MIMIC : A VOICE-ADAPTIVE PHONETIC-TREE SPEECH SYNTHESISER

Aimin Chen Saeed Vaseghi Charles Ho

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

ABSTRACT

This paper presents Mimic : a decision-tree based concatenative voice adaptive text to speech synthesiser. Mimic integrates text to speech synthesis (TTS) with speech recognition and speaker adaptation. Speech is synthesised from concatenation of triphone synthesis units. The triphone units are obtained from clusters of training examples modelled, labelled and segmented using clustered HMMs and Viterbi segmentation. The prosodic structure of pitch, duration and energy contours are captured using statistical training methods. The concept of a decision-tree based *statistical micro-prosody* model is introduced as a hierarchical method of modelling prosodic parameters. The voice adaptation component involves the adaptation of the spectral parameters as well as pitch, duration, and energy.

1. INTRODUCTION

Trainable voice-adaptive text to speech synthesisers (TTS) will have many applications beyond automatic directory enquiry [1-6]. Applications of voice adaptive TTS include; interpreted telephony, low bit rate speech coding comprising speech recognition and speaker parameterisation at the transmitter and voice-adaptive TTS at the receiver, broadcast studio and multi-media technology, voice dubbing and imitation, and personalised-voice for application's such as aid for the disabled. A wider application of TTS requires advances in two areas; (a) improving the prosodic quality of speech and (b) making TTS voice-adaptive.

Although in recent years there has been significant improvement in the quality of TTS, their naturalness still falls short of that of human speech. This is mainly due to the lack of natural prosody; the so called super segmental interrelation between concatenated segments of speech. The prosodic parameters are the pitch, duration, energy and stress which itself is a function of energy and duration. There are basically two approaches to the synthesis of prosody; rule base linguistics methods and statistical methods. The statistical methods of deriving prosody from training data, combined with the linguistic rules, could be as promising as the use of statistical techniques in speech recognition. In this paper we introduce and focus on the concept of micro-prosody; these are inter-phonetic relations between pitch duration and energy.

Voice conversion is another aspect of TTS described here. The goal of voice adaptive TTS is to employ a speech synthesiser in tandem with speech recognition so that the system can mimic a speaker's voice.

This paper is organised as follows. First, the design of TTS synthesis unit inventory is described. Then the prosody model is presented. Next is a description of the voice conversion method, followed by evaluation and conclusion.

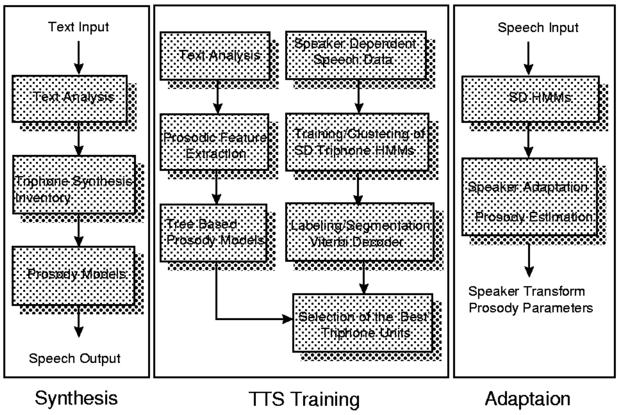
Text Input
Text Analysis
Triphone Synthesis Inventory
Prosody Models
Speech Output
Text Analysis
Speaker Dependent Speech Data
Prosodic Feature Extraction
Training/Clustering of SD Triphone HMMs
Tree Based Prosody Models
Labeling/Segmentation Viterbi Decoder
Selection of the 'Best' Triphone Units
Speech Input
SD HMMs
Speaker Adaptation Prosody Estimation
Speaker Transform Prosody Parameters
Synthesis
TTS Training
Adaptaion

Figure 1 - A voice adaptive TTS

2. DESIGN OF TTS SYNTHESIS TRIPHONE 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 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.

Speech is modelled with context dependent triphone units [1]. 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 labeling and segmentation of the training data. Speaker dependent HMMs used to segment the same data on which the models have been trained yields 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 data base. 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 quality segment may be based on maximising

$$x_{best} = \underset{x\in f_1(d)\cap f_2(e)\cap f_3(f_0)}{\arg\max} \quad p(x|\lambda, f_0, e, d) \qquad (1)$$

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.

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 [3-6]. 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

$$p\left(f0_n, \lambda_n | (\lambda_{n-1}, f0_{n-1}), (\lambda_{n+1}, f0_{n+1}), stress\right) \qquad (2)$$

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

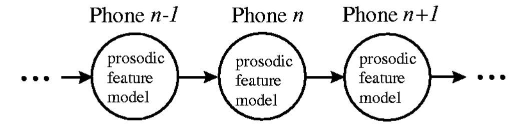

Figure 2 A chain model of prosodic feature trajectory.

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.

3.1 Micro-Prosody Adaptation

The mean and variance of the distribution of the micro-prosody parameters of the source speaker *s* is adapted to that of the target speaker *t* using the following relation [7]

$$f0^{t}_{a|b,c} = \alpha\, f0^{s}_{a|b,c} + \beta \qquad (3)$$

where the notation $f_{0a|b,c}$ denotes the pitch of the triphone *a* within the context of neighbouring phones *b* and *c*, and the adaptation coefficients α and β are given by

$$\alpha = \sqrt{\frac{\sigma_t^2}{\sigma_s^2}}, \qquad \beta = \mu_t - \alpha\mu_s \qquad (4)$$

where μ and σ^2 denote the context-dependent mean and variance of prosodic parameters. This relation is used for mapping of pitch, energy and duration parameters.

4. VOICE CONVERSION

Voice conversion is the mapping of the acoustic space of one speaker, the source speaker, to the acoustic space of another, the target speaker [3,4,7]. In [3] Abe, Nakamura etal describe the use of a vector quantiser code book as a one to one mapping function between the spectral vectors of the source and the target speakers. This approach was extended in [4] to a probabilistic Gaussian mixture model (GMM). In this paper these ideas are further extended to include hidden Markov models (HMMs) of context-dependent triphones. The factors that affect the voice characteristics of a speaker are gender, age, prosodic parameters and accent. Gender and age effect the vocal tract size and characteristics and also the pitch frequency. The simplest method for speaker adaptation involves frequency warping in which, given set of phonetic HMMs, for the input speech a phone-dependent ML warping parameter is estimated to map the frequency spectrum of the synthesiser’s voice to that of the input voice. A more detailed transformation has a full matrix linear transform for each triphone. The linear transformation matrices are estimated using an maximum likelihood criterion. The transforms are arranged in a phonetic-tree cluster structure, where the number of transform estimated at each level depends on the amount of training data from the target speaker.

4.1 Voice Spectral Mapping Functions

The mapping function converts the spectral envelop of the source speaker to that of the target speaker. Using least squared error optimisation the mapping function between the source spectrum $X(\omega)$ and the target spectrum $Y(\omega)$ for the k^{th} speech class is of the form

$$Y_k(\omega) = \frac{E[Y_k(\omega)]}{E[X_k(\omega)]} X_k(\omega) \qquad (5)$$

The expectation functions are obtained using VQ codebooks of the spectral envelopes of the source and the target speakers.

In [4] a Gaussian mixture model is described for mapping the source spectrum to the target spectrum. Extending the mapping function here to context-dependent phonetic HMMs, with *M*-mixture Gaussians per state model, the mapping between the corresponding states of phonetic HMMs yields

$$E[y|x] = \sum_{k=1}^{N_M}\sum_{i=1}^{N_S}\sum_{m=1}^{M} P(c_{kim}, s_{ki}, \lambda_k | x)[v_{kim} + \Gamma_{kim}\Sigma_{kim}^{-1}(x_{kim} - \mu_{kim})] \qquad (6)$$

where $\boldsymbol{v}$, $\boldsymbol{\mu}$, are the mean of $\boldsymbol{y}$ and $\boldsymbol{x}$, $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{x}$ and $\boldsymbol{\Gamma}$ the cross-covariance of $\boldsymbol{x}$ and $\boldsymbol{y}$. A drawback of Eq(6) is that it needs the cross-covariance of the source and target speakers.

An alternative method of spectral conversion is to use a linear speaker transform as in speaker adaptive speech recognition as

$$\boldsymbol{y}_k = \boldsymbol{A}\boldsymbol{x}_k \qquad (7)$$

Where the linear transformation is a full matrix. The solution for *A* can be obtained using a least squares or a probabilistic optimisation. Eq(7) can be extended to a decision tree structure of matrix transforms, where the number, and contextual resolution of transforms increase as more data becomes available.

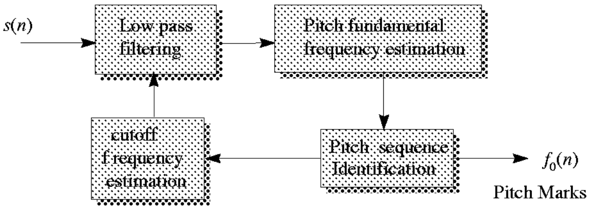
$s(n)$
Low pass filtering
Pitch fundamental frequency estimation
cutoff frequency estimation
Pitch sequence Identification
$f_0(n)$
Pitch Marks

Figure 3 A block diagram illustration of pitch estimation system.

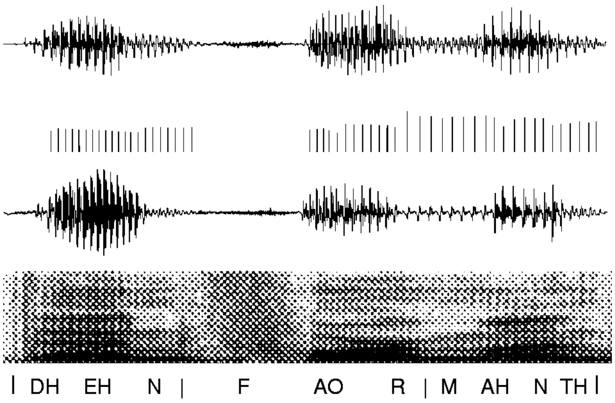

Figure 4 An original signal 'Then four months', the pitch mark sequence, the synthesised signal and its spectogram.

5. EVALUATIONS

The data base used for the initial training of Mimic 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 HMM training speech is segmented into frames of 25 ms length with 10 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 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 3.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 prosodic parameters of speech are modified using a harmonic synthesis model of speech segments. We asked our colleagues and visitors to our lab to compare the quality of TTS speech produce by Mimic against those of high quality TTS accessible on the internet. Mimic, without any manual intervention and tuning in its synthesis unit selection, was perceived to be as good as, and in many cases better than, those TTSs it was compared with. Particularly impressive is the ability of Mimic to retain good micro-prosodic aspects of the speech sound.

6. CONCLUSION

This paper introduced Mimic a voice adaptive decision-tree based context dependent TTS, that integrates speech recognition, text to speech synthesis and speaker adaptation. The concept of a tree based context dependent statistical micro-prosody model was presented. This model captures the statistical correlation of synthesis segments at phonetic level. The paper also described methods of speaker adaptation that more closely integrates speech recognition and synthesis.

REFERENCES

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

[2] 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.

[3] M. Abe, S. Nakamura, K. Shikano,H. Kuwabara (1996), Voice Conversion Through Vector Quantisation, Proc. IEEE Int. Conf. ICASSP-88, pages 655-658.

[4] Y. Stylianou, O Cappe (1998), Voice Conversion Based on Probabilistic Classification and Harmonic Noise Model, proc. ICASSP-98, pages 28-284.

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

[6] H. Shimodaira and M. Nakai, Prosodic phrase segmentation by pitch pattern clustering, Proc. ICASSP, pp.II-185 - II-188, 1994.

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