Neural-network-based F0 text-to-speech synthesiser for Mandarin

S.-H. Hwang
S.-H. Chen

Indexing terms: Mandarin speech synthesiser, Neural networks

Abstract: A neural-network-based approach to synthesising F0 information for Mandarin text-to-speech is discussed. The basic idea is to use neural networks to model the relationship between linguistic features, extracted from input text and parameters representing the pitch contour of syllables. Two MLPs are used to separately synthesise the mean and shape of pitch contour, using different linguistic features. A large set of utterances is employed to train these MLPs using the well known back-propagation algorithm. Pronunciation rules for generating F0 information are automatically learned and implicitly memorised by the MLPs. In the synthesis, parameters representing the mean and shape of the pitch contour of each syllable are generated using linguistic features extracted from the given input text. Simulation results confirmed that this is a promising approach for F0 synthesis. The resulting synthesised pitch contours of syllables match well with their original counterparts. Average root mean square errors of 0.94 ms/frame and 1.00 ms/frame were achieved.

1 Introduction

Speech is an effective and natural way for human beings to communicate with a computer. For two-way communication, the computer must speak like a person. Part of the requirement to reach this goal is to give the computer the ability to generate natural and fluent speech in response to any input text. A text-to-speech system is designed for this purpose. The synthesis of fundamental frequency (F0) is one of the most important things to influence the quality of the synthesised speech. This work studies the synthesis of F0 information for Mandarin text-to-speech.

In the past, the general approach was to invoke phonological rules for synthesis [1–8]. In this approach, input text is first analysed to extract linguistic features relevant to F0 synthesis, usually of different levels. These include lexical information, such as phonetic structure and accentuation of a word or syllable, syntactical structure, intonation pattern or declination effect for sentential utterance, semantic features etc. Phonological rules for synthesis are then used to generate F0 information. Fig. 1 is a schematic diagram of this approach. Phonological rules are inferred by observing a large set of utterances with the help of linguists. The relationship between the linguistic features of input texts and the F0 contour patterns of utterances is explored. Although this can be done by induction, it is generally difficult to explore the effect of mutual interaction of linguistic features at different levels. Hence, the inferred phonological rules for synthesis are always incomplete. Some synthesised speech therefore sounds monotonous and unnatural.

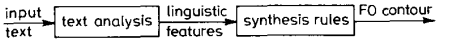


Fig. 1 *Schematic diagram of a general F0 synthesiser*

In an alternative approach, a statistical model [9] was used to synthesise F0 contours of syllables for Mandarin text-to-speech. The idea is to substitute a statistical model for explicit synthesis rules. The model is used to describe the relationship between F0 contour patterns of syllables and contextual linguistic features. In the training process, parameters of the model are empirically estimated from a large training set of sentential utterances. Phonological rules for synthesis are then automatically deduced and implicitly memorised in the model. In the synthesis process the best combination of F0 contour patterns of syllables is estimated based on the model, provided that linguistic features are given by analysing the input text. The primary advantage of the statistical approach is that pronunciation rules can be automatically extracted from the training data set through the training process and implicitly stored in the model. A major disadvantage of this approach is the need for a tremendously large set of utterances to properly train the statistical model as the number of parameters increases.

Motivated by the success of both the statistical approach [9] and NETalk [10], we propose a novel neural-network-based F0 synthesiser for Mandarin text-to-speech, taking neural networks as a mechanism for generating F0 information in response to the input linguistic features. This is similar to the statistical approach above. However, the requirement for large set of training utterances can be alleviated by taking advantage of the excellent interpolation property of neural networks. A similar idea was used in References 11 and 12. In Reference 11, two NETalk-like neural networks were used to

Paper 1421K (E5, C4), first received 9th August 1993 and in revised form 14th June 1994

The authors are with the Department of Communication Engineering and Center For Telecommunications Research, National Chiao Tung University, Hsinchu, Taiwan 300, Republic of China

This study was supported by the National Science Council, Republic of China. The database is provided by TL, MOTC, Republic of China.

generate the F0 value of a phoneme and the F0 fluctuations within that phoneme by using the linguistic features of several neighbouring phonemes. Promising results were obtained in the simulation using a small data set. In Reference 12, neural networks were used in Japanese text-to-speech to generate parameters of F0 contour, including the maximum, average and differential F0 of a 'mora'. Better performance than the conventional rule-based method was reported based on simulation results. This is the first study of a neural-network-based approach to synthesising F0 information for Mandarin speech. The study uses two multilayer perceptrons (MLPs) to generate the mean and shape of the pitch contour of syllables from linguistic features extracted from the input text. A large set of sentential utterances accompanying the texts is used to train the MLPs by adjusting their weights. The relationship between the F0 information of natural speech and linguistic features extracted from the text associated with the speech is then automatically explored and memorised in the weights of the MLPs. Several advantages can be found. First, as with the statistical approach, pronunciation rules are automatically inferred without the help of linguists and are implicitly contained in the MLPs. Secondly, the synthesised speech is usually more natural because the MLPs are trained using real speech. This mainly results from the strong learning capability of MLPs to make their outputs mimic the corresponding target patterns given in the training process.

2 Properties of the F0 information in Mandarin speech

F0 information is the most important prosodic information to synthesise in a text-to-speech system for producing natural and fluent speech. In a normal sentential utterance the F0 contour is, in general, following a specific pattern known as the intonation. For a declarative sentential utterance,the F0 contour is usually declining. But, as mentioned above, many other factors may also affect the pronunciation of F0 contour. The following discusses the properties of F0 contour in Mandarin speech in detail.

Mandarin Chinese is a tonal language, the basic pronunciation unit of which is the syllable. Each written character is pronounced as a syllable with a tone. It is therefore very natural to choose the syllable as the basic unit of synthesis in a Mandarin text-to-speech (TTS) system. Fig. 2 displays the phonetic structure of a syllable. It is composed of a vowel-final and an optional

initial	final		
(consonant)	(medial)	main vowel	(ending)

Fig. 2 *Phonetic structure of a syllable in Mandarin speech*

consonant-initial. There are 39 finals and 22 initials in total. Syllables with the same phonemic constituents (i.e. initial–final) and different tones have different lexical meanings. The tones of syllables are mainly characterised by their F0 contours. There are only five lexical tones, namely, high-level, mid-rising, mid-falling-rising, high-falling and neutral tones. They are commonly referred to as Tone 1–Tone 5. A previous study [6] concluded that the F0 contour of each of the first four tones can be simply represented by a standard pattern (Fig. 3). As for Tone 5, the pronunciation is usually highly context-dependent, so that its F0 contour shape is relatively arbitrary. Moreover, it is always pronounced short and light. It would therefore seem that the F0 contours of

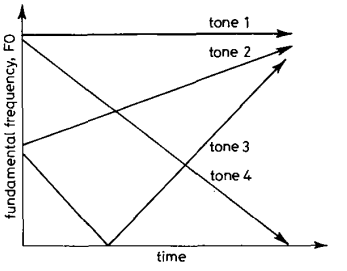


Fig. 3 *Basic F0 contour shapes of the first four tones*

sentential Mandarin speech are more regularly pronounced, so as to make synthesis for a TTS system much simpler. However, in practice, the contours of syllables are subject to various modifications in continuous speech. They are determined primarily by the tones and phrasal conditions of syllables and the coarticulation effect makes them affected by the F0 contours of neighbouring syllables. Influence also comes from neighbouring tones. This effect is known as snadhi rules. They are also greatly affected by the intonation pattern or declination effect of a sentence. Sometimes the semantics will change their shapes or mean levels. Moreover, emotion or the habit of speaking may also affect the pronunciation of F0 contour in running speech. Therefore, synthesis is not a trivial task.

3 The proposed approach

We now discuss the proposed neural-network-based approach of F0 synthesis for Mandarin text-to-speech. In this approach, the pitch contours of syllables are taken as basic synthesis units. These are first generated syllable by syllable and then concatenated to form the synthesised pitch contour for the given input sentential text. In our realisation, the pitch contour of a syllable is regarded as a pattern represented by certain parameters, which are generated instead of synthesising the pitch contour of each syllable frame-by-frame. The idea to adopt this approach is twofold. First, based on observing pitch contours of syllables in real Mandarin speech, we found that they are all smooth curves and suitable for parametric representation using curve-fitting techniques, such as polynomial expansion with few parameters. Secondly, it gives us an opportunity to isolate the influences of different linguistic features on the generation of pitch contour. As we decompose the pitch contour of a syllable into several components, we expect that ideally the coefficients of these components will be exclusively affected by different groups of linguistic features. Due to the fact that the pitch level of a syllable in Mandarin speech is more significantly affected by global linguistic features than by local features, and that the shape is more significantly affected by local features, the mean and shape of pitch contour are considered separately. Some global linguistic features, such as positional information in a sentence or clause, as well as some local features, such as tones of neighbouring syllables, are used to synthesise pitch means of syllables for mimicking the intonation pattern

of the input sentence. On the other hand, only local linguistic features extracted from neighbouring syllables are chosen to determine the pitch shape for each syllable to simulate the effects of coarticulation and sandhi rules. To be more specific, the pitch contour of a syllable is represented by a smooth curve formed by orthonormal polynomial expansion using coefficients up to the third order [13]. The zero-th-order coefficient represents the mean of the pitch contour and the other three coefficients represent its shape. The basic functions of the orthonormal polynomial expansion are normalised in length to [0, 1] and expressed as

$$\Phi_0\left(\frac{i}{N}\right) = 1 \tag{1}$$

$$\Phi_1\left(\frac{i}{N}\right) = \left[\frac{12N}{(N+2)}\right]^{1/2}\left[\left(\frac{i}{N}\right) - \frac{1}{2}\right] \tag{2}$$

$$\Phi_2\left(\frac{i}{N}\right) = \left[\frac{180N^3}{(N-1)(N+2)(N+3)}\right]^{1/2} \times \left[\left(\frac{i}{N}\right)^2 - \left(\frac{i}{N}\right) + \frac{N-1}{6N}\right] \tag{3}$$

$$\Phi_3\left(\frac{i}{N}\right) = \left[\frac{2800N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}\right]^{1/2} \times \left[\left(\frac{i}{N}\right)^3 - \frac{3}{2}\left(\frac{i}{N}\right)^2 + \frac{6N^2-3N+2}{10N^2}\left(\frac{i}{N}\right) - \frac{(N-1)(N-2)}{20N^2}\right] \tag{4}$$

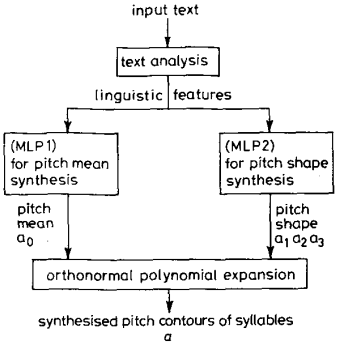
for $0 \leqslant i \leqslant N$, where $N+1$ is the length of the F0 contour and $N \geqslant 3$. These basis functions are, in fact, discrete Legendre polynomials. Using this representation, the pitch contour of a syllable can then be expressed as

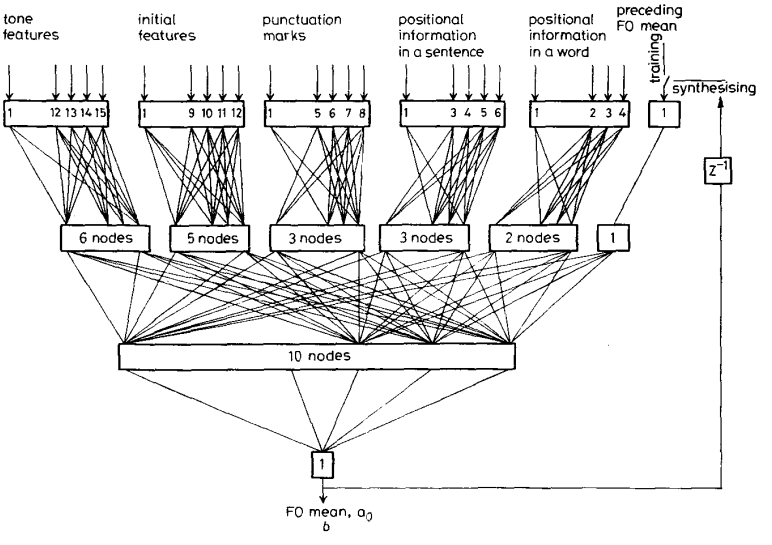
$$f\left(\frac{i}{N}\right) = \sum_{j=0}^{3} a_j \Phi_j\left(\frac{i}{N}\right) \tag{5}$$

for $0 \leqslant i \leqslant N$. These four coefficients are partitioned into two groups, a_0 and a_1, a_2, a_3, to be synthesised using separate neural networks with different input linguistic features.

Fig. 4 is a schematic diagram of the proposed F0 synthesis scheme. It is composed of two main parts, one being the text analysis for linguistic feature extraction and the other the neural networks for synthesising pitch contour. The input sentential text is first analysed to extract linguistic features relevant to the F0 synthesis. Two MLPs are then employed to generate parameters of the pitch contour of each syllable. The first MLP has a single output for generating the pitch mean and the second has three outputs for generating the other three parameters representing the pitch shape.

In text analysis, some linguistic features relevant to F0 synthesis for Mandarin text-to-speech are extracted from the given input sentential text. They include types of initials; tones of the processing syllable and/or the two contiguous syllables; punctuation marks before and after the processing syllable; and the locations of the processing syllable in the word it belongs to and in the input sentence. All these features have great influence on the F0





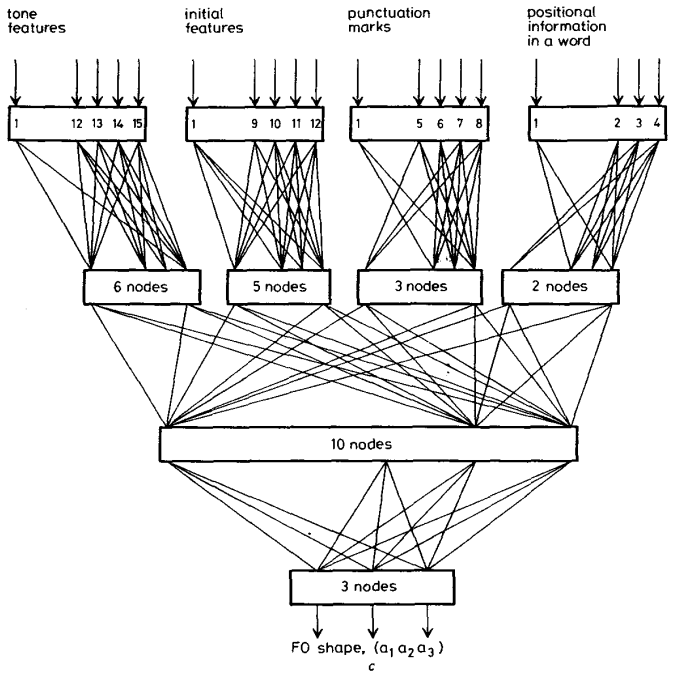


Fig. 4 *The proposed F0 synthesiser*
a Block diagram
b MLP for pitch mean synthesis
c MLP for pitch shape synthesis

pronunciation of Mandarin speech, and can be roughly categorised into global and local linguistic features. Linguistic features related to the intonation of a clause or a sentence are generally classified as global. Other features related to the phonetic structure of the processing syllable, as well as the contextual features extracted from neighbouring syllables, are local.

The coarticulation between the pitch contours of two contiguous syllables is affected by the type of initial of the latter syllable. As mentioned before, there is a total of 22 initials in Mandarin speech. These are broadly classified into six types, according to the manner of articulation, and were taken as input features (Table 1). Twelve binary linguistic features representing the initial types of two neighbouring syllables of the processing syllable are used.

The pitch contour pattern of the processing syllable is determined primarily by its tone. It is also greatly affected by the two neighbouring tones resulting from sandhi rules. Fifteen binary features representing the tones of the processing and the two neighbouring syllables are used.

Table 1: Six broad types of initials

Type	Initial
1	m, n, l, r, 'null'
2	h, sh, shi
3	b, d, g
4	tz, j, ji
5	p, t, k
6	ts, ch, chi, f, s

Punctuation marks are generally used as prosody control in pronunciation. From obsevations of real Mandarin speech, the level of pitch contour usually has large jumps at syntactical boundaries set by punctuation marks. Punctuation marks used in this study are classified into four categories, as shown in Table 2. Thus eight binary features indicating the punctuation marks before and after the processing syllable are used.

Table 2: Four broad types of punctuation mark

Type	Punctuation mark
1	Boundary of sentence
2	Comma
3	Pause, colon, semicolon
4	Query

Words are basic meaningful pronunciation units in Mandarin Chinese. There exist monosyllabic, disyllabic and polysyllabic words. In F0 pronunciation, intraword coarticulation is, in general, more significant than interword coarticulation. Besides, the rhythm of a sentential utterance seems to roughly match with words. Three types of positional information are specified to indicate whether the processing syllable is located at the beginning, intermediate or end position of a word. Monosyllabic words are specially indicated. Hence four binary features are used.

In running Mandarin speech the F0 contour of a sentential utterance will follow an intonation pattern. For a declarative utterance, the F0 contour usually declines. To

simulate the declination effect, the position of the processing syllable in the input sentential text is labelled as one of six levels and taken as a global linguistic feature. The labelling is a linear quantisation operation starting from the beginning syllable with step size equalling two syllables. So, six binary features are used. For a short sentence, only the first one or two of these six binary features are used, so as to make the synthesised pitch means stay in high-pitch frequency. For a long sentence, the synthesised pitch means for syllables following the 12th syllable will stay in low-pitch frequency and are smooth without abrupt change.

In this study, Features 1–4 and Features 1–5 are used, respectively, in the two MLPs for synthesising the shape and mean of pitch contour. One additional nonbinary input representing the pitch mean of the preceding syllable is used in the MLP for synthesising pitch mean to more accurately model the intonation. Both the MLPs are three-layer in structure, with two hidden layers. The first hidden layer of each is a concentration layer. Nodes on this layer are partitioned into several groups to be separately connected by different input features. The node number in each group is empirically determined to roughly consider the relative importance and variability of the input features. With this arrangement, input features can be more compactly encoded so that the complexities of the MLPs are reduced. The outputs of the MLPs are, respectively, the zero-th order and the following three coefficients of orthonormal polynomial expansion of the reproduction pitch contour of the processing syllable. Linear activation functions are used in all nodes of the output layer to linearly generate these coefficients.

Two phases are included in this F0 synthesis approach, the training phase and the synthesis phase. In the training phase, a large set of sentential utterances accompanying their texts is used to train the MLPs by using the well known back-propagation (BP) algorithm. The training process is done syllable by syllable. For each syllable of a training utterance, the linguistic features discussed previously are extracted from the corresponding text and taken as input features. Coefficients of orthonormal polynomial expansion extracted from the pitch contour $f(i/N)$ of the syllable in the training utterance are taken as targets. Formulations for coefficient extraction are expressed as

$$a_j = \frac{1}{N+1} \sum_{i=0}^{N} f\left(\frac{i}{N}\right) \Phi_j\left(\frac{i}{N}\right) \quad \text{for } 0 \leqslant j \leqslant 3 \tag{6}$$

The process of the BP training algorithm is to recursively adjust the weights of the MLPs with the goal of minimising the mean square error between the actual outputs and the desired targets. By sequentially feeding with training samples, the training process is continued until a convergence is reached. The criterion of convergence is simply defined as

$$\frac{SE_{i-1} - SE_i}{SE_{i-1}} < 0.001 \tag{7}$$

where SE_i is the mean square error of the ith iterative epoch. By this training process, the pronunciation rules of F0 information are expected to be automatically inferred and implicitly memorised in the MLPs. In the synthesis phase, the pitch contour for the input test sentential text is synthesised syllable by syllable. For each syllable, linguistic features are first extracted from the input text and then fed into the MLPs. The outputs are then used to generate the pitch contour of the processing syllable by orthonormal polynomial expansion. It is noted that the real value and the synthesised value of the pitch mean of the preceding syllable are used in the training and in the synthesis test, respectively, as the nonbinary input of the MLP which synthesises pitch mean.

4 Simulation results

The validity of the approach was examined by simulation. Two sets of reading utterances recorded by Telecommunication Laboratories (TL) were used. The first data set used for training consists of 30 522 syllables and the second one for testing contains 9014 syllables. The texts of these utterances include some phonetically balanced sentences and some paragraphic newspaper texts arbitrarily chosen from a large text file. All utterances are spoken naturally and fluently by a single male speaker. The speech signals were first 0–4.5 kHz lowpass filtered, sampled at 10 kHz and A/D converted into a 16 bit data format. They were then pre-emphasised and segmented into 10 ms frames. Preprocessing, including syllable segmentation, pitch detection and orthogonal transform, as then performed to extract one parameter vector for each syllable. The syllable segmentation was done manually by Telecommunication Laboratories, who provide the database with the help of observations of waveform and hearing. The pitch contour was detected by using the SIFT algorithm and then manually corrected. The first component of the parameter vector represents the mean of the pitch contour of the syllable, and the other three components represent the shape. Texts associated with all utterances were also analysed to extract various linguistic features. In the training, linguistic features and parameter vectors were respectively taken as the inputs and the desired output targets to properly train the MLPs using the BP algorithm. Over 400 epochs were needed to reach convergence for both MLPs. In the synthesis test, linguistic features were fed into the MLPs to generate output parameter vectors for synthesising pitch contours.

As discussed previously, the mean and shape of pitch contour were synthesised separately for each syllable. For the MLP synthesising pitch mean a total of 46 inputs, including 45 binary global and local linguistic features and one nonbinary feature feedback from the pitch mean of the preceding syllable, were used. The local linguistic features, such as tonality, initial type and positional information in a word, were used to simulate the local variation on the pitch level change, whereas the global linguistic features, such as positional information in a sentence, were used to simulate the intonation of a sentential utterance. Two typical examples of the inside and the outside tests are displayed in Figs. 5 and 6, respectively. An inside test means that the input text belongs to the training set, whereas the input text in an outside test is an untrained text. As seen in these two Figures, most synthesised pitch means of syllables match well to their original counterparts. It is worth noting that the intonation patterns of these two original pitch contours seem to be correctly synthesised. Table 3 lists the average root mean square error (RMSE) of the synthesised pitch

Table 3: Average RMSEs of the synthesised F0 means

	RMSE
Inside test	0.85 ms/frame
Outside test	0.90 ms/frame
Statistics of F0 mean	$\bar{a}_0$ = 8.09 ms/frame σ_{a_0} = 1.55 ms/frame

Note: a_0 is the pitch mean

means. Average RMSEs of 0.85 ms/frame and 0.90 ms/frame were achieved for the inside and the outside tests, respectively. These experimental results are reasonably good.

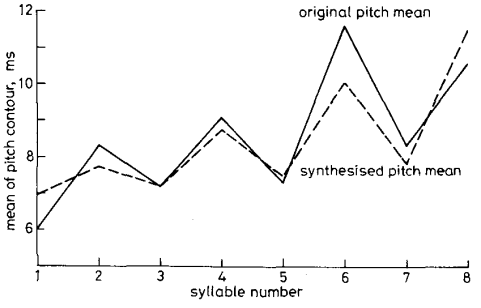


Fig. 5 *Original and synthesised pitch means for a typical sentence in the inside test*

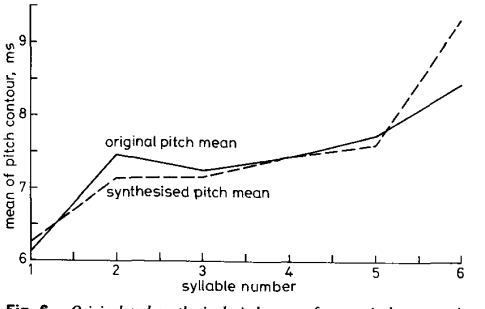


Fig. 6 *Original and synthesised pitch means for a typical sentence in the outside test*

For the MLP synthesising pitch shape, a total of 39 binary inputs composed of only local linguistic features were used to simulate the local variation on the shapes of pitch contours of syllables. Table 4 lists the average RMSE of the synthesised pitch shapes. Average RMSEs of 0.39 ms/frame and 0.43 ms/frame were achieved for the inside and the outside tests, respectively. These experiment results are also reasonably good. By combining the results shown in Tables 3 and 4, the overall average RMSEs of the synthesised pitch contours of syllables are 0.94 ms/frame and 1.00 ms/frame for the inside and the outside tests, respectively.

Table 4: Average RMSEs of the synthesised F0 shapes

	RMSE
Inside test	0.39 ms/frame
Outside test	0.43 ms/frame
Statistics of F0 shapes	$\sigma_b = 0.66$ ms/frame

Note: $\sigma_b = \left[\sum_{i=1}^{3} (a_i - \bar{a}_i)^2\right]^{0.5}$

(a_1, a_2, a_3) is the pitch shape

Figs. 7 and 8 display two typical synthesised pitch contours of sentential utterances in the inside and outside tests, respectively. As seen in these two Figures, both synthesised pitch contours resemble their original counterparts. Some pitch-level jumps appear at the boundaries of two connected pitch contours of contiguous syllables, but they are not significant because there is always an energy dip at the boundary of two contiguous syllables. Besides,

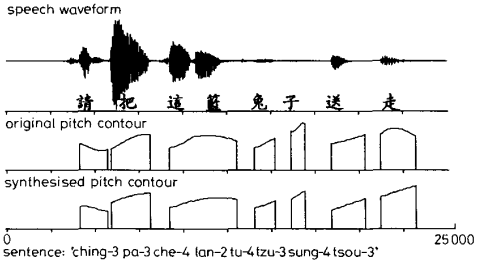


Fig. 7 *Original and synthesised pitch contours for a typical sentence in the inside test*

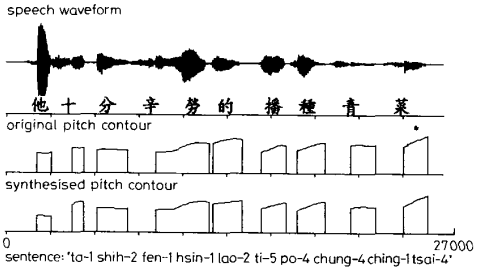


Fig. 8 *Original and synthesised pitch contours for a typical sentence in the outside test*

these pitch jumps are small, so they are usually imperceptible. By comparing all synthesised pitch contours with their original counterparts we found that, at some syntactic or semantic boundaries, the pitch-level change on the synthesised pitch contour cannot follow the abrupt jump of pitch level in the original utterance. This mainly results from the fact that both syntactic information and semantic information are not used in this approach.

Finally, a syllable-based Mandarin TTS system was implemented for the purpose of informally evaluating the quality of intonation of the synthesised speech. The basic speech synthesiser used in the TTS system is a linear prediction-based one with quality comparable to the CELP coder. All parameters were automatically generated by the system to synthesise the speech in response to an input text. Using an informal listening test, we found that most synthetic sentential utterances generated by this system sounded clear and natural. Only a few suffered from distortions of unnaturalness. These mainly resulted from the above-mentioned defect of synthesis occurring at some syntactic and semantic boundaries. Nevertheless, they were all still highly intelligible.

From the above simulation results, most synthesised pitch contours of syllables were confirmed to match well with their original counterparts. This shows that most pronunciation rules had been learned and memorised in the MLPs. This system was proved to be reasonably good. Of course, if more linguistic information could be incorporated into the system, further improvement to emulate a specific emotional status on the synthetic speech would be possible. Finally, an advantage of this neural-network-based approach over the statistical

approach [9] is discussed. In the statistical approach, the pitch contour of a syllable is vector quantised so that only a finite set of pitch contour patterns can be synthesised. In the neural-network-based approach, all parameters representing a pitch contour are unquantised, such that an infinite number of F0 contours can be generated by these MLP models.

5 Conclusion

In this paper, a novel-neural-network-based approach to synthesising F0 information for Mandarin text-to-speech has been discussed. It differs from a conventional rule-based approach by using MLPs to learn and memorise the pronunciation rules for generating F0 information. Simulation results confirmed that it performs well. Average RMSEs of 0.94 ms/frame and 1.0 ms/frame have been achieved for the synthesised pitch contour in the inside and the outside tests, respectively.

Some advantages of this approach are first, that the difficulty of analysing natural Mandarin language syntactically or semantically is avoided. Only simple linguistic features were used in our simulations. Secondly, the pronunciation rules of prosody are automatically inferred. Thirdly, the synthesised speech sounds more natural than that of a conventional rule-based approach.

6 References

1 HART, J.'t, and COHEN, A.: 'Intonation by rule: a perceptual guest', *J. Phonet.*, 1973, **1**, pp. 309–327

2 OLIVE, J.P., and NAKATANI, H.L.: 'Rule-synthesis of speech by word concatenation: a first step', *J. Acoust. Soc. Am.*, 1974, **55**, pp. 660–666

3 OLIVE, J.P.: 'Fundamental frequency rules for the synthesis of simple declarative English sentences', *J. Acoust. Soc. Am.*, 1975, **57**, pp. 476–482

4 FUJISAKI, H., HIROSE, K., TAKAHASHI, N., and MORIKAWA, H.: 'Acoustic characteristics and the underlying rules of intonation of the common Japanese used by radio and TV announcers', *IEEE Int. Conf. Acoust. Speech Signal Process*, 1986, pp. 2039–2042

5 ZHANG, J.: 'Acoustic parameters and phonological rules of a text-to-speech system for Chinese', *ICASSP*, Tokyo, Japan, April 1986, pp. 2023–2026

6 LEE, L.S., TSENG, C.Y., and OUH-YOUNG, M.: 'The synthesis rules in a Chinese text-to-speech system', *IEEE Trans.*, 1989, **ASSP-37**, pp. 1309–1319

7 SAGISAKA, Y.: 'On the prediction of global F0 shape for Japanese text-to-speech', 1990, **ICAS-SP**, pp. 325–328

8 CHAN, N.C., and CHAN, C.: 'Prosodic rules for connected mandarin synthesis', *J. Inform. Sci. Engineer*, 1992, **8**, pp. 261–281

9 CHEN, S.H., LEE, S.M., and CHANG, S.: 'A Chinese fundamental frequency synthesizer based on a statistical model', *ICSLP*, 1990, Kobe, Japan, pp. 829–832

10 SENOWSKI, T.J., and ROSENBERG, C.R.: 'NETalk: a parallel network that learns to read aloud', *Johns Hopkins University EECS Technical Report*, 1986

11 SCORDILIS, M., and GOWDY, J.: 'Neural network based generation of fundamental frequency contours', *ICASSP*, 1989, **1**, pp. 219–222

12 ABE, M., and SATO, H.: 'Two-stage F0 control model using syllable based F0 units', *ICASSP*, 1992, **2**, pp. 53–56

13 CHEN, S.H., and WANG, Y.R.: 'Vector quantization of pitch information in Mandarin speech', *IEEE Trans.*, 1990, **COM-38**, pp. 1317–1320

14 LIPPMANN, R.P.: 'An introduction to computing with neural nets', *IEEE ASSP Mag.*, 1987, **4**, pp. 4–22