Proceedings of the Institute of Acoustics

RECURRENT RADIAL BASIS FUNCTIONS FOR SPEECH PERIOD DETECTION

P. A. Moakes and S. W. Beet

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

ABSTRACT

This paper presents a radial basis function network as a one step ahead predictive speech signal filter. The prediction residual can be interpreted as a powerful pitch pulse detector which shows improved performance over a conventional autoregressive filter and allows further processing to make more accurate estimations of pitch pulse position, the pitch, and the regions of voiced and unvoiced speech. In noisy speech the introduction of recursive elements into the radial basis function network allows successful pitch estimation to be maintained.

1. INTRODUCTION

The aim of this paper is to present the application of a radial basis function network (RBFN) predictive filter to speech pitch period estimation. Speech production can be modeled using an auto-regressive all pole filter with an excitation signal comprising a series of quasi-periodic pitch pulses during voiced speech and white noise during unvoiced speech. The detection of the pitch pulse in voiced parts of speech is important for applications such as linear predictive coding (LPC) where reduced sensitivity to the fundamental frequency in the prediction residual during training provides a more accurate determination of the speech parameters.

The application of neural networks for the identification and interpretation of speech signals is of particular interest due to the non-linear and non-stationary nature of speech [4] and the ability of neural networks to model non-linear functions and time series [2,3]. However, neural network applications in speech signal processing have tended to focus on extracted feature spaces such as LPC coefficients for their inputs [5], due mainly to the importance of LPC parameters in vocal tract identification.

Work has been undertaken using the residuals of LPC prediction for the identification of non-linear speech elements [8] and this paper extends this concept to use RBFNs for the on-line non-linear prediction of speech signals in the time sample domain using minimal prior information about the signal. The prediction residual provides a powerful pitch pulse detector and the improvement in pitch pulse detection over a comparable linear system suggests that the non-linear model provides a more accurate representation of the speech.

2. RADIAL BASIS FUNCTION NETWORKS

RBFNs [1] 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.

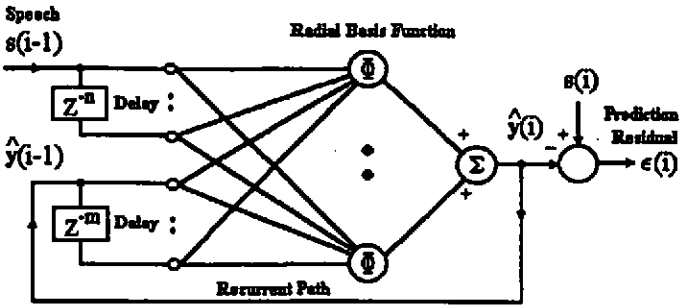

Figure 1: Recursive radial basis function network

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

$$\|x - c_j\|^2 = \sum_{d=1}^{n_c} (x_d - c_{dj})^2 \tag{1}$$

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

$$h_j = \Phi(\|x - c_j\|) \tag{2}$$

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. The RBFN centres are selected randomly within the bounds of the speech state-space and fixed to prevent the centres being biased by short term speech characteristics.

The output layer consists of a linear combiner which calculates the weighted sum of hidden layer nodes, giving an output at node i of :-

$$\hat{y}_i = \sum_{j=1}^{n_h} \eta_{ji} h_j \tag{3}$$

where η_{ji} are the node weights and n_h is the number of hidden nodes.

Recurrent RBFNs (RRBFNs) incorporate lagged network outputs as node inputs, hence noise corrupted input signals are augmented with prediction outputs which have a reduced noise content. The input vector x at sample k for a network with n lagged speech samples, s, and m lagged RBFN predictions, $\hat{y}$, is thus :-

$$x_k = [s_{k-1}, \ldots, s_{k-m}, \hat{y}_{k-1}, \ldots, \hat{y}_{k-n}] \tag{4}$$

Figure 1 shows the RRBFN structure.

3. NETWORK ADAPTATION

The response of the RBFN is linear with respect to the output weight for each non-linear node. This results in an output error surface with only one global minimum and allows a Kalman Filter (KF) approach to be used to update the hidden layer weights and reduce the mean squared prediction error. The KF equations for updating the hidden layer weights are :-

$$K_k = P_{k-1}\hat{\phi}_k \left[\lambda + \hat{\phi}_k^T P_{k-1}\hat{\phi}_k\right]^{-1} \tag{5}$$

$$P_k = \frac{1}{\lambda}\left[P_{k-1} - K_k\hat{\phi}_k^T P_{k-1}\right] \tag{6}$$

$$\Theta_k = \Theta_{k-1} + K_k\epsilon_k \tag{7}$$

where K is the KF gain and P is the prediction error covariance matrix. Θ is the vector of hidden layer weights, ϵ is the prediction error $s_k - \hat{y}_k$, and $\hat{\phi}$ is the vector of node outputs h_j.

λ is a forgetting factor which allows the KF to estimate system parameters which may be varying by exponentially windowing previous samples. A compromise of adaptive speed and previous sample bias must be achieved and Salgado et al. [6] suggest a value of $0.95 \leq \lambda \leq 0.99$ although this can be made adaptive based on the filter error information content. An optimum filter generates constant error information for a signal with a Gaussian noise driving source, but larger errors occur when the source signal changes, such as at pitch events [7]. The filter error information is the weighted sum of squares of the residual errors, V_k :-

$$V_k = \sum_{i=1}^{k} \lambda^{k-i}\epsilon_i^2 \tag{8}$$

which can be expressed recursively as :-

$$V_k = \lambda V_{k-1} + \epsilon_k^2\left(1 - \hat{\phi}_k^T K_k\right) \tag{9}$$

Applying a constraint of constant error information, $V_k = V_{k-1} = V_1$, allows a variable forgetting factor (VFF), λ_k, to be defined from equation (9) as :-

$$\lambda_k = 1 - \epsilon_k^2\left(1 - \hat{\phi}_k^T K_k\right)/V_1 \tag{10}$$

Large prediction errors occur at the instant of glottal closure resulting in a small value of λ_k. Since the effective memory of the system is $1/(1-\lambda_k)$ samples, the Kalman filter estimates are based on a shorter window of speech allowing rapid adaptation to the changing dynamics. This creates a sharp error at the point of closure which is rapidly eliminated by the changing forgetting factor, observed against the average filter error these peaks are candidates for the onset of the pitch pulse. Simple post-processing techniques can then be used to select the most likely pitch positions from these pulse candidates [9].

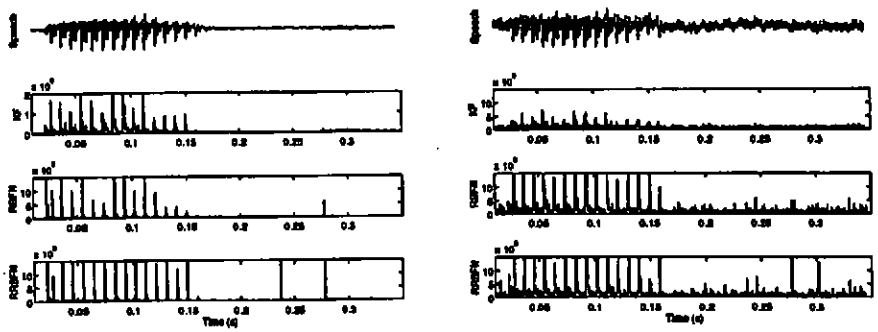

Figure 2: Filter residuals for "eight" at a) 21dB, b) 3dB

4. PITCH CANDIDATE DETECTION

An RRBFN with 20 hidden layer nodes was implemented as a one step ahead predictive speech filter with an input vector, x, comprising 6 speech samples and 3 lagged predictions. This was found to be the minimum network specification required to give good pitch detections in noise. The network was compared with a 20 node RBFN using only 6 speech samples and a Kalman filter where the network weights were connected directly to six speech samples. The network weights were updated using the KF equations, (5..7) and a constant forgetting factor of $\lambda = 0.95$ was found to give the best compromise of signal adaptation and pitch detection.

Lowering the SNR to 3dB considerably deteriorates the KF voice source estimate. The networks were tested using the utterance "eight" sampled at 20kHz with signal-to-noise ratios (SNRs) of 21dB and 3dB. The eight sample mean squared filter residuals were used to provide an estimate of the voice source signal and the results are shown in figure 2. At a SNR of 21dB both the RRBFN and the RBFN provide a powerful pitch pulse detector an order of magnitude better than the KF results, consequently the effects of noise are more significant in the interpretation of the KF prediction error. The RBFNs also detect the onset of the fricative /t/ after the stop, but does not show the noise source of the fricative.

The RBFN, however, produced a clear voice source estimate with a noise floor equal to that of the KF and allowed accurate pitch detection to be maintained. Because the RBFN prediction is based on only six lagged noisy speech samples, the voice source estimate still contains a large noise presence. The RRBFN, though, was able to reduce the noise floor between pitch events and produced sharper residual peaks as the recursive elements provided a signal estimate with reduced noise content. A consequence of this was that estimation errors which were fed back

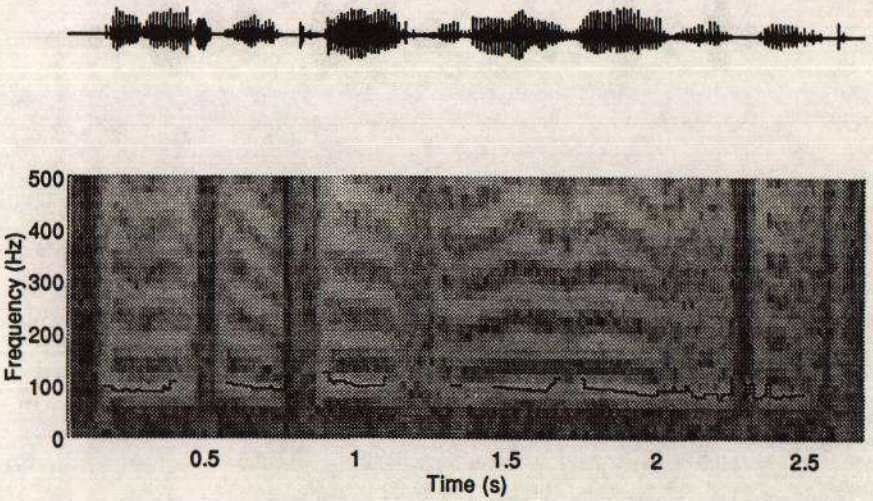

Figure 3: Pitch track for a male TIMIT speaker

produced significant peaks in the non-voiced areas of speech which were later eliminated by a pitch post-processing algorithm.

The RRBFN was extended to include the variable forgetting factor of equation (10). ϵ^2 was replaced by the voice source estimate and V_1 was replaced by the mean squared filter residual over several pitch periods. A lower limit was set for λ such that $\lambda_k = max\,[\lambda_k, 0.8]$ and a gain factor, $\gamma = 0.15$, was introduced to equation (10) to prevent large prediction errors erasing the filter memory, giving :-

$$\lambda_k = 1 - \gamma\epsilon_k^2 \left(1 - \hat{\phi}_k^T K_k\right) / V_1 \tag{11}$$

This approach produced a lower noise floor on the filter residual prior to a pitch event where the error information is constant and was found to produce the most reliable pitch candidates for the post-processing algorithm.

5. PITCH POST-PROCESSING AND TRACKING

An RRBFN pitch detector and associated post-processing algorithms were applied to real speech obtained from the DARPA TIMIT speech database. Pitch candidate selection was based on thresholding the voice source estimate at twice its standard deviation over several pitch pulses, producing a single pulse as the error exceeds the threshold.

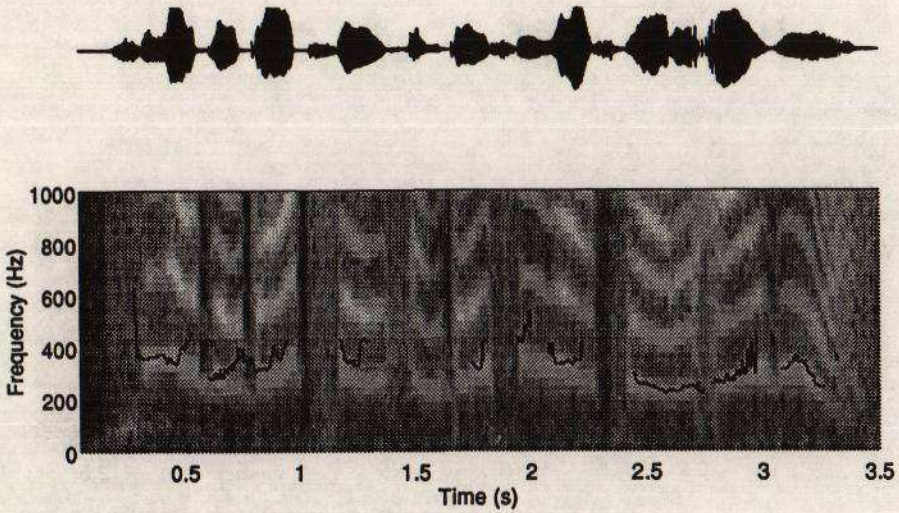

Figure 4: Pitch track for a female TIMIT speaker

Candidate pulses were then filtered using statistical methods to select the most likely pitch positions. A window of 15 pitch candidates was selected and the median and standard deviation of the estimated pitch periods calculated. The median is preferred because it is less influenced by extreme pitch estimation errors, producing more consistent traces when it is used to calculate the fundamental frequency. When the median pitch period exceeds the standard deviation this indicates a consistent pitch period within the candidate window. The speech is considered voiced and candidates in the window with a pitch within one standard deviation of the window median are selected as the pitch pulses. Discarded pulses are eliminated from the window and the remaining pitch period estimates are adjusted accordingly.

Figure 3 shows the pitch frequency track obtained in the above way for the TIMIT phrase "Don't ask me to carry an oily rag like that" spoken by a male. The track is plotted for what is determined as voiced speech and overlaid onto the FFT derived spectrogram. The algorithm provides a very clear indication of the areas of voiced speech, with no obvious misclassification of unvoiced speech as voiced. The largest errors occur in quiet speech where the SNR is lowest and the speech dynamics are changing, producing pitch estimates with variable statistics.

The pitch tracks lie along the fundamental resonance in the spectrogram which is further evidence for the correct determination of the pitch period, and closer inspection reveals that

the selected pitch pulses do occur at the instant of glottal closure. The track is not smooth because the median pitch value is used to calculate the frequency instead of the mean. Results show that the mean value produces smoother plots, but these were severely affected by poor pitch estimates and the resulting tracks are not as accurate.

In an attempt to stretch the validity of the algorithm this experiment was repeated using the phrase "She had your dark suit in greasy wash water all year" spoken by a female and the result is shown in figure 4. Again there is good identification of voiced speech and the pitch tracks appear to follow closely the fundamental spectral resonance. However, there is a greater tendency for the pitch track to incorporate incorrect pitch estimates, raising the estimated fundamental frequency. This is largely because voice source thresholding is now being performed over three times the number of pitch pulses as that of male speech. This could be overcome by adapting the pitch selection and post-processing algorithms to account for the reduced pitch period, although this makes the algorithm speaker dependent.

6. DISCUSSION

This paper has demonstrated the ability of RBFNs to estimate the non-linear system dynamics of speech. The prediction residual provides a powerful pitch pulse predictor and the improvement in pitch detection over a comparable linear predictor supports the proposition that a non-linear model provides a more accurate description of the speech signal. Although signal noise corruption causes significant deterioration of this result, incorporating recursion into the structure provides a reduced noise signal estimate which improves prediction.

The resulting front end speech processor has proved to be an excellent source of pitch candidates for pitch post-processing, achieving good performance in a voiced/unvoiced classifier and pitch tracking algorithm. The pitch pulses are suitable for pitch synchronous estimation, although it is preferable to use the initial voice source estimate as a more accurate guide to the areas of consistent dynamics within speech. The addition of a smoothing algorithm to the pitch tracks would provide a suitable estimate of pitch frequency for LPC synthesis.

The algorithm has performed well in both male and female speech, with only limited *a priori* information. Although in this paper the stages of processing have been implemented sequentially - prediction, detection, selection - the algorithm can be implemented on-line with statistical calculations being based on short term characteristics over eight samples and long term characteristics over a few pitch pulses. This will enable further work to concentrate on the incorporation of the pitch into the prediction model.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] BROOMHEAD, D.S. and LOWE, D. : 'Multivariable functional interpolation and adaptive networks', *Complex Systems*, 1988, **2**, (3), pp. 321–355.

[2] CHEN, S. , BILLINGS, S.A. , COWAN, C.F.N. , and GRANT, P.M. : 'Practical identification of NARMAX models using radial basis functions', *Int. J. Control*, 1990, **52**, (6), pp. 1327–1350.

[3] LOWE, D. and WEBB, A. : 'Adaptive networks, dynamical systems, and the predictive analysis of time series' *in* 'Proc. First IEE Int. Conf. on Artificial Neural Networks', 1989, pp. 95–99.

[4] MCLAUGHLIN, S. and LOWRY, A. : 'Nonlinear dynamical systems concepts in speech analysis' *in* 'Proc. EUROSPEECH 93', 1993, pp. 377–380.

[5] MOON, S. and HWANG, J.-N. : 'Coordinated training of noise removing networks' *in* 'Proc. IEEE ICASSP 93', 1993, vol. I, pp. 49–54.

[6] SALGADO, M.E. , GOODWIN, G.C. , and MIDDLETON, R.H. : 'Modified least squares algorithm including exponential setting and resetting', *Int. J. Control*, 1988, **47**, (2), pp. 477–491.

[7] TING, Y.T. and CHILDERS, D.G. : 'Tracking spectral resonances' *in* 'Proc. IEEE 4th ANN Workshop on Spectrum Estimation', 1988, pp. 49–54.

[8] TOWNSHEND, B. : 'Nonlinear prediction of speech' *in* 'Proc. IEEE ICASSP 91', 1991, vol. I, pp. 425–428.

[9] YANG, G. and LEICH, H. : 'A reliable postprocessor for pitch determination algorithms' *in* 'Proc. EUROSPEECH 93', 1993, vol. 3, pp. 2025–2028.