SPEECH RECOGNITION USING RECONSTRUCTED PHASE SPACE FEATURES

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}@marquette.edu

ABSTRACT

This paper presents a novel method for speech recognition by utilizing nonlinear/chaotic signal processing techniques to extract time-domain based phase space features. By exploiting the theoretical results derived in nonlinear dynamics, a processing space called a reconstructed phase space can be generated where a salient model (the natural distribution of the attractor) can be extracted for speech recognition. To discover the discriminatory power of these features, isolated phoneme classification experiments were performed using the TIMIT corpus and compared to a baseline classifier that uses MFCC features. The results demonstrate that phase space features contain substantial discriminatory power, even though MFCC features outperformed the phase space features on direct comparisons. The authors conjecture that phase space and MFCC features used in combination within a classifier will yield increased accuracy for various speech recognition tasks.

1. INTRODUCTION

Conventional speech signal processing techniques are predicated on linear systems theory where the fundamental processing space is the frequency domain [1]. Traditional acoustic approaches assume a source-filter model where the vocal tract is modeled as a linear filter. Cepstral analysis is then performed to separate the frequency domain characteristics of the vocal tract from the excitation source. The typical feature vector used by speech recognizers that results from this signal processing procedure are Mel frequency cepstral coefficients (MFCC). Although, these features have demonstrated excellent performance over the years, they are, nevertheless, rooted in the strong linearity assumptions of the underlying physics.

As an alternative to these traditional techniques, interest has emerged in studying speech as a nonlinear system [2-6]. Under this framework the analytical focus shifts from the frequency domain to a different processing space called a reconstructed phase space. A reconstructed phase space is created by establishing vectors in $\mathbb{R}^m$, whose the elements are time-lagged versions of the original time series as given in (1).

$$\vec{x} = \{x[n], x[n-\tau], \ldots, x[n-(m-1)\tau]\}, \quad (1)$$

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

where x is the original time series, m is the embedding dimension, τ is the time lag, and n is the time index. Geometric structures emerge in this processing space that are called attractors. An example of reconstructed phase space plot for a typical speech phoneme is illustrated in Figure 1 ($\tau = 6$, $m = 2$), and its characteristic attractor is clearly revealed. Reconstructed phase spaces have been proven to be topologically equivalent to the original system and therefore are capable of recovering the nonlinear dynamics of the generating system [7, 8]. This implies that the full dynamics of the system are accessible in this space, and for this reason, a phase space reconstruction and the features extracted from it can potentially contain more and/or different information than a spectral representation.

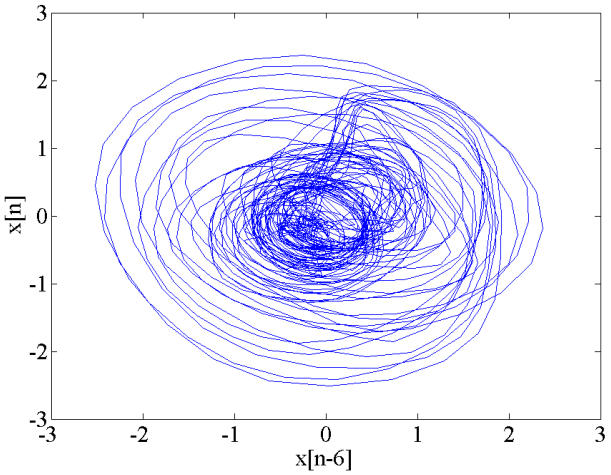

Figure 1: Reconstructed phase space plot of the phoneme '/ow/'

2. RECONSTRUCTED PHASE SPACE PARAMETERS

When presented with experimental data in the absence of expert or a priori knowledge, a key question arises when creating a phase space; how to discover the correct time lag and embedding dimension to ensure a proper reconstruction of the dynamics [9]. In order to tackle these questions, heuristics have been developed for providing guidance for a choice of the time lag and embedding dimension. A desirable property of time lags is to have as little information redundancy between the lagged

versions of the time series as possible. In order to achieve this property, the first zero of the autocorrelation function and the first minimum of the automutual information curve give an indication of what time lags are advantageous [9]. After determining the time lag, the embedding dimension can be chosen. To ascertain the embedding dimension, it is beneficial to discover the percentage of false crossings of trajectories of the attractor that occur for a particular embedding dimension. False crossings are an indication that the attractor is not completely unfolded, and therefore, the embedding dimension is too small. An algorithm called false nearest neighbors can be used to accomplish this task [9]. Given that speech is the signal of interest here, the algorithms were run over a sample of phonemes taken for the TIMIT speech corpus. Results are given in Figure 2. The graphs show that an appropriate choice of time lag is six or seven with an embedding dimension of five. Using these results, all subsequent analysis is carried out using a time lag of six and an embedding dimension of five.

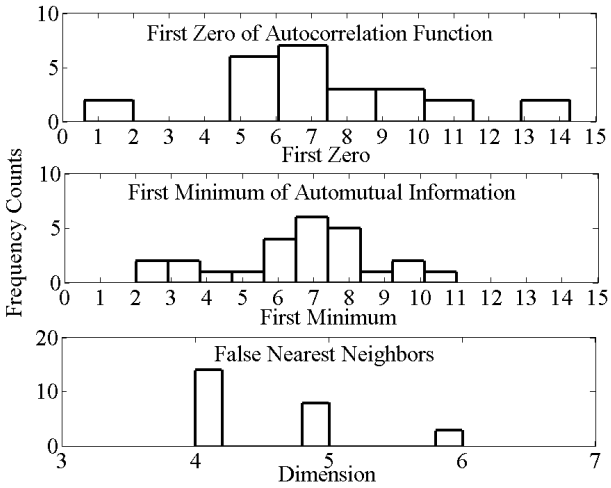

Figure 2: Histogram plots for first zero of autocorrelation, first minimum of automutual, and false nearest neighbors for a sample of speech phonemes

3. FEATURE SELECTION

Previous work in applying reconstructed phase spaces for signal processing applications focused on its use for control, prediction, and noise reduction [10]. A few studies have been performed to explore the uses of reconstructed phase spaces for feature extraction. The features that were extracted from the phase space and applied to speech recognition/classification tasks were Lyapunov exponents and correlation dimension. These features, when used in unison with cepstral features, were reported to improve speech recognition accuracy [5, 6]. These quantities are important, because they are invariant in both the original and reconstructed phase space [9]. Furthermore, they are also invariant to different initial conditions. Despite their significance, there are several issues that hinder the measurement of these quantities for experimental data. The most important drawback is their sensitivity to noise [11]. This obstacle reduces their potential for discriminability among phonemes. Additionally, the automatic computation of these quantities via a numerical algorithm can be cumbersome and even arbitrary. The overall effectiveness of such invariant measures with respect to pattern recognition tasks remains an open research question.

Another set of features that can be obtained from a reconstructed phase space relate to a quantity known as the natural measure or natural distribution of an attractor [9, 10]. The natural distribution is the fraction of time that the trajectories of an attractor spend in a particular neighborhood of the phase space, or simply, the distribution of points in the phase space as $n \to \infty$. This distribution is also independent of initial conditions and invariant, if the time series is of infinite length and the initial conditions are in the basin of attraction or attracting set [9, 10, 12]. Given experimental data, an estimation of this distribution can be performed with a Gaussian Mixture Model (GMM).

For implementation, the feature vector is given as,

$$\vec{x} = \frac{1}{\sigma_r}\left(\{x[n], x[n-\tau],..., x[n-(m-1)\tau]\} - \vec{\mu}\right), \quad (2)$$

where $\vec{\mu}$ is the mean vector (centroid of the attractor) and σ_r is the standard deviation of the radius in the phase space defined by

$$\sigma_r = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\vec{x} - \vec{\mu}\right|^2}. \quad (3)$$

The $\vec{\mu}$ and σ_r in (2) zero-mean the attractor in the phase space and normalize the amplitude variation from phoneme to phoneme. Upon examination of (2), it is apparent that the natural distribution is contained in set of feature vectors that represent the time evolution of the system at each sample point, which captures the attractor structure in the reconstructed phase space. This distribution model, consequently, endeavors to discriminate phonemes according to the similarity of their characteristic attractor structure. This model captures the position of the points in the reconstructed phase space, but not the flow or trajectory of the attractor. Such trajectory information could also have discriminatory power in classifying phonemes. In order to capture the flow as the attractor evolves, first difference information can be included in the feature vector as given by

$$\vec{x} = \frac{1}{\sigma_r}\left(\begin{Bmatrix} x[n], x[n-\tau],..., x[n-(m-1)\tau], \\ x[n]-x[n-1], x[n-\tau]-x[n-\tau-1], \\ ..., x[n-(m-1)\tau]-x[n-(m-1)\tau-1] \end{Bmatrix} - \vec{\mu}\right). \quad (4)$$

This feature vector contains the information for both the position of the embedded data points (natural distribution) and trajectory or flow of the attractor over time (first difference). This embedding, which includes the first difference elements, is also a valid reconstructed phase space according to the theory, because the first difference is merely a linear combination of time-delayed versions of the original time series [10].

4. MODELING TECHNIQUE

Statistical models are built over the reconstructed phase space

features using HTK [13]. The model is a one state Hidden Markov Model (HMM) with a Gaussian Mixture Model (GMM) observation PDF. The number of mixtures necessary to achieve a high-quality estimate far exceeds the commonly used number for MFCC features (usually 8-16 mixtures). The reason for this is that the complexity of the attractor pattern requires a large number of mixtures to adequately model it. Data insufficiency issues do not impede the estimation of a GMM with a large number of mixtures, however, because there is typically over one hundred times more data than in the MFCC case (one feature vector for each sample point). An example of the modeling technique applied to the reconstructed phase space features is demonstrated in Figure 3 for the phoneme '/aa/'. The attractor is the dotted line, while the solid lines are one standard deviation of each mixture in the model. The plot visibly demonstrates the ability of a GMM to capture the characteristic attractor structure of speech phonemes.

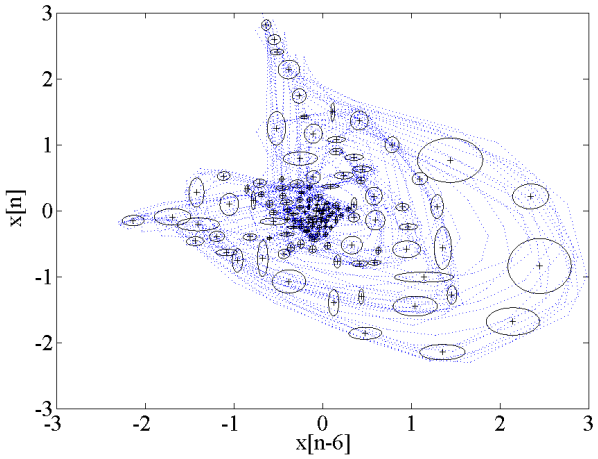

Figure 3: GMM modeling of the reconstructed phase space features for the phoneme '/aa/'

5. EXPERIMENTS

To investigate the discriminatory power of reconstructed phase space features, isolated phoneme classification experiments are performed. The motivation for performing isolated classification experiments versus continuous recognition is to determine how the features performed using only the available acoustic data in a phoneme segment, allowing one focus in on how the features compare on an acoustic level task. Using the expertly labeled, time-stamped phoneme boundaries present in TIMIT, all the phoneme exemplars from the "SI" and "SX" sentences are extracted according to these segmental boundaries. The phonemes are then folded according to the conventions discussed in [14] to give 39 classes. Using the predefined training partition present in the TIMIT corpus, parameter estimation (training) is carried out using these isolated phoneme segments. The testing is subsequently performed using isolated phonemes segments taken from the predefined testing partition.

In order to determine how many mixtures are necessary to model the reconstructed phase space features, classification experiments are performed beginning at one mixture, using binary splitting to increment the number of mixtures. Figure 4 illustrates the test classification accuracy as the number of mixtures is incremented. As evident from Figure 4, the approximate position of the elbow of the plot is at 128 mixtures. Therefore, a 128 mixture GMM properly captures the complexity of the distribution of phoneme attractors.

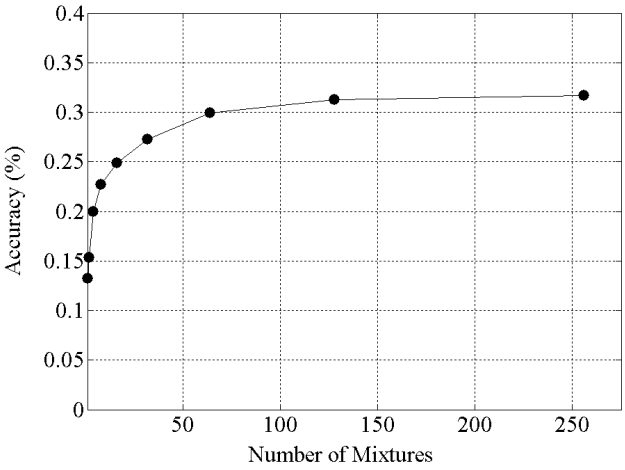

Figure 4: Classification accuracy vs. number of mixtures

Using a 128 mixture GMM, classification accuracies are compared to a baseline classifier that uses MFCC features. The parameters used for the MFCC feature extraction are 12 MFCC, log energy, deltas, and delta-deltas. A one state HMM with a 16 mixture GMM observation PDF is built over the MFCC features using HTK [13]. Results from the baseline are compared to the reconstructed phase space features and are displayed in Table 1.

Baseline	**Accuracy (%)**
12 MFCC feature set (16 mixtures)	50.34
12 MFCC, log energy, delta, delta-delta (16 mixture)	54.86
Reconstructed Phase Space	
$m = 5, \tau = 6$ (128 mixtures)	31.23
$m = 5, \tau = 6$, with first difference (128 mixtures)	38.06

Table 1: Accuracy for isolated phoneme classification

As evident from Table 1, the phase space features achieve approximately 75% the accuracy of the baseline features. Also, the first difference elements, which contain trajectory information, provide additional discriminatory power that resulted in 7 % increase in accuracy.

6. DISCUSSION AND CONCLUSIONS

The results confirm that phase space features contain significant discriminatory ability, even though MFCC features outperformed the phase space features on direct comparisons. Furthermore, the results show that the phase space features

generalize to a speaker independent task. This result is particularly interesting, because the source excitation was not removed from the phase space features as it is in the MFCC case via liftering.

The authors conjecture that phase space and MFCC features used in combination within a classifier will boost accuracy for various speech recognition tasks. There are several reasons for this conjecture. The phase space features are extracted in a different domain than MFCC features. Additionally, the MFCC features frame the data, filter out the excitation information, and discard the phase information. However, the reconstructed phase space features do not require explicit frames and retain all the signal information. It is reasonable, therefore, to presume that the information content is not equivalent between the two feature spaces.

In the light of the experimental results presented above, additional classification improvement can be made using the reconstructed phase space features. The experiments presented above utilized a five ($m = 5,\ \tau = 6$) and ten dimensional embedding ($m = 5,\ \tau = 6$ with first difference). Increased accuracy may result by employing a larger dimensional phase space to further expand the characteristic phoneme attractor structure. A larger phase space may produce larger differences between phoneme attractor characteristics, which may be overlapping in a smaller dimensional phase space. Analysis will also be performed to determine the effect that speaker variability has on the attractor structures of similar phonemes. Moreover, the phase space reconstruction methods employed in this work are general to any time series signal, and are not tailored for speech specifically. Investigation into the implementation of speech models, similar to the source-filter model employed in the linear regime, but applied to the reconstructed phase space, could yield valuable results.

A superior modeling technique will also be utilized for better performance. In future work, the authors plan on substituting the simple one state HMM (128 mixture GMM) with a fully connected HMM (one mixture GMM per state). The fully connected HMM model would capture the deterministic flow of the attractor in the phase space through the transition matrix probabilities in a convenient statistical framework. Work will also be performed to better model the direct trajectory information captured by using a better derivative estimate than a first difference approximation, such as the linear regression methods used in computing deltas and delta-deltas on spectrally based features. Also, based on the increase made through the inclusion of the first difference, higher order difference information (such as the second difference) may also boost results.

Other supplementary work will consist of building a continuous speech recognizer using phase space features. Several different issues arise, when developing a recognizer using these features. The primary issue is that the reconstructed phase space feature vector speed is over one hundred times faster than in the MFCC case (one feature vector for each time sample in the phase space case versus one feature vector every ~20 ms in the MFCC case). In order to account for the speed differential, a large HMM model (~100-150 state) might be used to capture the rapid changes of the attractor during a phoneme utterance and to account for the time duration differences. This feature vector speed differential also makes fusion between the reconstructed phase space features and the MFCC features a complex problem.

In conclusion, reconstructed phase space analysis is an attractive research avenue for increasing speech recognition accuracy. The methods have a strong theoretical justification provided by the nonlinear dynamics literature, and represent a fundamental philosophical shift from the frequency domain to the time domain, presenting an entirely different way of viewing the speech recognition problem, and offering an opportunity to capture the nonlinear characteristics of the acoustic structure. The initial experiments presented here affirm the discriminatory strength of this approach, and future work will determine their overall feasibility for both isolated and continuous speech processing applications.

7. REFERENCES

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

[2] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, 1999.

[3] G. Kubin, "Nonlinear Speech Processing," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds.: Elsevier Science, 1995.

[4] A. Kumar and S. K. Mullick, "Nonlinear Dynamical Analysis of Speech," *Journal of the Acoustical Society of America*, vol. 100, pp. 615-629, 1996.

[5] V. Pitsikalis and P. Maragos, "Speech analysis and feature extraction using chaotic models," presented at IEEE ICASSP, Orlando, Florida, 2002.

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

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

[8] 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.

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

[10] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, vol. 7, Cambridge: Cambridge University Press, 1997.

[11] E. Kostelich and T. Schreiber, "Noise reduction in chaotic time series: a survey of common methods," *Physical Review E*, vol. 48, pp. 1752-1763, 1993.

[12] 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.

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

[14] 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.