VOWEL CLASSIFICATION BY GLOBAL DYNAMIC MODELING

Xiaolin Liu, Richard J. Povinelli, Michael T. Johnson
Department of Electrical and Computer Engineering, Marquette University, WI USA
{xiaolin.liu, richard.povinelli, mike.johnson}@marquette.edu

ABSTRACT

A novel approach is presented in this paper for vowel classification by analyzing the dynamics of speech production in a reconstructed phase space. The proposed approach has the advantage of capturing nonlinear dynamics that exist in speech production. Global flow reconstruction is used to generate a compact and quantitative description of the structure and trajectory of vowel attractors in a reconstructed phase space. A distance measure is defined to quantify the dynamic similarity between phoneme attractors. Templates of the dynamics for each vowel class are selected by cluster analysis. Classifying out-of-sample vowel phonemes is done using a nearest neighbor classifier. Experiments are conducted on both speaker dependent and independent vowel classification tasks using the TIMIT corpus. The preliminary experimental results show that vowel classification by nonlinear dynamics analysis can produce very similar result when compared with a classifier using Mel frequency cepstral coefficient (MFCC) features.

1. INTRODUCTION

Traditionally, speech production has been modeled as a linear process. State-of-the-art speech recognition techniques typically use MFCC features, which are rooted in linear systems theory**.** However, recent work has suggested that nonlinearities may exist during speech production [1]. Conventional linear spectral methods cannot properly model nonlinear correlation within the signal. Therefore, methods that preserve nonlinearities may be able to achieve high classification accuracy.

This paper explores an approach to phoneme classification that captures nonlinear dynamic structures. Specifically, a study of vowel classification, which tends to have lower classification accuracies than other phoneme categories [2], is conducted. Instead of spectral analysis, the proposed approach analyzes speech dynamics using phase space reconstruction [3], which is a technique to recover a system's dynamics from observations of a single state variable. The reconstructed phase space is a plot of time-lagged signal vectors as illustrated in Figure 1. The pattern traced out in the plot is called an attractor. Takens showed that given a large enough reconstructed phase space dimension, the reconstruction is topologically equivalent to the original system [3]. Therefore, any nonlinearity existing in speech production can be captured in a reconstructed phase space.

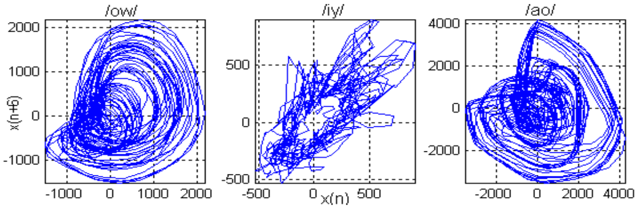

Figure 1. Attractors of phoneme /ow/, /iy/ and /ao/.

We have previously shown that vowels have deterministic attractors [4]. Thus, capturing and recognizing those dynamic structures may provide for an alternative method for vowel classification.

In this paper, global flow/vector-field reconstruction [5-7] is introduced to describe vowel dynamics in a global and quantitative manner. Dynamic similarity between vowels is quantified by defining a distance measure and thus vowels can be classified by a nearest neighbor classifier. The proposed technique is compared to a MFCC feature based classifier on vowels using the same speech data.

2. GLOBAL FLOW RECONSTRUCTION OF VOWEL DYNAMICS

Using global flow reconstruction [5-7] to quantitatively describe vowel dynamics can be regarded as a "dynamical inverse problem", which is to reconstruct an empirical dynamical system model equivalent to the one that originally generated the observed data.

Multivariate orthonormal polynomial fitting has been used for global flow reconstructions [5-8]. Gram-Schmidt orthonormalization [8] and least-square fitting [6] with monomials are two mathematically equivalent methods to construct polynomials. The method of least-square fitting with multivariate monomials [6] is used for global model reconstruction in this paper. The least-square fit is solved with singular value decomposition (SVD) because of its numerical stability [6].

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0113508.

The time-lagged signal vector is given in a form of

$$X = \left\{ x_i , x_{i+\tau} , \ldots , x_{i+(d-1)\tau} \right\},$$

where d is the embedding dimension and τ is the embedding delay. A trajectory matrix is generated by compiling these vectors.

$$\begin{pmatrix} x_1 & x_{1+\tau} & \cdots & x_{1+(d-1)\tau} \\ x_2 & x_{2+\tau} & \cdots & x_{2+(d-1)\tau} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-(d-1)\tau} & x_{n-(d-1)\tau+\tau} & \cdots & x_n \end{pmatrix}.$$

The global model is given in a sum of all possible monomials up to order P

$$F(X) = \sum_k \alpha_k x_1{}^{k_1} x_2{}^{k_2} x_3{}^{k_3} \ldots x_d{}^{k_d} ,$$

where x_i is the component of i^{th} dimension, $k = (k_1, k_2, k_3, \ldots)$ runs over all powers with the constraint $\sum_i k_i \le P$, the order of the polynomial. Least-square fitting is given in a matrix form to minimize the one-step prediction error for each dimension.

$$Error = \sum_{n=1}^{N-1} \left| X^{n+1} - F(X^n) \right|^2 = \left\| A \cdot \alpha - b \right\|^2 ,$$

where A can be decomposed into a product of orthogonal and diagonal matrices.

$$A = U \cdot diag(w_i) \cdot V^T ,$$

giving a solution in the form

$$\alpha_i = \sum_{j,k} V_{ij} \frac{1}{w_j} U_{kj} b_k .$$

To ensure the classification performance arises solely from measuring dynamic structures and not amplitude variation, the vowel signals are standardized to zero mean and unit variance to remove amplitude information.

The remaining question is if a global model can accurately describe the dynamic structures necessary for vowel classification. Ultimately, this question is answered through the classification results presented later. However, we can get a qualitative sense of the capability of global modeling by examining Figure 2. Here we can see a strong resemblance between the original attractors and their synthetic counterparts generated from the learned global models.

Synthetic trajectories are generated by iterating the global model with a seed value. A seed value is usually selected as a point from the original attractor to minimize the possibility of iterations going to infinity, the so-called blow-up. This is because the global model has no information about the neighborhood of the attractor. When a trajectory is outside of the attractor it may quickly diverge towards infinity.

However, stable synthetic attractors have been acquired for all the tested vowels for several seed values. It can be observed that the synthetic attractors have a good representation of the original attractor, at least in terms of describing the skeleton of dynamic structures. Two examples are shown in Figure 2 for the lowest three Broomhead-King coordinate projections [6].

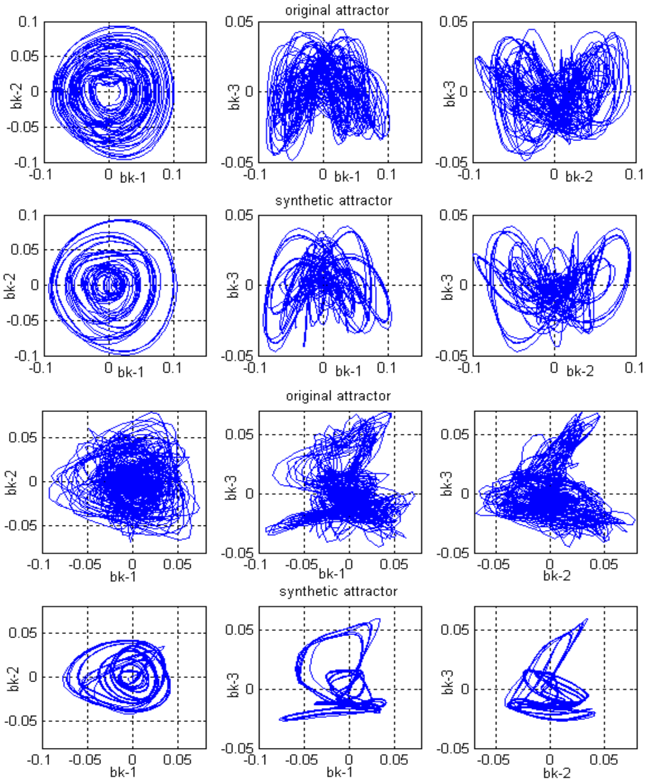

Figure 2. Dynamics representation by global models: original attractor /ow/ and /ay/: row 1 & 3; synthetic attractor: row 2 & 4.

As described in [5, 6], it is observed that minor noise contamination of the original time series helps to stabilize the generation of a synthetic attractor. The noise broadens the attracting neighborhood of the attractors, thus allowing iteration error at each step to vary within a larger range, thereby increasing the possibility of completing the iteration without blow-up.

In addition, there are cases where a stable iteration map cannot be acquired. However, the global model may still provide a good description of the dynamic structure of the original attractor. Figure 3 illustrates a global model that produces both stable and unstable synthetic maps with similar dynamic structures for different seed values. This is significant because we are not seeking to generate synthetic time series to infer quantifiable invariants of unknown dynamics, rather to ensure that dynamic structure can be well described by a global model for classification purposes.

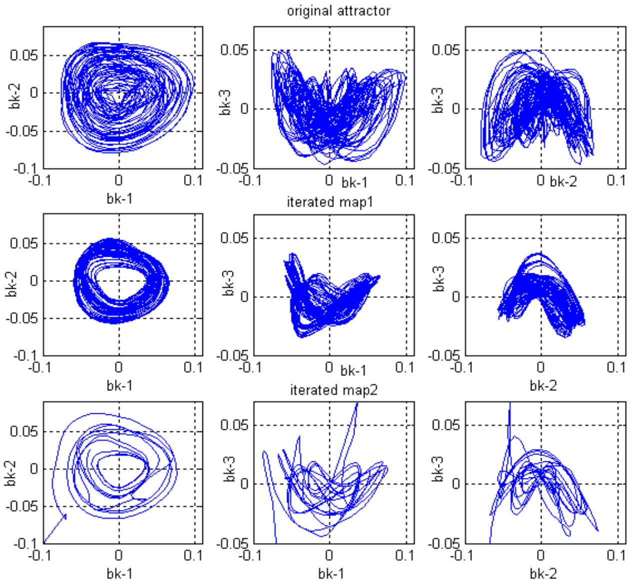

Figure 3. Global modeling of an attractor (Row 1): a stable iteration map (Row 2) and an unstable iteration map (Row 3).

3. MEASURING THE DYNAMIC SIMILARITY OF TWO ATTRACTORS

A distance measure is defined by manipulating global models to quantify dynamic similarity of time series. The basic idea is derived from the mathematical concept of distance between two time series, e.g. X_1 and X_2, of one-step cross prediction according to the estimated autoregressive models f_{X_1} and f_{X_2} [9].

$$(X_1, X_2) = \sum_{j}^{N-K} \left| f_{X_1}(x_1^j, \ldots, x_1^{j+K\tau}) - f_{X_2}(x_1^j, \ldots, x_1^{j+K\tau}) \right| ,$$
$$+ \sum_{j}^{N-K} \left| f_{X_1}(x_2^j, \ldots, x_2^{j+K\tau}) - f_{X_2}(x_2^j, \ldots, x_2^{j+K\tau}) \right|$$

where K is the order of autoregressive model and τ is the delay of input.

This distance measure is extended to a multi-dimensional vector space, where each one-step prediction from the global model F_1 and F_2 generates a vector, instead of a scalar. Dynamic similarity of two speech time series can be described by comparing the evolution of two vectors produced by the one-step cross prediction procedure. With a consideration of a vector having both a length and a direction, the quantity that describes the similarity of the dynamics between two speech time series is defined as the remainder of projecting the shorter vector to the longer vector, as shown in Figure 4.

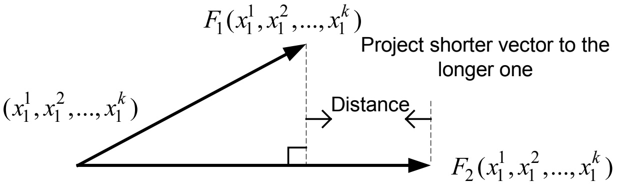
$F_1(x_1^1, x_1^2, \ldots, x_1^k)$
Project shorter vector to the longer one
$(x_1^1, x_1^2, \ldots, x_1^k)$
Distance
$F_2(x_1^1, x_1^2, \ldots, x_1^k)$

Figure 4. Distance to quantify similarity of the dynamics of two speech time series.

Cluster analysis is applied to the distance matrix acquired by calculating the distance between two arbitrary examples within a class. A cluster tree of one vowel class created by Ward's amalgamation method is shown in Figure 5. In all the vowel classes tested, the resulting created clusters have consistent appearance of the attractors. This again indicates that global modeling and the proposed distance measure are capturing the dynamic structure of vowel attractors. Templates to represent dynamic patterns of a vowel class are created by arbitrarily selecting one model from each cluster. This process reduces the risk of a mislabeled training example effecting the classification of an entire set as in the case of when a nearest neighbor classifier is used. Cluster analysis turns a nearest neighbor classifier into a nearest template classifier.

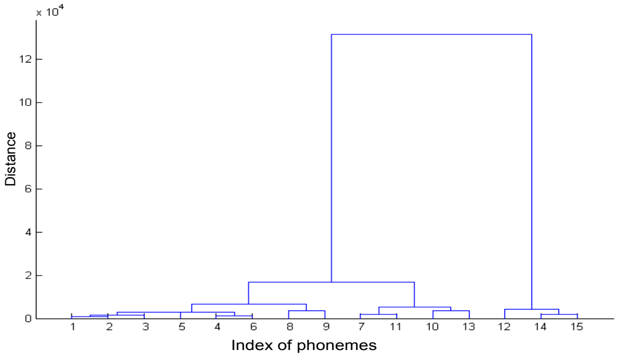

Figure 5. Example of cluster tree of an /ao/ phoneme class

4. EXPERIMENTS

Both speaker dependent and speaker independent experiments have been conducted using the TIMIT corpus. The training and testing sets consist of randomly selected 24 male speakers with three speakers from each of eight dialect regions. Speaker dependent experiments are carried out with the testing set. The parameters for the experiments are: 5-dimension embedding, 4th order of polynomial fitting, and an embedding delay of 6, which is derived by examining the first minimum of the average mutual information of speech time series.

The speaker-dependent classification results are compared with a Naive-Bayes classifier using 12-cepstral coefficients as features to a Gaussian Mixture Model. Experiment results are listed in Tables 1-3.

	/ao/	/ay/	/ey/	/ix/	/iy/	/ow/	/oy/
/ao/	121	9	0	2	0	24	1
/ay/	50	46	0	11	0	9	6
/ey/	8	4	19	49	14	8	1
/ix/	7	1	10	359	35	7	4
/iy/	4	1	11	236	129	2	0
/ow/	38	2	1	6	1	44	12
/oy/	19	2	0	0	0	11	7
Overall accuracy = 725/1331 = 54.5%							

Table 1. 24-Speaker independent test by dynamics modeling

	/ao/	/ay/	/ey/	/ix/	/iy/	/ow/	/oy/
/ao/	131	6	0	1	0	18	1
/ay/	34	67	0	16	0	4	1
/ey/	8	4	18	56	11	6	0
/ix/	3	5	10	339	58	8	0
/iy/	9	1	10	185	177	1	0
/ow/	53	3	2	9	1	35	1
/oy/	17	4	1	1	0	10	6
Overall accuracy = 773/1331 = 58.1%							

Table 2. 24-Speaker dependent test by dynamics modeling

	/ao/	/ay/	/ey/	/ix/	/iy/	/ow/	/oy/
/ao/	90	9	0	1	0	37	20
/ay/	2	97	4	8	0	0	11
/ey/	0	3	53	24	21	0	2
/ix/	0	11	66	252	59	21	14
/iy/	0	1	52	69	257	1	3
/ow/	17	7	0	6	0	49	25
/oy/	5	2	0	2	0	18	12
Overall accuracy = 810/1331 = 60.1%							

Table 3 24-Speaker dependent test using cepstral features

It is interesting to see that dynamics analysis produces similar results for both speaker dependent and independent test, which normally does not hold for the spectral analysis based phoneme recognitions.

Vowel classification by the dynamics analysis also produces very close results with the cepstral coefficients based classifier in the speaker dependent test. It can be observed that in both cases, the distribution of misclassified examples share some degree of consistency. This probably can be thought of as how the features extracted from time domain by dynamics analysis are related with the features extracted by spectral analysis. For each individual class, the classifier may have different sensitivity with these two kinds of features. It is expected that an effective combination of features from both domains will produce better results than either domain individually.

5. CONCLUSIONS

In this paper, speech is treated as a time series generated by a dynamical system. By globally modeling speech attractors and computing distance between them, we explore a new processing domain for speech recognition. Future work to improve classification results may include refining modeling technique, template selection, and distance measures. It is expected that the advantages of dynamic analysis, such as being able to capture signal nonlinearities, can be combined with traditional features and methods for both isolated and continuous speech processing applications. These preliminary results clearly indicate the potential of dynamics analysis for speech processing.

6. REFERENCE

[1] H. M. Teager and S. M. Teager, “Evidence for nonlinear sound production mechanisms in the vocal tract,” proceedings of NATO ASI on Speech Production and Speech Modelling, 1990, pp. 241-261.

[2] 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.

[3] F. Takens, “Detecting strange attractors in turbulence,” proceedings of Dynamical Systems and Turbulence, Warwick, 1980, pp. 366-381.

[4] X. Liu, R. J. Povinelli, and M. T. Johnson, “Detecting determinism in speech phonemes,” proceedings of IEEE Signal Processing Society 10th Digital Signal Processing Workshop, 2002.

[5] R. Brown, "Calculating lyapunov exponents for short and/or noisy data sets," *Physical Review E*, vol. 47, pp. 3962-3969, 1993.

[6] T. Serre, Z. Kollath, and J. R. Buchler, "Search for low - dimensional nonlinear behavior in irregular variable stars - the global flow reconstruction method," *Astronomy & Astrophysics*, 1996.

[7] G. Gouesbet and C. Letellier, "Global vector field reconstruction by using a multivariate polynomial l2-approximation on nets," *Physical Review E*, vol. 49, pp. 4955-4972, 1994.

[8] M. Giona, F. Lentini, and V. Cimagalli, "Functional reconstruction and local prediction of chaotic time series," *Physical Review A*, vol. 44, pp. 3496–3502, 1991.

[9] J. L. Hernandez, R. Biscay, J. C. Jimenez, P. Valdes, and R. Grave de Peralta, "Measuring the dissimilarity between eeg recordings through a non-linear dynamical system approach," *International Journal of Bio-Medical Computing*, vol. 38, pp. 121-9, 1995.