RADIAL BASIS FUNCTION NETWORKS FOR NOISE REDUCTION OF SPEECH

P A Moakes and S W Beet

University of Sheffield, UK

ABSTRACT

This paper compares eigen-decomposition, linear and non-linear Radial Basis Function network based filter prediction, and signal embedding for the removal of additive white gaussian noise from speech. The embedding based predictor achieves improved noise reduction in comparison with eigen-decomposition but does not perform as well as a non-linear prediction. However, signal embedding offers a compromise between retaining the high frequency spectral information found in non-linear prediction while achieving low frequency noise reduction comparable to eigen-decomposition.

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 as demonstrated in Potts and Broomhead [8]. 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, as investigated in Moakes and Beet [6]. 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 which is investigated in this paper.

Transmitted speech generally contains noise and the performance of automatic speech recognisers and speech vocoders based on linear prediction co-efficients drops drastically if tested with noise contaminated speech due to the mismatch of noise conditions in the reference and test patterns. Our aim is to reduce the noise content of transmitted speech, preserve or enhance intelligibility, and introduce as little distortion as possible. This can be achieved by pre-processing to remove non-speech signals and provide a better set of speech parameters for synthesis and recognition.

Digital signal processing based on subspace methods, namely principal component analysis and eigen-decomposition, offer high resolution spectral analysis (Slyh and Moses [9]), parameter estimation of noisy signals, and speech enhancement (Dendrinos et al. [3]). Spectral subtraction techniques offer similar results but require signal analysis in the frequency domain and the estimation of the background noise characteristics (a supervised approach). Most time domain speech enhancement has concentrated on linear prediction approaches, and Kalman filter based noise reduction techniques in particular have proved beneficial to a range of speaker verification methods.

To allow for, and even exploit, the non-linearity and non-stationarity of the speech, and make use of the dynamics of the underlying physical mechanism of speech production, new methods are required. Signal sub-space noise reduction and non-linear embedding have been applied to speech noise reduction in phonemes in [6] where embedded speech dynamics are the input to a neural network predictive speech filter. Here we extend this work to an on-line adaptive noise reduction technique for use with longer speech segments and compare our results with existing eigen-decomposition techniques and both linear and non-linear auto-regressive moving average (ARMA) filters. The non-linear approaches use a novel on-line adaptive Radial Basis Function network (RBFN) technique which has recently shown excellent noise robustness for speech analysis and pitch detection in Moakes and Beet [7].

ALTERNATIVE APPROACHES

Eigen-decomposition

One of many noise robust approaches for speech prediction is the backward linear predictor [9] where future samples are used to estimate the auto-regressive (AR) filter parameters. Using the filter on N windows of a speech signal s, n samples wide, the prediction equations can be written in matrix form as

$$\begin{bmatrix} s_1 & s_2 & . & s_n \\ s_2 & s_3 & . & s_{n+1} \\ \vdots & \vdots & & \vdots \\ s_N & s_{N+1} & . & s_{N+n} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix} = \begin{bmatrix} -s_0 \\ -s_1 \\ \vdots \\ -s_{N-1} \end{bmatrix} \quad (1)$$

or $Wg = -h$ where W has Hankel structure and g is a column vector defining the filter impulse response.

Linear prediction methods are attractive for their computational simplicity, but have a poor performance at low signal to ratios (SNRs), hence the noise must be separated from W. W can be decomposed as

$$W = \sum_{i=1}^{n} \sigma_i u_i v_i^T \quad (2)$$

using singular value decomposition (SVD), where u and v are the left and right singular vectors and σ are the singular values arranged in non-increasing order. If W is considered to comprise a signal subspace W_s of dimension $r : r < n$ and a noise subspace W_n which are mutually orthogonal, the signal can be recreated using the first r components of the SVD with the greatest information content. The result is a lower rank, minimum norm approximation of W, $\tilde{W}_r$, with a reduced noise content and signal space dimension [3] which can be written

$$\tilde{W}_r = \sum_{i=1}^{r} \sigma_i u_i v_i^T \quad (3)$$

r is dependent upon the spectral content of the signal frame and the noise level and the objective is to select r such that $\tilde{W}_r$ is closest to W_s in a minimum norm sense giving a reconstructed signal closest to the original signal. Optimal selection of r is a major problem in which significantly large, $\sigma_1 \ldots \sigma_r$, and insignificantly small, $\sigma_{r+1} \ldots \sigma_n$, singular values must be distinguished. Konstantinides and Yao [5] introduced an unsupervised matrix rank determination method using the statistical analysis of the singular values which is adopted in this paper.

$\tilde{W}_r$ often has a non-Hankel structure which can be restored by averaging the multiple occurrences of each sample along the anti-diagonals, regenerating the noise free estimate of the signal. This enhancement is strongly non-linear but fully automatic with results comparable to those of least mean square filters and spectral subtraction.

There are several problems with this approach which we aim to overcome with a non-linear technique. The signal has to be orthogonal to the noise, the noise has to be white, and the smallest eigenvalue of the signal must be greater than the largest eigenvalue of the noise.

Signal Embedding

Speech can be considered to be generated from a state vector ε_k and an iterative mapping, $\varepsilon_{k+1} = \varphi(\varepsilon_k)$, governing the evolution on the state space $\mathbf{S}$ which determines the system dynamics [6]. Since speech is a dissipative system the dynamics are confined to a sub-space of $\mathbf{S}$ with fewer degrees of freedom than $\mathbf{S}$ which requires less information to describe its state (Broomhead and King [1]).

To model the dynamics, an embedding space is created using the method of delays to form the matrix W of equation (1) [8]. W comprises data vectors which are generated by passing an n-element sliding window over the speech to form the vector $\mathbf{w}_k$ as

$$\mathbf{w}_k = \varphi_n(\varepsilon_k) = (s_k, s_{k+1}, \ldots, s_{k+n-1})^T \quad (4)$$

and the trajectory matrix is constructed as

$$W = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)^T \quad (5)$$

W can be decomposed into a trajectory matrix in a subspace containing the embedded manifold $\mathbf{M}$ using SVD, where $W = U\Sigma V^T$ and the embedding transformation is achieved by projecting the speech time series through the first r principal right eigenvectors, V_r [8]. The embedded signal m_k is thus given by

$$m_k = V_r^T \mathbf{w}_k \quad (6)$$

The optimal value of r can be found from the signal correlation dimension or the use of Lyapunov exponents. Here it is equivalent to the estimated optimal rank r of $\tilde{W}_r$ in eigen-decomposition. Embedding speech in this way preserves the non-linearity of the signal, which is ordinarily lost during SVD smoothing and linear prediction, since the original speech is only projected in the direction of the linear principal components, not filtered.

Radial Basis Function Networks

RBFNs (Broomhead and Lowe [2]) are two layer networks comprising a hidden layer and an output layer. The hidden layer contains nodes which calculate the Euclidean distance between a centre c_j and an input vector x. The result is passed through a non-linear function to generate the node output, h_j, which can be written

$$h_j = \Phi(\|x - c_j\|) \tag{7}$$

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 at the input vector.

Network approximation 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 in [4]. This fuzzy clustering approach avoids the sensitivity of κ-means clustering to the initial centre positions, preventing false minima being found.

The output layer comprises a linear combiner which calculates the weighted sum of hidden layer nodes, giving an output of

$$\hat{y}_i = \sum_{j=1}^{n_h} \eta_{ij} h_j \tag{8}$$

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. In this on-line application, network training is achieved by presenting input data x_k and desired output values y_k to the network sequentially. Weight adaptation is achieved using a Kalman filter technique with a forgetting factor [7] to minimise the mean square prediction error.

An RBFN can approximate arbitrarily well any continuous function on a compact domain if sufficient basis functions are used. The changing complexity of the speech dynamics implies that the number of nodes required will vary depending upon the speech frame, so Kolmogorov's theorem is used to set the initial number of nodes arbitrarily at $2m + 1$, where m is the dimension of the input vector x. However, this figure assumes independence in the dimensions of the input vector and the optimal number of nodes is selected by experiment.

Non-linear ARMA Filters

An AR predictive filter comprises a weighted sum of delayed speech samples. The weights can be adapted using a Kalman filter to minimise the mean square prediction error. A more accurate model of speech incorporates the pitch as a source signal, u, in an ARMA model and requires the estimation of u_k at each sample. Ting and Childers in [10] introduced an unsupervised approach to estimate the source signal from an error driven variable forgetting factor, λ_k, in the Kalman filter equation. When the Kalman filter is used to adapt RBFN weights a similar approach to pitch detection is seen to work with improved results [7].

The linear ARMA model output is a weighted sum of the elements of an input vector x_k comprising n delayed speech samples s_k and m delayed estimate of the source $\hat{u}_k$. The input vector is given by

$$x_k = [s_{k-1}, \ldots, s_{k-n}, \hat{u}_{k-1}, \ldots, \hat{u}_{k-m}] \tag{9}$$

The non-linear ARMA model comprises an RBFN, described above, with the input vector x_k. The weights of both filters are adapted using identical Kalman filter equations to estimate the current speech sample s_k with minimum mean square error, where $\hat{u}_k$ is calculated using with variable forgetting factors λ_k.

INVESTIGATION

The performance of the noise reduction approaches was tested using an utterance of the word "five" sampled at 20kHz with white noise added at various SNRs. The speech was assumed to contain not more than 5 formant frequencies, corresponding to a required minimum filter order of $n = 10$. Thus the prediction matrix W was constructed using $N = 300$ and $n = 30$ to ensure over-modelling and matrix redun-

dancy. This matrix was used for both eigen-decomposition and embedding.

The ARMA filters had an input vector of dimensions $n = 10$, corresponding to the minimum filter order, and $m = 3$ to allow spectral compensation for the voice source dynamics. Hence an RBFN of size $n_h = 27$ was initially selected to satisfy Kolmogorov's theorem which was reduced to an optimal 18 nodes by experiment. An error driven forgetting factor, λ_k, was used to accommodate pitch effects and to estimate u_k.

For embedding, an order of $r = 6$ was selected based on the observation that the optimal order of eigen-decomposition never exceeded $r = 6$. The embedded speech, m_k, was used as the input vector s_k to an RBFN ARMA filter to predict the speech sample s_{k+n}. $\hat{u}_k$ was again estimated using λ_k. Initially the network had 19 nodes to satisfy Kolmogorov's theorem and was reduced to an optimal 14 nodes by experiment.

The noise filtered speech estimate is taken as the *aposteriori* filter outputs and figure 1 shows the spectrograms of the filter outputs and reference signals, figures 1a and 1b, at 0dB SNR. Eigen-decomposition, figure 1c, extracts the principal signal frequencies along with some tonal noise described in [3] and high frequency signal structure is lost. This approach to noise reduction may be satisfactory for certain signals but the high frequency structure is important in speech characterisation. Dendrinos et al. [3] claim overall intelligibilty is not impaired below 10dB, however, in that case a supervised block order selection technique was used and spectral content was not discussed.

The spectral content of both ARMA filters, figures 1d and 1e, is very similar with the linear filter having slightly better high frequency and non-speech noise reduction with a corresponding loss of signal information. The embedding approach, figure 1f, presents an encouraging mixture of both the eigen-decomposition and ARMA properties. It contains a spectral content similar to the ARMA filters with improved non-speech noise reduction and better defined high frequency structure.

A similar appraisal of the filter error spectra shows that the non-linear ARMA loses the least signal information of all approaches and the error spectrum is generally white. The greatest amount of information is lost using eigen-decomposition, but surprisingly the linear ARMA model also loses a great deal of the signal structure. The embedding approach loses some low frequency information, but at high frequencies it models the speech better than the non-linear ARMA model. An objective term for the quality of the reconstructed signal is the ratio of the energy of the clean signal to the energy of the reconstruction error signal given in dB [3] which can be used to select the optimal order r for eigen-decomposition. Figure 2 shows the enhanced speech SNR gain attained by the above methods at a range of signal SNRs.

Clearly the eigen-decomposition approach achieves poor noise reduction. This is due mainly to the failure to incorporate pitch into the signal model resulting in r becoming unnecessarily large at pitch pulses to compensate. The linear and non-linear ARMA approaches present very similar and superior results, with embedding falling between ARMA and eigen-decomposition, highlighting the compromising nature of this approach.

DISCUSSION

Embedding creates a rotation of the speech into a more stable data space with fewer degrees of freedom. This allows a reduction of the RBFN size and complexity required to approximate the data. It also allows the removal of noise from the RBFN input vector, ensuring that the adaptation is consistent and preventing noise becoming part of the speech model. The clustering of RBFN centres contributes significantly to the success of this approach and the results of fuzzy node clustering for the embedding RBFN are shown in figure 3 for speech with an SNR of 0dB.

Clearly the effects of noise on the embedding of the speech in the first two principal dimensions are minimal and the RBFN nodes are evenly distributed to adequately sample the whole state space. In the fourth and fifth dimensions the added noise leads to a deviation of the noisy dynamics from the clean manifold. The RBFN nodes are not as evenly distributed in these dimensions resulting in inadequate modelling which has the effect of reducing the influence of noisy dimensions on the speech prediction.

This does not, however, exclude the modelling

of higher order speech components as these can be accommodated in the weight adaptation. It allows a compromise such that no information is lost outright as observed during the eigen-decomposition approach. However, equal significance in modelling is not given to noisy signal dimensions which affects the ARMA filtering results.

Eigen-decomposition has the greatest potential to produce more accurate signal estimates because it does not involve prediction, only interpolation. However, optimal order selection is hindered by the pitch dynamics which lead to model instability as the system compensates. Signal embedding offers a more robust approach to reconstruction order selection, removing the dependence upon accurate signal sub-space rank estimation. The incorporation of pitch dynamics into the embedding model improves this stability and allows improved spectral estimation comparable to ARMA filters with improved non-speech noise reduction.

Since the underlying speech dynamics are not known *a priori* there are limits to the amount of noise that can be eliminated by embedding. All speech contains an element of noise whose source can be external or a non-deterministic component of speech. This implies that the embedding approach can eliminate non-deterministic components of speech even from clean speech. This can reflect badly in *aposteriori* error based SNR assessment and spectrograms despite the improvement in capturing the dynamics, observed in the improved spectral content.

CONCLUSIONS

This work supports the theory that low-dimensional attractors can offer a noise robust approach to speech prediction. This dynamical approach to speech processing is a viable alternative to ARMA speech analysis methods and provides a better quality estimated speech signal than eigen-decomposition.

ACKNOWLEDGEMENTS

The authors wish to thank DRA Malvern for the CASE Studentship associated with this work and the valuable suggestions of Neil Thacker.

REFERENCES

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

2. Broomhead, D.S. and Lowe, D., 1988. "Multivariable functional interpolation and adaptive networks". *Complex Systems*, 2, 321–355.

3. Dendrinos, M., Bakamidis, S., and Carayannis, G., 1991. "Speech enhancement from noise: a regenerative approach". *Speech Communication*, 10, 45–57.

4. Huntsberger, T.L. and Ajjimarangsee, P., 1990. "Parallel self-organising feature maps for unsupervised pattern recognition". *Int. J. General Systems*, 16, 357–372.

5. Konstantinides, K. and Yao, K., 1994. "Statistical analysis of effective singular values in matrix rank determination". *IEEE Trans.*, ASSP-36, 757–763.

6. Moakes, P.A. and Beet, S.W., 1994. "Analysis of non-linear speech generating dynamics". In *Proc. Int. Conf. Spoken Language Processing*, Yokohama, Japan, pages 1039–1042.

7. Moakes, P.A. and Beet, S.W., 1994. "Non-linear speech analysis using recurrent radial basis function networks". In Vlontzos, J., Hwang, J.-N., and Wilson, E., editors, *Proc. IEEE Workshop on Neural Networks for Signal Processing IV*, Ermioni, Greece, pages 319–328.

8. Potts, M.A.S. and Broomhead, D.S., 1991. "Time series prediction with a radial basis function neural network". In *SPIE Proc. Adaptive Signal Processing*, San Diego, CA, USA, pages 255–266.

9. Slyh, R.A. and Moses, R.L., 1989. "Formant estimation using singular value decomposition". In *Proc. IEEE MWSCAS 89*, Champaign, IL, USA, pages 109–112.

10. Ting, Y.T. and Childers, D.G., 1990. "Speech analysis using the weighted recursive least squares algorithm with a variable forgetting factor". In *Proc. IEEE ICASSP 90*, pages 389–392.