Nonlinear Speech Synthesis *

Stephen McLaughlin
Signals and Systems Group,
Department of Electronics and Electrical Engineering,
University of Edinburgh, The King's Buildings,
Edinburgh, EH9 3JL, Scotland, UK

Email : sml@ee.ed.ac.uk
Tel : (+44)-131-650-5578, Fax : (+44)-131-650-6554

ABSTRACT

This paper examines the how and the why of nonlinear speech synthesis. It discusses why nonlinear speech synthesis should be considered, reviews the recent history and describes in detail a variety of approaches to the problem. It argues that while modern concatenative speech synthesisers produce speech which is intelligible, however they are very inflexible and often lack a human quality. The paper does not suggest that nonlinear speech synthesisers are ready to replace conventional approaches, but rather that they offer some potential advantages but there is a considerable amount of research still to be carried out.

1 Introduction

Speech synthesis is a complex task that aims to produce naturally–sounding speech. While working systems that produce intelligible speech have existed since the 1970's, the final aim of producing a synthesiser that is indistinguishable from a human speaker has still to be realised. There remain a number of problems at all stages of the process, including the actual generation of the speech signal itself with the required intonation. This paper is structured as follows, a brief review of conventional linear based approaches is followed by a quick review of nonlinearities which exist in speech generation. Then an example of nonlinear techniques applied to epoch marking is presented followed by two sections on nonlinear speech synthesis. Finally some conclusions are drawn.

2 Conventional Speech Synthesis Approaches

Conventionally the main approaches to speech synthesis depend on the type of modelling used. This may be a model of the speech organs themselves (articulatory synthesis), a model derived from the speech signal (waveform synthesis), or alternatively the use of pre–recorded segments extracted from a database and joined together (concatenative synthesis).

Modelling the actual speech organs is an attractive approach, since it can be regarded as being a model of the fundamental level of speech production. An accurate articulatory model would allow all types of speech to be synthesised in a natural manner, without having to make many of the assumptions required by other techniques (such as attempting to separate the source and vocal tract parts out from one signal) [1–3]. Realistic articulatory synthesis is an extremely complex process, and the data required is not at all easy to collect. As such, it has not to date found any commercial application and is still more of a research tool.

Waveform synthesisers derive a model from the speech signal as opposed to the speech organs. This approach is derived from the linear source–filter theory of speech production [4]. The simplest form of waveform synthesis is based on linear prediction (LP) [5]. The resulting quality is extremely poor for voiced speech, sounding very robotic.

Formant synthesis uses a bank of filters, each of which represents the contribution of one of the formants. The best known formant synthesiser is the Klatt synthesiser [6], which has been exploited commercially as DECTalk. The synthesised speech quality is considerably better than that of the LP method, but still lacks naturalness, even when an advanced voice–source model is used [7].

Concatenation methods involve joining together pre–recorded units of speech which are extracted from a database. It must also be possible to change the prosody of the units, so as to impose the prosody required for the phrase that is being generated. The concatenation technique provides the best quality synthesised speech available at present. It is used in a large number of commercial systems, including British Telecomm's Laureate [8] and the AT&T Next–Gen system [9]. Although there is a good degree of naturalness in the synthesised output, it is still clearly distinguishable from real human speech, and it may be that more sophisticated parametric models will eventually overtake it.

Techniques for time and pitch scaling of sounds held in a database are also extremely important. Two main techniques for time–scale and pitch modification in concatenative synthesis can be identified, each of which op-

* This work was supported by BT, EPSRC and the Royal Society.

erates on the speech signal in a different manner. The pitch synchronous overlap add (PSOLA) [10] approach is non–parametric as opposed to the harmonic method, which actually decomposes the signal into explicit source and vocal tract models. PSOLA is reported to give good quality, natural–sounding synthetic speech for moderate pitch and time modifications. Slowing down the speech by a large factor (greater than two) does introduce artifacts due to the repetition of PSOLA bells. Some tonal artifacts (*e.g.* whistling) also appear with large pitch scaling, especially for higher pitch voices, such as female speakers and children.

McAulay and Quatieri developed a speech generation model that is based on a glottal excitation signal made up of a sum of sine waves [11]. They then used this model to perform time–scale and pitch modification. Starting with the assumption made in the linear model of speech that the speech waveform $x(t)$ is the output generated by passing an excitation waveform $e(t)$ through a linear filter $h(t)$, the excitation is defined as a sum of sine waves of arbitrary amplitudes, frequencies and phases. A limitation of all these techniques is that they use the linear model of speech as a basis.

3 Nonlinearities in speech

There are known to be a number of nonlinear effects in the speech production process. Firstly, it has been accepted for some time that the vocal tract and the vocal folds do not function independently of each other, but that there is in fact some form of coupling between them when the glottis is open [12] resulting in significant changes in formant characteristics between open and closed glottis cycles [13]. More controversially, Teager and Teager [14] have claimed (based on physical measurements) that voiced sounds are characterised by highly complex air flows in the vocal tract involving jets and vortices, rather than well behaved laminar flow. In addition, the vocal folds will themselves be responsible for further nonlinear behaviour, since the muscle and cartilage which comprise the larynx have nonlinear stretching qualities. Such nonlinearities are routinely included in attempts to model the physical process of vocal fold vibration, which have focussed on two or more mass models [2, 3, 15], in which the movement of the vocal folds is modelled by masses connected by springs, with nonlinear coupling. Observations of the glottal waveform have shown that this waveform can change shape at different amplitudes [16] which would not be possible in a strictly linear system where the waveform shape is unaffected by amplitude changes.

In order to arrive at the simplified linear model, a number of major assumptions are made:

- the vocal tract and speech source are uncoupled (thus allowing source–filter separation);
- airflow through the vocal tract is laminar;
- the vocal folds vibrate in an exactly periodic manner during voiced speech production;
- the configuration of the vocal tract will only change slowly;

These imply a loss of information which means that the full speech signal dynamics can never be properly captured. These inadequacies can be seen in practice in speech synthesis where, at the waveform generation level, current systems tend to produce an output signal that lacks naturalness. This is true even of concatenation techniques which copy and modify actual speech segments.

4 Poincare maps and epoch marking

The section discusses how nonlinear techniques can be applied to pitch marking of continuous speech. We wish to locate the instants in the time domain speech signal at which the glottis is closed. A variety of existing methods can be employed to locate the epochs. These are Abrupt change detection [17], Maximum Likelihood epoch detection [18] and Dynamic programming [19]. All of the above techniques are sound and generally provide good epoch detection. The technique presented here should not be viewed as a direct competitor to the methods outlined above. Rather it is an attempt to show the practical application of ideas from nonlinear dynamical theory to a real speech processing problem. The performance in clean speech is comparable to many of the techniques discussed above.

In nonlinear processing a d–dimensional system can be reconstructed in an m–dimensional state space from a single dimension time series by a process called embedding. Takens' theorem states that $m \geq 2d+1$ for an adequate reconstruction [20], although in practice it is often possible to reduce m. An alternative is the singular value decomposition (SVD) embedding [21], which may be more attractive in real systems where noise is an issue.

A Poincaré map is often used in the analysis of dynamical systems. It replaces the flow of an n–th order continuous system with an $(n-1)$–th order discrete time map. Considering a three dimensional attractor a Poincaré section slices through the flow of trajectories and the resulting crossings form the Poincaré map. Re–examining the attractor reconstructions of voiced speech shown above, it is evident that these three dimensional attractors can also be reduced to two dimensional maps.[1] Additionally, these reconstructions are pitch–synchronous, in that one revolution of the attractor is equivalent to one pitch period. This has previously been used for cyclostationary analysis and synchronisation [22]; here we examine its use for epoch marking.

[1] Strictly these attractor reconstructions are discrete time maps and not continuous flows. However it is possible to construct a flow vector between points and use this for the Poincaré section calculation.

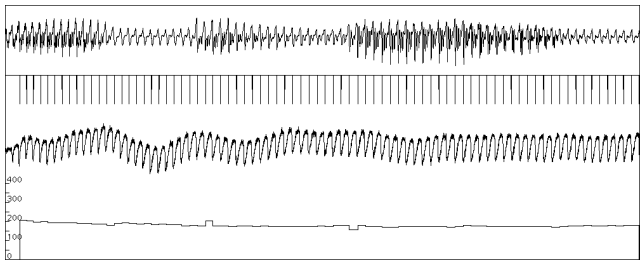

Figure 1: *Results for the voiced section of "came along" from the Keele database for a female speaker. From top to bottom: the signal; the epochs as calculated by the algorithm; the laryngograph signal; the pitch contour (Hz) resulting from the algorithm.*

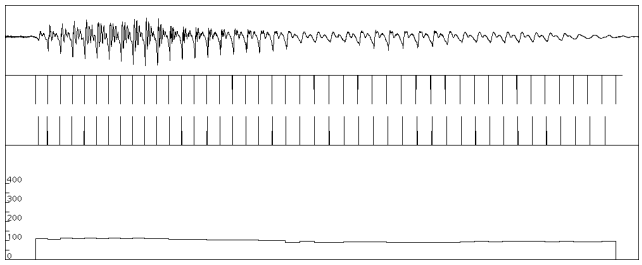

Figure 2: *Results for the voiced section of "raining" from the BT Labs database for a male speaker. From top to bottom: the signal; the epochs as calculated by the algorithm; the processed laryngograph signal; the pitch contour (Hz) resulting from the algorithm.*

The basic processing steps required for a waveform of N points are as follows:

1. Mark y_{GCI}, a known GCI in the signal.
2. Perform an SVD embedding on the signal to generate the attractor reconstruction in 3D state space.
3. Calculate the flow vector, $\mathbf{h}$, at the marked point $\mathbf{y}_{\mathrm{GCI}}$ on the attractor.
4. Detect crossings of the Poincaré section, Σ, at this point in state space by signs changes of the scalar product between $\mathbf{h}$ and the vector $\mathbf{y}_i - \mathbf{y}_{\mathrm{GCI}}$ for all $1 \leq i \leq N$ points.
5. Points on Σ which are within the same portion of the manifold as $\mathbf{y}_{\mathrm{GCI}}$ are the epochs.

When dealing with real speech signals a number of practical issues have to be considered. The input signal must be treated on a frame–by–frame basis, within which the speech is assumed stationary. Finding the correct intersection points on the Poincaré section is also a difficult task due to the complicated structure of the attractor. Two different data sets were used to test the performance of the algorithm, giving varying degrees of realistic speech and hence difficulty.

1. Keele University pitch extraction database [23]. This database provides speech and laryngograph data from 15 speakers reading phonetically balanced sentences.
2. BT Labs continuous speech. 2 phrases, spoken by 4 speakers, were processed manually to extract a data set of continuous voiced speech. Laryngograph data was also available.

The signals were up–sampled to 22.05 kHz, the BT data was originally sampled at 12 kHz and the Keele signals at 20 kHz. All the signals had 16 bit resolution.

Fig. 1 shows the performance of the algorithm on a voiced section taken from the phrase "a traveller came along wrapped in a warm cloak", spoken by a female speaker. There is considerable change in the signal, and hence in the attractor structure, in this example, yet the epochs are sufficiently well located when compared against the laryngograph signal.

In Fig. 2, which is a voiced section from the phrase "see if it's raining" spoken by a male speaker, the epochs are well located for the first part of the signal, but some slight loss of synchronisation can be seen in the latter part.

5 Nonlinear Synthesis Approaches

5.1 Neural network synthesis background

Kubin and Birgmeier reported an attempt made to use a RBF network approach to speech synthesis. They propose the use of a nonlinear oscillator, with no external input and global feedback in order to perform the mapping

$$x(n) = \mathcal{A}(\mathbf{x}(n-1)) \tag{1}$$

where $\mathbf{x}(n-1)$ is the delay vector with non–unit delays, and $\mathcal{A}$ is the nonlinear mapping function [24].

The initial approach taken [25] used a Kalman–based RBF network, which has all of the network parameters trained by the extended Kalman filter algorithm. The only parameter that must be specified is the number of centres to use. This gives good prediction results, but there are many problems with resynthesis. In particular, they report that extensive manual fine–tuning of the parameters such as dimension, embedding delay and number and initial positions of the centres are required. Even with this tuning, synthesis of some sounds with complicated phase space reconstructions does not work [24].

In order to overcome this problem, Kubin resorted to a technique that uses all of the data points in the training data frame as centres [24]. Although this gives correct resynthesis, even allowing the resynthesis of continuous speech using a frame–adaptive approach, it is unsatisfactory due to the very large number of varying

parameters, and cannot be seen as actually learning the dynamics of the speech generating system.

Following their dynamical analysis of the Japanese vowel /a/, Tokuda *et al.* constructed a feed–forward neural network to perform synthesis [26]. Their structure has three layers, with five neurons in the input layer, forty neurons in the hidden layer, and one in the output layer. The time delay in the input delay vector is set at $\tau = 3$ and the weights are learnt by back propagation. Using global feedback, they report successful resynthesis of the Japanese vowel /a/. The signal is noisy, but preserves natural human speech qualities. No further results in terms of speech quality or resynthesis of other vowels are given.

'An alternative neural network approach was proposed by Narashimhan *et al.* This involves separating the voiced source from the vocal tract contribution, and then creating a nonlinear dynamical model of the source [27]. This is achieved by first inverse filtering the speech signal to obtain the linear prediction (LP) residual. Next the residue waveform is low–pass filtered at 1 kHz, then normalised to give a unit amplitude envelope. This processed signal is used as the training data in a time delay neural network with global feedback. The NN structure reported is extremely complex, consisting of a 30 tap delay line input and two hidden layers of 15 and 10 sigmoid activation functions, with the network training performed using back propagation through time. Finally, the NN model is used in free–running synthesis mode to recreate the voiced source. This is applied to a LP filter in order to synthesise speech. They show that the NN model successfully preserves the jitter of the original excitation signal.

5.2 RBF network for synthesis

A well known nonlinear modelling approach is the radial basis function neural network. It is generally composed of three layers, made up of an input layer of source nodes, a nonlinear hidden layer and an output layer giving the network response. The hidden layer performs a nonlinear transformation mapping the input space to a new space, in which the problem can be better solved. The output is the result of linearly combining the hidden space, multiplying each hidden layer output by a weight whose value is determined during the training process.

The general equation of an RBF network with an input vector $\mathbf{x}$ and a single output is

$$\mathcal{F}(\mathbf{x}(n)) = \sum_{j=1}^{P} w_j \phi(\|\mathbf{x} - \mathbf{c}_j\|) \tag{2}$$

where there are P hidden units, each of which is weighted by w_j. The hidden units, $\phi(\|\mathbf{x} - \mathbf{c}_j\|)$, are radially symmetric functions about the point $\mathbf{c}_j$, called a centre, in the hidden space, with $\|.\|$ being the Euclidean vector norm [28]. The actual choice of nonlinearity does not appear to be crucial to the performance of the network. There are two distinct strategies for training an RBF network. The most common approach divides the problem into two steps. Firstly the centre positions and bandwidths are fixed using an unsupervised approach, not dependent on the network output. Then the weights are trained in a supervised manner so as to minimise an error function.

Following from the work of Kubin *et al.*, a nonlinear oscillator structure is used. The RBF network is used to approximate the underlying nonlinear dynamics of a particular stationary voiced sound, by training it to perform the prediction

$$x_{i+1} = \mathcal{F}(\mathbf{x}_i) \tag{3}$$

where $\mathbf{x}_i = \{x_i, x_{(i-\tau)}, \ldots, x_{(i-(m-1)\tau)}\}$ is a vector of previous inputs spaced by some delay τ samples, and $\mathcal{F}$ is a nonlinear mapping function. From a nonlinear dynamical theory perspective, this can be viewed as a time delay embedding of the speech signal into an m–dimensional state space to produce a state space reconstruction of the original d–dimensional system attractor. The embedding dimension is chosen in accordance with Takens' embedding theorem [20] and the embedding delay, τ, is chosen as the first minimum of the average mutual information function [29]. The other parameters that must be chosen are the bandwidth, the number and position of the centres, and the length of training data to be used. With these set, the determination of the weights is linear in the parameters and is solved by minimising a sum of squares error function, $E_S(\hat{\mathcal{F}})$, over the N samples of training data:

$$E_s(\hat{\mathcal{F}}) = \frac{1}{2}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2 \tag{4}$$

where $\hat{x}_i$ is the network approximation of the actual speech signal x_i. Incorporating Equation 2 into the above and differentiating with respect to the weights, then setting the derivative equal to zero gives the least-squares problem [30], which can be written in matrix form as

$$(\mathbf{\Phi}^T \mathbf{\Phi})\mathbf{w}^T = \mathbf{\Phi}^T \mathbf{x} \tag{5}$$

where $\mathbf{\Phi}$ is an NxP matrix of the outputs of the centres; $\mathbf{x}$ is the target vector of length N; and $\mathbf{w}$ is the P length vector of weights. This can be solved by standard matrix inversion techniques.

Two types of centre positioning strategy were considered:

1. Data subset. Centres are picked as points from around the state space reconstruction. They are chosen pseudo–randomly, so as to give an approximately uniform spacing of centres about the state space reconstruction.

2. Hyper–lattice. An alternative, data independent approach is to spread the centres uniformly over an m–dimensional hyper–lattice.

5.3 Synthesis

From analysis, an initial set of parameters with which to attempt resynthesis were chosen. The parameters were set at the following values:

Bandwidth = 0.8 for hyper–lattice, 0.5 for data subset; Dimension = 7; Number of centres = 128; Hyper–lattice size = 1.0; Training length = 1000;

For each vowel in the database, the weights were learnt, with the centres either on a 7D hyper–lattice, or chosen as a subset of the training data. The global feedback loop was then put in place to allow free–running synthesis. The results gave varying degrees of success, from constant (sometimes zero) outputs, through periodic cycles not resembling the original speech signal and noise–like signals, to extremely large spikes at irregular intervals on otherwise correct waveforms [31].

These result implied that a large number of the mapping functions learnt by the network suffered from some form of instability. This could have been due to a lack of smoothness in the function, in which case regularisation theory was the ideal solution. Regularisation theory applies some prior knowledge, or constraints, to the mapping function to make a well–posed problem [32].

The selection of an appropriate value for the regularisation parameter, λ is done by the use of cross–validation [30]. After choosing all the other network parameters, these are held constant and λ is varied. For each value of λ, the MSE on an unseen validation set is calculated. The MSE curve should have a minimum indicating the best value of λ for generalisation. With the regularisation parameter chosen by this method, the 7D resynthesis gave correct results for all of the signals except KH /i/ and KH /u/ when using the data subset method of centre selection. However, only two signals (CA /i/ and MC /i/) were correctly resynthesised by the hyper–lattice method. It was found that λ needed to be increased significantly to ensure correct resynthesis for all the signals when the hyper–lattice was used. Achieving stable resynthesis inevitably comes at some cost. By forcing smoothness onto the approximated function there is the risk that some of the finer detail of the state space reconstruction will be lost. Therefore, for best results, λ should be set at the smallest possible value that allows stable resynthesis. The performance of the regularised RBF network as a nonlinear speech synthesiser is now measured by examining the time and frequency domains, as well as the dynamical properties. In addition to comparing the output of the nonlinear synthesiser to the original speech signal, the synthetic speech from a traditional linear prediction synthesiser is also considered. In this case, the LP filter coefficients were found from the original vowel sound (analogous to the training stage of the RBF network). The estimate $(F_s + 4)$ [33] was used to set the number of filter taps to 26. Then, using the source–filter model, the LP filter was excited by a Dirac pulse train to produce the desired length LP synthesised signal. The distance between Dirac pulses was set to be equal to the average pitch period of the original signal. In this way, the three vowel sounds for each of the four speakers in the database were synthesised.

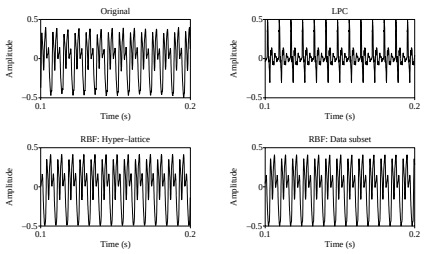

Figure 3: *Time domain examples of the vowel /u/, speaker MC. Top row: original signal (left) and linear prediction synthesised signal (right); Bottom row: RBF network synthesised signal, hyper–lattice (left) and data subset (right).*

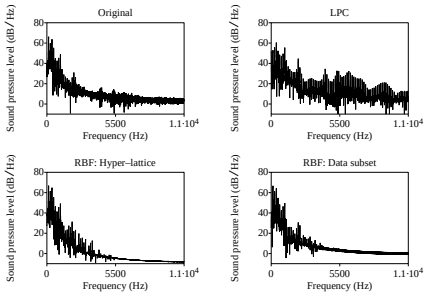

Figure 4: *Spectrums for examples of the vowel /u/, corresponding to the signals in Figure 3.*

Figure 3 shows the time domain waveforms for the original signal, the LP synthesised signal and the two RBF synthesised signals, for the vowel /u/, speaker MC. Figure 4 shows the corresponding frequency domain plots of the signals, and the spectrograms are shown in Figure 5. In these examples, the regularisation parameter λ was set at 0.01 for the hyper–lattice, and 0.005 for the data subset. In the linear prediction case, the technique attempts to model the spectral features of the original. Hence the reasonable match seen in the spectrum (although the high frequencies have been over–emphasised), but the lack of resemblance in the time domain. The RBF techniques, on the other hand, resemble the original in the time domain, since it is from this that the state space reconstruction is formed, although the spectral plots show the higher frequencies have not been well modelled by this method. This is because the networks have missed some of the very fine

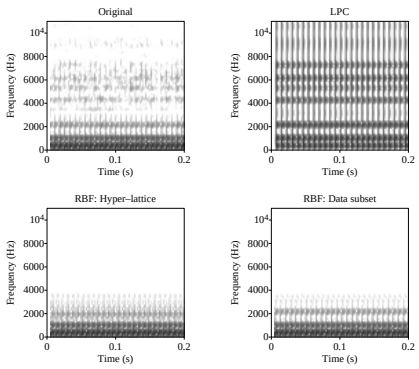

Figure 5: *Wide–band spectrograms for examples of the vowel /u/, corresponding to the signals in Figure 3.*

variations of the original time domain waveform, which may be due to the regularisation.

Further spectrogram examples for different vowels and speakers follow the same pattern, with the size of λ being seen to influence the quality of the signal at high frequencies.

5.4 Jitter and shimmer

Jitter and shimmer measurements were made on all of the original and RBF synthesised waveforms, using epoch detection[2] over a 500 msec window. Jitter is defined as the variation in length of individual pitch periods and for normal, healthy speech should be between 0.1 and 1% of the average pitch period [34]. Table 1 shows the results of the average pitch length variation, expressed as a percentage of the average pitch period length. Results for both centre placing techniques are presented, with the jitter measurements of the original speech data. The hyper–lattice synthesised waveforms contain more jitter than the data subset signals, and both values are reasonable compared to the original.

Shimmer results (the variations in energy each pitch cycle) for the original and synthesised waveforms are also displayed in Table 1. It can be seen that in general there is considerably less shimmer on the synthesised waveforms as compared to the original, which will detract from the quality of the synthetic speech.

6 Incorporating Pitch into the Nonlinear Synthesis Method

The approach adopted here is to model the vocal tract as a forced nonlinear oscillator and to embed an observed scalar time-series of a vowel with pitch information into a higher dimensional space. This embedding, when carried out correctly, will reconstruct the data onto a higher dimensional surface which embodies the dynamics of the

[2]Using Entropic Laboratory's ESPS Epoch function.

Data type	MC (male)	CA (female)	Average (female)
Hyper–lattice jitter (%)	0.470	1.14	0.697
Data subset jitter (%)	0.482	0.663	0.521
Original jitter (%)	0.690	0.685	0.742
Hyper–lattice shimmer (%)	1.00	1.33	0.922
Data subset shimmer (%)	0.694	7.65	2.34
Original shimmer (%)	4.21	7.06	5.17

Table 1: *Percentage jitter and shimmer in original and synthesised waveforms (hyper–lattice and data subset), averaged over the vowels /i/, /a/ and /u/ for each speaker, and as an average over the database.*

vocal tract, see for example, [35, 36] for issues regarding embedding).

Previous studies, discussed above, have successfully modelled stationary (i.e. constant pitch) vowel sounds using nonlinear methods, but these have very limited use since the pitch cannot be modified to include prosody information. The new approach described here resolves this problem by including pitch information in the embedding. Specifically, a non-stationary vowel sound is extracted from a database and, using standard pitch extraction techniques1, a pitch contour is calculated for the time series so that each time domain sample has an associated pitch value. In the present study measurements of rising pitch vowel sounds, where the pitch rises through the length of the time series, have been used as the basis for modelling; see, for example, figure 1.

The time series is then embedded in an m-dimensional space, along with the pitch contour, to form an (m+1)-dimensional surface. A mixed embedding delay between time samples (greater than unity) is used to capture the variable time scales present in the vowel waveform. The (m+1)-dimensional surface is modelled by a nearest neighbour approach, which predicts the next time series sample given a vector of previous time samples and a pitch value (it is envisaged that more sophisticated modelling techniques will be incorporated at a later date).

Synthesis is then performed by a modification of the nonlinear oscillator approach [37], whereby the input signal is removed and the delayed synthesiser output is fed back to form the next input sample. In contrast to previous techniques, the required pitch contour is also passed into the model as an external forcing input. Our results show that this method allows the vowel sound to be generated correctly for arbitrary specified pitch contours (within the input range of pitch values), even though the training data is only made up of the rising vowel time series and its associated pitch contour. In addition, sounds of arbitrary duration can be readily synthesised by simply running the oscillator for the required length of time. Typical synthesis results are shown. It can be seen that the sinusoidal pitch contour of the synthesised sound is quite different from the rising pitch profile of the measured data; the duration

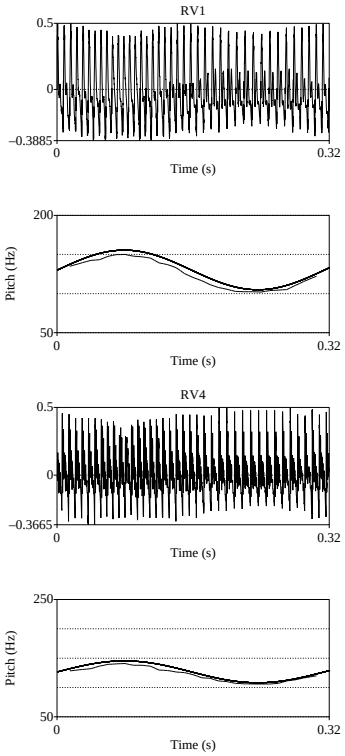

Figure 6: *Synthesised vowel sounds together with desired and measured pitch profile*

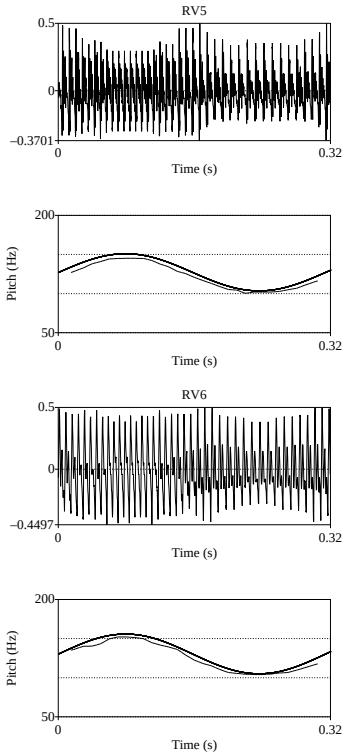

Figure 7: *Synthesised vowel sounds together with desired and measured pitch profiles*

of the synthesised data is also somewhat longer than that of the measured data. The small offset evident between desired and synthesised pitch contours is attributed to minor calibration error. The initial results presented here are encouraging. Indeed, perhaps somewhat surprisingly so. Specifically, good synthesis results are obtained using a simple nearest neighbour embedding model with only sparse data (typically around 1000 data points embedded in a space of dimension 17, corresponding to a very low density of around only 1.5 data points per dimension). A limited measured pitch excitation data: a simple rising pitch profile with a small number of data points at each specific pitch value.

7 Conclusions

In view of these observations, it seems likely that the data-based model of the vowel dynamics possesses an important degree of structure, perhaps reflecting physiological considerations, that requires further investigation. It is also clear that whilst encouraging there is still some way to go in overcoming the limitations of the approach. It is clear that Speech is a nonlinear process and that if we are to achive the holy grail of truly natural sounding synthetic speech that this must be accounted for. It is also clear that nonlinear synthesis techniques offer some potential to achive this although a great deal of research work remains to be done.

8 Acknowledgements

The contributions of my colleague Iain Mann to this work are gratefully acknowledged.

References

[1] B. Gabioud, *Fundamentals of Speech Synthesis and Speech Recognition*, ch. Articulatory Models in Speech Synthesis, pp. 215 – 230. John Wiley & Sons, 1994.

[2] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two–mass model of the vocal chords," *Bell System Technical Journal*, vol. 51, pp. 1233 – 1268, July–August 1972.

[3] T. Koizumi, S. Taniguchi, and S. Hiromitsu, "Two–mass models of the vocal cords for natural sounding voice synthesis," *Journal of the Acoustical Society of America*, vol. 82, pp. 1179 – 1192, October 1987.

[4] G. Fant, *Acoustic Theory of Speech Production*. Mouton, 1960.

[5] J. Markel and A. Gray, *Linear Prediction of Speech*. Berlin: Springer–Verlag, 1976.

[6] D. H. Klatt, "Software for a cascade/parallel formant synthesiser," *Journal of the Acoustical Society of America*, vol. 67, pp. 971 – 995, 1980.

[7] M. Edgington, A. Lowry, P. Jackson, A. Breen, and S. Minnis, "Overview of current text–to–speech techniques: Part II – prosody and speech generation," *BT Technical Journal*, vol. 14, pp. 84 – 99, January 1996.

[8] J. Page and A. Breen, "The laureate text–to–speech system, architecture and applications," *BT Technical Journal*, vol. 14, pp. 57 – 67, January 1996.

[9] M. Beutnagel, A. Conkie, J. Schroeter, and A. Syrdal, "The at&t next–gen tts system," in *Joint Meeting of ASA, EAA, and DAGA*, (Berlin, Germany), March 1999.

[10] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text–to–speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453 – 467, 1990.

[11] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 34, pp. 744 – 754, August 1986.

[12] T. Koizumi, S. Taniguchi, and S. Hiromitsu, "Glottal source–vocal tract interaction," *Journal of the Acoustical Society of America*, vol. 78, pp. 1541 – 1547, November 1985.

[13] D. M. Brookes and P. A. Naylor, "Speech production modelling with variable glottal reflection coefficient," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 671 – 674, 1988.

[14] H. M. Teager and S. M. Teager, "Evidence of nonlinear sound production mechanisms in the vocal tract," in *Proceedings of the NATO Advanced Study Institute on Speech Production and Modelling*, (Bonas, France), pp. 241 – 261, July 1989.

[15] I. Steinecke and H. Herzel, "Bifurcations in an asymmetric vocal–fold model," *Journal of the Acoustical Society of America*, vol. 97, pp. 1874 – 1884, March 1995.

[16] J. Schoentgen, "Non–linear signal representation and its application to the modelling of the glottal waveform," *Speech Communication*, vol. 9, pp. 189 – 201, 1990.

[17] R. J. DiFrancesco and E. Moulines, "Detection of glottal closure by jumps in the statistical properties of the speech signal," *Speech Communication*, vol. 9, pp. 401 – 418, December 1990.

[18] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 37, pp. 1805 – 1815, December 1989.

[19] D. Talkin, "Voicing epoch determination with dynamic programming," *Journal of the Acoustical Society of America*, vol. 85, Supplement 1, p. S149, 1989.

[20] F. Takens, "Detecting strange attractors in turbulence," in *Proceedings of Symposium on Dynamical Systems and Turbulence* (A. Dold and B. Eckmann, eds.), pp. 366 – 381, Lecture Notes in Mathematics, 1980.

[21] D. S. Broomhead and G. P. King, *Nonlinear Phenomena and Chaos*, ch. On the Qualitative Analysis of Experimental Dynamical Systems, pp. 113 – 144. Bristol: Adam Hilger, 1986.

[22] G. Kubin, "Poincaré sections for speech," in *Proceedings of the 1997 IEEE Workshop on Speech Coding*, (Pocono Manor, USA), September 1997.

[23] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH'95*, vol. 1, pp. 837 – 840, September 1995.

[24] G. Kubin, "Synthesis and coding of continuous speech with the nonlinear oscillator model," in *International Conference on Acoustics, Speech and Signal Processing*, (Atlanta, Georgia), pp. 267 – 270, May 1996.

[25] M. Birgmeier, *Kalman–trained Neural Networks for Signal Processing Applications*. PhD thesis, Technical University of Vienna, Vienna, 1996.

[26] I. Tokuda, R. Tokunaga, and K. Aihara, "A simple geometrical structure underlying speech signals of the Japanese vowel /a/," *International Journal of Bifurcation and Chaos*, vol. 6, no. 1, pp. 149 – 160, 1996.

[27] K. Narashimhan, J. C. Principe, and D. Childers, "Nonlinear dynamic modeling of the voiced excitation for improved speech synthesis," in *International Conference on Acoustics, Speech and Signal Processing*, (Phoenix, Arizona), pp. 389 – 392, March 1999.

[28] B. Mulgrew, "Applying radial basis functions," *IEEE Signal Processing Magazine*, vol. 13, pp. 50 – 65, March 1996.

[29] A. M. Fraser and H. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review A*, vol. 33, pp. 1134 – 1140, 1986.

[30] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[31] I. Mann, *An Investigation of Nonlinear Speech Synthesis and Pitch Modification Techniques*. PhD thesis, University of Edinburgh, 1999.

[32] S. Haykin and J. Principe, "Making sense of a complex world," *IEEE Signal Processing Magazine*, vol. 15, pp. 66 – 81, May 1998.

[33] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice–Hall, 1978.

[34] J. Schoentgen and R. de Guchteneere, "An algorithm for the measurement of jitter," *Speech Communication*, vol. 10, pp. 533 – 538, 1991.

[35] J. Stark, D. Broomhead, M. Davies, and J. Huke, "Takens embedding theorems for forced and stochastic systems," in *Proceedings of 2nd World Congress of Nonlinear Analysis*, 1996.

[36] J. Stark, "Delay embeddings for forced systems: Deterministic forcing," *Journal of Nonlinear Science*, vol. 9, pp. 255–332, 1999.

[37] H. Haas and G. Kubin, "Multi–band nonlinear oscillator model for speech," in *32nd Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 338 – 342, 1998.