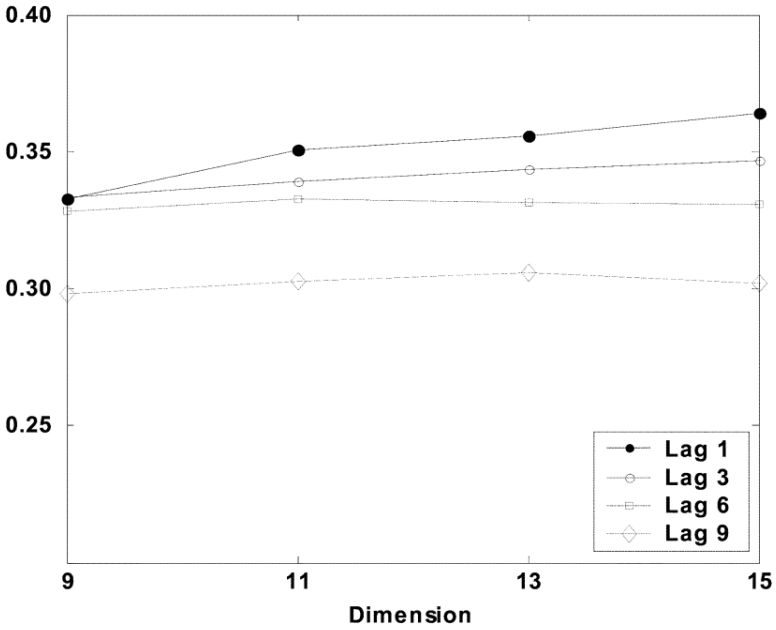

Fig. 7. TIMIT accuracy versus both lag and dimension.

Using this peak dimension, the dimension is held constant and the classification task is implemented with lag varying across a range of 1 to 10. Results are shown in Fig. 6.

While the dimension at which accuracy begins to asymptote follows roughly with the heuristic expectations, this is less the case with respect to time lag identification. It can be seen that the accuracy is highest for a lag of 1, with a decline followed by a second lower peak value at about lag 5, near the lag 6 value chosen according to the automutual information criteria. It is interesting to note though, that the shape of the accuracy curve of Fig. 6 and the automutual information histogram of Fig. 2 are both of a bimodal character, with peaks at lags of 1 and 6 in the automutual histograms and peaks of 1 and 5 for classification accuracy.

Overall results shown as a function of both lag and dimension, across lags 1, 3, 6 and 9 and dimensions 9, 11, 13, and 15, are given in Fig. 7. The overall accuracy of the system, using a lag of 1 and a dimension of 11, is 35.06%. In comparison, the baseline classification system, using a 39-element observation vector and a 16-mixture GMM, is 54.86%, indicating that the RPS method is still significantly behind the standard spectral approach.

Based on these studies, we see that the accuracy curves are smooth and relatively monotonic with respect to both lag and dimension, indicating that small adjustments in these parameters should be expected to lead to small changes in results, a conclusion which, although expected for linear system models, is not at all guaranteed for nonlinear

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

models such as these. This is an important characteristic for the RPS method, as it has already been seen that determination of lag and dimension is generally not exact and it is essential from a practical perspective that the approach be robust with respect to these parameters.

SECTION VI.

Variability Analysis and Accuracy by Phoneme Class

Dynamical systems theory shows that given sufficient dimension the RPS of a signal is a complete representation of the underlying system, including both spectral and higher-order characteristics. This does not, however, guarantee that the differences in attractor structure between phonemes, as captured by our statistical RPS models, are proportional to perceptual differences or will lead to optimal classification accuracy. Intraclass and interclass variability among attractors is a function of a number of factors, including not only lag and dimension as already discussed, but also parameters such as fundamental frequency (which affects RPS structure more than it affects cepstral features) and speaker differences, which have not been previously analyzed for this type of time-domain representation.

The affect of fundamental frequency on attractor structure[19] has been examined by using a variable-lag rather than fixed-lag RPS representation, where the lag was adjusted in proportion to the ratio of each phoneme exemplar's f_0 to the mean f_0 over the entire training set. This process essentially normalizes the periodicity of each attractor. Applied to classification of TIMIT vowels, the result was a small increase in accuracy, suggesting that while there is some variability due to f_0, the effect is not large.

The variability of attractor structure across speakers has been examined previously,[19] by comparing classification accuracy as a function of the number of speakers in a speaker-dependent task. The results showed that while accuracy is higher for the single-speaker case, it asymptotes relatively quickly and does not continue to degrade as larger numbers of speakers continue to be included. This result, combined with the overall accuracy results discussed in the previous section, demonstrates that the basic attractor structure for each phoneme class is consistent.

To investigate the relationships between perceptual and phonetic-acoustic differences and attractor structure, the class confusion matrices from the above classification experiments can be studied. The confusion matrices (available in[40]) indicate

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

that the vast majority of errors are between phonetically similar classes, with number of errors correlated with degree of phonetic similarity. The accuracy within each phoneme class is given in Fig. 8 as a function of dimension and in Fig. 9 as a function of lag. It can be seen that each class has a relatively flat accuracy curve, as was the case for the overall data set as well.

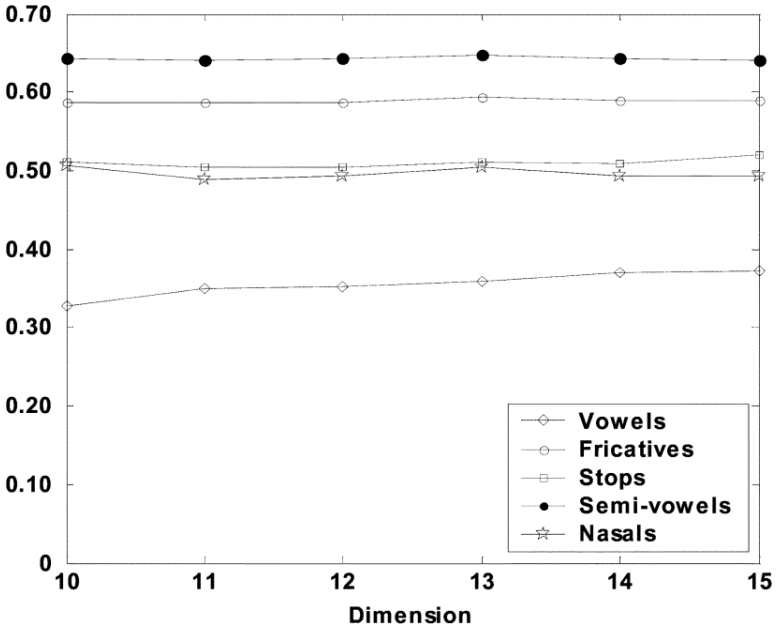

Fig. 8. TIMIT accuracy versus dimension at lag 1, by phoneme class.

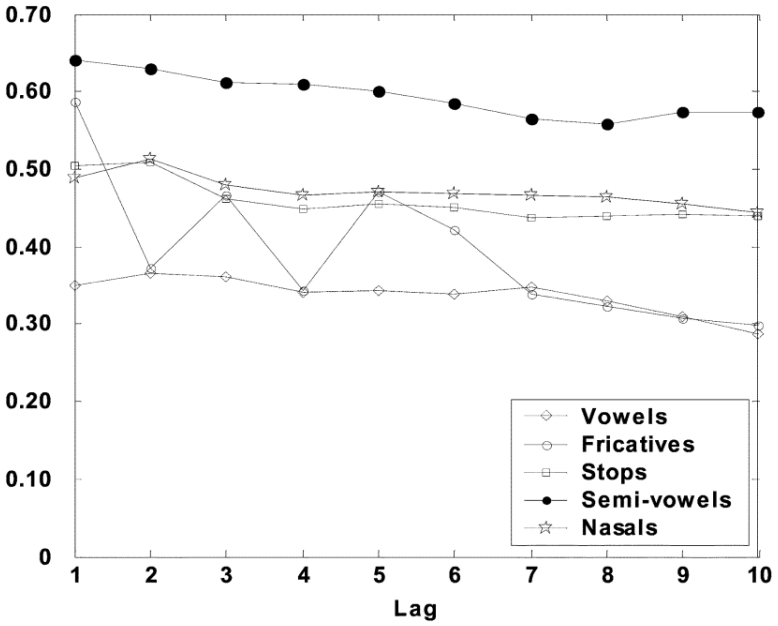

Fig. 9. TIMIT accuracy versus lag at dimension 11, by phoneme class.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

TABLE I Comparative Accuracy, by Phoneme Class

	RPS $\tau 1, d11$	39-MFCC+Δ,ΔΔ
Semivowels and glides	69.38	83.81
Affricates and fricatives	58.72	71.78
Stops	50.32	57.07
Nasals	48.94	66.67
Vowels	34.95	59.92
Overall accuracy	35.06	54.86

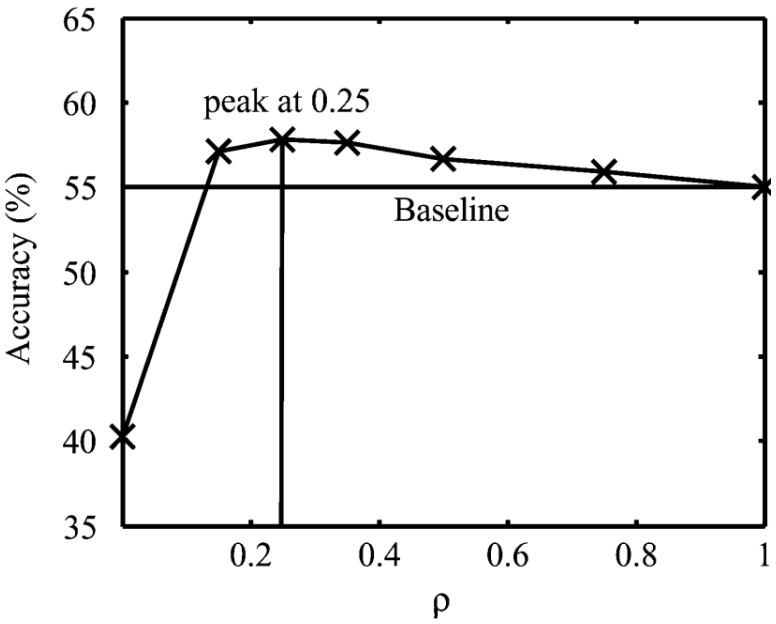

Fig. 10. TIMIT accuracy versus stream weight factor r.

The one exception is that accuracy of the fricative class is significantly affected by the selected time lag, whereas accuracy for the other classes changes only minimally. At lag 1, semivowels and glides have the highest within-class accuracy, followed by fricatives, stops, nasals, and vowels, respectively.

The results of comparisons to the MFCC based models are shown in Table I. Phoneme accuracies in percent are given for both the RPS and the MFCC models. The traditional frequency domain approach outperforms the time-domain RPS model across all of the phoneme classes, although to varying degrees. Relative to the baseline values, the RPS method performs the best on the affricates and fricatives class, and performs the worst on nasals and vowels.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

SECTION VII.

Composite RPS/MFCC Classifier

An analysis of the error patterns between the RPS classifier and the MFCC classifier indicated that many of the errors were disjoint, suggesting the possibility that the two methods could be combined to increase overall accuracy. A composite system [40] was built using the stream weight mechanism in HTK, with the time-rate mismatch between RPS points and cepstral coefficients handled by replicating the cepstral coefficients from each analysis frame for each sample. The overall likelihood score for a phoneme is then given by

$$\hat{c} = \arg max_{i=1\ldots C} \left\{ \sum_{n=1}^{N} ((1-\rho)\log \hat{p}_{\mathrm{RPS},i}(\mathbf{x}_n) + \rho \log \hat{p}_{\mathrm{MFCC},i}(\mathbf{MFCC}_n)) \right\} \quad (12)$$

where $1-\rho$ and ρ are the stream weights and $\hat{p}_{\mathrm{RPS}}$ and $\hat{p}_{\mathrm{MFCC}}$ are the GMM distributions for the RPS and the MFCC features, respectively. The RPS parameters for the composite system are a time-lag and dimension of $\tau = 6, d = 10$, where the first five dimensions are time-delay reconstructions and the next five are delta coefficients.[40]

Resulting accuracy as a function of the stream weight factor ρ is given in Fig. 10. Peak accuracy is 57.85%, an improvement of about 3% absolute error compared to the baseline system's 54.86% accuracy. Confidence interval analysis of these results indicates statistical significance level of above 0.999. The exact value of the maximizing stream weight factor should not be interpreted as indicative of relative feature strength in combination, since the differing distribution characteristics and the time-rate differentials have substantial impact on the optimal parameter value.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

SECTION VIII.

Discussion and Continuing Work

A new approach to speech representation and classification has been introduced, based on statistical models of phase spaces reconstructed from the time domain waveform. Investigation of the impact of RPS dimension and lag values indicates that representation capability as measured by recognition accuracy is relatively robust with respect to variation of those parameters, given a minimum dimension value of at least 5 or 6. Overall results indicate that statistical RPS models are able to differentiate isolated phonemes in a speaker independent task, and to increase classification accuracy when used in combination with frequency domain features. From a representation perspective, an RPS is able to capture aspects of the underlying speech production system that cannot be fully captured by spectral information, and the results presented here support further investigation of potential features and models stemming from this avenue of research.

References

[1]M. Banbrook, S. McLaughlin, "Is speech chaotic?", *Proc. IEE Colloq. Exploiting Chaos in Signal Processing*, pp. 8/1-8/8, 1994.

[2]M. Casdagli, "Chaos and deterministic versus stochastic nonlinear modeling", *J. R. Statist. Soc. B*, vol. 54, pp. 303-328, 1991.

[3]H. M. Teager, S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract", *Proc. NATO ASI Speech Production Speech Modeling*, pp. 241-261, 1990.

[4]H. Sheikhzadeh, L. Deng, "Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization", *IEEE Trans. Acoust. Speech Signal Processing*, vol. 2, pp. 80-91, 1994.

[5]A. Petry, D. Augusto, C. Barone, "Speaker identification using nonlinear dynamical features", *Chaos Solitons Fractals*, vol. 13, pp. 221-231, 2002.

[6]A. Kumar, S. K. Mullick, "Nonlinear dynamical analysis of speech", *J. Acoust. Soc. Amer.*, vol. 100, pp. 615-629, 1996.

[7]W. B. Kleijn, K. K. Paliwal, "Nonlinear speech processing" in Speech Coding and Synthesis, New York:Elsevier, 1995.

[8]S. S. Narayanan, A. A. Alwan, "A nonlinear dynamical systems analysis of fricative consonants", *J. Acoust. Soc. Amer.*, vol. 97, pp. 2511-2524, 1995.

[9]G. Wei, Y. J. Lu, J. Z. Oyang, "Chaos and fractal theories for speech signal processing", *ACTA Electron. SINICA*, vol. 24, 1996.

[10]Q. Ding, Z. Zhuang, L. Zhu, Q. Zhang, "Application of the chaos fractal and wavelet theories to the feature extraction of passive acoustic signal", *Acta Acustica*, vol. 24, pp. 197-203, 1999.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.

[11]J. Kadtke, "Classification of highly noisy signals using global dynamical models", *Phys. Lett. A*, vol. 203, pp. 196-202, 1995.

[12]J. Kadtke, M. Kremliovsky, Nonlinear Classification of Biologic Signals Using Global Dynamical Models, CA, San Diego: Inst. Pure Appl. Phys. Sci., Univ. California, 1996.

[13]F. M. Roberts, R. J. Povinelli, K. M. Ropella, "Identification of ECG arrhythmias using phase space reconstruction", *Proc. Principles and Practice of Knowledge Discovery in Databases (PKDD'01)*, pp. 411-423, 2001.

[14]R. J. Povinelli, F. M. Roberts, K. M. Ropella, M. T. Johnson, "Are nonlinear ventricular arrhythmia characteristics lost as signal duration decreases?", *Proc. Comput. Cardiol.*, pp. 221-224, 2002.

[15]R. J. Povinelli, J. F. Bangura, N. A. O. Demerdash, 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 Trans. Energy Conversion*, vol. 17, pp. 39-46, 2002.

[16]K. M. Indrebo, R. J. Povinelli, M. T. Johnson, "A combined sub-band and reconstructed phase space approach to phoneme classification", *Proc. ISCA Tutorial and Research Workshop on Non-Linear Speech Processing (NOLISP)*, pp. 107-110, 2003.

[17]X. Liu, R. J. Povinelli, M. T. Johnson, "Detecting determinism in speech phonemes", *Proc. IEEE Signal Processing Soc. 10th Digital Signal Processing Workshop*, pp. 2.3, 2002.

[18]X. Liu, R. J. Povinelli, M. T. Johnson, "Vowel classification by global dynamic modeling", *Proc. ISCA Tutorial and Research Workshop on Non-Linear Speech Processing (NOLISP)*, pp. 111-114, 2003.

[19]J. Ye, M. T. Johnson, R. J. Povinelli, "Study of attractor variation in the reconstructed phase space of speech signals", *Proc. ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP)*, pp. 5-10, 2003.

[20]J. Ye, M. T. Johnson, R. J. Povinelli, "Phoneme classification over reconstructed phase space using principal component analysis", *Proc. ISCA Tutorial and Research Workshop on Non-Linear Speech Processing (NOLISP)*, pp. 11-16, 2003.

[21]J. Ye, R. J. Povinelli, M. T. Johnson, "Phoneme classification using naive bayes classifier in reconstructed phase space", *Proc. IEEE Signal Processing Soc. 10th Digital Signal Processing Workshop*, pp. 2.2, 2002.

[22]A. C. Lindgren, M. T. Johnson, R. J. Povinelli, "Speech recognition using reconstructed phase space features", *Proc. Int. Conf. Acoustics Speech Signal Processing*, pp. 61-63, 2003.

[23]M. T. Johnson, A. C. Lindgren, R. J. Povinelli, X. Yuan, "Performance of nonlinear speech enhancement using phase space reconstruction", *Proc. Int. Conf. Acoustics Speech Signal Processing*, pp. 872-875, 2003.

[24]N. H. Packard, J. P. Crutchfield, J. D. Farmer, R. S. Shaw, "Geometry from a time series", *Phys. Rev. Lett.*, vol. 45, pp. 712-716, 1980.

[25]H. Whitney, "Differentiable manifolds", *Ann. Math.*, vol. 37, pp. 645-680, 1936.

[26]F. Takens, "Detecting strange attractors in turbulence", *Proc. Dynamical Systems and Turbulence*, pp. 366-381, 1980.

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

[28]H. Kantz, T. Schreiber, Nonlinear Time Series Analysis, U.K., Cambridge: Cambridge Univ. Press, 1997.

[29]J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, V. Zue, "TIMIT acoustic-phonetic continuous speech corpus", *Linguistic Data Consort.*, 1993.

[30]V. Zue, S. Seneff, J. Glass, "Speech database development at MIT TIMIT and beyond", *Speech Commun.*, vol. 9, pp. 351-356, 1990.

[31]Y. Gong, J.-P. Haton, "Stochastic trajectory modeling for speech recognition", *Proc. ICASSP*, pp. 57-60, 1994.

[32]L. Deng, Z. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the hidden vocal-tract-resonance dynamics", *J. Acoust. Soc. Amer.*, vol. 108, pp. 3036-3048, 2000.

[33]M. Ostendorf, V. V. Digalakis, O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition", *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 360-378, 1996.

[34]S. Young, J. Odell, D. Ollason, V. Valtchev, P. Woodland, The HTK Book, 1997.

[35]K.-F. Lee, H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models", *IEEE Trans. Acoust. Speech Signal Processing*, vol. 37, pp. 1641-1648, 1989.

[36]R. Badii, G. Broggi, B. Derighetti, M. Ravini, "Dimension increase in filtered chaotic signals", *Phys. Rev. Lett.*, vol. 60, pp. 979-982, 1988.

[37]S. H. Isabelle, A. V. Oppenheim, G. W. Wornell, "Effects of convolution on chaotic signals", *Proc. ICASSP*, pp. 133-136, 1992.

[38]M. B. Kennel, R. Brown, H. D. I. Abarbanel, "Determining minimum embedding dimension using a geometrical construction", *Phys. D*, pp. 3403-3411, 1992.

[39]H. D. I. Abarbanel, Analysis of Observed Chaotic Data, NY, New York: Springer, 1996.

[40]A. C. Lindgren, *Speech recognition using features extracted from phase space reconstructions*, 2003.

IEEE Transactions on Speech and Audio Processing, Vol 13, No. 4 (July 2005): 458-466. DOI.