TRAINABLE SPEECH SYNTHESIS WITH TRENDED HIDDEN MARKOV MODELS

John Dines and Sridha Sridharan

Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
j.dines@qut.edu.au s.sridharan@qut.edu.au

ABSTRACT

In this paper we present a trainable speech synthesis system that uses the trended Hidden Markov Model to generate the trajectories of spectral features of synthesis units. The synthesis units are trained from a transcribed continuous speech corpus, making the speech more natural than that produced by conventional diphone synthesisers which are generally trained from a highly articulated speech database and require a large investment of time and effort in order to train a new voice. The overall system has been incorporated into a PSOLA synthesiser to produce speech that is natural sounding and preserves the identity of the source speaker.

1. INTRODUCTION

Trainable speech synthesis (more accurately described as *voice synthesis* if we are attempting to retain the training speaker's characteristics) is the technique where by a customised voice for a Text-To-Speech (TTS) system is automatically learned via a set of training data. This type of synthesis relies on obtaining representative models of context-dependent speech units from the training set.

In our research we have used the trended Hidden Markov Model to represent the basic synthesis unit [4]. The standard Hidden Markov Model (HMM) that has been used widely in speech recognition and more recently in speech synthesis applications [6, 12] assumes that the modeled data is independent and identically distributed (IID). We know this assumption to be incorrect for plosives and longer segments of speech sounds, thus, a non-stationary model of spectral trajectories is required in order to accurately reconstruct such speech segments. This is achieved by including a linearly varying function of time in our formulation of the synthesis model — hence our use of the trended HMM.

This paper presents our initial work and findings on the application of trended Hidden Markov Models to trainable speech synthesis. Trended HMMs have been trained from a Bark scaled Line Spectral Frequency (LSF) parameterisation. Synthesis models have been trained using isolated words for comparison between trended and stationary HMMs in Modified Rhyme Testing (MRT). Models were also trained from a large continuous speech corpus and used to resynthesise speech using prosody information from the original speech. Preliminary results have shown significant improvement of trended HMMs over stationary HMMs.

The paper is organised as follows. Section 2 describes the training speech database and automatic labeling process. Section 3 describes the automatic training of the synthesis units and the incorporation of the trended Hidden Markov Models into a PSOLA synthesiser frame work. Finally, Sections 4 and 5 present the evaluation of synthesiser performance and discussion respectively.

This work was supported by the CSIRO Division of Telecommunications and Industrial Physics.

2. SPEECH DATABASE AND SEGMENTATION

Trainable speech synthesis requires that we have a phonetically aligned database from which we can train the speech production models. A Hidden Markov Model alignment tool was used for this purpose. The speech analysis comprised a 10th order MFCC parameterisation plus 0th order coefficient, delta and delta^2 terms (making a total of 33 coefficients). This was used to train 42 left-right, single-mixture, continuous density monophone HMMs including silence and short-pause models. All models comprised of three states except the plosives which only had two and the fricatives which only had one (as recommended in [7]). Context dependent models (cross-word triphones) were cloned and retrained from the monophone models. State tying was performed using the decision tree technique resulting in approximately 3000 unique models. The number of mixtures was incremented to four for each distribution.

Speaker independent models were initially trained using the male speakers of the TIMIT phonetically balanced speech corpus. The speaker dependent database from which synthesis models were trained comprised the journalistic speakers of the WSJ1 continuous speech recognition corpus. 1200 sentences were aligned for training of the synthesis models and a further 40 were used as test sentences. Maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR) adaptation of the speaker independent models was performed before alignment of the speech was carried out. Phonetic transcriptions of the corpus were obtained using the provided text transcriptions and the CMUDICT 4.0 Amercian English lexicon [1].

3. SPEECH SYNTHESIS

In this section we describe the design and implementation of the speech synthesiser. In particular we give an outline of the trended

HMM theory and then detail the training of these models and the overall speech synthesis system.

3.1. Trended Hidden Markov Models

Proposed by Deng [3], the trended Hidden Markov Model attempts to capture the non-stationary statistics of speech signals by representing the feature vector observations, $O_t, t = 1, 2, \cdots, T$ within state i, as a time varying function plus a stationary residual, Eq. (1).

$$O_t = \sum_{m=0}^{M} \mathbf{B}_i(m) f_m(t - \tau_i) + R_t(\Sigma_i) \quad (1)$$

where the left-hand term is the state-dependent polynomial regression function of order M with $\mathbf{B}_i(m)$ as the polynomial coefficients and $f_m(t - \tau_i)$ the mth order polynomial trend function. The right-hand term is the residual, R_t, with covariance Σ_i. The purpose of τ_i is to normalise the time at which the regression begins (to a value of $zero$).

We have used the Legendre family of orthogonal polynomials, Eq. (2), as used in [4], to estimate the polynomial coefficients $\mathbf{B}_i(m)$. Model parameters are estimated via a two stage process involving a segmentation step followed by a maximisation step, similar to the K-means algorithm used to initialise standard HMMs.

$$\begin{aligned} f_0(t) &= 1 \\ f_1(t) &= \sqrt{3}(2x - 1) \\ f_2(t) &= \sqrt{5}(6x^2 - 6x + 1) \\ f_3(t) &= \sqrt{7}(20x^3 - 30x^2 + 12x - 1) \\ f_4(t) &= 3(70x^4 - 140x^3 + 90x^2 - 20x + 1) \end{aligned} \quad (2)$$

where $x = t/T_0$ such that the polynomial $f_m(t)$ is defined on $[0, T_0]$ where T_0 is the duration of the state.

It has been shown [3, 4] that the trended Hidden Markov Model is able to better fit models to speech data then its stationary counterpart (see Figure 1). From this we would intuitively expect that its use in a HMM based synthesiser would result in better quality speech production.

3.2. Training of Synthesis Units

Bark scaled Line Spectral Frequency (LSF) parameterisation of the training speech with its phonetically aligned transcription (see Section 2) is sufficient to train a speaker dependent voice using trended Hidden Markov Models.

Initially, models were trained from a small database of isolated words which were later used in Modified Rhyme Tests (see Section 4). Phonemes were clustered according to the phonetic classes of adjacent speech units (eg. Consonant-ae+Nasal). Trended and stationary HMMs were trained using the tokens from each cluster.

A more comprehensive speech synthesis system was built by training models from a single journalistic speaker of the WSJ1 database using 1200 of the 1240 avaliable sentences. The remaining 40 sentences were retained for testing purposes. Given the large quantity of triphone combinations that are encountered in the English language the first task was to cluster similar models in order to reduce

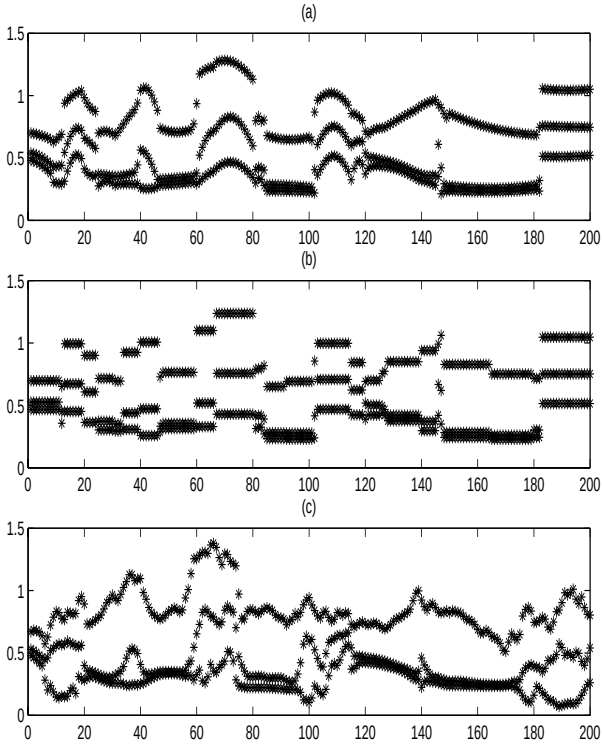

Figure 1: LSF parameter sequences for (a) trended (3rd order) (b) stationary HMMs compared with (c) original speech

the overall size of the model set and also to ensure that there was sufficient training data for each model. This was done by training a set of standard triphone HMMs and using the decision tree clustering technique to cluster whole models. This also enables the synthesis of unseen models. A total of approximately 2000 unique models resulted from over 10,000 observed triphones. State tying was not performed as the current formulation of trended HMMs does not enable embedded re-estimation of sub-word model parameters which is a necessary condition for the training of state tied HMMs.

Single mixture trended Hidden Markov Models were then trained using the clustered models determined from the previous step. The number of states was set to correspond with the stationary HMMs used for alignment. In order to improve synthesis quality the number of training tokens used for each model was limited, as it may be expected that some alignment errors would result from the automatic alignment process, and such tokens should be omitted from the training process. Hence, a maximum of twenty-four training tokens were selected based on the average log-likelihood per frame scored by the speech alignment tool.

3.3. Synthesiser Output

Synthesis of LSF parameters from model statistics is achieved by calculating the polynomial values over time given the state durations, Eq. (3). Speech synthesis was carried out using a PSOLA

system with a simple binary excitation scheme.

$$O_t = \sum_{m=0}^{M} \mathbf{B}_i(m) f_m(t - \tau_i), \qquad t = \tau_i, \cdots, \tau_i + T_\mathrm{o} \tag{3}$$

Synthesis of words for the Modified Rhyme Test was carried out using a monotone pitch contour. Line Spectral Frequency and energy features were synthesised from the model statistics only. State durations were set according to the mean state occupations measured during training. Synthesis of sentences from the WSJ1 corpus used the prosody and energy information extracted from the original test sentences.

In order to reduce spectral discontinuities at state and phoneme boundaries dynamic features may be used, as developed in [9]. It was found that the trended HMM synthesis models had relatively smooth transitions between state and even phoneme boundaries due to better modelling of LSF trajectories, hence, this technique was not as essential as for stationary HMMs. This is demonstrated in Figure 2 which shows the evolution of the smoothed spectrum with time.

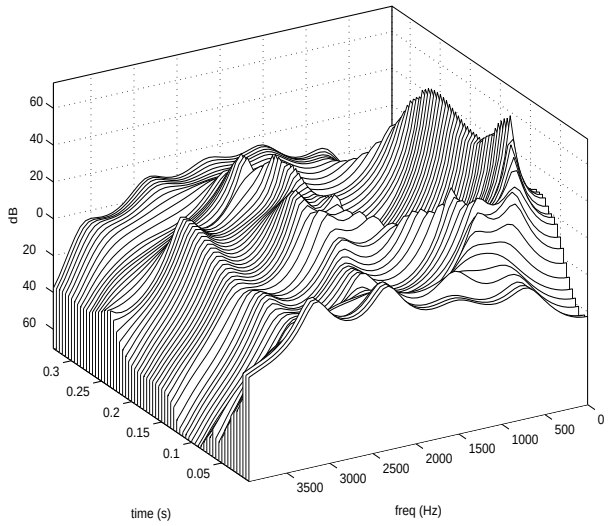

Figure 2: 3-D smoothed spectrum synthesised using 3rd order trended HMMs

It has been observed, as in other trainable speech synthesis systems, that the formant bandwidths tend to be wider and the spectral peaks less significant in the synthesised speech. This results from the fact that the trajectories are "averaged" during regression carried out for multiple training tokens. In [10] it is suggested that this may be rectified by constraining the Line Spectrum Pair difference parameters. In our current work this was partially compensated for by post-filtering the speech with a spectral sharpening filter, Eq. (4), as developed in [11].

$$H(z) = \frac{1 - A(z/\beta)}{1 - A(z/\gamma)} \tag{4}$$

where $A(z) = a_1(z^{-1}) + a_2(z^{-2}) + \cdots + a_m(z^{-m})$ is the mth order linear prediction filter and $0 < \beta < \gamma < 1$.

4. EVALUATION

Informal Modified Rhyme Testing was carried out using five untrained listeners. The subjects listened to a total of 50 words synthesised using both trended and stationary HMMs. An error rate of 29.6% and 43.2% was achieved for the trended and stationary HMMs respectively. The most frequent errors are detailed in Table 1. It can be seen that errors mostly occurred with plosive and nasal sounds. It was found that, in general, plosives were synthesised poorly by both trended and stationary HMMs due to the highly non-stationary behaviour of the phoneme (especially the energy information), though the trended HMM did provide some improvement. Also, nasal sounds were often mistaken due to flattening of the formant structure (which is discussed in more detail later on). A mechanised quality was particularly apparent in sustained vowels due to the binary excitation scheme, but generally the trended HMMs were judged to produce more natural vowel sounds with fewer spectral discontinuities.

Trended		Static	
Phone	% of Errors	Phone	% of Errors
/d/	21.6	/d/	16.6
/g/	13.5	/t/	12.9
/k/	13.5	/g/	11.1
/m/	8.1	/k/	11.1
/b/	8.1	/m/	9.2

Table 1: Most frequent errors in MRT.

Although these results may compare poorly with other speech synthesis systems that are widely available. It is worth noting that some errors could be attributed to inappropriate vowel durations under some phonetic contexts. A more advanced duration modeling scheme would be expected to improve results for both trended and stationary HMMs. The system is currently very basic and most of the errors that were recorded were localised to specific phoneme classes which our future research can concentrate on improving.

Testing of the models trained from the WSJ1 corpus was carried out by resynthesising speech from the 40 test sentences using the trained synthesis models. Observations by untrained listeners further support the findings of the MRT testing, in particular, the "formant flattening". It may be postulated that this is due to the wider variety of contextual variations that were present in the continuous speech database compared with the models trained from isolated words, and vowel, nasal and glide sounds are particularly influenced by the context in which they were uttered. Despite these apparent short-falls the synthesised sentences still maintained reasonable intelligiblity and naturalness.

Speech synthesis examples may be found at: *http://www.eese.qut.edu.au/~speech/demos/index.html*.

5. DISCUSSION

This paper has presented work on a trainable speech synthesis system that utilises the trended Hidden Markov Model. This has the advantage over previous stationary HMM synthesis techniques in that it is better able to synthesise spectral trajectories of speech

parameters, especially in vowel and plosive sounds. The use of trended HMMs was also found to reduce spectral discontinuities at state and phoneme boundaries when compared with stationary HMMs. The synthesised speech is natural sounding and maintains the key characteristics of the training speaker.

It has been noted that the synthesis system tends to exhibit "flattened" formant profiles which result in a degradation in intelligibility, especially in vowel, glide and nasal segments. (see Figure 3). Investigation of this phenomenon has shown it to be a consequence of pooling many phonetic occurrences into a single trajectory model, as explained in Section 3.3. Other research in this area has displayed similar behaviour [5, 10]. Experiments were carried out in order to verify this "pooling effect" by generating synthesis models that were trained from a single occurrence of each context dependent phoneme with the highest log-likelihood per frame. The results yielded almost identical perceptual quality of synthesised speech. It may be postulated that this is because the most likely phoneme is that which is closest to the centroid of our pool of training examples.

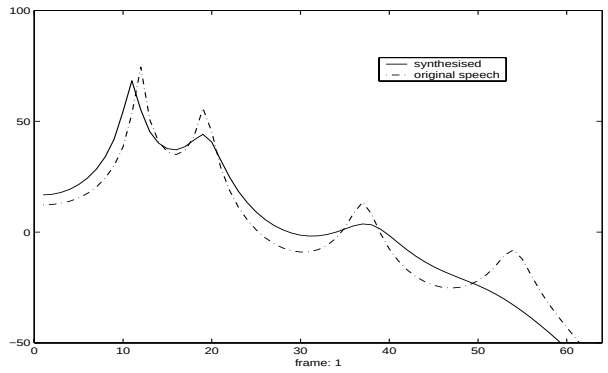

Figure 3: Example of formant flattening in the vowel /ax/.

In order to overcome this problem it must be recognised that coarticulation effects span more then just the adjacent phonetic contexts, especially in rapidly spoken continuous speech as is often encountered in the WSJ1 corpus. These coarticulation effects may be modeled by taking in to account greater phonetic context (eg. quinphones), but this would further increase problems with sparsity of training data. A more practical solution would utilise multiple mixtures of trend functions for each state of our models, which should be sufficient to capture the significant contextual variations of spectral trajectories encountered in natural speech. This is the direction that will be taken in our future work.

It was also observed that in some cases, models produced poorly synthesised speech (esp. plosives), probably due to excessive time modification of these sounds coupled with insufficient modeling of their dynamic behaviour. Future work needs to incorporate an improved duration modelling scheme that will better constrain such phenomena. Some instances may also have been due to inappropriate selection of training tokens for that synthesis unit. A more robust training unit selection algorithm that utilises multiple selection criteria — such as phone duration, voicing and pitch — in addition to the log-likelihood of the phone need to be implemented.

In addition to addressing the challenges that have been discussed above, we shall also endeavour to implement a full synthesis of speech by incorporating our automatically trained voices into the Festival Text-To-Speech system.

6. ACKNOWLEDGEMENTS

Experiments were carried out using a TD-PSOLA synthesiser based upon the Festival Text-To-Speech Synthesis System [2] and OGIresLPC plug-in residual excited LPC synthesiser [8].

7. REFERENCES

[1] CMUDICT 4.0 American English Lexicon. Centre for Speech Technology Research, University of Edinburgh, UK, 1996,1997.

[2] A. W. Black, P. Taylor, and R. Caley. *The Festival Speech Synthesis System*. Centre for Speech Technology Research, University of Edinburgh, UK, 1999.

[3] L. Deng. A generalised hidden markov model with state-conditioned trend functions of time for the speech signal. *Signal Processing*, 27(1):65–78, April 1992.

[4] L. Deng, M. Aksmanovic, X. Sun, and C. F. J. Wu. Speech recognition using Hidden Markov Models with polynomial regression functions as nonstationary states. *IEEE Transaction on Speech and Audio Processing*, 2(4):507–520, October 1994.

[5] R .E. Donovan. *Trainable Speech Synthesis*. PhD thesis, Cambridge University Engineering Department, June 1996.

[6] R. E. Donovan and P. C. Woodland. Automatic speech synthesis parameter estimation using HMMs. In *Proc. ICASSP-95*, pages 640–643, Detroit. MI, 1995.

[7] H. Hon, A. Acero, J. Liu, and M Plumpe. Automatic generation of synthesis units for trainable text-to-speech systems. In *Proc. ICASSP-98*, volume 1, pages 293–296, Seattle, WA, May 1998.

[8] M. Macon, A. Cronk, J. Wouters, and A. Kain. OGIresLPC: Diphone synthesiser using residual excited linear prediciton. Technical report, CSLU, Oregon Graduate Institute, Portland, OR, October 1997.

[9] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. Speech synthesis using HMMs with dynamic features. In *Proc. ICASSP-96*, 1996.

[10] B. Pellom and J. Hansen. Trainable speech synthesis based on trajectory modeling of line spectrum pair frequencies. In *IEEE Nordic Signal Processing Symposium*, pages 125–128, Vigso, Denmark, June 1998.

[11] A. Schaub and P. Straub. Spectral sharpening for speech enhancement / noise reduction. In *Proceedings of the IEEE*, volume 2, pages 993–996, 1991.

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *Proc. ICASSP-00*, 2000.