A statistical model based fundamental frequency synthesizer for Mandarin speech

Sin-Horng Chen and Saga Chang
Department of Communication Engineering and Center for Telecommunications Research, National Chiao Tung University, Hsinchu, Taiwan 300, Republic of China

Su-Min Lee
Telecommunication Laboratories, Directorate General of Telecommunications, Ministry of Communications, Chung-Li, Taiwan 320, Republic of China

(Received 7 May 1990; accepted for publication 23 March 1992)

A novel method based on a statistical model for the fundamental-frequency ($F0$) synthesis in Mandarin text-to-speech is proposed. Specifically, a statistical model is employed to determine the relationship between $F0$ contour patterns of syllables and linguistic features representing the context. Parameters of the model were empirically estimated from a large training set of sentential utterances. Phonologic rules are then automatically deduced through the training process and implicitly memorized in the model. In the synthesis process, contextual features are extracted from a given input text, and the best estimates of $F0$ contour patterns of syllable are then found by a Viterbi algorithm using the well-trained model. This method can be regarded as employing a stochastic grammar to reduce the number of candidates of $F0$ contour pattern at each decision point of synthesis. Although linguistic features on various levels of input text can be incorporated into the model, only some relevant contextual features extracted from neighboring syllables were used in this study. Performance of this method was examined by simulation using a database composed of nine repetitions of 112 declarative sentential utterances of the same text, all spoken by a single speaker. By closely examining the well-trained model, some evidence was found to show that the declination effect as well as several sandhi rules are implicitly contained in the model. Experimental results show that 77.56% of synthesized $F0$ contours coincide with the VQ-quantized counterpart of the original natural speech. Naturalness of the synthesized speech was confirmed by an informal listening test.

PACS numbers: 43.72.Ja, 43.70.Kv

INTRODUCTION

Synthesis of fundamental frequency is an important part of a text-to-speech system. A general approach to the synthesis of fundamental frequency is to invoke some phonological rules for synthesis, which emulate the pronunciation rules of human beings (Klatt, 1987; O'Shaughnessy, 1987; Willems *et al.*, 1988; Fujisaki *et al.*, 1986; Fujisaki and Kawai, 1988; Lee *et al.*, 1989). Figure 1 shows a schematic diagram of this approach. First, a text is analyzed to extract some linguistic features relevant to $F0$ synthesis. Owing to their influence on the pronunciation of $F0$ contours in natural sentential utterances, a variety of linguistic features on different levels may be extracted. They include lexical information such as phonetic structure and accentuation of a word or syllable, syntactical structure representing intonation pattern or declination effect of sentence, semantic features, etc. Then, phonological rules for synthesis are used to generate the $F0$ contour. Traditionally, phonological rules are inferred from observing a large set of utterances with the help of linguists. Due to the fact that it is, in general, difficult to explore the effect of mutual interaction of linguistic features on different levels, most $F0$-by-rule algorithms consider the influence of each feature independently and then add them up ('t Hart and Cohen, 1973; Olive and Nakatani, 1974; Olive, 1975; Fujisaki *et al.*, 1986; Lee *et al.*, 1989). For example, in one approach to synthesizing English intonation ('t Hart and Cohen, 1973), two falling declination lines gradually converging toward the end of each sentential utterance were chosen to bound the $F0$ contour. Then, by identifying

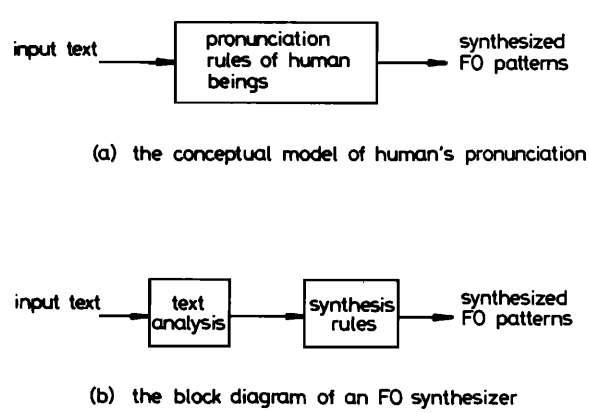

FIG. 1. The schematic diagram of a general $F0$ synthesizer to emulate human's pronunciation of $F0$.

 0001-4966/92/070114-07$00.80

stressed syllables and major phrasal boundaries in the input text, the $F0$ contour of a syllable is assigned to follow one of the declination lines, except during stressed syllables, where it switches levels, rising at the start of a phrase and falling at its end.

For Mandarin Chinese, due to the fact that it is a tonal language, $F0$ contours of sentential speech are usually more regularly pronounced so that the synthesis of fundamental frequency in Mandarin seems to be a much simpler task than that in a nontonal language such as English. To be more specific, some distinctive properties of Mandarin Chinese related to $F0$ synthesis are discussed in what follows. First, each Chinese character is pronounced as a syllable. It is therefore very natural to choose the syllable as the basic synthesis unit. Second, each syllable has a tone associated with it. Syllables with the same phonetic structure but different tones mostly have different lexical meanings. Third, there are only five lexical tones, namely, high-level, mid-rising, mid-falling-rising, high-falling, and neutral tones. For simplicity, they are commonly labeled in sequence from tone 1 to tone 5. Fourth, tones of syllables are mainly discriminated by their $F0$ contour shapes.

A previous study (Chao, 1968) concluded that the $F0$ contour of each of the first four tones can be simply represented by a single standard pattern, as shown in Fig. 2. 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. Fifth, although the so-called standard tone patterns do exist, $F0$ contours of syllable are subject to various modifications in continuous speech. They are primarily determined by the tones and phrasal conditions of syllables. They are also greatly affected by the intonation pattern or declination effect of sentence. Both the semantics and the emotional status of speaker may further change their shapes or levels. Therefore, $F0$ contour patterns cannot be simply synthesized by using the standard tone patterns only. Synthesis rules for tone pattern modification are still needed. Nevertheless, much simpler synthesis algorithm can be used for Mandarin speech by taking advantage of its simple tone structure (Lee *et al.*, 1989; Zhang, 1986; Yang and Xu, 1988). For example, in a rule-based approach (Lee *et al.*, 1989), basic $F0$ contour patterns were initially assigned to syllables in a sentence according to their tones. Then, several concatenation rules known as *sandhi rules* were employed to modify the pattern of each syllable according to the tones of the preceding and/or the following syllables. Some other modification rules using syntactic and semantic information were then applied to modify the $F0$ contour. Finally, stress rules for phrases and intonation pattern for declination effect were used to further modify the $F0$ contour patterns.

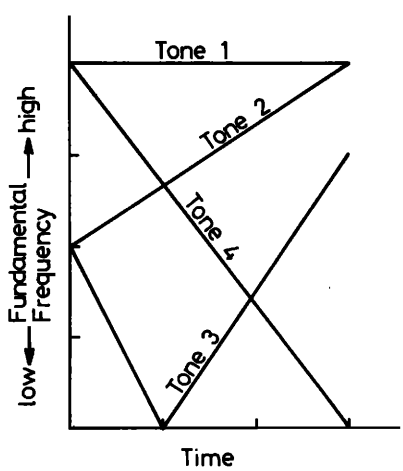

FIG. 2. Standard $F0$ contour patterns of tone 1–tone 4.

Although $F0$-by-rule approaches are reasonable and popular in Mandarin text-to-speech (Lee *et al.*, 1989; Zhang, 1986; Yang and Xu, 1988), several drawbacks can be found. First, phonological rules are always incomplete. The resulting synthetic speech is hence not very fluent. Moreover, the naturalness level of a synthesized speech may become unacceptable. Second, the effect of mutual interaction among linguistic features on different levels is largely ignored in the inference of phonological rules. Third, rules must be elaborately inferred by manually analyzing a large set of utterances with the help of linguists.

In this paper, a novel approach based on a statistical model is proposed for synthesizing $F0$ contour to generate natural Mandarin speeches. The basic idea is to substitute a statistical model for explicit synthesis rules. The statistical model is used to describe the relationship between $F0$ contour patterns of syllable and linguistic features representing the context. 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 memorized in the model. In the synthesis process, the best combination of $F0$ contour patterns is estimated on the basis of the model provided that linguistic features are given by analyzing the input text. The use of a statistical model in this approach can be regarded as providing a stochastic grammar for reducing the perplexity (Jelinek *et al.*, 1977) of syllable's $F0$ contour pattern. The number of candidates of $F0$ contour pattern at each decision point of synthesis can hence be greatly reduced.

Obviously, when more features are included in the model, we can expect better control of tone modification. However, due to the fact that the data available for training is very limited in this study, only some relevant linguistic features are extracted from the context of the processing syllable, for the purpose of demonstrating the feasibility of this approach. Improvement by incorporating syntactic and/or semantic information with a more sophisticated text analysis are left for future study.

The remainder of this paper explains the proposed method of $F0$ synthesis, examines its performance by simulations, and gives some conclusions based on experimental results.

I. THE PROPOSED F0 SYNTHESIZER

A. Problem statement

The problem of $F0$ synthesis in Mandarin text-to-speech, using syllables as the basic synthesis unit, can be

simply stated as finding the best sequence of $F0$ contour patterns, $C_1\ C_2 \cdots C_n$, given an input sentential text formed by syllable sequence, $W_1\ W_2 \cdots W_n$. By taking advantage of simple tone structure of Mandarin speech, $F0$ contours for syllables can be efficiently vector-quantized by using a codebook of proper size without causing much degradation in perception (Chen and Wang, 1990). In $F0$ synthesis, it is usually more convenient to construct one codebook for each tone. The problem can then be simplified as finding the best combination of $F0$ contour patterns (reproduction codewords) from codebooks of five tones. Statistically, we seek the pattern sequence with maximum probability, i.e.,

$$C_1^{T_1^*}\ C_2^{T_2^*} \cdots C_n^{T_n^*} = \arg \max_{\substack{C_i^{T_i} \in B_{T_i} \\ 1 \le i \le n}} p(C_1^{T_1}\ C_2^{T_2} \cdots C_n^{T_n} | W_1\ W_2 \cdots W_n), \tag{1}$$

where T_i is the tone syllable W_i, $C_i^{T_i}$ is a codeword in the codebook B_{T_i} of tone T_i, and $p(X \mid Y)$ is the conditional probability of X given Y.

To solve the problem using a joint conditional probability shown in Eq. (1) is impractical because training data will be insufficient given an almost unlimited combination of syllable sequences and variable lengths of sentential text. A simplified model must be adopted to make the problem mathematically tractable. Two directions of model simplification are followed in this paper. First, based on the assumption of short left-to-right coarticulation, a Markovian statistical model is adopted to decompose the joint conditional probability. It is worthwhile noting that the declination effect can be partially compensated by the Markov model. Then, by assuming that the $F0$ contour pattern of a syllable is mainly determined by syllabic context, the model is further simplified by substituting some relevant contextual features extracted from neighboring syllables. Although linguistic features on various levels can be incorporated into the model, due to the fact that it is in general difficult to analyze a natural Mandarin text syntactically or semantically, only three sequences of contextual features, tonality, stress, and continuity, were used in this paper.

To be more specific, a first-order Markov model,

$$p(C_i^{T_i} | C_{i-1}^{T_{i-1}}\ T_{i-1}\ T_i\ T_{i+1}\ G_i\ G_{i+1}\ A_i),$$

is employed to define the dependence of the $F0$ contour pattern of the ith syllable on that of the $(i\text{-}1)$th syllable as well as on several contextual features. Here, G_i is a bit used to indicate whether the ith syllable starts with voiced stop consonant or not, and A_i is a bit used to indicate whether the ith syllable is stressed or not. The selection of these contextual features is based on the following considerations:

(1) Right-to-left (R–L) coarticulation: The anticipation of an ensuing tone will make the articulator move toward a state appropriate for the ensuing tone (Lee *et al.*, 1989).

(2) Left-to-right (L–R) coarticulation: Some features of a tone may persist in ensuing tones (Lee *et al.*, 1989).

(3) Continuity property: When the ensuing syllable starts with a voiced, nonstop consonant, strong coarticulation will occur to make the $F0$ contours of the two successive syllables connected together and continuous across the boundary. This mainly results from the property of Mandarin Chinese that all syllables are ended with voiced phonemes (i.e., vowels or nasals).

(4) Stressed syllable: The $F0$ contour pattern of a stressed syllable is quite different from that of an unstressed one. In general, when a polysyllabic phrase is emphasized, the first syllable of the phrase is stressed.

Based on the simplified model, the problem of $F0$ synthesis can then be expressed as

$$C_1^{T_1^*}\ C_2^{T_2^*} \cdots C_n^{T_n^*} = \arg \max_{\substack{C_i^{T_i} \in B_{T_i} \\ 1 \le i \le n}} p(C_1^{T_1} | T_1\ T_2\ G_2\ A_1) \cdot p(C_2^{T_2} | C_1^{T_1}\ T_1\ T_2\ T_3\ G_2\ G_3\ A_2)$$
$$\cdots p(C_i^{T_i} | C_{i-1}^{T_{i-1}}\ T_{i-1}\ T_i\ T_{i+1}\ G_i\ G_{i+1}\ A_i) \cdots p(C_n^{T_n} | C_{n-1}^{T_{n-1}}\ T_{n-1}\ T_n\ G_n\ A_n). \tag{2}$$

The problem is further simplified by making an assumption of stationarity for conditional probabilities in Eq. (2). But, to partially compensate for the declination effect, two separate conditional probability density functions are added to the first and the last syllables of sentential text, respectively.

B. System description

The block diagram of the proposed $F0$ synthesizer is depicted in Fig. 3. It consists of two main parts: training and synthesis. In the training, a large set of sentential utterances with their associated texts are used to estimate parameters of the statistical model used. First, the acoustic signal of each

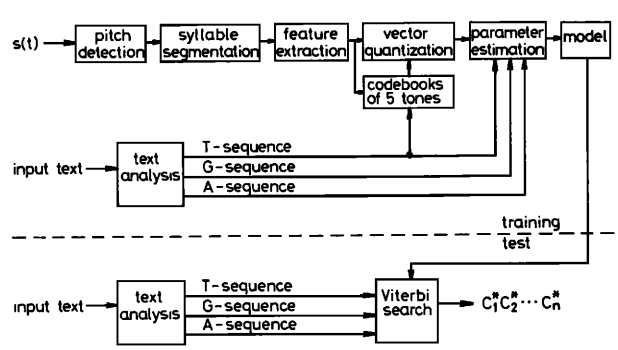

FIG. 3. The block diagram of the proposed $F0$ synthesizer.

sentential utterance is preprocessed to extract $F0$ contour patterns of syllables. The preprocessing includes four operations: pitch extraction, syllable segmentation, feature extraction, and vector quantization. In the preprocessing, the fundamental frequency of each frame is extracted by the cepstrum method (Noll, 1967). Then, the $F0$ contour is manually segmented into syllable periods. Each $F0$ contour segment is then orthogonally transformed to obtain a four-dimensional feature vector using the following four orthonormal polynomials (Chen and Wang, 1990):

$$\phi_0\left(\frac{i}{N}\right)=1, \tag{3a}$$

$$\phi_1\left(\frac{i}{N}\right)=\left(\frac{12N}{(N+2)}\right)^{1/2}\left[\left(\frac{i}{N}\right)-\frac{1}{2}\right], \tag{3b}$$

$$\phi_2\left(\frac{i}{N}\right)=\left(\frac{180N^3}{(N-1)(N+2)(N+3)}\right)^{1/2}\times\left[\left(\frac{i}{N}\right)^2-\left(\frac{i}{N}\right)+\frac{N-1}{6N}\right], \tag{3c}$$

$$\phi_3\left(\frac{i}{N}\right)=\left(\frac{2800N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}\right)\times\left[\left(\frac{i}{N}\right)^3-\frac{3}{2}\left(\frac{i}{N}\right)^2+\frac{6N^2-3N+2}{10N^2}\left(\frac{i}{N}\right)-\frac{(N-1)(N-2)}{20N^2}\right], \tag{3d}$$

for $0\leqslant i\leqslant N$, where $N+1$ is the length of the segment to expand. All feature vectors of $F0$ contour segments in an utterance are then vector-quantized to generate a $F0$ contour pattern sequence, $C_1^{T_1}\ C_2^{T_2}\ \cdots\ C_n^{T_n}$, by using codebooks of five tones (Chen and Wang, 1990). After preprocessing, the text associated with each utterance is analyzed to obtain three linguistic feature sequences: $T_1\ T_2\ \cdots\ T_n$ (T sequence), $G_2\ G_3\ \cdots\ G_n$ (G sequence), and $A_1\ A_2\ \cdots\ A_n$ (A sequence). Finally all these feature sequences are combined and parameters of the model are finally empirically estimated.

In the synthesis, the objective is to find the best sequence of $F0$ contour patterns for syllables using the well-trained model for a given input text. The input text is initially analyzed in the same way as in the training to extract the three linguistic feature sequences, i.e., tone (T), voiced nonstops (G), and stress (A) sequences. A Viterbi search is employed to find out the best sequence of $F0$ contour patterns with the highest probability conditioned on these three linguistic feature sequences. When a tie happens, a random decision with equal probability is made.

II. SIMULATIONS

Performance of the proposed $F0$ synthesizer was examined by simulations. The database used in our simulations consists of 9 repetitions of utterances of 112 phonetically balanced sentences (Yu and Liu, 1989) spoken by a single male speaker. Each sentence consists of about eight syllables. The speaking rate is about four–five syllables per second, a rate commonly used in conversational Mandarin. Recordings of these utterances were made in a computer room. Recordings of the first four repetitions occurred over several days; recordings of the last five repetitions were accomplished within 1 day.

Speech signals were low-pass filtered at 4-kHz bandwidth, sampled at 10 kHz, and A/D-converted into 16-bit data format. They were then pre-emphasized and segmented into 25.6-ms frames. Then, preprocessing, including pitch extraction, syllable segmentation, and orthogonal transform, was performed to extract a feature vector for each syllable. Feature vectors of the first three segments of utterances in the database were then used to train codebooks of five tones using the well-known LBG algorithm (Linde *et al.*, 1980). The size of the codebooks were determined by considering the following two factors: (1) the decreasing rate of the achievable mean-square error becomes saturated as the codebook size is increased; and (2) no apparent perceptual distortions suffered on the reconstructed sentential utterances generated by an LPC synthesizer with $F0$ information being replaced by the vector-quantized version. By trial and error, codebook sizes were set to 7, 11, 10, 18, and 5 for the five tones, respectively. Then, all $F0$ feature vectors in the training set were VQ quantized.

Texts associated with all utterances were also analyzed to extract linguistic features of T, G, and A sequences. Then, the statistical model was trained using all nine segments of utterances in the training set. Parameters of the model were empirically estimated based on relative frequencies of occurrence of $F0$ contour patterns. By closely examining the well-trained model, we found that the declination effect as well as some sandhi rules shown by Lee *et al.* (1989) were automatically inferred and implicitly included in the model. Fig. 4 displays the distributions of $F0$ contour patterns (codewords) for the first four tones in the three submodels used for the first syllable, the intermediate syllables, and the last syllable of sentential utterance. As shown in Fig. 4, the distribution of $F0$ contour pattern for each tone concentrates on low-order codewords in the submodel of the first syllable, and on high-order ones in that of the last syllable. Because all codewords of each tone are labeled in an increasing order according to descending $F0$ mean, we can therefore conclude that on an average, the $F0$ level of the first syllable is the highest and that of the last syllable is the lowest. This shows that the use of three separate submodels can partially compensate for the declination effect. A well-known sandhi rule for tone 3 is shown in Fig. 5. Figure 5(a) depicts the distribution of $F0$ contour pattern of a tone 3 followed by another tone 3. The two most frequently occurring patterns, codewords 2 and 6, are displayed in Fig. 5(b). Due to the fact that their shapes strongly resemble the standard mid-rising pattern of tone 2 shown in Fig. 2, the sandhi rule that a tone 3 preceding another tone 3 will be changed to a tone 2 is confirmed to be inferred and stored in the model. Some other sandhi rules, such as the decreases in the $F0$ levels of tone 3 and tone 1 following a tone 4, have also been found by examining the corresponding conditional distributions of $F0$ contour patterns.

The performance of synthesizing $F0$ contour patterns using the well-trained model was also examined. A coincidence rate or hit rate, defined as the percentage of the synthe-

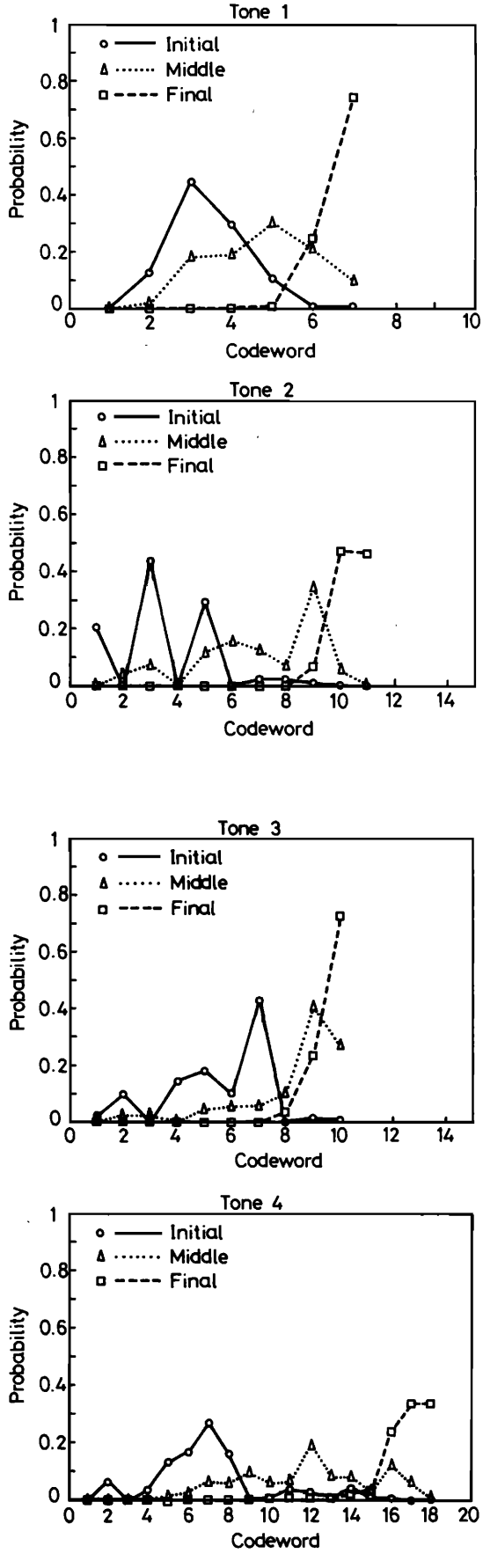

FIG. 4. Distributions of tone 1–tone 4 in the three submodels used, respectively, for the first syllable, intermediate syllables, and the last syllable of sentence.

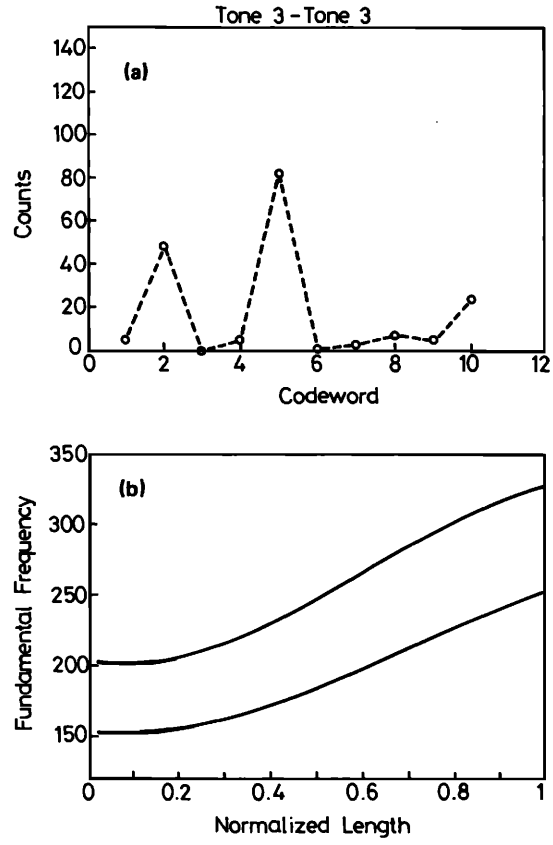

FIG. 5. (a) The statistics of $F0$ contour pattern of tone 3 followed by a tone 3, and (b) the orthogonally expanded contours of the two most frequently occurring patterns.

sized $F0$ contour patterns (codewords) of syllables that coincide with their VQ-quantized counterparts in original utterances, is employed to quantitatively evaluate the performance of the system. Experimental results are listed in Table I. As shown in Table I, the performance of the synthesizer gradually improved as more linguistic features were incorporated into the statistical model. A coincidence rate of 77.56% was achieved when all three linguistic feature sequences were used. The results in Table I also show that the G sequence is slightly more effective than the A sequence.

The consistency (or invariability) of pronouncing $F0$ contour patterns in the training database was also computed as a reference for performance evaluation. It is defined as

TABLE I. Performance of $F0$ synthesis using various linguistic features.

Linguistic features	Coincidence rate
T	54.25%
$T\&A$	58.61%
$T\&G$	74.14%
$TG\&A$	77.56%
Pronunciation consistency of database	82.96%

follows. First, for each syllable in each sentence, the most frequently occurring $F0$ contour pattern (codeword) in the nine segments of utterances is taken as the standard pattern. Comparing with the standard patterns, the coincidence rate of $F0$ contour patterns of the training set was calculated and taken as a measure of the consistency of pronunciation. Due to the fact that it measures the inherent invariability of human pronunciation disregarding the context, it is taken as an upper bound of performance that a synthesizer can reach. Pronunciation consistency was 82.96% for the training database. As a measure of system performance, the coincidence rate of 77.56% compares very favorably with pronunciation consistency.

The performance of the synthesizer can also be evaluated on the basis of the perplexity (Jelinek *et al.*, 1977) of the training database. Perplexity is a quantitative measure of the variability of $F0$ contour patterns in the training corpus as the model is given. Table II lists the perplexities and the corresponding coincidence rates for models using various linguistic features. As shown in Table II, the perplexity decreases when more linguistic features were incorporated into the model. And the performance of the synthesizer, as expected, becomes better as the perplexity decreases. Note that the coincidence rate decreases to 18.1% when only the tone of the processing syllable is used for $F0$ synthesis. This confirms the effectiveness of the present approach.

Figures 6 and 7 displays two typical synthesized $F0$ contours for sentences. As seen in these figures, both synthesized contours strongly resemble their original counterparts. Some distortions appear at the boundaries of two connected $F0$ contours of syllable, but they are not serious because the jumps at these discontinuous boundaries are small. By comparing all synthesized $F0$ contours with their original counterparts, we found that, at some syntactic boundaries, the change of a synthesized $F0$ contour pattern cannot follow the abrupt jump of $F0$ level in the original utterance. This is mainly the consequence of ignoring both the syntactic and the semantic information in this study.

Finally, an LPC synthesizer was implemented for the purpose of informally evaluating the quality of the synthesized speech. By replacing the original $F0$ contours with the synthesized $F0$ contour patterns, most synthetic sentences generated by the LPC synthesizer sounded clear and natural. Only few suffered from distortions of unnaturalness. They mainly resulted from the above-mentioned defect of synthesis occurred at syntactic boundaries. Nevertheless, they were all still highly intelligible.

TABLE II. Perplexities of training corpus evaluated based on statistical models using various linguistic features.

Linguistic features	Perplexity	Coincidence rate
T	2.37	54.26%
T&A	2.23	58.61%
T&G	1.87	74.14%
TG&A	1.78	77.56%
Without model	5.80	18.1 %

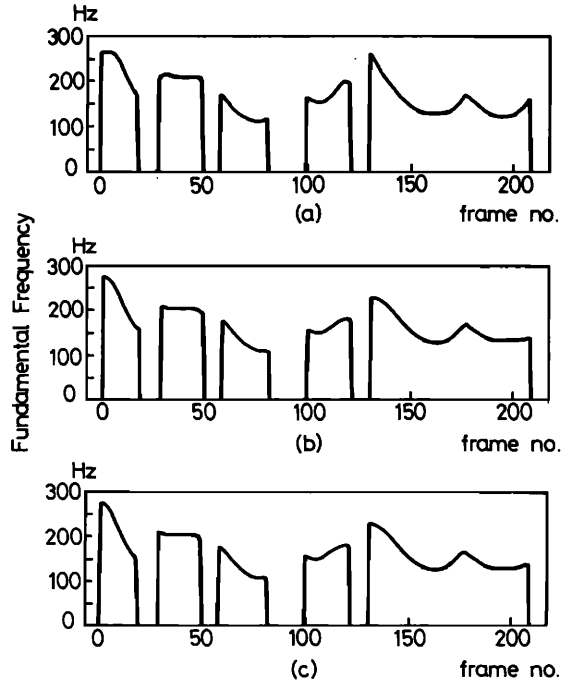

FIG. 6. A typical example: (a) the original, (b) the vector-quantized, and (c) the synthesized $F0$ contours of "TZ4 SHIN1 JE3 TSUNG2 SHAN4 RU2 LIOU2."

III. CONCLUSIONS

A novel statistical model approach to $F0$ synthesis for generating natural Mandarin speech was presented in this

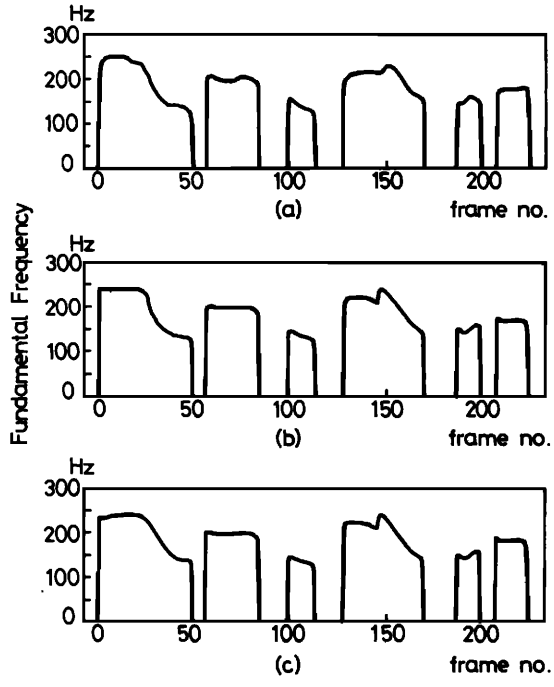

FIG. 7. Another example: (a) the original, (b) the vector-quantized, and (c) the synthesized $F0$ contours of "TSA1 LENG3 SHUNG1 KE3 TZ1 RUEN4 PI2 FU1."

paper. Its effectiveness has been verified by simulations using a set of utterances of declarative sentences. The resulting coincidence rate of the synthesized $F0$ contour pattern is high compared to the pronunciation consistency of the training corpus. Naturalness of most synthesized utterances was confirmed by an informal listening test.

Although only some relevant contextual features were involved in the current study, the performance of the system is reasonably good. The extension of the system to incorporate more linguistic features into the statistical model may be easily accomplished. Improvement of the system by incorporating additional syntactical and/or semantical information with a more sophisticated text analysis is worth further studying.

Another way to improve the quality of the synthesized speech is to increase the codebook size for the five tones. But, this would increase the complexity of the model and make the size of inventory required to reliably train the model increase. A promising way to solve this problem is to split VQ for mean and shape of $F0$ contour pattern. Besides, they can be separately synthesized using different linguistic features. Linguistic features on higher levels, such as syntactic and semantic information, can be incorporated in the synthesis of $F0$ mean to emulate the intonation of sentences. On the other hand, contextual features extracted from neighboring syllables can be used to synthesize $F0$ shape for simulating sandhi rules.

A preliminary outside test using ten sentences was performed. The coincidence rate decreased to about 40%. By closely examining the synthesis process, it was found that the total probabilities obtained in Viterbi search were drastically lowered for most sentences. This shows that the constituents of context between the training and the outside-test sentences are inconsistent and hence result in a low coincidence rate in the outside test. We can therefore conclude that the model currently trained is still not suitable for applying to the case of untrained text. Enlarging the training inventory to include more constituents of context is a straight forward way to extend the method to the case of unlimited text, and is hence worthwhile for future study.

ACKNOWLEDGMENTS

This study was supported by the National Science Council, Republic of China.

Chao, Y. R. (**1968**). *A Grammar of Spoken Chinese* (University of California, Berkeley Press, Berkeley, CA).

Chen, S. H., and Wang, Y. R. (**1990**). "Vector quantization of pitch information in Mandarin speech," IEEE Trans. Commun. **COM-38**, 1317–1320.

Fujisaki, H., Hirose, K., Takahashi, N., and Morikawa, H. (**1986**). "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, 2039–2042.

Fujisaki, H., and Kawai, H. (**1988**). "Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese," IEEE Int. Conf. Acoust. Speech Signal Process. S, 663–666.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. (**1977**). "Perplexity—a measure of difficulty of speech recognition tasks," J. Acoust. Soc. Am. Suppl. 1 **62**, S63.

Klatt, D. H. (**1987**). "Review of text-to-speech conversion for English," J. Acoust. Soc. Am. **82**, 737–793.

Lee, L. S., Tseng, C. Y., and Ouh-Young, M. (**1989**). "The synthesis rules in a Chinese text-to-speech system," IEEE Trans. Acoust. Speech Signal Process. **ASSP-37**, 1309–1319.

Linde, Y., Buzo, A., and Gray, R. M. (**1980**). "An algorithm for vector quantization design," IEEE Trans. Commun. **COM-28**, 84–95.

Noll, A. M. (**1967**). "Cepstrum pitch detection," J. Acoust. Soc. Am. **41**, 293–309.

Olive, J. P., and Nakatani, L. H. (**1974**). "Rule-synthesis of speech by word concatenation: a first step," J. Acoust. Soc. Am. **55**, 660–666.

Olive, J. P. (**1975**). "Fundamental frequency rules for the synthesis of simple declarative English sentences," J. Acoust. Soc. Am. **57**, 476–482.

O'Shaughnessy, D. (**1987**). *Speech Communication* (Addison-Wesley, New York), Chap. 9.

't Hart, J., and Cohen, A. (**1973**). "Intonation by rule: a perceptual guest," J. Phonet. **1**, 309–327.

Willems, N., Collier, R., and 't Hart, J. (**1988**). "A synthesis scheme for British English intonation," J. Acoust. Soc. Am. **84**, 1250–1261.

Yang, S., and Xu, Y. (**1988**). "An acoustic-phonetic oriented system for synthesizing Chinese," Speech Commun. **7**, 317–325.

Yu, S. M., and Lin, C. S. (**1989**). "The construction of phonetically balanced Chinese sentences," Telecommunication Laboratory Tech. Report, Taiwan, R.O.C.

Zhang, J. (**1986**). "Acoustic parameters and phonological rules of a text-to-speech system for Chinese," IEEE Int. Conf. Acoust. Speech Signal Process, 2023–2026.