Vowel Classification By Global Dynamic Modeling

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

Abstract - **An 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 ability of capturing nonlinearities that may exist in speech production. Global flow reconstruction is used to generate a 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 similar result when compared with a classifier using Mel frequency cepstral coefficient (MFCC) features.**

I. Introduction

Traditionally, speech production has been modeled as a linear process. State-of-the-art speech recognition techniques typically use MFCC features, which are based 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 (RPS) 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 RPS dimension, the reconstruction is topologically equivalent to the original system [3]. Therefore, any nonlinearity existing in speech production may be captured in an RPS.

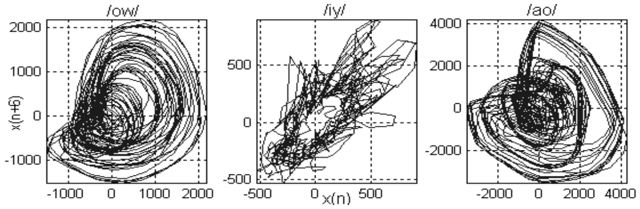

Figure 1 – Attractors of phoneme /ow/, /iy/ and /ao/.

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 approach. The proposed technique is compared to a MFCC feature based classifier.

II. 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.

Several approaches have been taken to global flow reconstruction. These include multivariate orthonormal polynomial fitting [5-8], Gram-Schmidt orthonormalization [8], and least-square fitting with monomials [6]. Because of its numerical stability, we use Serre's singular value decomposition (SVD) approach to determining the coefficients of the multivariate monomials [6].

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

$$\mathbf{x}_n = \begin{bmatrix} x_n & x_{n-\tau} & \cdots & x_{n-(d-1)\tau} \end{bmatrix} \quad n = \left(1+(d-1)\tau\right)\ldots N\ ,$$

where d is the embedding dimension and τ is the embedding delay. The trajectory matrix $\mathbf{X}$ is from the time-lagged signal vectors $\mathbf{x}$. The global model is the sum of all possible monomials up to order P

$$F(\mathbf{x}_n) = \sum_k \alpha_k {x_n}^{k_n} {x_{n-\tau}}^{k_{n-\tau}} \ldots {x_{n-(d-1)\tau}}^{k_{n-(d-1)\tau}}\ ,$$

where $k = (k_n, k_{n-\tau}, \ldots, k_{n-(d-1)\tau})$ represents all monomials with the constraint

$$\sum_{i=n}^{n-(d-1)\tau} k_i \le P\ ,$$

the order of the polynomial. The prediction error is defined as follows:

$$error = \sum_{n=1+(d-1)\tau}^{N} \left|\mathbf{x}_{n+1} - F(\mathbf{x}_n)\right|^2 = \left\|A \cdot \alpha - b\right\|^2 ,$$

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

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

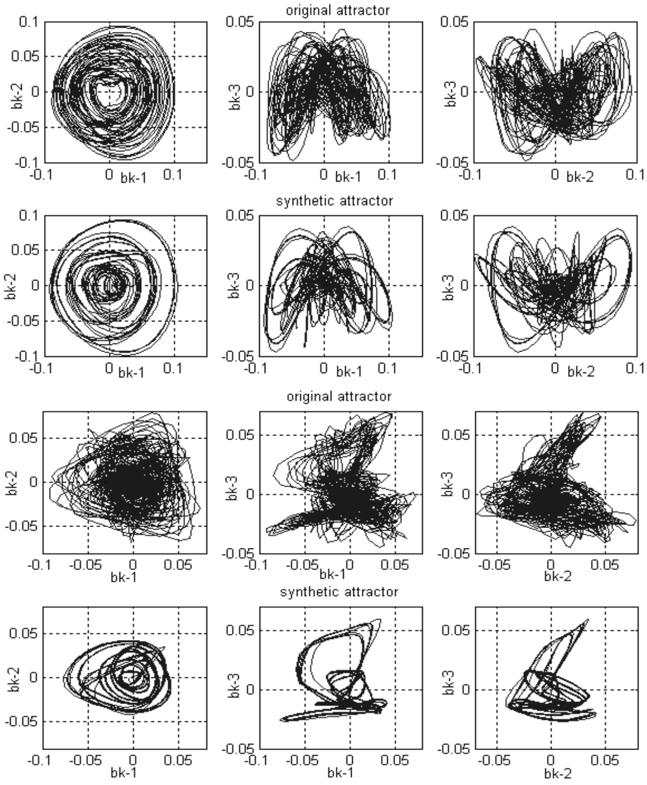

Figure 2 – Dynamics representation by global models: original attractor /ow/ and /ay/: row 1 & 3; synthetic attractor: row 2 & 4.

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

The ability of this approach to model the dynamics of vowels is first illustrated qualitatively as illustrated in Figure 2. A strong similarity between the original attractors and their synthetic counterparts generated from the learned global models is seen.

Synthetic trajectories are generated by iterating the global model with a stable initial condition. An initial condition is selected from the original attractor to minimize the possibility of iterations diverging to infinity. 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].

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.

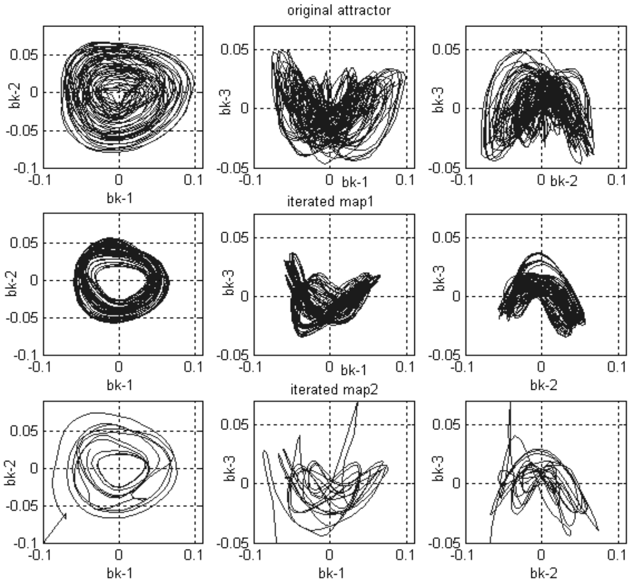

Figure 3 – Global modeling of an attractor (Row 1): a stable iteration map (Row 2) and an unstable iteration map (Row 3).

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.

III. Measuring the Dynamic Similarity of Two Attractors

One-step cross prediction error is used as a distance measure to quantify the dynamic similarity of the attractors.

$$dist(\mathbf{X},\mathbf{Y}) = \sum_{1+(d-1)\tau}^{N} f\left(F_{\mathbf{X}}\left(\mathbf{x}_n\right), F_{\mathbf{Y}}\left(\mathbf{x}_n\right)\right) + \sum_{1+(d-1)\tau}^{N} f\left(F_{\mathbf{X}}\left(\mathbf{y}_n\right), F_{\mathbf{Y}}\left(\mathbf{y}_n\right)\right),$$

where d is the embedding dimension, τ is the embedding delay, and $f(\mathbf{x},\mathbf{y})$ is a projection difference measure defined in Figure 4.

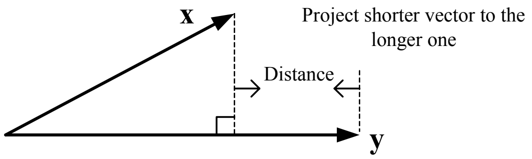

Figure 4. Distance to quantify similarity of the dynamics of two vowel attractor.

An analysis of vowel attractor patterns indicates that there are subclasses of attractors associated with each vowel. To determine these subclasses a clustering algorithm (Ward's amalgamation method) is applied to the matrix of within class attractor distances. An example cluster tree is shown in Figure 5. A random exemplar from each cluster is selected as the prototype for that subclass.

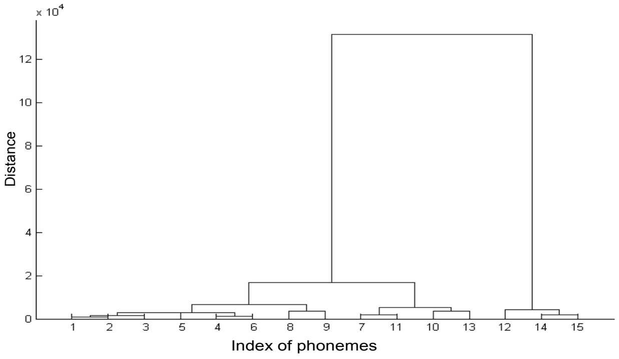

Figure 5. Example of cluster tree of an /ao/ phoneme class

IV. 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 polynomial fitting, and an embedding delay of 6, which was determined 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

Table 1. 24-Speaker independent test by dynamics modeling. Overall accuracy = 725/1331 = 54.5%

	/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

Table 2. 24-Speaker dependent test by dynamics modeling. Overall accuracy = 773/1331 = 58.1%

	/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

Table 3 24-Speaker dependent test using cepstral features. Overall accuracy = 810/1331 = 60.1%

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

Vowel classification by the dynamics analysis also produces comparable results to 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.

V. Conclusions

In this paper, speech is treated as a signal 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.

ACKNOWLEDGEMENTS

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

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, pp. 2.3.

[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*, pp. 811-833, 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.