

Improved modelling of speech dynamics using non-linear formant trajectories for HMM-based speech synthesis

Hongwei Hu, Martin J. Russell

School of Electronic, Electrical & Computer Engineering
University of Birmingham, Birmingham, B15 2TT, UK

hwh400@bham.ac.uk, m.j.russell@bham.ac.uk

Abstract

This paper describes the use of non-linear formant trajectories to model speech dynamics. The performance of the non-linear formant dynamics model is evaluated using HMM-based speech synthesis experiments, in which the 12 dimensional parallel formant synthesiser control parameters and their time derivatives are used as the feature vectors in the HMM. Two types of formant synthesiser control parameters, named piecewise constant and smooth trajectory parameters, are used to drive the classic parallel formant synthesiser. The quality of the synthetic speech is assessed using three kinds of subjective tests. This paper shows that the non-linear formant dynamics model can improve the performance of HMM-based speech synthesis.

Index Terms: Dynamics, HMM synthesis, speaker adaptation

1. Introduction

The purpose of this study is to investigate the use of non-linear formant trajectories to improve the dynamics model for speech processing. From this perspective, the starting point of this research is a *linear/linear* multiple-level segmental HMM (MSHMM) [1], in which dynamics are modelled as piecewise constant, *linear* trajectories in a formant-based articulatory layer, as is shown in Figure 1. These linear formant trajectories are then transformed into the acoustic space using one or more *linear* articulatory-to-acoustic mappings. The intermediate layer presented in [1] is based on formant frequencies, ranging from the simplest form which just consists of the first three formant frequencies to a comprehensive representation using the 12 parallel formant synthesiser (PFS) control parameters [2]. Alternative intermediate-layer models of dynamics have also been studied in the past, such as [3, 4], which provide much complex models of dynamics. A major motivation for incorporating such a formant-based intermediate layer into a MSHMM is to allow speech dynamics to be modelled simply and directly in an articulatory related space without compromising the recognition performance. Indeed, both phonetic classification [1] and recognition [5] results on TIMIT show that, even with this simple linear/linear MSHMM system, speech recognition performance can achieve the upper bound of a fixed linear-trajectory acoustic segmental HMM. In other words, the incorporation of such an intermediate layer does not ‘hurt’.

One of the limitations of the above linear/linear MSHMM system is the use of linear trajectories to model dynamics. Although a piecewise linear model provides an adequate ‘passive’ approximation to the formant trajectories, it does not capture the active dynamics of the articulatory system. Moreover, no continuity constraints are considered across the segment boundaries. This motivates the work in [6], which demonstrates that

the phone recognition accuracy can be further improved by using non-linear formant trajectories to model dynamics. In [6], the non-linear trajectories are generated by using the speech parameter generation algorithm described in [7].

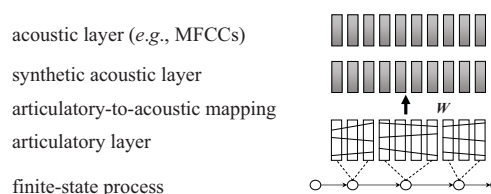


Figure 1: A *linear/linear* multiple-level segmental HMM.

This research builds on the work in [6] to evaluate the performance of the non-linear model of dynamics in a HMM-based speech synthesis paradigm. A simple HMM-based speech synthesis system is built, in which the HMMs are trained on 12 PFS control parameters and their time derivatives. It was hoped that the 12 PFS control parameters, which are conventionally used for formant synthesis, would also be useful for HMM-based speech synthesis, and the non-linear dynamics model presented in [6] for speech recognition would be beneficial for HMM synthesis as well. Both HMM-based speech synthesis [7] and the use of formant information for HMM-based speech synthesis [8] have been studied in the past. In either case, the speech parameter generation algorithm is used to produce smoothed cepstral or formant trajectories. The emphasis of this research, however, is to study the effect of the non-linear dynamics model for HMM-based synthesis, rather than to develop a state-of-the-art synthesis system. Compared with other HMM-based synthesis systems, a novelty of the HMM-based synthesis system developed in this research is the use of the 12 PFS control parameters, which combine both source and filter parameters in a compact form, to train the HMMs and the use of a parallel formant synthesiser to synthesize speech.

This research has implications for developing unified, trainable models which can support both recognition and synthesis within the framework of a MSHMM, whose intermediate layer is based on 12 PFS control parameters. In principle, these parameters are sufficient to create a ‘talker table’ to define a ‘voice’ for a formant synthesiser. If speaker adaptation techniques such as MLLR or MAP are used to operate in the formant domain to adapt the model to an individual’s speech, the resultant synthetic speech should sound like that individual. Moreover, adaptation in the articulatory layer should result in more interpretable changes in formant frequencies.

The rest of this paper is organized as follows. Section 2

briefly reviews the speech parameter generation algorithm to generate non-linear trajectories. Speech synthesis experiments and results are shown in Section 3 and 4. Conclusions and future work are presented in the final section.

2. Speech parameter generation algorithm

The non-linear formant trajectories used to model dynamics in this research are generated based on the speech parameter generation technique [7]. A detailed description of this algorithm appears in [7], where it is referred to as case 1, and a brief review is provided here for completeness.

Let $\mathbf{O} = \{\mathbf{o}_1, \dots, \mathbf{o}_T\}$ be a sequence of speech observations and $S = \{s_1, \dots, s_T\}$ a fixed state sequence of a HMM \mathcal{M} . Assume that the speech vector \mathbf{o}_t consists of static feature vector \mathbf{c}_t and dynamic feature vector $\Delta \mathbf{c}_t, \Delta^2 \mathbf{c}_t$. The delta and delta-delta coefficients are computed using the following equations, where θ is set to 1 in this paper.

$$\Delta \mathbf{c}_t = \frac{\mathbf{c}_{t+\theta} - \mathbf{c}_{t-\theta}}{2\theta}, \Delta^2 \mathbf{c}_t = \frac{\Delta \mathbf{c}_{t+\theta} - \Delta \mathbf{c}_{t-\theta}}{2\theta}. \quad (1)$$

Let \mathbf{W} be the linear transform matrix which transforms the sequence of static parameter vectors \mathbf{C} into an ‘augmented’ sequence of static plus dynamic vectors \mathbf{O} . Then the above equations can be written as

$$\mathbf{O} = \mathbf{W}\mathbf{C}. \quad (2)$$

For a given state sequence S , the speech parameter generation problem is to determine the parameter sequence \mathbf{C} which maximize $P(\mathbf{O}|S, \mathcal{M})$ with respect to \mathbf{C} under the constraints (2). By setting

$$\frac{\partial \log P(\mathbf{W}\mathbf{C}|S, \mathcal{M})}{\partial \mathbf{C}} = \mathbf{0}, \quad (3)$$

a set of linear equations are obtained. The non-linear PFS control parameters used in this research are obtained by solving equation (3). The speech parameter generation technique has now been widely used in HMM-based speech synthesis systems to improve the naturalness of the synthetic speech due to the smoothing effect of the algorithm.

3. Speech synthesis experiments on TIMIT

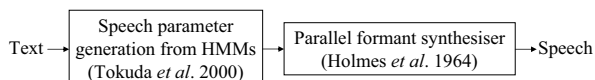


Figure 2: Diagram showing the HMM-based speech synthesis system developed in this research.

Figure 2 shows the HMM-based TTS synthesis system developed for the experimental purpose of this research. An orthographic English word string is converted into a sequence of 12 PFS control parameters based on a set of HMMs, which are trained on 12 PFS control parameters and their time derivatives. Synthesized speech is then produced by sending generated PFS control parameters to the parallel formant synthesiser.

3.1. Speech data and model set

The TIMIT corpus, downsampled to 8KHz for compatibility with the formant analyser, was used for all experiments. The training data included all male utterances from the TIMIT training set (3,252 utterances, 8 utterances discarded due to the data

corruption). Each training utterance was computed using the Holmes formant analyzer to generate corresponding 12 PFS control parameters, and then converted to HTK format and augmented with Δ and Δ^2 coefficients, resulting in a representation of 36 dimensional feature vectors with a 10 ms frame rate. 49 three-state, left-to-right (no-skip) single Gaussian monophone HMMs were built on 36 PFS control parameters using HTK.

3.2. Generation of formant synthesiser control parameters

Given a set of HMMs trained on 36 PFS control parameters and a phone sequence, the synthesis problem is to generate a sequence of 12 PFS control parameters from HMMs, which are then used to drive the parallel formant synthesiser. An appropriate set of 12 PFS control parameters is crucial for the quality of the synthetic speech. [9] has shown that given appropriate synthesiser control parameters, a parallel formant synthesiser can generate extremely high quality speech which is almost indistinguishable from natural speech. The synthesis problem can be solved in two steps: 1) to decide an optimal state sequence S given a HMM model set Λ and a phone sequence \mathcal{W} , and 2) based on the optimal state sequence S , generate a sequence of speech parameters \mathbf{o} , which maximizes the probability $p(\mathbf{o}|S, \Lambda)$. For 1), if the state duration for each state is known, then the state sequence can be obtained. Hence, the problem is to determine the state duration for each state. Two methods have been used in this research to generate 12 PFS synthesiser control parameters from HMMs.

In the first approach, two assumptions are made in order to simplify the synthesis problem. Firstly, the state sequence is decided based on the expected state duration, which is determined by the self-transition probability of each state. Secondly, the state mean vectors are assumed to be the output observation vectors. The resulting synthesiser control parameters generated in this way are referred to as piecewise constant PFS control parameters.

The second approach is based on the same state sequence as above. However, the dynamic constraints imposed by the dynamic features, *i.e.*, Δ and Δ^2 coefficients, are considered and the speech parameter generation algorithm described in Section 2 is applied to generate a sequence of non-linear, smooth PFS control parameters. Speech synthesiser control parameters generated in this way are referred to as smooth trajectory PFS control parameters.

3.3. Synthetic speech evaluation methods

In order to evaluate the quality of the synthetic speech and more importantly, to assess the effect of the non-linear formant trajectories, three types of subjective listening tests were conducted to evaluate different aspects of the synthetic speech. 20 native English speakers (10 females and 10 males) were invited to participate in the test.

3.3.1. Diagnostic Rhyme Test

The Diagnostic Rhyme Test was performed to test the intelligibility of consonants in words’ initial positions. 96 word-pairs, which are taken from [10], are used in the DRT to test 6 single acoustic features (or attributes), *i.e.*, *Voicing*, *Nasality*, *Sustension*, *Sibilantion*, *Graveness* and *Compactness*.

3.3.2. Naturalness test

The naturalness test aims to evaluate the overall quality of the synthetic speech in sentence level. Each subject is given 20

synthetic utterances and asked to give their subjective impression on the overall quality of the speech using the mean opinion score (MOS) method, *i.e.*, ‘Excellent/Good/Fair/Poor/Bad’. Subjects are prompted to take factors into account such as naturalness, listening effort, speaking style, comprehension problems, pronunciation, speaking rate and voice pleasantness *etc.* at the beginning of the test. Each subject is given a different set of 20 synthetic utterances.

3.3.3. Mimicry Test

The purpose of the Mimicry test is to test the perceptual similarity between an individual’s original speech and the synthetic speech, which is generated from a set of models adapted to that individual’s speech. In other words, it tests how likely the listener thinks the two utterances are spoken by the same speaker. In this experiment, the similarity is measured in 1 – 10 scales, where 10 means the two utterances are spoken by exactly the same speaker and 1 indicates that they are spoken by two completely different speakers. ‘Semi-synthetic’ (or copy synthesis) speech is used as an intermediate speech to make the comparison between original speech and synthetic speech, which is generated by passing the original speech through the Holmes formant analyzer to derive a sequence of 12 PFS control parameters, which are then sent back to the parallel formant synthesiser to generate synthetic speech.

Both MLLR and MAP adaptation techniques are used in the Mimicry test. In addition, the number of adaptation utterances varies. Since each subject speaks 10 sentences in the TIMIT corpus, the maximum number of adaptation utterances is 9 in this experiment. The number of adaptation utterances is chosen as 0 (without adaptation), 5, 9 and 9, where 9 means the model set is adapted to a different speaker using 9 adaptation utterances from that speaker.

4. Experiment results

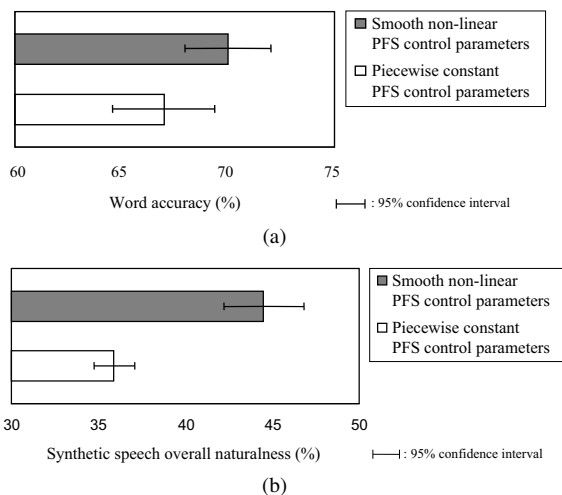


Figure 3: (a) DRT test results. (b) Naturalness test results.

It can be seen from Figure 3(a) that the intelligibility of non-linear trajectory based synthetic speech is higher than synthetic speech based on piecewise constant PFS control parameters, with word accuracy percentages of 70% and 67% re-

spectively. The naturalness test result follows the same trend, in which non-linear trajectory based synthetic speech gives a higher score, 45%, while piecewise constant control parameters based synthetic speech scores 35.9%, as is shown in Figure 3(b). A 1-tailed paired *t*-test was performed to determine if the non-linear dynamics model was effective and the results are summarized in Table 1. An alpha level of 0.05 was used for all statistical tests. Given the small values of *p*, there is strong evidence that, on average, the use of the non-linear formant trajectories does lead to improvements on intelligibility and naturalness of the synthetic speech. Moreover, a detailed analysis

	<i>M</i>	<i>SD</i>	<i>df</i>	<i>t</i>	<i>p</i>
DRT	2.9	5.2	19	2.49	0.01
Naturalness	4.55	4.29	19	4.75	< 0.001

Table 1: *t*-test results for the DRT test and naturalness test.

of the DRT results is shown in Figure 4, in which the average scores for each individual acoustic attribute are presented and comparison is made between two types of synthesiser control parameters. It can be learnt from these results that there is no benefit of using the speech parameter generation algorithm in discriminating voiced/unvoiced and nasal/oral sounds with this particular type of speech synthesis. The results for the rest of the acoustic attributes, *i.e.*, sustension, sibilation, graveness and compactness, follow the same trend, with non-linear, smooth synthetic speech giving higher scores than piecewise constant synthetic speech. In addition, a paired *t*-test discovers that only the differences between these two types of synthetic speech for attributes ‘sustension’ and ‘compactness’ are significant.

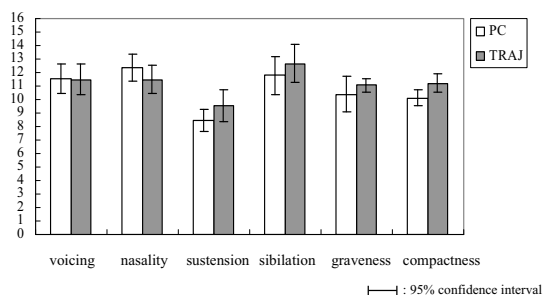


Figure 4: DRT test results shown in individual acoustic attribute, where ‘PC’ stands for piecewise constant synthesiser control parameters and ‘TRAJ’ means non-linear, smooth synthesiser control parameters.

The Mimicry test results are summarized in Figure 5. Figure 5(a) shows the perceptual similarity between an original utterance and a semi-synthetic speech. It can be seen that, even for the first group (S-S), *i.e.*, same speaker saying the same utterance, the similarity score between original speech and semi-synthetic speech drops to 56.8%, which suggests the semi-synthetic speech lost some information about speaker’s characteristics. In addition, it can be noticed that there is no significant differences between the test results for groups S-D and D-D, with scores of 33.5% and 33.6% respectively. This indicates that the perceptual similarity between an original utterance and a semi-synthetic utterance drops sharply when the utterances are different (*i.e.*, the content of the sentence), even though they are from the same speaker.

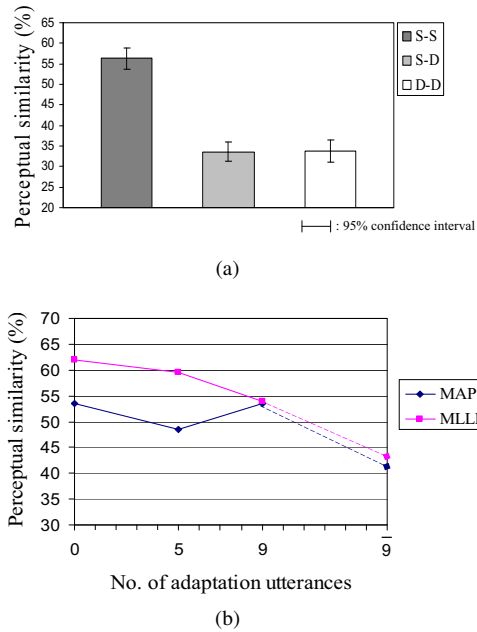


Figure 5: (a) Perceptual similarity between original speech and semi-synthetic speech. S-S: same speaker - same utterance; S-D: same speaker - different utterances; D-D: different speaker - different utterances. (b) Perceptual similarity between semi-synthetic speech and synthetic speech based on HMMs adapted to the same or different speaker. 9: HMMs adapted to a different speaker with 9 adaptation utterances.

Figure 5(b) shows the perceptual similarity between semi-synthetic and synthetic speech based on HMMs adapted to the same and different speakers with different number of adaptation utterances. Number ‘0’ means there is no adaptation, *i.e.*, speaker independent models are used to generate synthetic speech. Figure 5(b) shows that the perceptual similarity falls when the model set is adapted to a different speaker (denoted by 9), which is a reasonable result. However, the results of the mimicry test using MLLR and MAP adapted to the same speaker are unexpected. The perceptual similarity score decreases as the number of adaptation utterances increases when MLLR is used. When the MAP adaptation technique is used, the similarity falls from 53.5% to 48.5% as the number of adaptation utterance turns from 0 to 5, and then returns to 53.5% when 9 adaptation utterances are used.

5. Conclusions and future work

This paper shows unambiguously that the use of non-linear formant trajectories to model dynamics can improve both intelligibility and the overall quality of the HMM-based synthetic speech. The non-linear model of dynamics, generated using speech parameter generation algorithm, is particularly useful in improving the overall quality of the synthetic speech with longer contexts, *e.g.*, at sentence level, as shown in the Naturalness test. This is largely due to the smoothing effect of the algorithm, which in practice reduces the discontinuities, particularly between the boundaries of adjacent phones, which typically occur in the synthetic speech based on piecewise constant synthesiser control parameters.

In order to produce adaptable voices, conventional speaker adaptation techniques, including MLLR and MAP, were employed to generate adapted synthetic speech. However, these techniques proved to be unsuccessful in the Mimicry test. A possible explanation is that, while as few as 3 adaptation utterances can result in improvement in the recognition accuracy, the amount of the adaptation data might be too small for HMM-based speech synthesis and for this particular type of Mimicry test. However, this is due to the limitation of the TIMIT corpus. For speech recognition, MLLR is generally performs better with small amounts of adaptation data and MAP gradually catches up when more adaptation data is available. This rule of thumb seems still applicable in the similarity test, in which MLLR-based synthetic speech outperforms MAP-based synthetic speech, with at most a 10% difference between the similarity scores (when 5 adaptation utterances are used), as can be seen in Figure 5(b). Compared to other HMM-based speech synthesis systems, *e.g.*, HTS developed at Nagoya Institute of Technology [11], the overall quality of the synthetic speech is fairly poor. Although the purpose of this work is to develop dynamics model and naturalness is not the highest priority, the overall quality of the synthetic speech can surely be improved by adding prosodic structure and using context-dependent models.

Future work includes building a TTS synthesis system based on a set of multiple-level segmental HMMs, whose intermediate layer is based on 12 PFS control parameters, and generating formant synthesiser control parameters from these MSHMMs. Comparison can then be made between the synthetic speech generated by using HMMs and MSHMMs.

6. References

- [1] Russell, M.J. and Jackson, P.J.B., “A multiple-level linear/linear segmental HMM with a formant-based intermediate layer”. *Computer Speech and Language*, vol. 19, pp. 205–225, 2005.
- [2] Holmes, J.N., Mattingly, I.G. and Shearme, J.N., “Speech synthesis by rule”, *Language & Speech*, 7, pp. 127–143, 1964.
- [3] Richards, H.B. and Bridle, J.S., “The HDM: a segmental Hidden Dynamic Model of coarticulation”, *Proc. ICASSP*, pp. 357–360, 1999.
- [4] Deng, L., “A dynamic, feature-based approach to the interface between phonology and phonetics for speech modelling and recognition”, *Speech communication*, 24(4), pp. 288–323, 1998.
- [5] Russell, M.J., Zheng, X. and Jackson, P.J.B., “Modelling speech signals using formant frequencies as an intermediate representation”, *IET Signal Processing*, 1, pp. 43–50, 2007.
- [6] Hu, H. and Russell, M.J., “Speech recognition using non-linear trajectories in a formant-based articulatory layer of a multiple-level segmental HMM”, *Proc. INTERSPEECH*, pp. 2422–2425, 2008.
- [7] Tokuda, K., Yoshimura, T., Masuko T., Kobayashi, T. and Kitamura T., “Speech parameter generation algorithms for HMM-based speech synthesis”, *Proc. ICASSP*, vol.3, pp. 1315–1318, 2000.
- [8] Acero, A., “Formant analysis and synthesis using hidden Markov models”, *Proc. EUROSPEECH1999*, Budapest, 1999.
- [9] Holmes, J. N., “The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer”, *IEEE Transactions on Audio and Electroacoustics*, 21, pp. 298–305, 1973.
- [10] Pratt, R. L., “The assessment of speech intelligibility at RSRE”, *Proceedings of the Institute of Acoustics*, 6, pp. 439–443, 1984.
- [11] available online: <http://hts.sp.nitech.ac.jp/>.