DYNAMICAL MODELLING OF VOWEL SOUNDS AS A SYNTHESIS TOOL

M.Banbrook, S.McLaughlin.

Department of Electrical Engineering, University of Edinburgh,
King's Buildings, Mayfield Road, Edinburgh, EH9 3JL, UK.
tel : +44 131 6505655 fax : +44 131 6506554
email : mb@ee.ed.ac.uk

ABSTRACT

Speech synthesis can be produced using many varied techniques from formant/parametric synthesis to concatenation approaches. This paper presents a novel technique[1], based on the nonlinear dynamics of speech rather than than the time or frequency domain representations. It is demonstrated that the technique can be implemented effectively and used to produce high quality synthesised speech[2].

1. INTRODUCTION

Speech synthesisers today still lack the qualities that are needed to make them sound natural. Some of the short comings are at the phonetic transcription and intonation stage but there are also problems with the actual underlying sounds that the synthesisers reproduce. This can most commonly be found by attempting to reproduce sustained vowels which often results in very mechanical sounds which lack emotion.

Exactly what is missing from these sounds is not clear but one problem is the tendency to reproduce exactly the same sound each time it is required. If the dynamics of the signal, rather than the signal itself, could be used to model the speech then this problem would be avoided since the resulting output would change quite drastically depending on the starting conditions. This technique depends greatly on the dynamics of the signal and therefore it is important to investigate the nonlinear dynamical properties of speech before attempting any synthesis. This is an area that has received much interest recently with many authors reporting differing evidence for and against the existence of low dimensional attractors for speech [4, 5, 1, 6]. In a previous paper [3] the authors have shown that there is evidence that speech is a low dimensional, nonlinear, non-chaotic system, and as such it should be feasible to use the dynamics as a synthesis tool.

[1] Patent application number GB 9600774.5

[2] This work funded by BT Labs Martlesham, EPSRC and the Royal Society

2. THE ALGORITHM

This section gives a description of the underlying theory and a description of implementation details for production of a complete synthesis technique.

2.1. General Overview

Before describing all the details of how the synthesiser operates it is instructive to consider a brief, and somewhat generalised, overview of a typical implementation. The synthesiser consists of a number of building blocks : a basic parser which converts input words into their constituent phonemes giving both duration and co-articulation intervals, a general controller which decides which template to use and how to morph between different templates to create co-articulation and volume changes, two routines which synthesise the next point depending on whether the segment is voiced or not. In this particular example the synthesiser simply copies from the template when the segment is not voiced.

2.2. Production of voiced segments

In previous analyses [3] the short term prediction properties and the Lyapunov spectra for isolated vowels were explored . The short term prediction properties show that a locally linear model performs better than a globally linear model and suggesting that the system can be considered as nonlinear[4]. Further more the results suggest that the system is low dimensional and that the vowel sounds contain varying amounts of intrinsic fricative noise which must be considered in the calculation of Lyapunov spectra. A similar analysis on fricatives shows them to be modelled well by high dimensional, or stochastic signals. A complete noise robust analysis of the Lyapunov spectra show that the system has no positive exponents and therefore is not sensitive to initial conditions (i.e. chaotic).

The underlying dynamics of the system are obtained by using the state space representation of the time domain signal. This is achieved by embedding the signal using time delay embedding [2] and the nearest point to the initial start point for the synthesiser is located, as shown in Figure 1. It is then

possible to estimate the dynamics of how points evolve onto the next step for a small localised area around that point, which naturally includes the start point itself. This estimate of the dynamics is then applied to the synthesised start point to produce a point one time step ahead. The new synthesised point can then be viewed as a new start point and the process repeated, thus building up as long or as short a segment as is required.

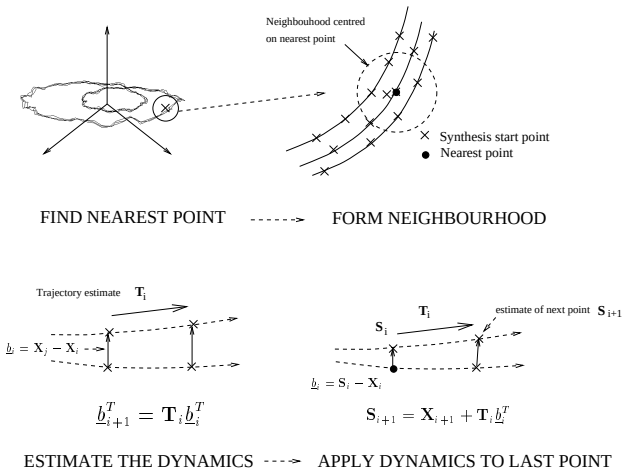

Figure 1: Steps in the synthesis of a vowel

Since the dynamics are applied to the displacement vector, $\underline{b}_i$ rather than the actual vector, $\underline{x}_i$, it is important to ensure that $\underline{b}_i > 0$; if $\underline{b}_i = 0$ then it follows that $\underline{b}_{i+1} = 0$ and the synthesiser is merely copying the embedded template which is no different from copying the time domain signal. By thus ensuring that the chosen 'nearest point' is never actually coincident with an actual point on the stored template then the synthesised points will always produce a unique trajectory which is dependant on its exact starting position and the chosen 'nearest points'.

2.3. Morphing

When the synthesiser needs to move from one phoneme onto another then essentially a full set of intermediate attractors, or templates, between the two phonemes are required. For example, to move from /i/ to /A/ (SAMPA representation) it is necessary to pass through /I/ and /E/ as seen on the formant chart, Figure 2. Using only the stored attractors results in a discontinuity each time the template jumps from one phoneme to another; in practice this jump is found to be too great and causes audible effects in the synthesised material. A natural extension would be to store a larger range of intermediate sounds but this is not really practical. Instead the intermediate attractors need to be constructed from the knowledge of the attractors that lie either side.

A simple and very effective approach is to produce an intermediate attractor $\underline{t}_i$ which is defined by

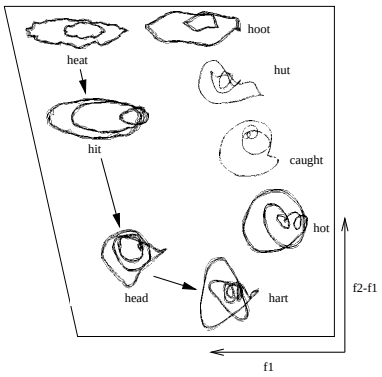

Figure 2: Formant chart of phonemes

$$\underline{t}_i = (\underline{a}_i - \underline{e}_i)d + \underline{e}_i, \qquad (1)$$

such that $\underline{t}$ is the set of points that lie a particular fraction, d, of the euclidean distance between corresponding points on the two main attractors, $\underline{a}$ and $\underline{e}$, as shown in Figure 3.

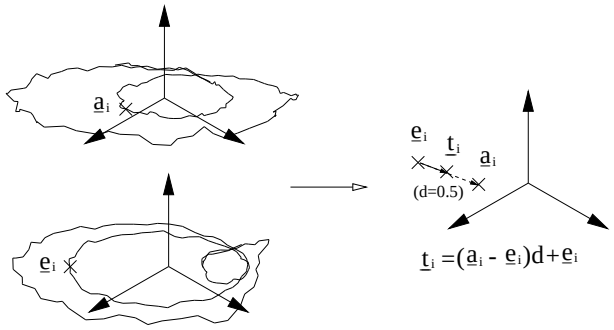

Figure 3: Morphing between phonemes

Once the intermediate attractor has been defined then the synthesis can be performed as before using the dynamics calculated from the new intermediate attractor. By using only small incremental changes to d, the synthesised points never stray far from the template attractor that is being used and therefore no noticable discontinuity occurs.

In order for this approach to work successfully it must be possible to locate 'corresponding' points on each attractor. If each cycle of the attractor was a constant then corresponding points are simply those with equal time indexes. Unfortunately most people cannot hold perfect, steady pitch and therefore the recorded samples will contain a small amount of variation in the average pitch level and in the constancy of the pitch. It is therefore necessary to perform a normalisation of the recorded samples before applying them to the synthesiser as described in the next section.

One further point that is worth noting is that the system should use all the information available. This means that

where intermediate attractors exist in the database, as in the example of moving from /I/ to /a/, then these stored attractors are used as intermediate targets themselves. Thus to morph from /I/ to /a/ the system would morph from /I/ to /i/ and then /i/ to /e/ and then /e/ to /a/. This is found to be much more effective than trying the morph from /I/ to /a/ directly.

2.4. Additional System Features

Other features are required to form a complete synthesis system such as pitch normalisation for the templates; volume and pitch adjustment to allow for intonation and prosodic information; ordinary concatenation rules to allow for unvoiced speech and plosives; and some form of high level parser to produce the phonetic description. All these features are beyond the scope of this paper but suffice to say that all such features are possible and are included in the synthesiser appraised in the next section.

3. SYSTEM APPRAISAL

The synthesiser presented in this paper is meant as a demonstration that the underlying waveform synthesis technique works and does offer possible advantages over conventional techniques. It is important to bear this in mind when appraising the system since the synthesiser is a very basic one, using no complicated phonetic description or intonation and a very simple resampling technique to produce fundamental frequency shifts. These shortcomings are naturally reflected in the resulting speech and so it is important to focus on the underlying aim which is to produce elongated segments of speech which sound natural, that is that they do not pocess the 'buzziness' common in other techniques, and that simple coarticulation between phonemes is possible.

The test words chosen to examine the effectiveness of the technique are a simple set of numbers. These are used because they contain a range of important properties : elongated vowels as in eighteen where the /i/ vowel is sustained, examples of vowel to plosive and plosive to vowel transitions, examples of fricative to vowel transition and examples of vowel to nasal and nasal to vowel transitions. It should be stressed that there is a need for dipthongs, such as /eI/ in "eight", where two vowels slide continuously into one another. Such a procedure is not possible in a straight concatenation approach but is feasible using the synthesis technique described earlier.

The first word synthesised is "eight" which is created as shown in figure 4, with the resulting waveform and spectrogram analysis shown in Figures 5 and [IMAGE A238G01.GIF] which also shows an example of a real "eight" for comparison.

Clearly the real example is rather longer in duration than the synthesised example but it still serves as a very useful benchmark for the synthesised version. Several important

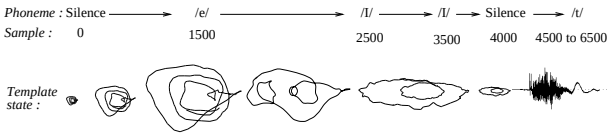

Figure 4: Steps in the generation of the word "eight"

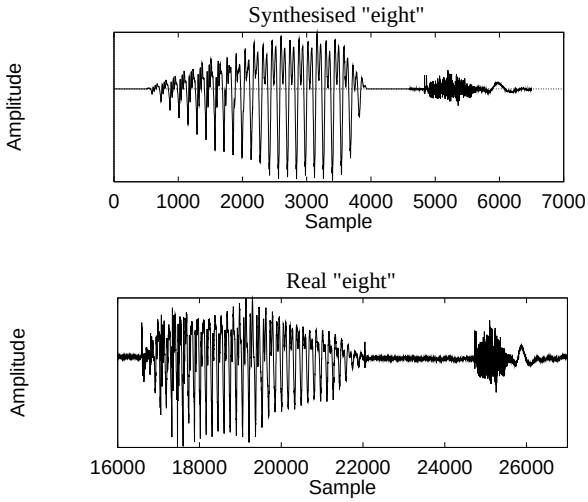

Figure 5: Time domain plots of a synthesised "eight" and a real "eight"

features can be seen from the spectrogram plots [IMAGE A238G01.GIF]: the basic formant structure has been accurately reproduced; during the morphing period the formants of the two phonemes appear to overlap rather than gradually moving from one set to the other, as seen in the real "eight"; there are no excessive dicontinuities such as would be audible as clicks.

Of course the proof of the pudding is in the eating and therefore to really find out how good the synthesised version is it must be listened to [SOUND A238S01.WAV].Listening to the synthesised eight two things become apparent; firstly the word is too short and secondly the reproduction of the diphone, /eI/, is extremely smooth and realistic. This would suggest that the overlapping of the formants, in this case, works quite well but it should be stressed that further work needs to done to show whether this is generally true.

The second word tested is the word "one" [SOUND A238S02.WAV]. This proved an extremely difficult word to reproduce since the move from silence to /u/ and then from /u/ to /Q/ seemed to produce an unexpected /m/ sound at the start of the word creating "mone" rather than "one". The solution to this was to include an extra target template between /u/ and /Q/ which makes sense since the formant trapezium traversal requires both /U/ and /O/ be passed through to get to /Q/. The full description of the steps used is shown in Figure 6 and the resulting waveforms and spectrograms in Figures 7 and [IMAGE A238G02.GIF].

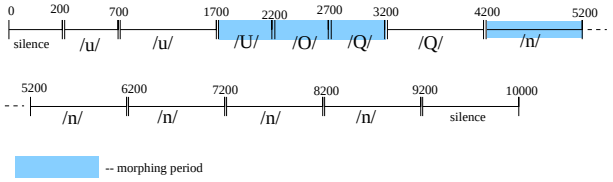

Figure 6: Steps in the generation of the word "one"

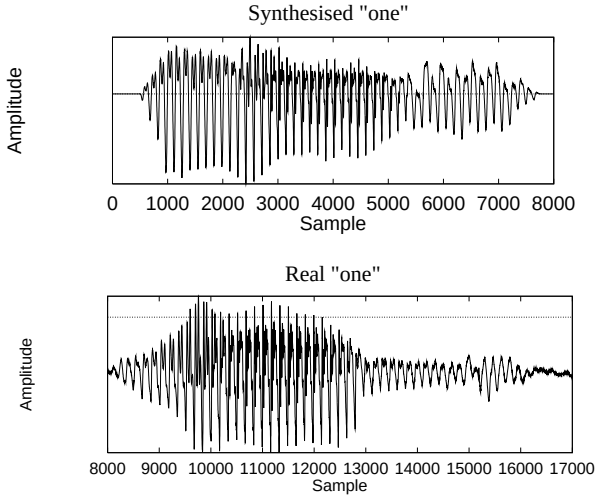

Figure 7: Time domain plots of a synthesised "one" and a real "one"

This time the two waveforms look substantially different especially in terms of the amplitude envelope. This was found to be necessary since simply emulating the real world resulted in a poor synthesised word. By allowing a much quicker volume increase at the start the spurious /m/ problem was reduced and it was found that the nasal /n/ needed to be quite loud to be heard significantly. Even though the waveforms do look quite different they do in fact yield similar spectrograms : the formant structure is clear in both. The main difference between the two is at the end where the synthesised nasal seems to contain a small amount of high frequency formant structure not present in the real one. The cause of this is not known and further work is required on the subject but it is possible that the reason is that the nasals are of a different dimension or contain chaotic properties which would show through on a full analysis similar to that performed on the vowels.

Other examples of synthesised numbers are given on the CD ([SOUND A238S03.WAV],[SOUND A238S04.WAV],[SOUND A238S05.WAV) with one in particular which is worth mentioning. The word "three" requires a small amount of rolling 'r' in order to sound like "three" and not "thwee". The closest phoneme to this is /U@/ as in hurt where the 'r' is allowed to roll slightly. Unfortunately, as is evident from the synthesised "three", the database was not constructed with this in mind and so the sound is closer to "her" than "hurt" resulting in a word that sounds more like "thwee".

Since one of the aims of the synthesiser is to reproduce high quality elongated vowels then a natural progression is to add an elongated "een" on the "eight" to produce "eighteen". This entails the synthesiser being able to start from a predetermined point in state space, as defined by the end of /t/, and then produce a non-repeating /i/. The cd image [IMAGE A238G03.GIF] shows both the spectrogram and the waveform for a portion of the synthesised "eighteen" [SOUND A238S06.WAV]. It is clear that the vowel is not simply repeating since variations are obvious in both the time domain and the spectrogram and indeed upon listening the word sounds natural and contains none of the characteristic buzz evident from other forms of synthesis.

4. CONCLUSION

In this paper a novel synthesis technique has been proposed which makes use of the local, low dimensional, nonlinear dynamics of vowels to produce a synthesiser capable of high quality, natural speech including elongated vowels. The underlying theory of the synthesiser is explained and a working demonstration is detailed along with a number of example synthesised words. From the demonstration it is clear that technique has much potential although a large amount of further work is required to integrate the technique into a fully operational system.

5. REFERENCES

1. H.F.V. Boshoff and M. Grotepass. The fractal dimension of fricative speech sounds. In *COSMIG '91*, pages 12–16. IEEE, 1991.
2. D.S. Broomhead and G.P. King. *Nonlinear phenonema and chaos*, chapter On the qualitative analysis of experimental dynamical systems, pages 113–144. Malvern science series. Adam Hilger, Bristol, 1986.
3. M.Banbrook and S.McLaughlin. Speech characterisation by nonlinear methods. submitted for rewiew to IEEE Trans. on Speech and Audio Processing, 1996.
4. M.Casdagli. Chaos and deterministic versus stochastic non-linear modelling. *Journal of the Royal Statistical Society B*, 54(2):303–328, 1991.
5. S. McLaughlin and A. Lowry. Nonlinear dynamical systems concepts in speech analysis. In *EUROSPEECH '93*, pages 377–380, 1993.
6. S.S. Narayanan and A.A. Alwan. A nonlinear dynamical system analysis of fricative consonants. *The Journal of the Acoustical Society of America*, 97(4):2511–2524, April 1995.