JOINT FREQUENCY DOMAIN AND RECONSTRUCTED PHASE SPACE FEATURES FOR SPEECH RECOGNITION

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

Department of Electrical and Computer Engineering
Marquette University, Milwaukee, WI USA
alindgren@mrcday.com, {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 features. This work examines the incorporation of trajectory information into this model as well as the combination of both MFCC and RPS feature sets into one joint feature vector. The results demonstrate that integration of trajectory information increases the recognition accuracy of the typical RPS feature set, and when MFCC and RPS 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.

1. INTRODUCTION

In our previous work [1, 2], we demonstrated the use of reconstructed phase space (RPS) features for speech recognition tasks. We formulated the RPS 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 [3] on an isolated phoneme classification task over the TIMIT corpus. The purpose of this work is two-fold. First, we explore the incorporation of trajectory information into our feature vector using delta coefficients, high-dimensional RPSs, and first difference coefficients. Second, we extend the nonlinear methods we developed, in order to combine the new features with the traditional MFCC feature set to achieve a boost in accuracy over what each feature vector could possibly do in isolation. With this objective in mind, we briefly describe the methodology that was established in our previous work.

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

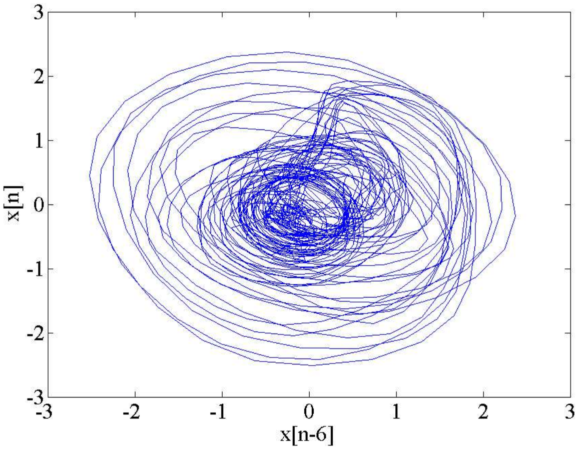

Figure 1: Reconstructed phase space plot of the phoneme '/ow/'

$$\mathbf{x}_n = [\ x_n \quad x_{n-\tau} \quad x_{n-2\tau} \quad \ldots \quad x_{n-(d-1)\tau}\ ] \qquad (1)$$
$$n = 1 + (d-1)\tau, 2 + (d-1)\tau, 3 + (d-1)\tau, \ldots N$$

where, τ 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 [4, 5]. 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. A typical RPS plot of a speech phoneme is given below, where $d = 2$ and $\tau = 6$. As evident from the figure, geometric structure appears in the RPS that takes the form of a bounded subset of orbits as $t \rightarrow \infty$. These geometric structures or bounded subsets of orbits are known as attractors and are revealed in Figure 1. In order to create a RPS representation of a time series, the correct choice of time lag and embedding dimension must be used 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 automutual information curve [6]. Such criteria endeavor to

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

0-7803-8484-9/04/$20.00 ©2004 IEEE

reduce the information redundancy between the lagged versions of the time series. By examining these criteria, we established that $\tau = 6$ was an appropriate value for subsequent analysis [1, 2]. To establish the embedding dimension, a well-known algorithm called false nearest neighbors [6] was used, which tabulates the percentage of false crossings to determine when the attractor is unfolded. By examination of a large sample of training set phonemes from TIMIT, it was determined that $d = 5$ and $d = 10$ were a suitable values for most of the subsequent analysis.

The feature set that was extracted [1, 2] is used in the estimation of a quantity known as the natural distribution or natural measure of an attractor [7, 8]. 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 RPS neighborhood regions goes to zero $(V^d \to 0)$. For experimental data, an estimate of the natural distribution can be performed with a Gaussian Mixture Model (GMM) built over the feature vectors, which are the normalized RPS data points, given by

$$\mathbf{x}_n^{(d,\tau)} = \frac{\mathbf{x}_n - \boldsymbol{\mu}_\mathbf{x}}{\sigma_r} \tag{2}$$

where $\mathbf{x}_n$ are vectors that constitute the RPS, $\boldsymbol{\mu}_\mathbf{x}$ is the mean vector (centroid of attractor), and σ_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 σ_r normalizes out amplitude variation from phoneme to phoneme.

It is clear from Equation (2) that the natural distribution estimate, which is obtained by building a Gaussian Mixture Model over the normalize RPS vectors, endeavors to capture the time evolution of the attractor as the distinguishing characteristic of speech phonemes. This estimate attempts to discriminate phonemes on the premise that the natural distribution and its attractor structure (or part of it), 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 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. The RPS feature vector then with $\tau = 6$, $d = 5$, is denoted by $\mathbf{x}_n^{(5,6)}$.

2. TRAJECTORY INFORMATION

Now that we have established the formulation of the RPS feature vector provided in [1], consideration is given to how to integrate the attractors' trajectory into the feature vector. While this natural distribution estimate does capture the position of the points in the RPS, it does not capture the flow or trajectory as the attractor evolves as illustrated Figure 2.

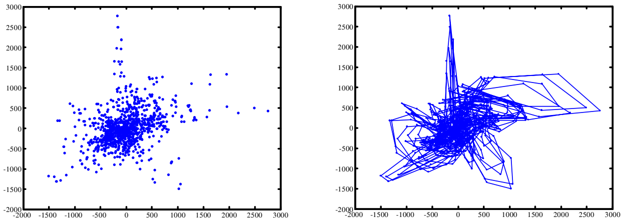

Figure 2: RPS of a typical speech phoneme demonstrating the natural distribution as well as trajectory information

The trajectory information also can have discriminatory ability and can be appended to the feature vector given in Equation (2) using both first difference and delta coefficients. The feature vectors that contain the trajectory information are given by

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

A Gaussian Mixture Model built over these features vectors jointly models both the natural distribution as well as the trajectory information. 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, which are also covered under the theory of RPS reconstruction. Given this fact, one would correctly interpret a GMM built over such a feature vector as again an estimate of the natural distribution, but over a different RPS, which was created using both the original vectors as well as trajectory dimensions. For comparison purposes then, a feature vector that is 10 dimensional, $\mathbf{x}_n^{(10,6)}$, is also used to determine whether an intelligent choice of RPS dimensions (first difference and deltas in this case) has any impact on recognition accuracy as compared to an any arbitrary/native RPS such as $\mathbf{x}_n^{(10,6)}$.

3. JOINT FEATURE VECTOR

The RPS 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 RPS feature set should increase classification accuracy, given that the information content between the two is not identical. The joint feature vector is given in Equation (5), where $\mathbf{x}_n^{(d,\tau,\&\boldsymbol{\Delta})}$ is given in Equation (4) and $\mathbf{O}_t$ is the typical MFCC feature set (12 MFCCs, energy, deltas, and delta-deltas).

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

The $\mathbf{x}_n^{(d,\tau,\&\boldsymbol{\Delta})}$ feature vector was chosen for this purpose, since it achieved the best performance as demonstrated in Section 5.

There are two central issues that arise when assembling the joint feature vector: probability scaling and feature vector rate mismatch. The first issue arises due to the fact that the two feature sets each reside in their own unique feature space with distinct characteristics and likelihoods. This difficulty will be addressed in the next section. The second issue is the result of the fact that there is a RPS feature for each time sample except endpoints, while there is one MFCC feature vector each analysis window; meaning that, there are approximately 160 RPS features for every 1 MFCC feature vector with an analysis window of 160 time samples. There are several possible ways to address this issue; in this work we simply replicate (or zero-order holding) the MFCCs for every RPS derived feature vector in the spectral analysis window.

4. MODELING TECHNIQUE

Statistical modeling of the RPS features was done using the HTK toolset [9]. The model choice for both the RPS derived features and MFFC features sets was a simple one state HMM with a GMM state distribution [1, 2]. For the task of isolated phoneme classification undertaken here, this model choice is justified because this task requires a less complex model than that used during continuous recognition. The number of mixtures for the RPS 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 is quite high, because a large number is required to properly capture the complex natural distribution and attractor structure. An example of GMM modeling of the RPS features is shown in Figure 3. As evident, the GMM clusters accurately adjust to the attractor shape in the RPS.

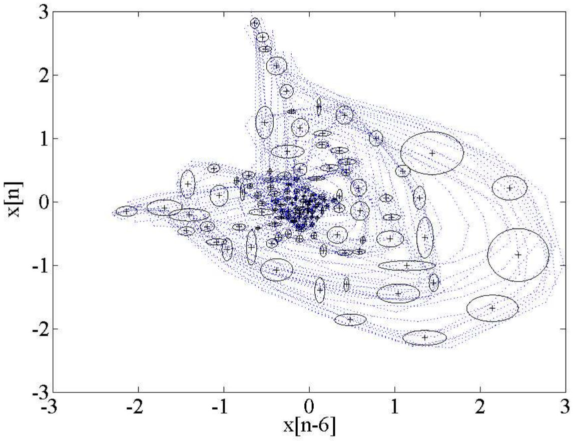

Figure 3: *GMM clusters and modeling of the RPS features*

As aforementioned, the joint feature vector must be modeled appropriately, because its components (RPS and MFCC) have completely different characteristics and time scales. To address this issue, the joint feature vector is modeled using two different streams, which can be implemented easily in the HTK architecture. One stream is for the natural distribution and other stream is for the MFCC features. The stream model of the GMMs is given by the equation

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

where $0 \leq \rho \leq 1$. The ρ 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 features and 16 for the MFCC feature set). Upon inspection of Equation (6), it is apparent $\rho = 1$ is equivalent to the baseline MFFC feature set system, while $\rho = 0$ is equivalent to $\mathbf{x}_n^{(d,\tau,\&\boldsymbol{\Delta})}$ feature set.

5. EXPERIMENTS

In order to investigate the performance of the RPS feature vectors, isolated phoneme classification experiments were performed over the TIMIT corpus. Phonemes were extracted from the ”SI” and ”SX” sentences using the preexisting phonetic transcriptions and time stamps. The original set of 64 phonetic units is grouped to form 48 phoneme classes and then the accuracies of the 48-class set are folded into 39 classes for testing using the conventions discussed in [10]. To discover the proper stream weight, the testing accuracy was found as a function of ρ. As illustrated in Figure 4, the peak accuracy occurs at $\rho = 0.25$.

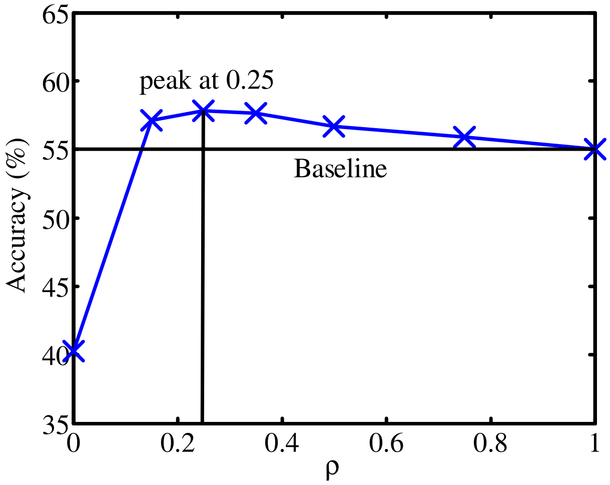

Figure 4: *Testing accuracy vs. stream weight for the joint feature vector*

The first experiments examined the question of how the RPS feature vectors that incorporate trajectory information compare to the RPS feature vector without trajectory information. The feature set that contained the deltas performed better then both the 5-d and 10-d feature sets. The second set of experiments explored how the joint vector results compared to the baseline. When comparing all of the feature vectors together, the joint feature vector delivered the best performance achieving 2.99 % improvement over the baseline.

	Feature Set	**Test Set Accuracy**
RPS feature sets	$\mathbf{x}_n^{(5,6)}$ - RPS features capturing natural distribution	31.43% (15017)
	$\mathbf{x}_n^{(10,6)}$ - RPS features capturing natural distribution	34.02% (16353)
	$\mathbf{x}_n^{(5,6,\&fd)}$ - RPS features capturing natural distribution with first difference trajectory information appended	38.06% (18296)
	$\mathbf{x}_n^{(5,6,\&\boldsymbol{\Delta})}$ - RPS feature capturing natural distribution with delta trajectory information appended	39.19% (18840)
Baseline	$\mathbf{c}_t$ - 12 MFCC features	50.34% (26372)
	$\mathbf{O}_t$ - 12 MFCCs, energy, delta, delta-deltas	54.86% (24199)
Joint feature	$\mathbf{y}_n$, $\rho = 0.25$ - RPS feature capturing natural distribution with delta trajectory information appended & 12 MFCCs, energy, delta, delta-deltas	**57.85%** **(27810)**

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

6. DISCUSSION AND CONCLUSIONS

The first set of experiments demonstrate that the incorporation of trajectory information significantly boosts the accuracy of the RPS 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 feature vector ($\mathbf{x}_n^{(10,6)}$) was inferior to the feature vectors that contained trajectory information. The results also demonstrate that using RPS 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. Additional future work will investigate the effects of amplitude scaling issues, higher dimensional RPS features, and the use of the RPS features in a continuous speech recognizer. Overall, the results show that the RPS features are an interesting technique to explore for increasing speech recognition accuracy.

7. REFERENCES

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

[2] A. C. Lindgren, “Speech recognition using features extracted from phase space reconstructions,” Master’s thesis, Marquette University, 2003.

[3] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, vol. IEEE Press, New York, second edition, 2000.

[4] T. Sauer, J. A. Yorke, and M. Casdagli, “Embedology,” *Journal of Statistical Physics*, vol. 65, no. 3, pp. 579–616, 1991.

[5] F. Takens, “Dynamical systems and turbulence,” in *Lecture Notes in Mathematics*, D A Rand and L S Young, Eds., vol. 898, pp. 366–81. Springer, Berlin, 1981.

[6] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data*, Springer-Verlag, New York, 1996.

[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, no. 4, pp. 649–52, 1997.

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

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

[10] 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.