ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

Decision Tree Micro-prosody Structures for Text to Speech Synthesis

Aimin Chen Shu Lian Wong Saeed Vaseghi Charles Ho

The Queen's University of Belfast, N. Ireland. Email (a.chen, s.vaseghi, ch.ho@ee.qub.ac.uk)

10.21437/Eurospeech.1999-366

ABSTRACT

This paper explores the use of micro-prosody in improving the quality of synthesised speech in concatenated text to speech synthesis (TTS) systems. Micro-prosody are defined as prosodic signals within context-dependent triphone units and across neighbouring triphones. Micro-prosody parameters are modelled using a Markovian model whose state distributions depend on the current linguistic-prosodic state as well as the current and the neighbouring phones. The use of various speech unit selection criteria in the design of the TTS sound inventory and their effects in reducing the variance of micro-prosodic parameters in concatenated speech and on the TTS output speech are explored. The effect of the variability of the prosodic parameters of speech in the recorded samples from a given speaker, and the influence of accents, such as the US and the UK accented English, on speech prosody variability and on the design of TTS are considered.

1. INTRODUCTION

This paper explores the signals that convey prosody and presents a decision-tree based Markovian method for the modelling and synthesis of prosodic signals at a subword level. Prosodic parameters are super-segmental parameters that through the tonal quality of speech and stress convey meaning, intention, emphasis, and some of the talker's speaking characteristics [2-5]. Prosodic parameters are signalled with variations of pitch trajectory, duration, energy contour, and stress. The speech units selected to form the inventory of a concatenative TTS are taken from words in different sentences, in different contexts, and even from different recording sessions. The unnatural quality of synthesised speech is mainly due to the lack of correct interrelation between successive concatenated speech segments and the lack of appropriate contextual prosody.

In this paper we present experimental results of synthesising prosody based on the concept of decision tree clustered Markovian micro-prosody model. Micro-prosody are defined as prosodic signals within context dependent triphone units and across neighbouring triphones. Although prosody is defined as super-segmental parameters, much of it is signalled at a subword level by changing the trajectories of such parameters as the pitch signal. Micro-prosody parameters are modelled using a Markovian model whose state distributions depend on the current linguistic-prosodic state as well as the current and the neighbouring phones. For example the probability of the fundamental pitch frequency of the n^{th} phone in a sequence, $F_0(n)$, can be modelled by a conditional probability distribution given the values and the distributions of the pitch frequencies of the neighbouring (n-1)th and (n+1)th phones, $F_0(n-1)$ and $F_0(n+1)$, and the stress level at time n.

The experiments in this paper have focused on : (a) Identification of micro-prosodic signals; we show several typical forms of the variation of pitch trajectory for signalling different types of prosody, (b) parameterisation, statistical analysis and modelling of micro-prosody of natural speech, (c) the use of prosodic models in selecting speech units for the TTS inventory in order to minimise subsequent processing, and (d) synthesising prosody for TTS speech. The statistical models that define the correct prosodic trajectories are trained from naturally spoken speech, and are then used to maintain the expected relations between the prosodic parameters of successive triphone units in synthesised speech. The databases used for the initial training of micro-prosodic statistics are six hours recordings each of two males and a female speakers speaking in a natural clear conversational manner in UK and US accent English. We present results for micro-prosody statistics for natural speech, and for synthesised speech before and after prosody modification.

2. DESIGN OF SYNTHESIS TRIPHONE WAVEFORM INVENTORY

The speech unit for synthesis is chosen so as to reduce the subsequent signal processing required to improve the TTS quality. The automatic design of the TTS synthesis unit inventory [1] involves the following steps :

1. The choice of the synthesis unit; phone, syllable, etc.

2. Statistical modelling of the synthesis units.
3. Labelling and segmentation of the training database.
4. Selecting the best synthesis unit examples from the many available in the training database.
5. Training prosodic models for the TTS inventory.

Speech is modelled with context dependent triphone units. The use of triphones, in addition to capturing the contextual correlation of successive speech units, alleviates the distortion effects of any timing errors in unit segmentation process. In general the quality of TTS improves with increased contextual resolution. Particularly the naturalness of synthesised speech improves substantially when different synthesis units for word internal and cross word triphones are used. The first stage in the design of a concatenative TTS is the modelling, segmentation and labelling of the training speech units, and the selection of the best examples for TTS inventory. With the 45 phone set of the English BEEP dictionary there are theoretically more than 90,000 triphones. Due to phonological constraints, many of these do not occur and a total of about 20000 was observed in training data. A decision-tree clustering method is employed to cluster the triphone HMMs, and to estimate the models and the synthesis units for unseen triphones.
The triphone HMMs are then used for the labelling and segmentation of the training data. Speaker dependent HMMs used to segment the same data on which the models have been trained yield highly accurate segmentation and estimation of the timing boundaries of the triphones.
In general for each triphone there are a number of examples in the training database. These examples are ranked in terms of their power, duration, and their likelihood from their respective HMMs. The best example for each triphone are selected to form the triphone inventory. The criterion for selecting the best segment may be based on maximising

$$x_{best} = \underset{x \in f_1(d) \cap f_2(e) \cap f_3(F_0)}{\arg\max} \; p(x|\lambda, F_0, e, d) \qquad (1)$$

the probability of a segment given the HMM λ and pitch F_0, energy e, and duration d. In Eq(1) $x \in f_1(d) \cap f_2(e) \cap f_3(F_0)$ selects an intersection of the examples with preferred values of prosody parameters. For example the functions of duration and energy, $f_1(d)$ and $f_2(e)$, may be selected to favour units around or on the positive side of the mean value.

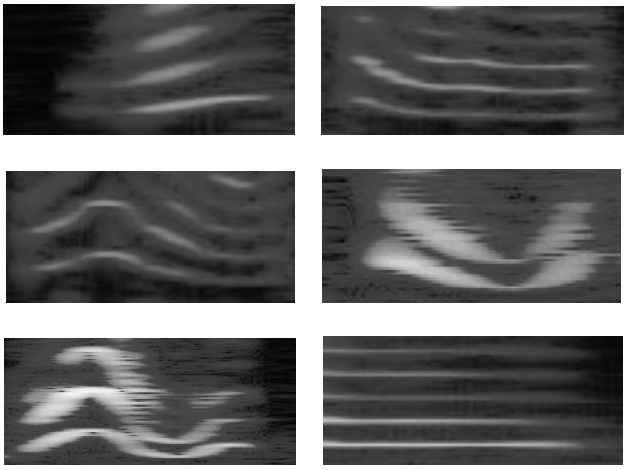

Figure 1 - Illustrations some elementary form of pitch trajectory that convey micro-prosody.

3. STATISTICAL MICRO-PROSODY TREE MODEL

Prosodic parameters span the duration of a word, a phrase or a sentence, and are used in speech to convey tonal quality, intention, and meaning [2-5]. Prosodic parameters include pitch, energy, and duration, these parameters are also affected by the level of word stress. The triphone segments in a TTS synthesis unit inventory are taken from various words spoken in different contexts and sentences, and even in different recording sessions. Hence the sequence of triphones in a concatenative synthesised speech sentence usually lack the correct interrelation between pitch, loudness, duration and stress. The prosodic parameters need to be processed to maintain a natural sounding relation between the prosody of successive triphones. The synthesis of the prosodic parameters, due to the lack of an effective computational model of prosody, remains the most challenging aspect of the design of TTS.
This section presents the concept of decision tree *statistical micro-prosody* model. Micro-prosody are defined as prosodic relations between successive phonetic segments. Micro-prosody parameters are considered as signals whose states depend on the current and the neighbouring phones, for example the probability of pitch frequency can be modelled as shown in figure2 as

$$p\left(F_{0_n}, \lambda_n | (\lambda_{n-1}, F_{0_{n-1}}), (\lambda_{n+1}, F_{0_{n+1}}), stress\right) \qquad (2)$$

where the prosody of a phone is affected by the neighbouring phones, their prosodic conditioning and the stress.

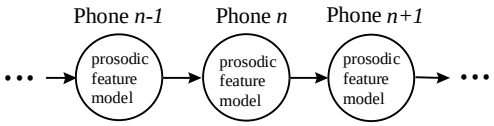

Figure 2 A chain model of prosodic feature trajectory.

For modelling and training of prosodic parameters a hierarchical decision tree-based prosody clustering structure is used in which linguistic knowledge and statistical training methods are combined. At the lowest level for each triphone a set of parameters are estimated to maintain the correct 'micro-prosodic' relationship between the energy, the duration and the pitch of successive triphones in a sentence. For example for triphone '*b*' with a left phonetic context of '*a*' and a right context of '*c*', '*a-b+c*', we estimate triphone level prosodic parameters such as pitch($b|a,c$) energy($b|a,c$), and duration($b|a,c$). For context dependent parameters the mean and variance of the prosodic parameters and their ratios such as [e_b/e_a, e_b/e_c, d_b/d_a, d_b/d_c, F_{0b}/F_{0c}, F_{0b}/F_{0c}] are estimated.
These statistics are then used to maintain the correct relation between prosody of successive triphone units in synthesised speech.

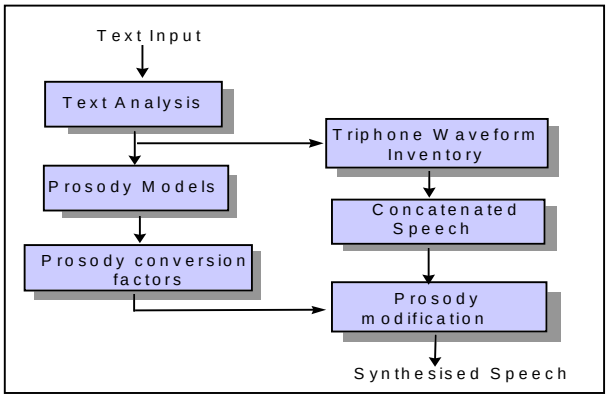

Figure 3 - Prosody modification

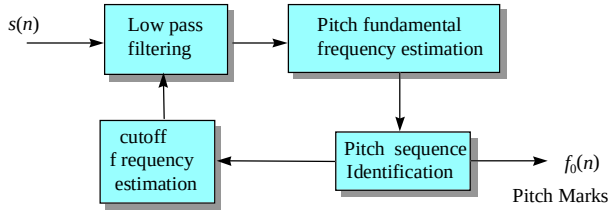

Figure 4 A block diagram illustration of pitch estimation system.

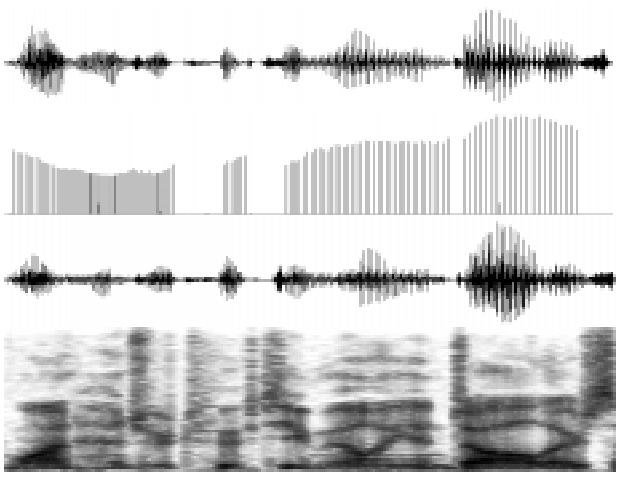

Figure 5 An original signal "One hundred fifty million dollars", the pitch mark sequence, the synthesised signal and its spectogram.

4. EVALUATIONS

The database used for the initial training of TTS is six hour recording of a person's voice speaking in a natural clear conversational manner. The speech is modeled using context dependent triphone HMMs. For training HMMs speech is segmented into frames of 25 ms length with 15 ms overlap between successive frames, and each frame is represented by 13 cepstral coefficients and the first and second derivatives. Decision tree clustering is used to limit the number of triphone HMMs to about a total of 9000 word internal and cross-word triphones and to synthesis the unseen triphones. A decision tree clustering method was also used to model the space of the prosody parameters. To derive prosodic models estimates of duration, energy and the fundamental pitch frequency are needed. The pitch frequency and the rate of change of pitch for each phone was estimated using a closed loop harmonic analysis system shown in figure 4. The speech units for the synthesis inventory are selected to reduce the subsequent signal processing steps needed for high quality synthesis. The distance from the HMMs and prosodic models are used to rank and select the best speech units.
For text to speech synthesis, the text is first analysed and then synthesised using the inventory and the prosody model. The micro-prosodic parameters of speech are then modified using prosodic models derived from training speech.

4.1 Experiments on Speech Unit Selection Criteria for the TTS Inventory

Experiments demonstrate that the process of selection of speech units for the TTS inventory can have a significant bearing on the success of the subsequent prosody modification process, and ultimately on the quality of synthesized speech. Ideally we want the statistics of temporal variations of concatenated speech units to be similar to that of natural speech. In the experiments the selection of speech units were based on the following criteria : (1) pitch trajectory statistics, (2) energy statistics, (3) duration statistics and finally, (4) HMM likelihood. Experiment show that the most important criteria in TTS unit selection is the pitch variations. For each triphone unit, the speech segment selected for the inventory has a pitch trajectory nearest to the mean of the pitch trajectories of the examples available for that triphone. This selection criterion will minimise the overall variance of the pitch trajectories and produce concatenated examples with less discontinuity in the pitch frequency of successive units. This approach produces significant improvement in speech quality and takes some of the processing burden from subsequent pitching process.

4.2 The Influence of Intra-Speaker Variability and Accents on TTS Quality

We experimented on 3 different speakers; two males, one with a UK English accent and the other with a US accent, and one female with a US accent. The standard UK English accent (the so called received pronunciation) generally demonstrates a greater variability in prosodic parameters of speech compared to the US accent. This larger variance is manifested on average in about 5% decrease in UK automatic speech recognition rate compared to the US English. The variability in speech is not only a function of the accent but also speaker characteristics.

The variability in the speech of a given speaker, intra-speaker variability, has a significant effect on the quality of concatenated TTS based on automatic segmentation of speech. In our experiments we achieved better TTS quality with US English, and the best synthesis quality was obtained with the female voice whose pitch frequency had remarkably low variance compared to that of the male speech data bases.

5. CONCLUSION

In this paper we explored the use of the micro-prosody concept in improving the quality of synthesised speech. Experiments show that with current technology the quality of TTS depends on three factors : (i) intra-speaker; variability (that is the consistency of pitch and other parameters) in prosodic parameters of the recorded speech used for the TTS inventory, (ii) criteria for selection of speech units for the TTS inventory, (iii) the use of micro-prosody models, and (iv) the use of linguistic prosody models. Good quality TTS can be obtained through careful selection of speech units and the use of micro-prosody concept. Improvement in synthesised speech quality can be obtained by ensuring that the statistics of the variations of the prosodic parameters in synthesised speech are similar to and as smooth as that of the natural speech.

REFERENCES

[1] R.E. Donovan. (1996) Trainable Speech Synthesis, PhD Thesis, Cambridge University.

[2] J. Lopez-Gonzalo, M. Rodriguez-Garcia, L. Hernandez-Gomez and J. M. Villar, Automatic prosodic modeling for speaker and task adaptation in text-to-speech, Proc. ICASSP, pp.927-930,1997.

[3] K. Ross and M. Ostendorf. (1996) Prediction of abstract prosodic labels for speech synthesis, Computer Speech and Language,10, pp.155-185.

[4] C.W. Wightman and M. Ostendorf. (1994) Automatic Labeling of Prosodic Patterns, IEEE Trans on Speech and Audio Processing, Vol. 2, No.4, pp. 469-481.

[5] L. M. Arslan, D. Talkin (1998), Speaker Transformation Using Sentence HMM Based Alignment and Detailed Prosody Modification, IEEE Proc. ICASSP98.

[6] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, M.Plumpe. (1997) Recent improvements on Microsoft's trainable text-to-speech system - Whistler, ICASSP.

[7] R.E. Donovan, P.C. Woodland, (1995) Improvements in an HMM-based speech Synthesiser, EUROSPEECH'95.4th European Conference on Speech Communication and Technology., Madrid ,Sep. , pp.573-576.

[9] H. Ohmura (1994), Find pitch contour extraction by voice fundamental wave filtering method, Proc. ICASSP, pp.II-189 - II-192, .

[10] Silverman, K., (1987) The Structure and Processing of Fundamental Frequency Contours. PhD Thesis, University of Cambridge.