# JOINT FREQUENCY DOMAIN AND RECONSTRUCTURED PHASE SPACE DERIVED FEATURES FOR SPEECH RECOGNITION

*Andrew C. Lindgren, Michael T. Johnson, Richard J. Povinelli*

Department of Electrical and Computer Engineering
Marquette University, Milwaukee, WI USA
{andrew.lindgren, mike.johnson, richard.povinelli}@mu.edu

## ABSTRACT

A novel method for speech recognition is presented, utilizing nonlinear/chaotic signal processing techniques to extract time-domain based, reconstructed phase space derived features. By exploiting the theoretical results derived in nonlinear dynamics, a distinct signal processing space called a reconstructed phase space can be generated where salient features (the natural distribution and trajectory of the attractor) can be extracted for speech recognition. To discover the discriminatory strength of these reconstructed phase space derived features, isolated phoneme classification experiments are executed using the TIMIT corpus and are compared to a baseline classifier that uses Mel frequency cepstral coefficient features (MFCCs). The results demonstrate that reconstructed phase space derived features contain substantial discriminatory power, and when the two feature sets are combined, improvement is made over the baseline. This result suggests that the features extracted using these nonlinear techniques contain different discriminatory information than the features extracted from linear approaches alone. Because they attack the speech recognition problem in a radically different manner, these reconstructed phase space derived features are an attractive research opportunity for improving speech recognition accuracy.

## 1. INTRODUCTION

In our previous work [1, 2], we demonstrated the use of reconstructed phase space (RPS) derived features for speech recognition tasks. We formulated the RPS derived feature vector, built statistical models over those features for classification, and compared our nonlinear methods to a baseline recognizer that used the traditional MFCC feature set on an isolated phoneme classification task over the TIMIT corpus. The purpose of this work is to extend the nonlinear methods we developed, in order to combine the nonlinear based RPS derived features with the traditional MFCC feature set to achieve a boost in accuracy over what each feature vector could possibly do in isolation in a speech recognizer. With this objective in mind, we briefly describe the methodology that was established in our previous work, and simultaneously develop the process by which this new joint feature vector is created.

The central premise of the nonlinear techniques presented here is that RPSs retain the nonlinear dynamics of a speech time series. A RPS is produced by establishing vectors in $\mathbb{R}^d$ whose elements are time-lagged versions of the original time series. If the original time series is given by $x[n]$ or $x_n$, where $n = 1,2,3\ldots N$, then its corresponding RPS representation is given by

$$\begin{aligned}&\mathbf{x}_n = \begin{bmatrix} x_n & x_{n-\tau} & x_{n-2\tau} & \cdots & x_{n-(d-1)\tau} \end{bmatrix}\\ &n = 1+(d-1)\tau, 2+(d-1)\tau, 3+(d-1)\tau \ldots N,\end{aligned} \tag{1}$$

where $\tau$ is the time lag and $d$ is the embedding dimension. RPSs have a strong theoretical justification provided in the nonlinear dynamics literature, and have been proven to be topologically equivalent to the original phase space of the generating system [3, 4]. Given this fact, the features extracted from RPSs may contain more and/or different discriminatory information than the typical spectral features, which are rooted in linearity assumptions of the underlying signal [5]. A typical RPS plot of a speech phoneme is given below, where $\tau = 6,\ d = 2$.

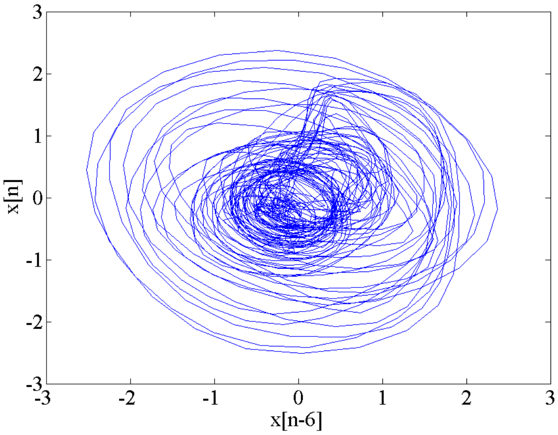

*Figure 1: Reconstructed phase space plot of the phoneme '/ow/'*

As evident from the figure, geometric structure appears in the RPS that takes the form of a bounded subset of orbits as $t \to \infty$. These geometric structures or bounded subsets of orbits are known as attractors and are revealed in Figure 1.

This material is based on work supported by the National Science Foundation under Grant No. IIS-0113508 and the Department of Education GAANN Fellowship.

## 2. RECONSTRUCTED PHASE SPACE DERIVED FEATURES

### 2.1. Time lag and embedding dimension determination

In the absence of a priori information about an experimental time series, the question of the correct choice of time lag and embedding dimension must be addressed to ensure proper reconstruction of the dynamics of the system. Two common methods frequently discussed in the literature to guide the choice of time lag are the first zero of the autocorrelation function and the first minimum of the automutal information curve. Such criteria endeavor to reduce the information redundancy between the lagged versions of the time series. Choosing too small of time lag causes the attractor to be compressed, and choosing too large of time lag causes the attractor structure too spread out as shown below.

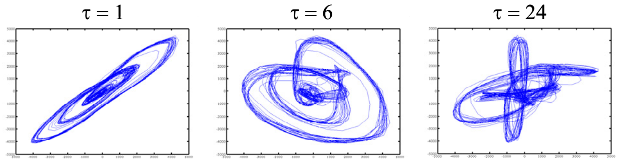

*Figure 2: Time lag comparison in RPS for a typical speech phoneme*

By examination of the autocorrelation and automual information criteria, and visual inspection, it was determine that $\tau = 6$ is an appropriate value for subsequent analysis.

For the embedding dimension choice, a well-known algorithm called false nearest neighbors can be used, which tabulates the percentage of false crossings to determine when the attractor is unfolded. Trajectories that cross indicate that the attractor is not completely unfolded, and consequently, the embedding dimension should be increased. A histogram plot (shown below) of 500 speech phonemes taken from TIMIT demonstrates at what embedding dimension the attractor is unfolded.

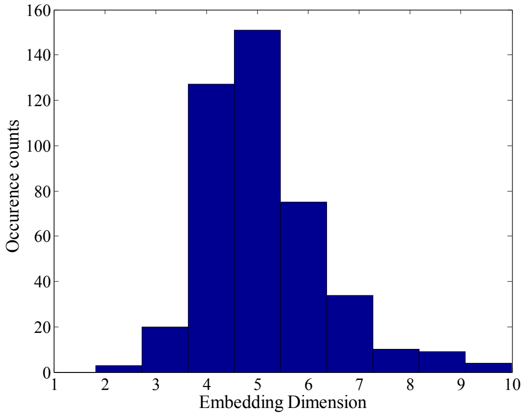

*Figure 3: False nearest neighbors for 500 random speech phonemes*

The mode of the distribution in the figure is at $d = 5$, and therefore, $d = 5$ is the embedding dimension chosen for most of the subsequent analysis.

### 2.2. Feature selection

The feature set that is finally extracted from the RPS representation relates to a quantity known as the natural distribution or natural measure of an attractor [6, 7]. The natural distribution is defined as the fraction of time that the trajectories spend in a particular neighborhood of the RPS as $t \to \infty$, and the size of the neighborhoods goes to zero. For experimental data, an estimate of the natural distribution can be performed with a Gaussian Mixture Model (GMM) built over the feature vectors given by,

$$\begin{aligned} \mathbf{x}_n^{(d,\tau)} &= \frac{\mathbf{x}_n - \boldsymbol{\mu}_\mathbf{x}}{\sigma_r}, \\ n &= 1+(d-1)\tau, 2+(d-1)\tau, 3+(d-1)\tau \ldots N, \end{aligned} \tag{2}$$

where $\mathbf{x}_n$ are vectors that constitute the RPS, $\boldsymbol{\mu}_\mathbf{x}$ is the mean vector (centroid of attractor), and $\sigma_r$ is the standard deviation of the radius in the RPS defined below,

$$\begin{aligned} \boldsymbol{\mu}_\mathbf{x} &\triangleq \frac{1}{N-(d-1)\tau} \sum_{n=1+(d-1)\tau}^{N} \mathbf{x}_n \\ \sigma_r &\triangleq \sqrt{\frac{1}{N-(d-1)\tau} \sum_{n=1+(d-1)\tau}^{N} \left\| \mathbf{x}_n - \boldsymbol{\mu}_\mathbf{x} \right\|^2}\,. \end{aligned} \tag{3}$$

The $\boldsymbol{\mu}_\mathbf{x}$ serves to zero-mean each phoneme attractor, while $\sigma_r$ normalizes out amplitude variation from phoneme to phoneme.

It is clear from Equation (2) that the natural distribution features endeavor to capture the time evolution of the attractor in the RPS as the distinguishing characteristic of speech phonemes. This feature set affirms that the natural distribution and its attractor structure (or part of it anyway), remains consistent for utterances of the same phoneme, while differing in an appreciable way among utterances of different phonemes. It is reasonable to assert this, because it makes sense to consider the fact that the system dynamics of the speech production mechanism, as captured through the natural distribution, would represent a particular phoneme utterance, and that some portion of the dynamics would approximately remain constant for a particular utterance of the same phoneme.

While this feature set does capture the position of the points in the RPS, it does not capture the flow or trajectory as the attractor evolves as illustrated in Figure 4. The trajectory information also can have discriminatory ability and can be appended to feature vector using both first difference and delta coefficients typically implemented when computing spectral features. The RPS derived feature vectors that contain the trajectory information are given in Equation (4),

$$\begin{aligned} \mathbf{x}_n^{(d,\tau,\&fd)} &= \left[ \mathbf{x}_n^{(d,\tau)} \quad \middle| \quad \mathbf{x}_n^{(d,\tau)} - \mathbf{x}_{n-1}^{(d,\tau)} \right] \\ \mathbf{x}_n^{(d,\tau,\&\Delta)} &= \left[ \mathbf{x}_n^{(d,\tau)} \quad \middle| \quad \frac{\sum_{\theta=1}^{\Theta} \theta \left( \mathbf{x}_{n+\theta}^{(d,\tau)} - \mathbf{x}_{n-\theta}^{(d,\tau)} \right)}{2\sum_{\theta=1}^{\Theta} \theta^2} \right]. \end{aligned} \tag{4}$$

It should be pointed out that the feature vectors in Equation (4) also constitute a valid RPS, since the trajectory information is a simply a linear combination of time-delayed versions of the signal.

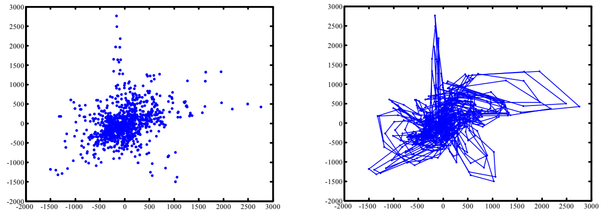

*Figure 4: RPS of a typical speech phoneme demonstrating the natural distribution and the trajectory information*

### 2.3. Joint feature vector

The RPS derived features can also be used in unison with the MFCC feature set to create a joint or composite feature vector. The reason for creating the joint feature vector is that the MFCC feature set has been successful for speech recognition in the past, and utilizing them with the RPS derived feature set will increase classification accuracy, if the information content between the two is not identical. The joint feature vector then is given by

$$\mathbf{y}_n = \left[ \mathbf{x}_n^{(d,\tau,\&\Delta)} \quad | \quad \mathbf{O}_t \right], \tag{5}$$

where $\mathbf{x}_n^{(d,\tau,\&\Delta)}$ is given in Equation (4) and $\mathbf{O}_t$ is the typical MFCC feature set (12 MFCCs, energy, deltas, and delta-deltas).

There are two central issues that arise when assembling the joint feature vector: probability scaling and feature vector time speed mismatch. The first issue arises due to the fact that the two feature sets each reside in their own unique feature space, which must be modeled differently. This difficulty will be addressed in the next section. The second issue is the result of the fact that there is one RPS derived feature for almost every time sample, while there is one MFCC feature vector for every analysis window; meaning that, there are approximately 160 RPS derived features for every 1 MFCC feature vector if one assumes an analysis window of 160 time samples. This time speed mismatch issue is solved by simply replicating the MFCCs for every RPS derived feature vector in the spectral analysis window.

## 3. MODELING TECHNIQUE

Statistical modeling of the RPS derived features was done using HTK. The model choice for both the RPS derived features and MFFC features sets was a simple one state HMM with a GMM state distribution. The model choice is justified, since the task is isolated phoneme classification, which requires a less complex model than that used during continuous recognition. The number of mixtures for the RPS derived features is set at 128. This number was derived empirically by examination of the accuracy versus number of mixtures curve described in [1]. The number of GMM mixtures necessary to achieve a high quality distribution estimate of these feature sets is quite high, because a large number is required to properly capture the complex attractor structure present in the RPS. An example of GMM modeling of the RPS derived features is shown in Figure 5. As evident, the GMM clusters accurately adjust to the attractor shape in the RPS.

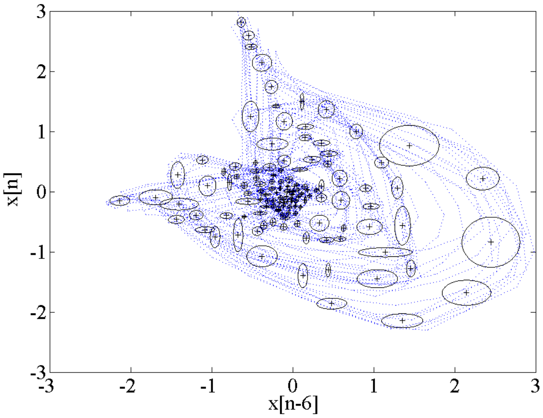

*Figure 5: GMM clusters and modeling of the RPS derived features*

As aforementioned, the joint feature vector must be modeled appropriately, because its components (RPS and MFCC) have completely different characteristics. To address this issue, the joint feature vector is modeled using two different streams, which can be implemented effortlessly in the HTK architecture [8]. One stream is for the RPS derived features and one stream is for the MFCC features. The stream model of the GMMs is given in the equation below

$$\begin{aligned} \log\left|b(\mathbf{y}_n)\right| = (1-\rho)\log\left|\sum_{m=1}^{M_1} w_{m,1} N\left(\mathbf{y}_{n,1};\boldsymbol{\mu}_{m,1},\Sigma_{m,1}\right)\right| \\ +\rho\log\left|\sum_{m=1}^{M_2} w_{m,2} N\left(\mathbf{y}_{n,2};\boldsymbol{\mu}_{m,2},\Sigma_{m,2}\right)\right| \end{aligned} \tag{6}$$

where $0 \le \rho \le 1$. The $\rho$ in the equation above is the stream weight, which must be determined empirically to ensure that the evaluation of the two distributions is scaled properly, since the number of mixtures required for the two features sets vary drastically (128 for the RPS derived features and 16 for the MFCC feature set). $\rho = 1$ is equivalent to the baseline MFFC feature set system, while $\rho = 0$ is equivalent to $\mathbf{x}_n^{(d,\tau,\&\Delta)}$ feature set system.

## 4. EXPERIMENTS

In order to investigate the performance of the feature vectors, isolated phoneme classification experiments were performed over the TIMIT corpus. The motivation for conducting isolated phoneme classification versus continuous speech recognition was to keep the focus on the acoustic data alone, using only the available acoustic data in a given segment to make a classification decision on what phoneme was uttered. Phonemes were extracted from the “SI” and “SX” sentences of TIMIT using the preexisting phonetic transcriptions and time stamps. The number of phoneme classes used in training was 48, and then 39 classes were considered for testing using the conventions discussed in [9]. Training and testing sets were taken from the predefined training and testing partitions provided in TIMIT.

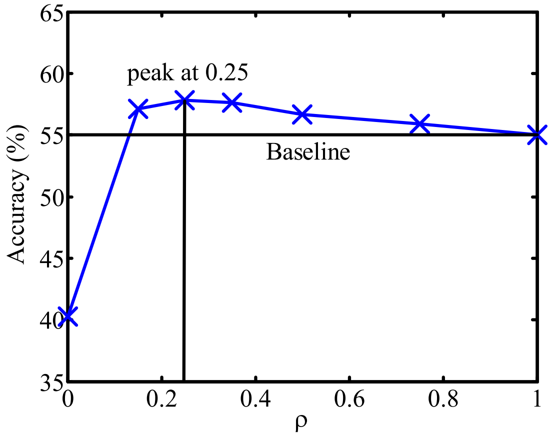

*Figure 6: Testing accuracy vs. stream weight for the joint feature vector*

| | Feature set | Test set accuracy |
|---|---|---|
| **RPS derived feature sets** | $\mathbf{x}_n^{(5,6)}$ -- RPS derived features capturing natural distribution $(d=5,\tau=6$, Total = 5 elements) | 31.43 %<br>(15017) |
| | $\mathbf{x}_n^{(10,6)}$ -- RPS derived features capturing natural distribution $(d=10,\tau=6$, Total = 10 elements) | 34.02 %<br>(16353) |
| | $\mathbf{x}_n^{(5,6,\&fd)}$ -- RPS derived features capturing natural distribution with first difference trajectory information appended $(d=5,\tau=6$ & first difference, Total = 10 elements) | 38.06 %<br>(18296) |
| | $\mathbf{x}_n^{(5,6,\&\Delta)}$ -- RPS derived feature capturing natural distribution with delta trajectory information appended $(d=5,\tau=6$ & $\mathbf{\Delta}_n$, Total = 10 elements) | 39.19 %<br>(18840) |
| **Baseline feature sets** | $\mathbf{c}_m$ -- 12 MFCC features (Total = 12 elements) | 50.34 %<br>(26372) |
| | $\mathbf{O}_m$ -- 12 MFCCs, energy, delta 12 MFCCs, delta energy, delta-delta 12 MFCCs, delta-delta energy (Total = 39 elements) | 54.86 %<br>(24199) |
| **Joint feature set** | $\mathbf{y}_n$ $\rho=0.25$ -- Joint feature vector: RPS derived feature capturing natural distribution with delta trajectory information appended & 12 MFCCs, energy, delta 12 MFCCs, delta energy, delta-delta 12 MFCCs, delta-delta energy (Total = 49 elements*)* | **57.85 %**<br>(27810) |

*Table 1: Performance comparison of the feature sets (48072 total testing examples)*

In order to determine the correct choice of $\rho$ (the stream weight), classification experiments over the testing set were run over a range of stream weights ($0 \le \rho \le 1$), and the results are given in Figure 6. The peak accuracy occurs at $\rho = 0.25$. Now that the $\rho$ is properly set, classification experiments over the testing set were performed using all of the different features sets (RPS alone, baseline alone, and joint). The results are summarized in Table 1. The joint feature vector delivered the best performance achieving 2.99 % improvement over the baseline.

## 5. DISCUSSION AND CONCLUSIONS

The results demonstrate that using RPS derived features in unison with traditional MFCC features yield improvement over the baseline alone. This result suggests that the nonlinear methods are capturing information that the MFCC features neglect that could aid in the discrimination of speech phonemes. In addition, it is clear that the incorporation of trajectory information significantly boosts the accuracy of the RPS derived features by more than 7 %. This shows that an intelligent choice of the embedding dimensions of the RPS can produce better accuracy as evident from the fact that the conventional $d=10$, RPS derived feature vector ($\mathbf{x}_n^{(10,6)}$) was inferior to the feature vectors that contained trajectory information. Additional future work will investigate the effects of amplitude scaling issues, higher dimensional RPS derived features, and the use of the RPS derived features in a continuous speech recognizer. Overall, the results show that the RPS derived features are an interesting technique to explore for increasing speech recognition accuracy.

## 6. REFERENCES

[1] A. C. Lindgren, M. T. Johnson, and R. J. Povinelli, "Speech recognition using phase space features," IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 2003 vol. I, 61-63.

[2] A. C. Lindgren, "Speech Recognition Using Features Extracted from Phase Space Reconstructions," Master's Thesis, Milwaukee, WI: Marquette University, 2003.

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

[4] F. Takens, "Dynamical systems and turbulence," in *Lecture Notes in Mathematics*, vol. 898, D. A. Rand and L. S. Young, Eds. Berlin: Springer, 1981, pp. 366-81.

[5] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, Second ed. New York, 2000.

[6] E. Ott, *Chaos in Dynamical Systems*. Cambridge: Cambridge University Press, 1993.

[7] Y. C. Lai, Y. Nagai, and C. Grebogi, "Characterization of natural measure by unstable periodic orbits in chaotic attractors," *Physical Review Letters*, vol. 79, pp. 649-52, 1997.

[8] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*: Microsoft Corporation, 2001.

[9] 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.