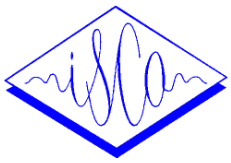

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ANALYSIS OF NON-LINEAR SPEECH GENERATING DYNAMICS

Paul A. Moakes and Steve Beet

Department of Electronic and Electrical Engineering, University of Sheffield
P.O.Box 600, Mappin Street, Sheffield S1 4DU, England.

ABSTRACT

This paper demonstrates that the non-linear system dynamics generating speech can be embedded in a low dimensional Euclidean space which resembles a manifold. Phoneme manifolds are extracted from the speech time series and compared using a radial basis function network to attempt speaker independent phoneme classification. The initial results are inconclusive but show promising results. Phoneme manifolds are also used for speech prediction resulting in a predictor which is able to achieve noise reduction equal to that of a non-linear predictor but exhibits an improvement in the signal quality.

1. INTRODUCTION

Speech can be considered as a nonstationary time series with inherently non-linear dynamics created by an underlying physical mechanism. Consequently geometric invariance measures may offer a method of estimating the underlying speech processes instead of working with the statistics of the observed signal. Deterministic mathematical models with few degrees of freedom can generate extremely complex behaviour, and if speech is deterministic then it may be possible to construct accurate low-complexity speech predictors and generators. Since the vocal tract dynamics change slowly over time compared with the speech signal, the ability to code and recreate speech using low dimensional attractors has great potential for speech compression and data reduction, it may also offer a degree of noise robustness.

Speech displays the periodic orbit of a classical chaotic system and the phase trajectories of voiced and voiceless phonemes are distinctly different [4, 8]. This has led to speculation on their use for voiced/voiceless classification. In this paper we extend this to phoneme classification by modelling the speech generating dynamics and using neural networks to compare phoneme models.

Lowe and Webb [6] have trained neural networks to model the dynamics of isolated vowels and fricatives, successfully extracting the qualitative statistics of the time series, and work has been undertaken using the residuals of LPC prediction for the identification of non-linear speech elements [11]. In this paper it is shown that working with the speech dynamics instead of the statistical properties of the time series provides a new approach to speech modelling and may yield improved results. This is investigated with the use of phoneme dynamics as the input to a neural network predictive speech filter, and comparisons are made to the performance of a non-linear filter in high levels of noise.

2. SPEECH DYNAMICS

Discrete time systems, including speech, comprise a state vector ε_k and an iterative mapping, φ, governing the evolution $\varepsilon_{k+1} = \varphi(\varepsilon_k)$ on the state space $\mathbf{S}$, this determines the system dynamics. Based on the observation that speech is a dissipative system, due to friction and other generating constraints, trajectory evolution occurs asymptotically on a sub-space of the state space [9]. In this paper it is assumed that the asymptotic dynamics are confined to an attractor $\mathbf{M}$, which is a sub-space of $\mathbf{S}$, with the structure of a smooth manifold which is locally Euclidean. This attractor has fewer degrees of freedom than $\mathbf{S}$ and therefore requires less information to describe its state [1].

Dynamical systems in state spaces of widely differing dimensions may belong to the same class if their asymptotic dynamics are confined to attracting manifolds of the same dimensionality. Physically different time series may therefore yield equivalent dynamics and all members of a given class may be represented by a general dynamic model. This is the approach to phoneme classification adopted here.

The minimum embedding dimension, d (the dimension of $\mathbf{M}$), can be calculated using the correlation dimension or the use of Lyapunov exponents. Voiced conversational speech has a dimension $d < 3$ and fricatives a dimension $d > 5$ [5, 7], suggesting that voiced speech is a deterministic but not chaotic process. $\mathbf{M}$ can be described by a two or three torus underlying attractor [7] with the frequencies associated with that attractor being related to vocal fold vibration and the correlation dimension representing the degrees of freedom in the vocal tract.

3. STATE SPACE RECONSTRUCTION

The speech time series, s_k, is generated by the evolution of ε_k on manifold $\mathbf{M}$. To model the dynamics

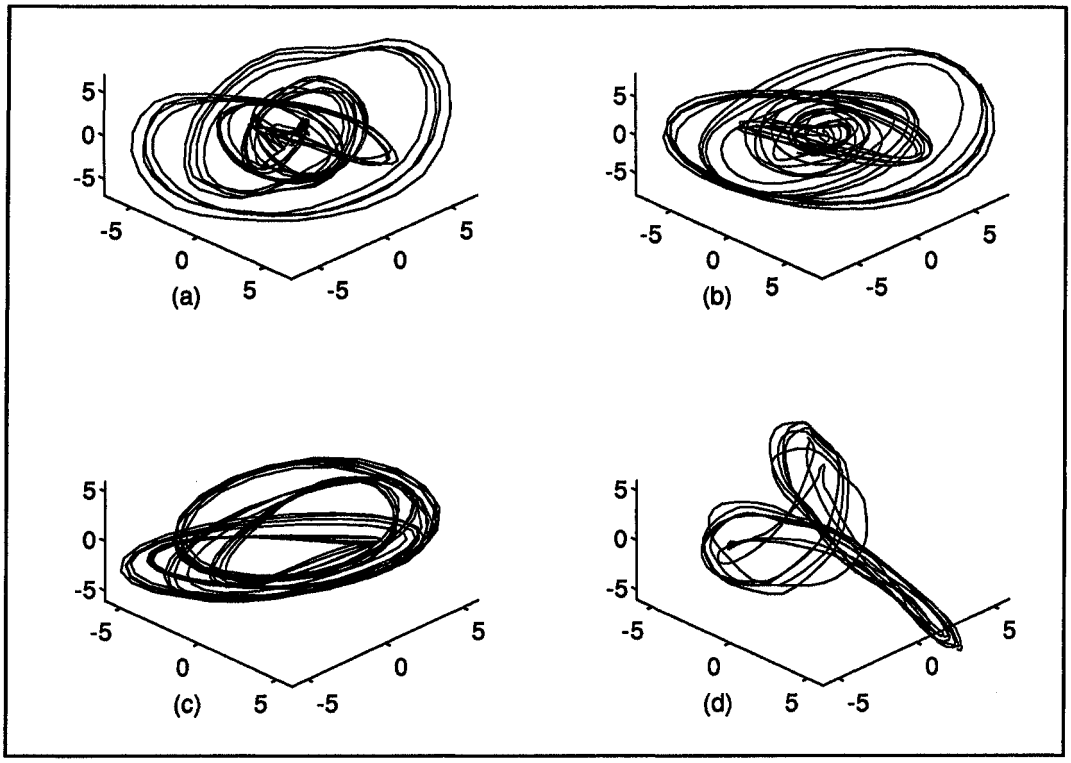

Figure 1: Trajectories of m_k on manifold **M** *for the phoneme /ae/ (a) spoken by male 1, (b) spoken by male 2, (c) spoken by female 1, and (d) shows the trajectory of the vowel /iy/ spoken by female 1.*

on **M** using only the information in s_k a diffeomorphic map, $\phi : \mathbf{S} \rightarrow \mathbf{R}^n$, is created, where $\mathbf{R}^n$ is an n-th order embedding space. To satisfy Takens' Theorem [10], n is chosen to be $n \gg 2d + 1$ such that no two vectors corresponding to distinct parts of **M** have the same elements.

The trajectory matrix W of the speech on $\mathbf{R}^n$ is found by passing an n-element sliding window over the speech to form the vector $\mathbf{w}_k$ [9] as

$$\mathbf{w}_k = \phi_n \varepsilon_k = (s_k, s_{k+1}, \ldots, s_{k+n-1})^T \qquad (1)$$

and the normalised trajectory matrix is constructed as

$$W = N^{-\frac{1}{2}}(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)^T \qquad (2)$$

where N is the number of trajectories in the matrix.

W can be decomposed into a trajectory matrix in a subspace containing the embedded manifold **M** using the singular value decomposition (SVD) $W = U\Sigma V^T$. V is the matrix of eigenvectors for W and the diagonal matrix Σ the eigenvalues in descending order. The embedding transformation of s_k onto **M**, $\varepsilon_k \rightarrow m_k$, is achieved by projecting the speech time series through the n x p submatrix V_s of V such that

$$m_k = V_s^T \mathbf{w}_k \qquad (3)$$

where p is the embedding dimension of m_k. The results of embedding phonemes from different speakers can be seen in figure 1. Visual inspection shows the similarity of manifolds for the phoneme /ae/ across all three speakers and the distinct difference of phoneme /iy/ even when spoken by the same speaker.

4. RADIAL BASIS FUNCTION NETWORKS

Radial basis function networks (RBFNs) [2] are two layer networks comprising a hidden layer and an output layer. The hidden layer contains nodes which perform a non-linear transformation of the input data. The Euclidean distance between a parameter vector (called a centre) and the input data is calculated, and the result is passed through a non-linear function to generate the node output.

The Euclidean distance, $\|x - c_j\|$, of a node can be written

$$\|x - c_j\|^2 = \sum_{v=1}^{n_c} (x_v - c_{vj})^2 \qquad (4)$$

where c_{vj} is the centre for input d on node j, x_v is element v of the input vector x, and n_c is the number of inputs to each node. The node output is given by

$$h_j = \Phi(\|x - c_j\|) \qquad (5)$$

where $\Phi(.)$ is a non-linear function. The thin-plate spline function, $\Phi(\nu) = \nu^2 \log(\nu)$, is chosen here for its non-localised response which accommodates the rapidly changing speech state-space.

Network approximation ability is dependent upon the RBFN centre locations. The centres are initially selected randomly within the bounds of the speech state-space and clustered to the training data using a variation of the Kohonen Self-Organising feature Map introduced by Huntsberger and Ajjimarangsee [3]. This fuzzy clustering approach avoids the sensitivity of κ-means clustering to the initial centre positions, preventing false minima being found.

The output layer consists of a linear combiner which calculates the weighted sum of hidden layer nodes, giv-

Test	Training Phoneme			
	/ae$_1$/	/ae$_2$/	/ae$_3$/	/ae$_4$/
/ae$_1$/	0.0111	0.1221	2.0509	0.6511
/ae$_2$/	0.0828	0.0116	1.5776	0.4190
/ae$_3$/	0.1892	0.1919	0.0073	0.2946
/ae$_4$/	0.0706	0.1125	1.2686	0.0055

Table 1: Average matching error, E, for phonemes from the same speaker in different contexts.

ing an output of

$$\hat{y}_i = \sum_{j=1}^{n_h} \eta_{ij} h_j \qquad (6)$$

where η_{ij} are the node weights and n_h is the number of hidden nodes. The response of the RBFN is linear with respect to the node output weights resulting in an output error surface with only one global minimum if the centres are fixed. In this case network training is achieved by presenting a set of input values, x, and desired network outputs, y, to the network and using the linear least squares (LLS) approach to find the optimal node weights.

5. PHONEME CLASSIFICATION

The phonemes /ae/, /iy/, /r/, and /s/ spoken by 2 male speakers and 2 female speakers in several contexts were extracted from the DARPA TIMIT speech corpus using the phonetic labelling. Each time series was normalised to have zero mean and unit variance.

To compare phoneme manifolds a RBFN was taught the manifold of a template phoneme. A window of $n = 50$ was used to construct the trajectory matrix W for the speech template using (1) and (2) and the embedded speech trajectory, m_k, on **M** was then calculated from (3) with a value of $p = 7$ chosen to satisfy Takens' Theorem for an estimated vowel correlation dimension of 3. A 20 node RBFN was used as a one step ahead trajectory predictor with $x = m_k$ and $y = m_{k+1}$. 500 x-y pairs were selected at random from the manifold and LLS used to select the optimal RBFN weights.

The trajectory m_k was calculated for the phoneme under test and the trained RBFN used to predict the trajectory. The covariance of the prediction error $\epsilon_j = m_{kj} - \hat{y}_{kj}$ was found for each embedding dimension, j, and the overall matching error, E, was calculated as

$$E = \sum_{j=1}^{p} \epsilon_j \Sigma_j \qquad (7)$$

where Σ_j is the normalised eigenvalue of dimension j from the SVD.

The comparison results are shown in tables 1-3. Although the results show a consistency of manifolds within phoneme classes amongst different speakers, table 2, it is clear from table 1 that manifolds can vary even for the same speaker. Therefore, although these results are promising, they are inconclusive and require further investigation over a larger range of utterances. In table 3 the closest manifold matches do occur within phoneme classes, which supports the view that this approach could be used for speaker and context independent phoneme classification. The largest matching

Test	Training Phoneme			
	/ae$_1$/	/ae$_2$/	/ae$_3$/	/ae$_4$/
/ae$_1$/	0.2769	0.4153	0.2044	0.5364
/ae$_2$/	0.6066	0.4910	0.5771	0.5377
/ae$_3$/	0.5319	0.4739	0.4162	0.6597
/ae$_4$/	0.5297	0.5785	0.2559	0.5562

Table 2: Average matching error, E, for phonemes from different speakers in different contexts.

Test	Training Phoneme			
	/ae/	/iy/	/r/	/s/
/ae/	0.5412	0.6315	0.6058	8.8876
/iy/	0.9048	0.6190	1.1740	14.5185
/r/	1.3705	1.5513	0.5106	21.5209
/s/	7.0633	6.8604	7.1234	5.3036

Table 3: Average matching error, E, for different phonemes from different speakers in all contexts.

errors occur between voiced and unvoiced speech, as suggested by Keller [4], and this approach appears adequate for voiced/unvoiced classification.

6. NOISE REDUCTION

To investigate the noise robustness of the dynamical approach to speech prediction a 20 node RBFN was trained as a one step ahead predictive speech filter using the phoneme trajectory m_k as the input vector, where $x = m_k$ and $y = s_k$. Training was achieved using 2000 x-y pairs selected at random from the manifold and the optimal RBFN weights found using the LLS approach. The speech signal was then recreated as the output of the RBFN over the full trajectory of m_k on **M**. For clean speech the output is given as time series A.

Unit variance white noise was then added to the speech and two approaches to noise reduction by estimating the time series from the noisy trajectories were investigated. In the first instance the RBFN was trained on the clean speech trajectory and produced time series B. In the second case the RBFN was trained using the noisy trajectory and produced time series C, without reference to the noise free signal. As a comparison, a RBFN was trained as a one step ahead predictive filter with an input of 50 lagged noisy speech samples. The trained RBFN produced time series D when presented with the whole noisy phoneme.

Figure 2 shows the resulting time series for the vowel /r/ and table 4 shows the time series signal to noise ratios (SNRs) based on the covariance of the error between the time series and the clean speech.

Phoneme	Time Series			
	A	B	C	D
/ae/	10.27	2.62	4.40	5.43
/iy/	20.76	8.37	13.48	12.12
/r/	13.11	1.75	9.46	7.99
/l/	19.91	8.32	8.85	8.68

Table 4: Comparison of predicted speech SNRs (dB) for clean speech (A) and speech with 0dB SNR (B-D).

The results show reasonably accurate prediction of clean speech with the exception of pitch events and the

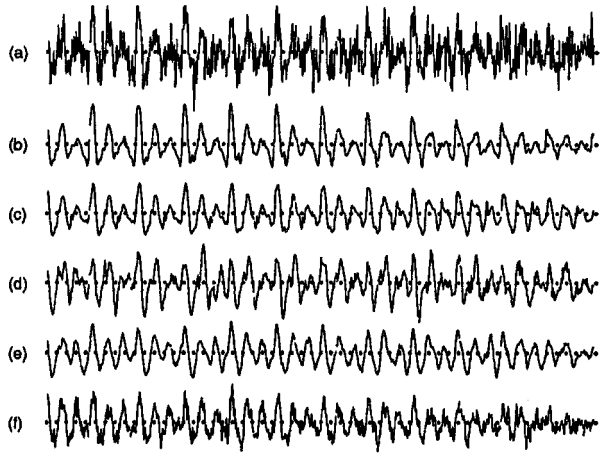

Figure 2: Reduced noise speech estimates, (a) noise corrupted vowel /r/, (b) clean speech, (c) time series A, (d) series B, (e) series C, (f) series D.

areas of phoneme onset and offset which reduce the overall SNR of the time series. This is due mainly to the assumption of dissipative and consistent dynamics in the system. The noise reduction performance of the RBFN trained and tested with noisy speech provides a better performance than the use of a RBFN trained with clean speech. This is because noisy orbits stray into regions unmodelled by clean speech dynamics, predictions then become unreliable and this is exaggerated by the unstable nature of phoneme manifolds.

The dynamics approach, time series C, is able to improve the speech SNR as well as the non-linear filter but, as can be seen in figure 2, produces a better signal quality with greater stability. This supports the theory that low-dimensional attractors offer an increased noise robustness and greater approximating capabilities than statistical time series analysis. Since the underlying dynamics are not known *a priori* there are limits to the amount of noise that can be eliminated, and all speech contains an element of noise whose source can be external or a non-deterministic component of speech. Therefore the dynamical approach to speech processing is a viable alternative to statistical speech analysis methods and provides a better quality estimated speech signal.

7. DISCUSSION

The nature of the manifold **M** is caused by the dominant frequencies in the vocal tract. Phoneme identification using the manifold is thus related to resonant frequencies in the vocal tract and could be viewed as an alternative approach to frequency domain based methods. This method therefore provides a time domain approach to determining the vocal tract dynamics and analysing the resonances of the system which, due to the low-dimensional embedding, also possesses an amount of noise robustness.

This work has demonstrated the existence of deterministic dynamics within speech and has shown that the use of speech dynamics for the identification and prediction of speech can offer an alternative to time series statistical analysis techniques. However, these results require verification with larger data sets. This would include the use of several phonemes to define the matching template and possibly the incorporation of new data as phonemes are identified. In speech time series prediction the algorithm must be extended to adequately model the pitch and to adapt the manifold to gradually changing dynamics. This would improve the SNR and the quality of the predicted speech.

8. ACKNOWLEDGEMENTS

The authors wish to thank DRA Malvern for the CASE Studentship associated with this work.

9. REFERENCES

[1] D.S. Broomhead and G.P. King, "Extracting qualitative dynamics from experimental data," *Physica D*, vol. 20, pp. 217–236, 1986.

[2] D.S. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks," *Complex Systems*, vol. 2, no. 3, pp. 321–355, 1988.

[3] T.L. Huntsberger and P. Ajjimarangsee, "Parallel self-organising feature maps for unsupervised pattern recognition," *Int. J. General Systems*, vol. 16, no. 4, pp. 357–372, 1990.

[4] E. Keller, "Phase representations of acoustic speech waveforms," in *Visual Representations of Speech Signals*, M. Cooke, S.W. Beet, and M. Crawford, Eds., Chichester: J. Wiley & Sons, 1993, pp. 285–292.

[5] A. Kumar and S.K. Mullick, "Attractor dimension, entropy and modelling of speech time series.," *Electronics Letters*, vol. 26, no. 21, pp. 1790–1791, October 1990.

[6] D. Lowe and A. Webb, "Adaptive networks, dynamical systems, and the predictive analysis of time series," in *Proc. First IEE Int. Conf. on Artificial Neural Networks*, London, 16-18 Oct. 1989, pp. 95–99.

[7] S. McLaughlin and A. Lowry, "Nonlinear dynamical systems concepts in speech analysis," in *Proc. EUROSPEECH 93*, Berlin, GERMANY, 21-23 Sept. 1993, pp. 377–380.

[8] S.S. Narayanan and A.A. Alwan, "Strange attractors and chaotic dynamics in the production of voiced and voiceless fricatives," in *Proc. IEEE ICASSP 93*, 1993, pp. 77–80.

[9] M.A.S. Potts and D.S. Broomhead, "Time series prediction with a radial basis function neural network," in *SPIE Proc. Adaptive Signal Processing*, San Diego, CA, USA, 1991.

[10] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*, D.A. Rand and L.-S. Young, Eds., Berlin: Springer-Verlag, 1981, pp. 366–381.

[11] B. Townshend, "Nonlinear prediction of speech," in *Proc. IEEE ICASSP 91*, Toronto, Canada, 1991, vol. I, pp. 425–428.