VOCAL TRACT MODELLING WITH RECURRENT NEURAL NETWORKS

T.L. Burrows and M. Niranjan

Cambridge University Engineering Department, Trumpington Street
Cambridge,CB2 1PZ, United Kingdom

ABSTRACT

In this paper, the speech production system is modelled using the true glottal excitation as the source and a recurrent neural network to represent the vocal tract. The hidden nodes have multiple delays of one and two samples, making the network equivalent to a parallel formant synthesiser in the linear regions of the hidden node sigmoids. An ARX model identification is carried out to initialise the neural network parameters. These parameters are re-estimated in an analysis-by-synthesis framework to minimise the synthesis (output) error. Unlike other analysis-by-synthesis speech production models such as CELP, the source and filter in this approach are decoupled, enabling manipulation of the source time-scale to achieve high quality pitch changes.

1. INTRODUCTION

Typical source-filter models of the vocal tract system use a linear filter to model the frequency response (formants) of the vocal tract, and an impulse train or turbulent excitation to model the source. The filter parameters are usually estimated by linear prediction analysis to minimise the equation error [1]. A linear filter has a limited performance on nonlinear speech data, and nonlinear neural network-based models have been shown to give improved synthesis performance [2]. With continuous data, the hidden nodes of small, single hidden layer networks operate predominantly out of saturation [3], motivating a linear initialisation of these networks to give improved convergence over random initialisations [4].

In CELP and multi-pulse coders, the excitation is typically computed in an analysis-by-synthesis framework to minimise the synthesis error and can therefore compensate for much of the information lost by the filter representation. However, performing time modifications to the source in order to change the pitch of the synthesised speech distorts the synthetic speech by changing the rate of articulation. This can be overcome by using a fixed excitation and re-estimating the filter parameters to minimise the synthesis error [5]. A fixed stylised glottal excitation and synthesis error minimisation is used in the JSRU parallel formant vocoder [6]. The quality of the synthesis produced is still as good as that of multi-pulse coders, but the vocoder has the advantage that the excitation can be manipulated without causing distortion of the synthetic speech. For a recurrent neural network model, minimisation of the synthesis error is achieved by back-propagation training.

In this paper, a nonlinear recurrent network-based model of the vocal tract is derived, in which the weights are initialised from a linear ARX model [7]. The network parameters are trained using the true glottal excitation as input and the effect of manipulating the pitch of this excitation on the quality of the synthesised speech is shown.

2. VOCAL TRACT MODELLING

A model of the vocal tract system approximates the transfer function between the acoustic waveform at the glottis (the glottal excitation waveform) and the acoustic waveform at the lips (the speech waveform). These can be measured as electrical signals by means of a laryngograph [8] and a microphone respectively. The laryngograph signal is free from the influences of the resonances of the vocal tract and can be used to determine the fundamental frequency, $f0$, of the glottal excitation. Typical examples of the glottal excitation and speech waveform are shown in Fig. 5, (a) and (b). The data was preprocessed by sampling at 16kHz, normalising the amplitude to lie in the range $[-1, 1]$ and removing the mean.

In this section, a recurrent network model of the vocal tract is described which allows the speech data, $y(t)$, to be synthesised from the glottal data, $x(t)$ with minimum synthesis error.

2.1. Structure of Recurrent Network

The structure of the recurrent network model is shown in Fig. 1, in which there are first and second order delays in the feedback around the hidden nodes of a single layer with nonlinearities $f(.) = \tanh(.)$. There are no cross-connections between hidden nodes and a single linear output node is used. The network synthesis, $\hat{y}(t)$, is given by (1).

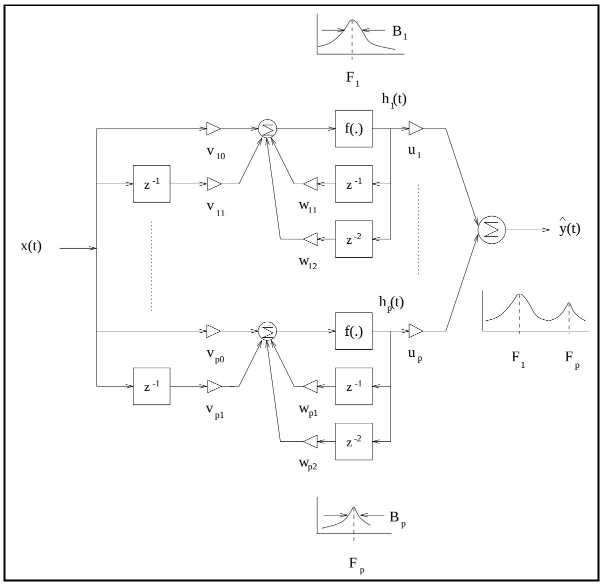


Figure 1: Structure of recurrent network.

$$\hat{y}(t) = \sum_{i=1}^{p} u_i h_i(t) \quad (1)$$

$$h_i(t) = f(\sum_{j=0}^{1} v_{ij} x(t-j) + \sum_{k=1}^{2} w_{ik} h_i(t-k)) \quad (2)$$

where p is the number of hidden nodes and $h_i(t)$ is the output of hidden node i. The structure allows an initialisation of the network from a parallel formant representation of the speech, derived by factorisation of a linear ARX model, as described below.

2.2. Initialisation of Network Parameters

The speech spectrum is considered to be composed of a sum of parallel formants (resonances of the vocal tract) which are represented by second order IIR filters. Within the linear region of the tanh function, each hidden node of the recurrent network represents such a formant, with resonant frequency F_i and bandwidth B_i, as shown in Fig. 1. The filters are derived from a partial fraction expansion of the synthesis transfer function $H(z)$, (3), of a linear ARX model of the vocal tract. Recombination of complex conjugate poles and pairs of real poles gives p quadratic factors (5), from which the network weights can be initialised. All output weights, u_i, are taken to be equal, thus assuming an equal contribution from each formant in (4).

$$H(z) = \frac{b_0 + b_1 z^{-1} + \ldots + b_{n_b-1} z^{-n_b+1}}{1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_{n_a} z^{-n_a}} \quad (3)$$

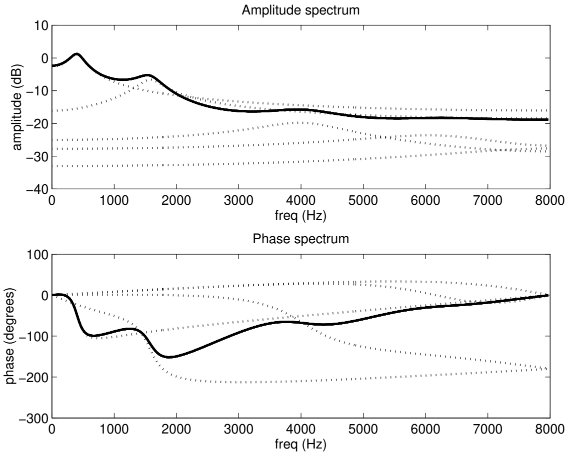


Figure 2: Spectra and formant factors (dotted curves) for a typical ARX model ($n_a = 10$, $n_b = 1$) .

$$H(z) = \sum_{i=1}^{p} u_i H_i(z) \quad (4)$$

$$H_i(z) = \frac{v_{i0} + v_{i1} z^{-1}}{1 - w_{i1} z^{-1} - w_{i2} z^{-2}} \qquad i = 1 \ldots p \quad (5)$$

The order, n_a, of the ARX model is given by $n_a = 2p$, where the number of hidden nodes equals the number of formants of voiced speech, typically 4 or 5. Any number of delays, n_b, can be used depending on whether an all-pole or pole-zero initialisation is preferred. Zeros are desirable for the modelling of nasals. To avoid direct terms in the factorisation, $1 \leq n_b \leq n_a - 1$. Fig. 2 shows the spectra of the formant factors for a typical ARX model ($n_a = 10$, $n_b = 1$). Fig. 3 shows a comparison of the learning curves for random initialisation and ARX initialisation from different order models, when trained over a short length of voiced speech. These curves show that ARX factorisation provides a good network initialisation with improved convergence over random initialisation. The reduction in synthesis error results in a better synthetic waveform as shown in Fig 4.

A problem with the initialisation and training procedure is maintaining stability. Stability of the initialisation factors, $H_i(z)$, and the linearised nodes after training, is ensured by reflecting unstable poles back inside the unit circle.

3. SYNTHESIS PERFORMANCE

For long lengths of speech, an ARX model was generated over consecutive 20ms frames of data at a frame rate of 10ms. From the ARX initialisation, a network was trained by back-propagation to minimise the normalised mean squared synthesis error (nmse) over the first 10ms of the current frame. The values of $h(t-k)$

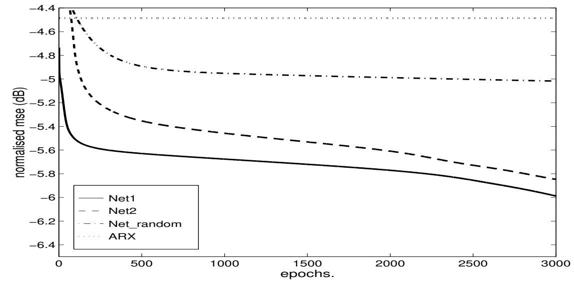


Figure 3: Learning curves for different initialisations. Net1:ARX initialisation ($n_a = 10$,$n_b = 9$). Net2:ARX initialisation ($n_a = 10$,$n_b = 1$). Net_random:random initialisation ($n_a = 10$,$n_b = 9$). ARX:synthesis error for actual ARX model ($n_a = 10$,$n_b = 9$).

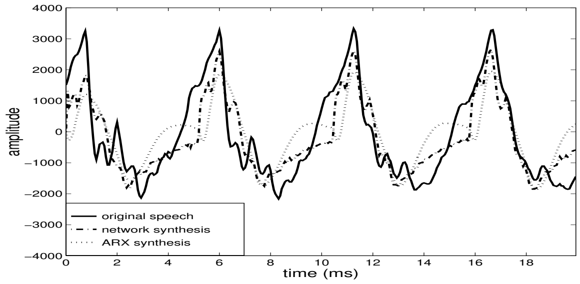


Figure 4: Synthesis for ARX initialisation model (n_a=10, n_b=9) and recurrent network after 3000 epochs of training from the ARX initialisation.

at the start of the next frame were passed on from the current frame. The trained networks were used to re-synthesise the speech from the glottal waveform. An example of the re-synthesis of the speech in Fig. 5(b) is shown in Fig. 5(c). This is a fragment from the utterance "and the concept of a base". For this utterance, the average (per frame) nmse in Table 1 shows that re-estimation of the network parameters consistently reduces the synthesis error below that of the ARX model. This is due to the fact that the ARX parameters, a and b, are derived by *equation error* minimisation in which the mean squared prediction error is minimised [7]. These parameters are not optimal for synthesis, unlike those of the network which are derived by back-propagation to minimise the *synthesis error* directly. The networks were trained for 200 epochs per frame to reduce training times. Examination of the learning curves for each frame showed that the networks did not fully converge within this time. A larger improvement over the ARX synthesis is expected for longer training times.

Table 1 also shows the effect of changing the order of the ARX model on the ARX and network synthesis. Audibly, a model order of $n_a = 10$, $n_b = 9$

ARX Model				
Order		Average nmse (dB)		
n_a	n_b	all	voiced	unvoiced
6	2	-1.33	-2.66	-0.33
6	5	-1.88	-3.84	-0.49
8	2	-1.29	-2.56	-0.33
8	7	-2.13	-4.37	-0.57
10	2	-1.23	-2.38	-0.35
10	9	-2.12	-4.51	-0.49

Recurrent Neural Network				
Order		Average nmse (dB)		
n_a	n_b	all	voiced	unvoiced
6	2	-2.25	-4.61	-0.63
6	5	-2.10	-4.44	-0.50
8	2	-2.37	-4.52	-0.86
8	7	-2.44	-4.62	-0.92
10	2	-2.03	-3.77	-0.76
10	9	-2.47	-5.02	-0.74

Table 1: Average normalised mean squared error per frame for ARX and network models for the utterance "and the concept of a base".

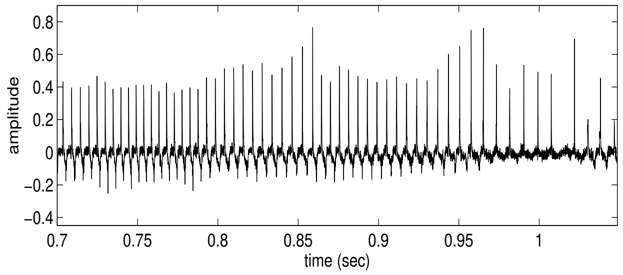
gave the best synthesis. Although changing the model order n_b does not require a change in the number of network parameters, a higher order n_b gave a better network initialisation and final performance. For unvoiced frames, in which the speech does not contain a strong formant structure, the use of a higher order n_b introduces additional zeros into the model which allows for the possibility of pole-zero cancellation to produce a flatter spectrum. Increasing n_a requires additional hidden nodes in the network.

4. PITCH MANIPULATION

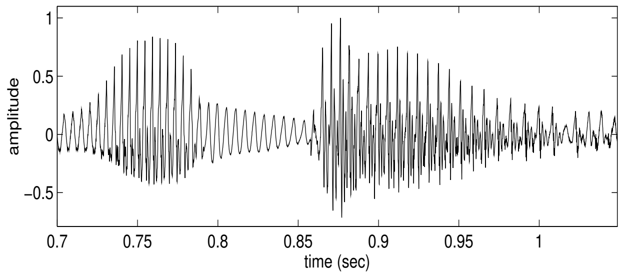
The trained networks were used to re-synthesise the speech from a glottal waveform in which the original pitch, $f0$, was altered by $\pm 25\%$. The effect on the synthesised speech is shown in Fig. 5(d). Examination of the spectra of these waveforms showed that the pitch of the synthesised speech was altered but the positions of the formants were maintained. The pitch changes were audible but no change in articulation rate was apparent.

5. CONCLUSION

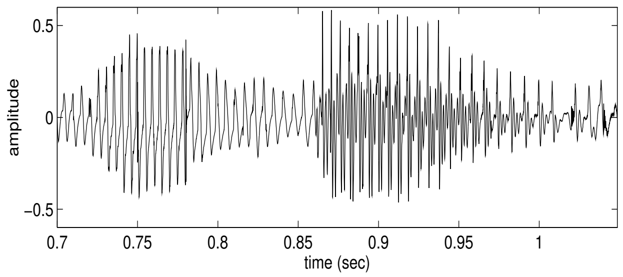
A single hidden layer recurrent network with first and second order feedback delays has been implemented to synthesise speech from actual glottal waveforms. From a linear ARX model of the vocal tract system, an initialisation for the network weights has been derived which has improved convergence properties over random initialisations. The network has better synthesis



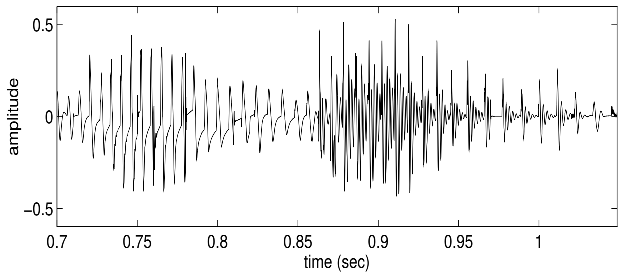
(a) Glottal excitation, pitch $f0$



(b) Original speech, pitch $f0$



(c) Synthesised speech, pitch $f0$



(d) Synthesised speech, pitch $0.75 \times f0$

Figure 5: Glottal excitation, original speech and synthesised speech for the fragment "... of a ba ..." taken from the phrase "and the concept of a base".

performance than the linear model because the weights are derived to minimise the synthesis error directly. This analysis-by-synthesis method is still valid if the actual glottal excitation is not available, since it may be replaced by a codebook of typical glottal excitation signals in a CELP-like framework. The current method for maintaining stability is sub-optimal and an improvement will be to incorporate the stability conditions into the training procedure. The effects of perceptual weighting of the synthesis error and quantisation of the network parameters on synthesis performance have not been investigated.

Using the true glottal excitation to derive the network parameters, a vocal tract model has been developed in which the source and filter are decoupled. This allows more information about the vocal tract to be carried by the network parameters. The coder is more flexible than other source-filter models such as multipulse coders, since the pitch of the synthesised speech can be altered by changing the time scale of the glottal excitation, without causing distortion of the synthetic speech. This allows new pitch contours to be applied to the synthesised speech, without loss of quality.

6. REFERENCES

[1] R. W. Schafer and L. R. Rabiner, *Digital Processing of Speech Signals.* Englewood Cliffs, N.J.: Prentice-Hall, 1978.

[2] L. Wu, M. Niranjan, and F. Fallside, "Fully vector-quantized neural network-based code excited nonlinear predictive speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 482–489, October 1994.

[3] T. L. Burrows and M. Niranjan, "The use of feed-forward and recurrent neural networks for system identification," Tech. Rep. CUED/F-INFENG/TR158, Cambridge University Engineering Department, 1993.

[4] P. Hirschauer, P. Larzabal, and H. Clergeot, "Design of neural estimators for multisensors : Second order backpropagation, initialisation and generalisation," in *Proc. ICASSP*, pp. 537–540, April 1994.

[5] M. Niranjan, "CELP coding with adaptive output error model identification," in *Proc. ICASSP*, pp. 225–228, April 1990.

[6] J. Holmes, *Speech Synthesis.* London: Mills and Boon Ltd., 1972.

[7] L. Ljung, *System Identification : Theory for the User.* Englewood Cliffs, N.J.: Prentice-Hall, 1987.

[8] G. Borden and K. Harris, *Speech Science Primer.* USA: Williams and Wilkins, 1984.