A Set of Corpus-Based Text-to-Speech Synthesis Technologies for Mandarin Chinese

Fu-Chiang Chou, Chiu-Yu Tseng, and Lin-Shan Lee, *Fellow, IEEE*

***Abstract*—This paper presents a set of corpus-based text-to-speech synthesis technologies for Mandarin Chinese. A large speech corpus produced by a single speaker is used, and the speech output is synthesized from waveform units of variable lengths, with desired linguistic properties, retrieved from this corpus. Detailed methodologies were developed for designing "phonetically rich" and "prosodically rich" corpora by automatically selecting sentences from a large text corpus to include as many desired phonetic combinations and prosodic features as possible. Automatic phonetic labeling with iterative correction rules and automatic prosodic labeling with a multi-pass top-down procedure were also developed such that the labeling process for the corpora can be completely automatic. Hierarchical prosodic structure for an arbitrary desired text sentence is then generated based on the identification of different levels of break indices, and the prosodic feature sets and appropriate waveform units are finally selected and retrieved from the corpus, modified if necessary, and concatenated to produce the output speech. The special structure of Mandarin Chinese has been carefully considered in all these technologies, and preliminary assessments indicated very encouraging synthesized speech quality.**

***Index Terms*—Automatic labeling, Mandarin Chinese, prosody, synthesis, text-to-speech.**

I. Introduction

TEXT-TO-SPEECH (TTS) synthesis technology for converting an arbitrary text into corresponding speech signals has been successfully developed for a long time. Significant improvements in this area have been observed in the past decades. In recent years, the overwhelming developments of Internet services as well as wireless personal communications have created a completely new environment for TTS applications. For example, it is highly desirable for people to listen to e-mails or Web pages read by TTS technology over mobile handsets at any time, from anywhere, if the TTS synthesized speech quality is good enough. In fact, the rapidly increasing demand for TTS technology also leads to higher requirements for the intelligibility and naturalness for TTS synthesized speech. When the contents of the texts to be read by TTS technology are of high importance (e.g., personal e-mails) or high diversity (e.g., Web pages) and the synthesized speech is to be listened to by large number of users, it is often found that the capabilities of the available technology are still limited. The ultimate goal of true naturalness of synthesized speech seems not easy to achieve today, specially for the case of general domain applications, although very often it is believed that this goal is already very close [1].

Among the many developments in TTS technology improvements, the new paradigm of "corpus-based approaches" is clearly important. In this paradigm, the synthesized speech is not obtained by concatenating modified versions of voice units pre-stored in a database. Instead, a large speech corpus (on the order of 10 h, or even much more, of speech) produced by a single speaker is collected. The corpus is designed so that almost all linguistic and prosodic features for the target language (either for general domain or specific domain) have been included. Parallel analysis of all prosodic and linguistic features of the speech signals as well as the corresponding texts can lead to a much better prosodic model. There can be many repetitions of a given voice unit in the corpus, but in different context with different prosodic features. During the synthesis process, the most appropriate units of variable lengths, with the desired prosodic features within the corpus, are automatically retrieved and selected on-line in real-time, and concatenated (with modifications when necessary) to produce the output speech. By doing this, very often longer units (especially commonly used words or even phrases) can be used in the synthesis if they appear in the corpus with desired prosodic features. Also, by doing this the need for signal modification to obtain the desired prosodic features for a voice unit, which usually degrades the naturalness of the synthesized speech, is significantly reduced. This is why much better performance can be achieved using this approach [2]–[4].

For TTS technology for Mandarin Chinese, great efforts have been made in the past decades as well, and many successful systems have been developed [5]. Considering first the text analysis part, one early system [6] simply synthesized the speech directly from phonetic transcription input without handling the text analysis. The earliest system including text analysis functionality may have appeared around 1989 [7]. By the mid-1990s there were many systems that could read Chinese text directly [8]–[11]. The major problems in text analysis include word segmentation (because there are no blanks between words in Chinese texts serving as word boundaries) and part-of-speech (POS) tagging. In most cases simple heuristic rules or statistical models were used for these two tasks, although some special algorithms were also applied [9], [12].

As for prosodic models in TTS for Mandarin Chinese, both rule-based [6], [13] and statistics-based [14], [15] approaches

Manuscript received April 3, 2001; revised May 9, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Juergen Schroeter.

F.-C. Chou is with Philips Speech Processing, Voice Control, Taipei, Taiwan, R.O.C. (e-mail: fuchiang.chou@philips.com).

C.-Y. Tseng is with the Institute of Linguistics, Academia Sinica, Taipei, Taiwan, R.O.C. (e-mail: cytling@gate.sinica.edu.tw).

L.-S. Lee is with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. (e-mail: lsl@iis.sinica.edu.tw).

Digital Object Identifier 10.1109/TSA.2002.803437

have been useful. The former needs expert knowledge and sophisticated handcrafted efforts, while the latter can achieve similar results by training with appropriately labeled speech corpora. Because both the available quantities and processing capabilities for speech corpora have been growing rapidly since 1990s, statistics-based approaches have been getting more and more attractive and popular in recent years. However, the lack of a good prosodic model for Mandarin Chinese remains a major difficulty. Although some important results have been obtained, primarily focused on the tone Sandhi and related problems [16], [17], the need for a useful prosodic labeling system for Mandarin Chinese like ToBI [18] for English is quite clear. Because Mandarin Chinese is a tonal language for which the tone has lexical meaning, the prosody is specially important for TTS synthesis, not only for the naturalness, but for intelligibility as well.

Regarding the choice of the voice units for synthesis in TTS for Mandarin Chinese, the syllable was popularly used starting in early years [6] due to the monosyllabic structure of Mandarin Chinese, i.e., each Chinese character is pronounced as a monosyllable, while a Chinese word is composed of one to several characters (or syllables). Such syllable-sized units can perfectly model the intra-syllabic coarticulation, but not that across syllabic boundaries. The di-phone units commonly used in most western languages were also found useful [9]. The corpus-based approaches mentioned above have led to many methods for selection of appropriate voice units from large speech corpora [19]–[21]. The essential points for these methods are nonuniform units, multiple candidates, and on-line selection. But as yet such methods have only been used by very few systems for Mandarin Chinese [22]. Considering basic synthesis methods in TTS for Mandarin Chinese, both the vocal tract model approaches [9], [23] and waveform-based approaches [8], [11], [24] have been used for some time. Waveform-based approaches produced more intelligible speech if only minor prosodic modifications and spectral transitions were needed. The vocal tract model, on the other hand, allowed more significant prosodic modifications and smoother spectral transitions, but produced slightly less intelligible speech.

In this paper, a new set of text-to-speech synthesis technologies for Mandarin Chinese is presented. Detailed methodologies were developed for designing "phonetically rich" and "prosodically rich" corpora by automatically selecting sentences from a large text corpus to include as many desired phonetic combinations and prosodic features as possible. Automatic phonetic labeling with iterative correction rules and automatic prosodic labeling with a multi-pass top-down procedure were also developed such that the labeling process for the corpora can be completely automatic. Hierarchical prosodic structure for an arbitrary desired text sentence is then generated based on the identification of different levels of break indices, and the prosodic feature sets and appropriate waveform units are finally selected and retrieved from the corpus, modified if necessary, and concatenated to produce the output speech. The special structure of Mandarin Chinese has been carefully considered in all these technologies, and preliminary assessments indicated very encouraging synthesized speech quality. In the following, the corpus design is presented in Section II, and the automatic phonetic and prosodic labeling for the speech corpus in Sections III and IV, respectively. Automatic generation of the prosodic structure for an arbitrary input text sentence is then described in Section V, while selection, modification and concatenation of the waveform units to produce the output speech is discussed in Section VI. The prototype system and preliminary performance assessments are finally given in Section VII. Section VIII is the concluding remarks.

TABLE I
(a) 21 INTIALs OF MANDARIN SYLLABLES AND
(b) 40 FINALs OF MANDARIN SYLLABLES

(a) 21 INTIAL's of Mandarin Syllables
p, p^h, m, f, t,t^h, n, l, k, k^h,x
tɕ, $tɕ^h$, ɕ, tʂ,$tʂ^h$, ʂ, ʐ, ts, ts^h,s
(b) 40 FINAL's of Mandarin Syllables
a, ai, au, an, aŋ,o, ou, ə, ən, əŋ
ɚ, i, ia, iɛ, iai,iau, iou, iɛn, in, iaŋ
iŋ, io, u, ua, uo,uai, uɛi, uan, un, uaŋ
uŋ, uoŋ, y, yɛ, yɛn,yn, ɛ, ɛi, ɨ, ɯ

II. Corpus Design and Labeling Systems

Good corpus-based TTS technology relies on the availability of a good corpus which carries all desired phonetic as well as prosodic features for the target language and the target task, such that good prosodic models, linguistic properties and synthesis units can be derived. If the size of the corpus could be infinite, no corpus design would ever be needed since everything could be included. However, when only a very limited size of corpus is achievable, careful design of the corpus becomes very important. In this research, two corpora were developed, one "phonetically rich" and one "prosodically rich." Both of them have texts selected by some automatic algorithms but with different selection criteria from the Academia Sinica Balanced Corpus [25], which is a "balanced" (in the sense of topic domains, styles, genres, media sources, etc.) Chinese text corpus.

The purpose of the "phonetically rich corpus" is to include as many phonetic combinations, including intra-syllabic and inter-syllabic structures, as possible in a corpus of acceptable size. Chinese syllables are conventionally described in INITIAL/FINAL format very similar to the CV structure for syllables in other languages. Here INITIAL is the initial consonant of a syllable, and FINAL is the vowel (or diphthong) part plus optional medials and nasal endings. The total number of INITIALs and FINALs are, respectively, 21 and 40. The 21 INITIALs and 40 FINALs for Mandarin syllables are listed in Table I(a) and (b), respectively, in International Phonetic Alphabet (IPA). More detailed discussions about these INITIALs and FINALs as well as Mandarin syllables are referred to earlier literature [26], [27]. The phonetically rich corpus thus should include all possible syllabic INITIAL–FINAL structures considering phonotactic constraints, and as many FINAL–INITIAL or FINAL–FINAL (for the case that the second syllable does not have an INITIAL) combinations as possible for cross-syllabic features. In addition, Mandarin Chinese is a tonal language and each syllable is assigned a tone. There are a total of four lexical tones plus a neutral tone. It is therefore desired to have all possible syllable-tone combinations, plus all possible tone concatenation combinations including the tone Sandhi variations. All these criteria were entered into a word/sentence selection algorithm [28] to be performed over the Academia Sinica Balanced Corpus. The result is a word database and a paragraph database. The former consists of the 1455 most frequently used 1-, 2-, 3-, and 4-syllabic lexical items covering 338 tonal combination patterns and 1351 inter-syllabic combinations, and the latter consists of 400 paragraphs made up of frequently used words covering 289 tonal combination patterns and 1434 inter-syllabic combinations. Not all desired combination patterns are present in these two databases, but it is believed that those which are absent should be rarely used, because they apparently did not appear in the Academia Sinica Balanced Corpus.

The purpose of the "prosodically rich corpus," on the other hand, is to include more prosodic behavior of Mandarin Chinese which may not be well covered by the above "phonetically rich corpus." This is much more challenging because the prosody of Mandarin Chinese has not yet been well studied. Four forms of intonation [29] or 13 types of intonation based on different speaker attitudes [30] were developed in the early literature, while later studies analyzed the intonation in Mandarin speech according to three modalities of utterances, i.e., interrogative, exclamatory, and declarative [31]. The design of the "prosodically rich corpus" was based on these three modalities of intonation [32]. The interrogative sentences were classified into four groups and the exclamatory sentences into three groups [33], both based on the existence of some words or word patterns in the sentences. So sentences of these modalities can be selected from the Academia Sinica Balanced Corpus using punctuation marks plus these words or word patterns. For example, at the time this research was performed, a total of 280 000 sentences in the Academia Sinica Balanced Corpus were well tagged with manual correction. The tags include 44 parts-of-speech and ten punctuation marks. Out of the 280 000 sentences, 8350 ending with question marks and 5471 with exclamation marks were first automatically extracted as candidates for interrogative sentences and exclamatory sentences. Further selection was then performed with these candidates based on specific words or word patterns. The result is a set of 550 interrogative sentences and 300 exclamatory sentences to be used in the "prosodically rich" corpus.

The selection of declarative sentences, on the other hand, was much more difficult, since no prior studies can be found in the literature. In this research, it was assumed that the prosodic structures of declarative sentences may have to do with the concatenation patterns of parts-of-speech and punctuation marks, which have been well tagged on the Academia Sinica Balanced Corpus. In that sense the bigram/trigram coverage for concatenation patterns of parts-of-speech and punctuation marks may be a reference parameter for selection. Therefore the declarative sentences were selected by an algorithm trying to maximize such bigram/trigram coverage. The selection algorithm is described here. The score for each bigram/trigram item is defined proportional to the inverse of its count in the corpus. In otherwords, the less frequently a bigram/trigram item appears in the corpus, the higher priority it has to be selected. In this way the bigram/trigram coverage can be maximized with minimum number of sentences. If the count is too small (less than a chosen threshold), then the score of that bigram/trigram item is set to zero. This is to avoid including unreliable items. The score of a sentence is then the sum of the scores of all bigram/trigram items in the sentence, normalized to the number of characters. The sentence with the highest score in the corpus was then selected first automatically. After a sentence is selected, the scores of all the bigram/trigram items in the selected sentence are automatically set to zero, i.e., these bigram/trigram items are not desirable any more. The scores of all sentences are then recalculated and the sentence with the highest score is selected. This is again the way to maximize the bigram/trigram coverage with minimum number of sentences. This process is repeated iteratively. The result is a set of 800 declarative sentences. These 800 sentences plus the 300 exclamatory sentences and 550 interrogative sentences selected based on words or word patterns as mentioned in the above together formed the sentence set for the "prosodically rich" corpus. Although there does not exist any direct proof that the sentences selected in this way really cover the prosody of Mandarin Chinese reasonably well, at least it was observed that the sentences in this database really include many different prosodic patterns.

While recording the speech corpora, for each of the "phonetically rich" and "prosodically rich" databases six speakers were asked to produce the speech in read speech mode, but as naturally as possible. The six speakers for either database include three males and three females in three age groups: 20–35, 35–50, and 50 and above. The six speakers producing the speech for the two databases were completely different, except one speaker in common. So there were 11 speakers in total, and there was one speaker who produced both the "phonetically rich" and "prosodically rich" corpora. The "phonetically rich corpus" includes a total of more than 18 h of speech, while the "prosodically rich corpus" includes more than 31 h of speech. The speech corpora of a total of roughly 50 h of speech here produced by 11 speakers were for various research purposes. For the research of corpus-based TTS synthesis technologies to be presented below, only the data produced by the single speaker who produced both the "phonetically rich" and "prosodically rich" corpora were used for consistency in prosodic and phonetic properties. So all the experimental data described below are based on the speech of this speaker of roughly 8 h, which is also the synthesis inventory used in the final TTS system. This speaker is male, who is a teacher in a university at Taipei, whose job is to teach students to speak in accurate Mandarin Chinese.

The collected speech corpora need a good phonemic transcription system and a good prosodic labeling system. A SAMPA-T phonemic transcription system following the general design principles of the SAMPA system [34] but considering the phonetic characteristics of Mandarin Chinese as well as a few local dialects spoken in Taiwan was developed [35]. In addition, a prosodic labeling system following the general design principles of the ToBI system [18] but considering the prosodic characteristics of Mandarin Chinese was developed, in which the prosodic features are represented by break indices and emphasis levels at the moment [36]. The break indices are marked at the end of each syllable with a value from 0 to 5 to characterize the following six categories of boundaries:

B0: reduced syllabic boundary;
B1: normal syllabic boundary;
B2: minor phrase boundary;
B3: major phrase boundary;
B4: breath group boundary;
B5: prosodic group boundary.

The emphasis levels range from zero to three. The break indices were extensively used in the following TTS technology, while the emphasis levels have not been used as yet.

III. Automatic Phonetic Labeling for the Speech Corpora

Two levels of labels for the speech corpora are needed for TTS synthesis, phonetic labels and prosodic labels. Both levels of labeling were made completely automatic in the proposed technologies [37]. In other words, no human corrections were needed at all. As will be clear later on, although some errors may be inevitable in the completely automatic processes, the final tests showed such errors are in fact negligible or acceptable, and the completely automatic labeling processes are actually adequate to a good extent. Automatic labeling is definitely needed not only because the quantity of the data is huge and manual labeling is simply impossible, but because automatic labeling is the only approach to achieve consistency in labeling, which is the key for good synthesis results. Automatic labeling also makes it possible to synthesize the voice of any new speaker very quickly, as long as the corpora for the desired new speaker are available. In this section automatic phonetic labeling is presented, and automatic prosodic labeling is to be described in the next section.

For automatic phonetic labeling, a phonemic transcription of the speech data is used as the input. The possible pronunciations of the words in each sentence were derived using a text analysis module (because the pronunciation may depend on the context), including establishing a set of possible pronunciations for each word, and converting the results into HTKs [39] net format. These net files include the homographs and the pronunciation variations. A Viterbi process is then performed to recognize and align the speech data based on the net. The units for the hidden Markov models (HMMs) used here are context independent INITIALs and FINALs. The INITIAL models have three states each and the FINAL models have five states each. The feature vectors include 12 dimensions of Mel-frequency cepstral coefficients (MFCC), one dimension of root mean square (RMS) power and their differential values. The frame rate is 5 ms for better alignment precision. In a conventional phonetic labeling procedure as illustrated in the left part of Fig. 1, the speaker independent (SI) HMMs were used to perform a rough alignment for the initial training of the speaker dependent (SD) HMMs. The parameters of the SD HMMs can then be further re-estimated with an embedded Forward–Backward algorithm. The Viterbi process then produces the final alignment with these final SD HMMs.

The results for the above conventional approach were not accurate enough for the Mandarin speech data used here, at least for the purposes of TTS synthesis. During the manual correction of the alignment results, it was found that most of the errors can be classified and corrected with some phonetic rules. An algorithm for postprocessing the output label files with such rules was therefore developed [38]. The adjusted results were then applied to re-estimate the parameters of the SD HMMs. This procedure was repeated iteratively to fine tune the models. The block diagram is illustrated in the right part of Fig. 1, in which the re-estimation process in the left part of the figure is replaced by an iterative procedure including Viterbi alignment, correction and training. The input of the block diagram is the speech signal and its transcription net file; while the output is the INITIAL/FINAL sequences with the corresponding boundary positions.

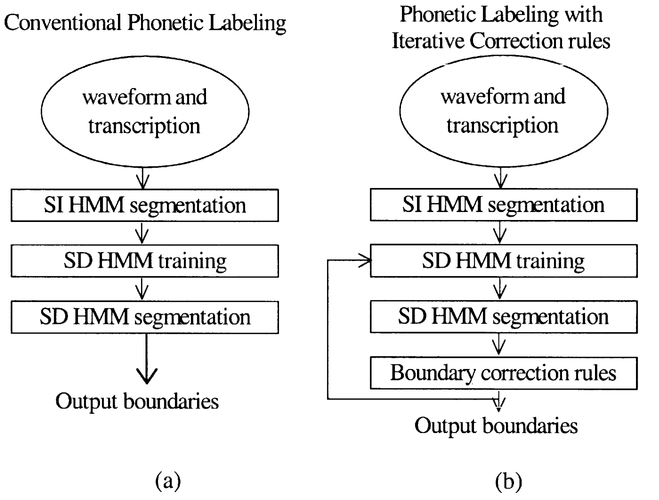
Conventional Phonetic Labeling
waveform and transcription
SI HMM segmentation
SD HMM training
SD HMM segmentation
Output boundaries
(a)
Phonetic Labeling with Iterative Correction rules
waveform and transcription
SI HMM segmentation
SD HMM training
SD HMM segmentation
Boundary correction rules
Output boundaries
(b)

Fig. 1. Automatic phonetic labeling with iterative correction rules. (a) Conventional approach and (b) the approach used here.

The correction rules mentioned above were based on a set of acoustic features and the phonetic classes of the voice segments in the local vicinity of the Viterbi alignment boundaries. The acoustic features used here include RMS power, voicing probability and subband energies derived from FFT. The window size used for obtaining these features varied from 5 ms to 20 ms for different features and different phonetic classes for the voice segments. The voice segments were categorized into seven phonetic classes: silence, nasal, liquid, fricative, plosive, affricate and vowel. Different rules were developed for different concatenations of these phonetic classes, such as: (nasal) + (vowel), (vowel) + (fricative), etc. Two examples which have been shown to be the most successful correction rules are described below. The first example is (silence) + (affricate). The RMS power with 5 ms window size was applied to locate the affricate. Because there is a short burst of energy when the sound is released, the boundary can be more precisely decided with the sharp increase of the RMS power. The second example is (vowel) + (fricative). The voicing probabilities derived by the ESPS tool were used here to locate the boundary more precisely. In the original alignment, the ending of the vowel was very often taken as a part of the following fricative. With the correction rule, the boundary can be shifted to the right position. These two example correction rules are very useful. There are some other rules in addition to these two, but there do not necessarily exist correction rules for all the combinations of

TABLE II
TEST RESULTS FOR THE AUTOMATIC PHONETIC LABELING PROCESS AND THE HUMAN LABELING RELIABILITY

		Conventional without Correction Rules	Including Iterative Correction Rules	Difference between Two Human Labelers
Mean Boundary Error		14.2 ms	8.3 ms	1.6ms
Alignment Accuracy	<10ms	66.3 %	78.4 %	94.6%
	<20ms	91.2 %	96.5 %	100%

phonetic classes. For some combinations the Viterbi alignment boundaries are already quite accurate and no more corrections are needed.

For evaluation of the performance of the labeling process, a small set of speech data produced by the single speaker mentioned previously was tested. It included 130 min (500 paragraphs) of speech for training and 30 min (100 paragraphs) of speech for testing. A set of manually labeled data generated by a single human labeler was used as the reference. The boundary errors are defined as the difference between the aligned boundaries and the reference boundaries. The alignment accuracy is then defined as the percentage of boundary errors within 10 ms and 20 ms. The evaluation results are listed in the first two columns of Table II. It can be found that without the boundary correction rules, the mean boundary error was 14.2 ms, and the alignment accuracies were 66.3% and 91.2% within 10 ms and 20 ms, respectively. By retraining the HMMs with the boundary correction rules, the mean boundary error was reduced to 8.3 ms, and the alignment accuracies within 10 ms and 20 ms improved to 78.4% and 96.5%, respectively. The results here are those for all the iterative correction rules being terminated after five iterations. This number of five iterations was determined empirically, simply based on the observation that no further improvements as compared to those in Table II could be obtained if any individual correction rule was iterated more times while all other rules were terminated after five iterations. In the test no classification errors were assumed, and the classification consistency was not even checked. This is because this is basically an alignment problem and it is reasonable to consider all the given phonetic transcriptions to be correct. Another good question is the reliability of the reference, or whether the labels generated by the single human labeler is accurate enough. A second human labeler was asked to do the same labels for the 30 min of testing speech. Comparison between the labels generated by the two human labelers gave the data in the right column of Table II. There apparently existed some nonzero mean boundary errors, but significantly smaller than that produced by the automatic labeling processes. Therefore the alignment accuracy of human labelers was apparently not perfect. More test results regarding how the overall synthesized speech quality is dependent on the techniques discussed here will be presented later on in Section VII.

IV. AUTOMATIC PROSODIC LABELING FOR THE SPEECH CORPORA

In this research, only break indices were automatically labeled at the end of each syllable. A feature vector p_i is generated at the end of each syllable, and a break index b_i (ranging from B0 to B5 as mentioned previously) should be labeled accordingly. Therefore the task here is to map a sequence of prosodic feature vectors $(p_1, p_2 \cdots p_n)$ for an utterance of n syllables into a sequence of prosodic labels $(b_1, b_2, \ldots b_n)$ [37], [40]. The feature vectors used here will be explained below. This process has been performed in earlier studies [41] with a hidden Markov model based on a simplified assumption that the current label is dependent only on the previous label. However, according to our experiences of manual prosodic labeling of Mandarin speech, it seemed that such dependency is more on the upper level unit rather than on the previous label, i.e., labeling of B4 (for a breath group) is more dependent on the locations of B5 indices (for prosodic groups), labeling of B3 more dependent on B4 indices, etc. Because of such possible hierarchical structure of prosody in Mandarin Chinese, we chose to use a multiple-pass top-down procedure for labeling of break indices. The algorithm includes two principal components. The feature extraction component transforms the various sources of information (the phonetic transcription, pitch-extraction results, etc.) into a time-ordered sequence of feature vectors $(p_1, p_2, \ldots, p_n)$. The feature vectors are then classified by decision trees in a multiple-pass hierarchical phrasing component. The multiple-pass procedure is simple. We only identify the location of one level of break indices each time and the sequence is from B5 to B1, i.e., first locating B5, then B4, then B3, etc. This procedure is illustrated in Fig. 2.

The features used to construct the feature vectors p_i and determine the break indices are listed in Table III. The first feature is the pause duration, if any. The next three classes of features are then derived from basic acoustic features: duration, energy, and F_0. All these parameters were normalized and the z-score values [i.e., shifted to be zero-mean and normalized to the standard deviation, $z_x = (X - m_x)/\sigma_x$] were used. Another class of important features is derived from the position of the boundaries for the various units. The information includes the duration of the upper level unit (for example the prosodic group defined by B5 when determining B4 for a breath group), and the distance of the current boundary from the beginning and the end of the upper level unit. These features are measured both in seconds and in number of syllables. The features mentioned above are all derived from the acoustic information. Additional features can be derived from the corresponding text transcription. The locations of punctuation marks can be directly copied from the text. They are very useful for identifying B4 and B5. The word boundary is also very helpful, which can be obtained using a

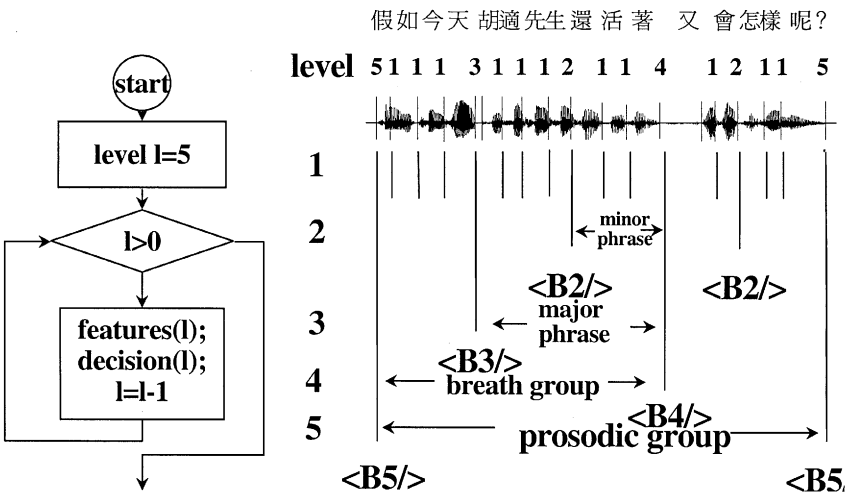

Fig. 2. Multipass procedure for automatic prosodic labeling.

TABLE III
THE SET OF FEATURES USED TO PERFORM THE AUTOMATIC PROSODIC LABELING

Class	Symbols	Descriptions
Pause	Pd	pause duration
Duration	Dp	normalized duration of the preceding syllable
	Df	normalized duration of the following syllable
	Dr	Df / Dp
Energy	Ep	normalized energy of the preceding syllable
	Ef	normalized energy of the following syllable
	Er	Ef / Ep
F_0	Fr	normalized F_0 reset
	Fb	normalized F_0 at the beginning of the following syllable
	Fe	normalized F_0 at the end of the preceding syllable
Position	Lt/Ln	Length (in seconds or number of syllables) of the upper level unit
	Bt/Bn	distance (seconds or syllables) from the Beginning of the upper level unit
	Et/En	distance (seconds or syllables) from the end of the upper level unit
	Eu	end of utterance
Text	Pm	punctuation mark
	Wb	word boundary

word segmentation program. This program identifies the word boundaries automatically from a string of Chinese characters (since there are no blanks between words in Chinese texts). In this research, the word segmentation program developed earlier in research of Chinese natural language processing was directly adopted. Most of the syllabic boundaries within a word are B1. Only these two features, punctuation marks and word boundaries, derived from text transcription were used in this research, although more information will probably be helpful.

Some investigation with respect to human labeling reliability was performed first [42], [43]. 100 paragraphs of speech corpora were used in this preliminary test. Table IV is the comparison of the break indices labeled by two different human labelers A and B. Table IV(a) presents the independent labeling results of the two labelers based on the proposed criteria; while Table IV(b) presents the labeling results of the same set of data after the two labelers compared their individual notes of labeling criteria used. We found that although the consistency between labelers was improved after the discussion, the less identifiable categories remained unchanged. Most of the inconsistency occurred in B1 versus B2 and B4 versus B5. Even after the discussion [in Table IV(b)], a total of 204 boundaries were labeled as B1 by labeler A, but as B2 by labeler B. Also, 48 boundaries were labeled as B5 by labeler A but as B4 by labeler B. This may imply that labeler A is more sensitive to global prosodic changes and labeler B is more sensitive to finer local prosodic changes. In any case, the reliability in human labeling is an important issue, as was previously discussed in other research works for other languages [42], [43]. For the tests for automatic prosodic labeling to be described below, only the indices labeled by the labeler B were used in both training and testing for consistency purposes.

The tests for automatic prosodic labeling were performed with a database of 599 paragraphs produced by the single male speaker, 399 for training and 200 for testing. In the first experiment, only features derived from acoustic information were used, and in the second experiment the second type of features derived from transcribed texts (punctuation marks plus word boundaries) were also included. Table V(a) and (b) are the confusion matrices for the automatically obtained labels with respect to manual labels for these two experiments. The average error rate is, respectively, 20.3% and 15.1%. The confusion can be effectively reduced with the text derived features, as can be observed from Table V(a) and (b). This verifies the text information is helpful for prosodic labeling. By comparing the labeling accuracy (the diagonal elements) between Tables IV(b) and V(b), it is interesting to see that human labeling is not necessarily always better than automatic labeling. For example, 81.7% of automatically labeled B2 indices were consistent with the human labeler B, but only 64.3% of B2 indices labeled by the human labeler A were. Similarly for B3 (81.2% for automatic labeling and 78.8% for human labeler A). Although such comparison may not be rigorous because the testing paragraphs are different, but this indicated the accuracy achieved by the automatic processes presented here is reasonable. The possible reason is that the machine may be able to learn the consistency for labeling, while human labelers really have different indi-

TABLE IV
THE BREAK INDICES LABELED BY TWO LABELERS (A AND B) (a) BEFORE AND (b) AFTER THE EXCHANGE OF THEIR INDIVIDUAL NOTES FOR LABELING

(a)

A B	B1	B2	B3	B4	B5
B1	2041	114	16	2	4
	93.8%	5.2%	0.7%	0.1%	0.2%
B2	205	394	87	2	0
	29.8%	57.3%	12.6%	0.3%	0.0%
B3	14	80	187	45	5
	4.2%	24.2%	56.5%	13.6%	15%
B4	0	1	67	163	108
	0.0%	0.3%	19.8%	48.1%	31.9%
B5	1	0	1	7	103
	0.9%	0%	0.9%	6.3%	92.0%

(b)

A B	B1	B2	B3	B4	B5
B1	2162	83	1	0	1
	96.2%	3.7%	0%	0%	0%
B2	204	422	30	0	0
	31.1%	64.3%	4.6%	0%	0%
B3	5	45	330	36	3
	1.2%	10.7%	78.8%	8.6%	0.7%
B4	0	1	46	124	48
	0%	0.5%	21.0%	56.6%	21.9%
B5	1	0	0	2	103
	0.9%	0%	0%	1.9	97.2%

TABLE V
CONFUSION MATRICES FOR AUTOMATIC LABELING OF BREAK INDICES WITH RESPECT TO MANUAL LABELS: (a) EXPERIMENT 1 WITHOUT USING THE TEXT DERIVED FEATURES, AND (b) EXPERIMENT 2 USING THE TEXT DERIVED FEATURES

Manual Labels	Automatic Labels					Total
	B1	B2	B3	B4	B5	
B1	3211 (83.2%)	581 (15.0)	65 (1.6%)	0 (0%)	0 (0%)	3857
B2	113 (11.1%)	809 (79.4%)	84 (8.2)	7 (0.7%)	5 (0.5%)	1018
B3	14 (2.2)	71 (11.3%)	487 (77.5%)	35 (5.6%)	21 (3.3%)	628
B4	0 (0%)	3 (0.1%)	48 (13.8)	171 (49.0%)	127 (36.3%)	349
B5	0 (0%)	0 (0%)	17 (4.2)	76 (19%)	304 (76.5%)	397
Total	3338	1464	701	289	457	6249

(a) experiment 1: (error rate 20.3%)

Manual Labels	Automatic Labels					Total
	B1	B2	B3	B4	B5	
B1	3466 (89.8%)	378 (9.8%)	13 (0.3%)	0 (0%)	0 (0%)	3857
B2	98 (9.6%)	832 (81.7%)	78 (7.7%)	8 (0.8%)	2 (0.1%)	1018
B3	15 (2.3%)	65 (10.3%)	510 (81.2%)	25 (4.0%)	13 (2.5%)	628
B4	0 (0%)	4 (1.1%)	34 (9.7%)	186 (53.2%)	125 (35.8%)	349
B5	0 (0%)	0 (0%)	12 (3.0%)	72 (18.1%)	313 (78.8%)	397
Total	3579	1279	647	291	453	6249

(b) experiment 2: (error rate 15.1%)

vidual perception sensitivities. More test results regarding how the overall synthesized speech quality is dependent on the techniques discussed here will be presented later on in Section VII.

V. AUTOMATIC GENERATION OF HIERARCHICAL PROSODIC STRUCTURE FOR TAGGED TEXT SENTENCES

For any input text sentence, converting it into the corresponding prosodic structure including various levels of groups, phrases, and break indices will be a key for synthesizing speech signals with good naturalness and intelligibility. In this research, this is achieved hierarchically, just as the process that identified the break indices from the upper level units one by one, as discussed previously [44], [45]. Because a Chinese sentence is a string of characters without blanks indicating the word boundaries, the input text sentence needs to be first segmented into words. The segmented words should then be tagged with the corresponding parts-of-speech. These processes of word segmentation and parts-of-speech tagging have been well studied [46], [47] and in this research such processes are adopted directly. Although syntactic structure of or even semantic knowledge about the sentences are certainly helpful in prosodic analysis, they require the input sentences to be properly parsed, and the cost is relatively high. In this research, it was found that prosodic phrasing simply based on statistical analysis of part-of-speech tags can in fact produce reasonably well synthesized speech [45]. This will be discussed in this section.

The analysis started with a total of 44 Parts-of-Speech (POSs) primarily derived from early studies on Chinese natural language processing [48]. Because it is possible that not as many as all the 44 POSs are directly relevant to prosodic structures studied here, it may be reasonable to cluster these 44 POSs into a smaller number of groups. Such clustering may also increase the number of samples in each group to be used in the following statistical analysis. Three different approaches for grouping these POSs were considered. The first approach, referred to here as syntactic grouping, used syntactic knowledge from human experts for clustering, and a total of 26 groups of POSs was derived. The second approach, referred to as text-corpus grouping, was based on the statistical behavior of the 44 POSs in the text corpus. In this approach, a feature vector was defined for each POS, whose components were the normalized frequency counts of all the preceding and following POSs in the texts of the speech

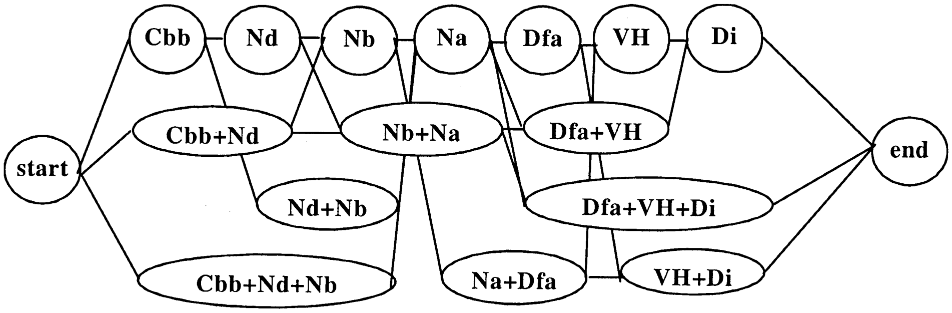

Fig. 3. An example of "minor phrase lattice."

database designed as mentioned previously. These feature vectors were then vector quantized and clustered, and a total of 18 groups was obtained. Neither of the above two approaches used prosodic information from the speech corpus. The third approach, referred to as speech-corpus grouping here, tried to use some prosodic information from the speech corpus. In this approach, a number from 0 to 5 (0 for B0, 1 for B1, … 5 for B5) was first assigned to each break index. For each POS, the mean values of these numbers were then evaluated for the two boundaries on both sides of all the words corresponding to the POS in the speech corpus. These mean values were finally added to the feature vectors of the POSs constructed in the second approach. The rest of the third approach is almost identical to the second approach, and a total of 18 groups was obtained. The relative effectiveness of these three approaches will be tested in the experiments to be discussed later on.

According to the six levels of break indices mentioned previously, B0 almost does not exist in the speech corpus of read speech used here, and B1 is marked at the end of each syllable (or character). So the first important lower level index is B2 and the unit between two B2 indices is the "minor phrase." After the POSs were classified as discussed above, the patterns for the groups of POSs within a "minor phrase" (for example: $Adj + N$, $Adv + V$, $N + N + N$, etc.) between two B2 indices recorded in the speech corpus labeled previously were collected and used to construct a "minor phrase table." Frequencies of occurrence in the speech corpus for all these patterns of "minor phrases" were also recorded in the table.

The identification of the B2 indices and the "minor phrases" for an arbitrary text sentence based on the above "minor phrase table" can be achieved as follows. The text sentence is first segmented into a sequence of words with POSs tagged. This POS sequence is then matched with those POS patterns in the "minor phrase table." Because very often there can be more than one way to segment the sentence into such "minor phrases" as those listed in the "minor phrase table," a lattice of possible "minor phrases" for the sentence can be constructed. A typical example of such a lattice can be found in Fig. 3. A dynamic programming procedure is then used to determine the best path in the lattice based on the scores obtained from the frequencies of occurrence of the "minor phrases." Longer "minor phrases" are preferred in this procedure and weights for higher priority are given. This is based on the experiences that longer phrases very often represent more probable structures. After the "minor phrases" and B2 indices have been determined from the best path obtained above in this way, the construction of "major phrases" between two B3 indices can be performed in a similar way with a second dynamic programming procedure based on a table of "major phrase" patterns and the associated frequency scores. This process can then be repeated to identify B4 indices and so on, so as to construct all levels of prosodic phrases hierarchically bottom-up. B5 indices, on the other hand, are directly identified from the punctuation marks.

Some experiments were performed to test the effectiveness of the above approaches. 599 paragraphs in the speech corpus were tested, 500 for training and 99 for testing. First, the three different POS grouping methods mentioned above were compared, and the results of labeling accuracy with respect to manual labeling for B2, B3, B4 all together are listed in Table VI(a). From the table, it can be seen that the speech-corpus grouping including the information derived from the speech corpus achieved the highest accuracy. This may imply that it is not easy to predict precisely the prosodic phrases from the syntactic structures only. In fact, similar to experiences with other languages, it was found in this research that the prosodic phrase breaks do not necessarily coincide with the syntactic phrase boundaries, and the relationship between prosody and syntax is still not yet well understood. The detailed confusion table for the obtained break indices was listed in Table VI(b). It can be seen from Table VI(b) that the B2 indices could be identified with the highest accuracy of 85.7%, and even for the worst case of B4 indices, an accuracy of 78.7% was achieved. Unfortunately, it is not easy to compare the performance with other reported results [49], [50] due to the differences in the languages and the corpora used. An important consideration in evaluating the approach is that the prosodic phrase structure of a given text sentence is not necessarily unique. A human speaker can easily produce a sentence in several different ways without altering the naturalness or the meaning. There are still many questions unanswered in this area. More test results regarding how the overall synthesized speech quality is dependent on the techniques discussed here will be presented later on in Section VII.

VI. Voice Unit Selection and Concatenation

When the prosodic phrase structure for an arbitrary input text sentence is obtained as described above, the next process is to select the appropriate voice units from the speech corpus and concatenate these units together to obtain the synthesized speech waveform [51]–[53]. This is discussed in this section.

TABLE VI
TESTING RESULTS FOR AUTOMATIC GENERATION OF BREAK INDICES FROM TAGGED TEXTS: (a) LABELING ACCURACY FOR DIFFERENT POS GROUPING APPROACHES, AND (b) DETAILED CONFUSION TABLE AMONG B2, B3 AND B4 FOR SPEECH-CORPUS GROUPING

(a)

POS Grouping Approaches	**Prosodic Labeling Accuracy**
(1) syntactic grouping: 26 groups	80.9%
(2) text-corpus grouping: 18 groups	78.3%
(3) speech-corpus grouping: 18 groups	83.1%

(b)

Manually Labeled \ **Automatically Labeled**	**B2**	**B3**	**B4**
B2	85.7%	10.5%	3.8%
B3	10.8%	81.2%	8.0%
B4	4.6%	16.7%	78.7%

TABLE VII
THE LINGUISTIC/SYMBOLIC FEATURES USED TO SPECIFY EACH INITIAL/FINAL UNIT IN THE SPEECH CORPORA

Symbol	Description	Total Number
SID	Syllable identity	408 syllables
PCP	Preceding phoneme class	8 types
PCF	Following phoneme class	12 types
LW	Location in the word	3 types
LP1	Location in the minor phrase	3 types
LP2	Location in the major phrase	3 types
LP3	Location in the breath group	3 types
TID	Tone identity	5(=4+1) tones
TIDP	Preceding tone identity	6(5+beginning)
TIDF	Following tone identity	6(5+ending)
BP	The preceding break index	6 types
BF	The following break index	6 types
PM	Punctuation mark	4 types

First, an indexing file was generated for each INITIAL/FINAL unit in the speech corpus to be used for selection and synthesis. Two sets of information were included in the file. The first is the linguistic/symbolic features such as the phonetic identities, tonal identities, and other position and contextual information including those with respect to different levels of break indices in the prosodic structure. These linguistic/symbolic features are listed in Table VII. Some of these features have to do with the segmental properties of the units, some with the prosodic parameters, and some with both. The other set of information is the acoustic and prosodic parameters including F_0 values, energy values, duration, etc. plus the cepstral parameters. Many of them were derived with ESPS tools. For the input text sentence to be converted into speech, after the prosodic structure and break indices are determined as discussed previously, an indexing file of linguistic/symbolic features just as those in Table VII was also generated for each INITIAL/FINAL unit in the desired text sentence. This sequence of indexing files of linguistic/symbolic features for the desired text sentence is thus the input to the voice unit selection process to be discussed below. The selected units can be an INITIAL/FINAL unit, a syllable, a syllable plus a preceding FINAL and/or a following INITIAL, or any longer units which may include a few syllables and so on.

For a sequence of input indexing files of linguistic/symbolic features for the desired text sentence, both a sequence of waveform units and a sequence of desired prosodic parameter sets are to be selected. This is because the speech corpus may not be large enough, and the best matched waveform unit may not have exactly the desired prosodic characteristics. In that case the prosodic characteristics of the selected waveform units may need to be modified based on the selected prosodic parameter sets. Therefore two selection processes have to be performed, one for the waveform units and one for the prosodic parameter sets. In both cases a lazy decision tree selection algorithm was developed, in which the selection is to trace along a path in a tree to find the set of best candidates based on some error measures [51], [52]. In both cases the error measures are evaluated based on the linguistic/symbolic features as listed in Table VII for the desired text sentence.

When the above two selection processes are completed, the selected candidates of waveform units are further verified by the selected prosodic parameter sets before concatenation. The verification process is described below:

1) Evaluate a distance measure between the selected waveform units and the selected prosodic parameter sets. Remove the waveform units with distance measures above a threshold unless there is only one unit left.

2) The selected candidate waveform units for the desired text sentence are listed to construct a synthesis unit lattice, i.e., there may be more than one waveform units for a desired voice segment. Choose a path in the lattice that minimizes the concatenation cost, which is defined as the sum of the cost functions calculated for all the concatenation points along the path.

3) Modify the units on the path obtained in step 2) which have the distance measures obtained in step 1) above some pre-defined threshold with TD-PSOLA based on the selected prosodic parameter sets. The sequence of the finally obtained waveform units are then concatenated and smoothed as the output speech.

The complete voice unit selection and concatenation process is shown in a diagram in Fig. 4.

There are three types of concatenation and smoothing processes for the waveform units obtained above:

1) Hard concatenation: Simply put two units together directly and no smoothing is needed. This is used when the beginning INITIAL in the second unit is a plosive or an affricate, as in an example shown in Fig. 5(a).

2) Soft concatenation: The concatenation is smoothed by including the transition parts from both sides. A certain amount of overlap between the two waveform units makes the transition smooth. This is used for the concatenation of any two syllables, if the hard concatenation condition mentioned above cannot be applied, as in an example in Fig. 5(b).

3) INITIAL–FINAL concatenation: This is used when no syllable waveform that matches the requirements can be found in the speech corpus. In this case an INITIAL and a FINAL are concatenated to construct a syllable, as in an example in Fig. 5(c). Overlapping and smoothing are needed.

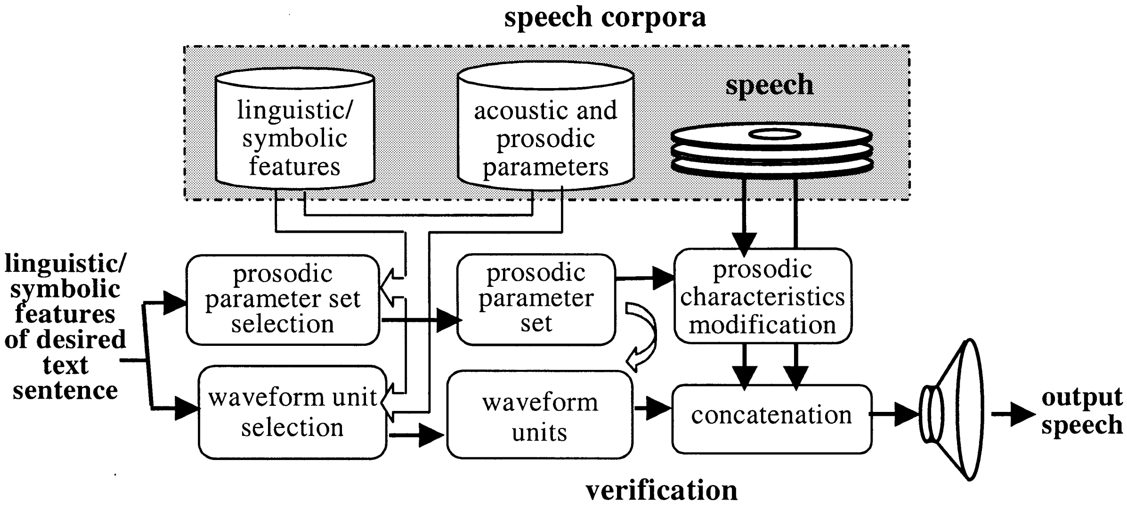
speech corpora
linguistic/ symbolic features
acoustic and prosodic parameters
speech
linguistic/ symbolic features of desired text sentence
prosodic parameter set selection
prosodic parameter set
prosodic characteristics modification
waveform unit selection
waveform units
concatenation
output speech
verification

Fig. 4. Voice unit selection and concatenation.

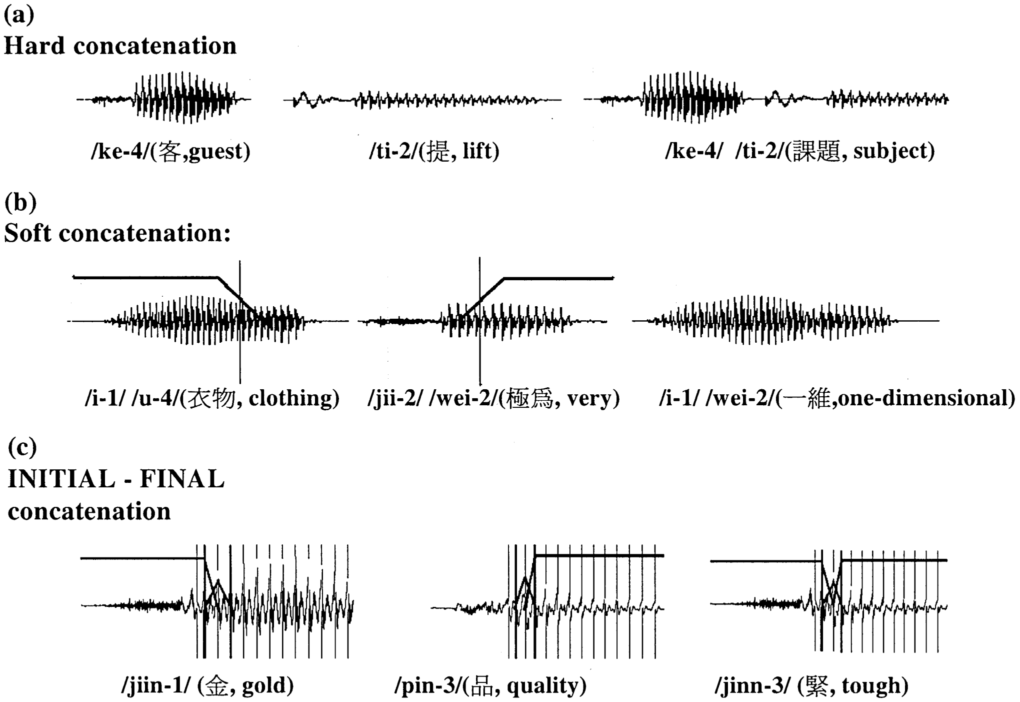

Fig. 5. Three types of unit concatenation and smoothing.

VII. Prototype System and Preliminary Performance Assessment

A prototype system based on the above corpus-based text-to-speech synthesis technologies for Mandarin Chinese has been implemented. The system can accept Chinese text input in BIG5 format, and output the speech from the speaker or in waveform files. The basic hardware requirement is a Pentium PC(300 MHz) with 64 Mbytes RAM, 1 Gbytes hard disk and a 16-bit sound card. The OS could be Windows95 or Windows NT. There are about 630 Mbytes of speech data. The extracted linguistic/symbolic features as well as acoustic and prosodic parameters are about 25 Mbytes and the lexicon is about 5 Mbytes. The lexicon, the linguistic/symbolic features, and the acoustic and prosodic parameters are loaded into the memory when the program is executed. The text input is converted into a word sequence with POSs tagged by a text analysis module. The rest is exactly the same as described above. The processing time needed for synthesizing a speech paragraph is in general shorter than the length of the speech paragraph, therefore real-time synthesis is realizable if a reasonable synchronous scheme is applied.

The performance assessment is a much more difficult part. The individual performance of each modular component [54], [55], such as the automatic phonetic/prosodic labeling and the automatic assignment of break indices for a desired text sentence, can be tested as reported in the previous sections. However, the assessment of the overall synthesized speech quality, or identifying how such quality is dependent on the approaches used in each individual modular component, is much more challenging. This is also true even for western languages, since it has to do with human perception and is therefore subjective [40], [43]. Substantial efforts have been made to develop good methodologies for such purposes. The primary approaches include subjective tests by enough number of listeners, side-by-side comparative tests among different systems, providing real-time access to many systems via Web and so on, with target criteria including intelligibility and naturalness [56]–[59]. There exist various difficulties here. First, very few systems for Mandarin Chinese are actually accessible for comparison and academic testing. Second, different corpus-based systems, even with Mandarin Chinese and accessible for testing, are based on corpora of different sizes, different design, produced by different speakers with different prosody, in addition to being processed by different set of technologies. So the comparison results, even if obtainable, may not be able to provide too much information. Thirdly, comparison with systems based on different approaches, for example rule-based approaches, tells even less since the systems are fundamentally quite different.

Even with the above difficulties, some preliminary performance evaluation was conducted on this prototype system to see how the technologies mentioned here jointly provide the synthesized speech quality. In a regular classroom setting without using earphones, synthesized speech from the system was played to a total pool of 40 subjects, all of them undergraduate university students. The subjects were asked to rate what they heard in the following experiments. A set of 20 paragraphs of texts were used. They were manually selected out of 120 paragraphs randomly extracted from daily newspapers and magazines published in Taiwan. The only criterion for the manual selection was to try to avoid too many repeated phrases or sentence structures in the 20 paragraphs. The length of these paragraphs ranged from 12 to 30 syllables.

TABLE VIII
SUBJECTIVE QUALITY RATING STATISTICS OBTAINED IN THE 3 EXPERIMENTS (1) (2) (3) WITH SIX SYSTEM CONFIGURATIONS (a), (b), (c), (d), (e), and (f)

Experiments	System Configurations	Subjective Quality Rating Statistics	
		Mean	Standard Deviation
(1)Automatic Phonetic Labeling	(a) Baseline	3.41	0.15
	(b) Plus Iterative Correction Rules	3.47	0.17
(2)Automatic Prosodic Labeling	(c) Hierarchical without Text Features	3.55	0.17
	(d) Hierarchical Plus Text Features	3.61	0.14
(3)POS Grouping in Automatic Prosody Generation	(e) Text-corpus Grouping	3.57	0.16
	(f) Speech-corpus Grouping	3.67	0.17

A series of three experiments (1) (2) (3) were performed, with purposes to evaluate how the overall synthesized speech quality is related to, respectively, the automatic phonetic labeling, automatic prosodic labeling and POS grouping techniques used in prosodic structure generation discussed here in this paper [40]. A "baseline system" primarily based on conventional corpus-based TTS technologies was first constructed. This included the conventional automatic phonetic labeling approach as mentioned in Section III and shown in Fig. 1(a) without iterative correction rules, the automatic prosodic labeling approach previously proposed based on hidden Markov models [41] without using the hierarchical approach, the prosodic structure generation based on 26 POS groups simply using syntactic knowledge from human experts as described in Section V, plus the speech corpora design proposed in this paper as described in Section II, and the unit selection and concatenation approaches developed in this paper as described in Section VI. Experiment (1) was to test the dependence of the overall synthesized speech quality on the iterative correction rules in automatic phonetic labeling [discussed in Section III and shown in Fig. 1(b)] alone, therefore the above "baseline system" [referred to as system configuration (a) here] was compared to a "system configuration (b)," in which everything was exactly the same as the "baseline system," except the iterative correction rules for automatic phonetic labeling were used. For each paragraph out of the 20 mentioned above, three versions of speech were played to the subjects. The first was the natural speech produced by the same male speaker who produced the speech for the TTS inventory. The subjects were told that this was the upper bound reference with a rating 5.0. The next two versions were those synthesized by the system configuration (a) (the "baseline system") and (b) (with iterative correction rules). The 40 subjects were divided into two groups of 20 subjects each, in which the two versions of synthesized speech were played in different order. The subjects were not informed which version was produced by which system configuration, but simply asked to provide a rating for speech quality from 1.0 to 5.0, with 5.0 representing the upper bound of the natural speech. In order to avoid too large variance in the scores provided by the subjects, the subjects were told that reasonable scores for the synthesized speech may be between 3.0 and 4.0, although they were actually allowed to give any scores between 1.0 and 5.0. The results are listed in the first two rows of Table VIII. It can be found that with the iterative correction rules the mean score was slightly improved, although to a very limited extent as compared to the standard deviation for the scores. As can be found that the standard deviation here is not very large, probably because the subjects were told that the reasonable scores were between 3.0 and 4.0. This may also be the reason why the mean score is not very far from 3.5. Note that the difference in the mean scores for configurations (a) and (b) may not be significant. All can be said is that configuration (b) seems to be slightly better in average.

Experiment (2) was conducted a few days later after the Experiment (1) described above, with a purpose of testing the dependence of the overall synthesized speech quality on the hierarchical automatic prosodic labeling approaches discussed in Section IV, based on the slightly better system configuration (b) obtained in Experiment (1). Two versions of speech were first played to the subjects as references. The first was the natural speech with given rating 5, and the second was that produced by system configuration (b) including a rating given by each individual subject himself a few days before. Next played were two versions of speech synthesized by system configurations (c) and (d), in which everything is exactly the same as in system configuration (b), except the prosodic labeling was done using the hierarchical approach proposed in this paper. System configuration (c) used only the parameters derived from the acoustic signals [just as those in Table V(a)], while system configuration (d) used the parameters derived from texts in addition [just as those in Table V(b)]. These two versions of speech were played to two groups of 20 subjects each with different order. They were asked to give the rating based on the two references. The results are listed in the next two rows of Table VIII. It can be found that the mean scores were again slightly improved with the hierarchical prosodic structure although the difference is not very significant, and the use of parameters derived from texts was even slightly better. Note that the standard deviation here is also relatively small, probably because the two reference scores [5.0 for natural speech and the scores the subjects themselves gave for system configuration (b)] were given before the tests.

Experiment (3) was conducted again a few days later after the Experiment (2), with a purpose of testing the dependence of the overall synthesized speech quality on the POS grouping techniques used in prosodic structure generation discussed in Section V, based on the slightly better system configuration (d) obtained in Experiment (2). Just as in Experiment (2), two versions of speech were first played to the subjects as references. The first was the natural speech with given rating 5.0, and the second was that produced by system configuration (d) including a rating given by each individual subject himself a few days before. Next played were two versions of speech synthesized by system configurations (e) and (f), in which everything was exactly the same as in system configuration (d), except the POS grouping used in the prosodic structure generation was different. System configuration (e) used the text-corpus grouping approach while system configuration (f) used the speech-corpus grouping approach as discussed in Section V and shown in Table VI(a). The two versions were again played in different order to two groups of subjects. The results are listed in the last two rows of Table VIII. It was found that the text-corpus grouping [used in system configuration (e)] did not necessarily provide better synthesized speech quality as compared to the syntactic grouping based on human knowledge [used in system configurations (a)–(d)]. However, the speech-corpus grouping [used in system configuration (f)] including information obtained from the speech corpus provided slightly better synthesized speech quality, although the difference is not very significant either. This is in good agreement with the results in Table VI(a). Again, it can be noted that the standard deviation here is relativity small. System configuration (f) represents the integration of all technologies proposed in this paper, and was shown in the tests here with improved synthesized speech quality as compared to system configuration (a), or the "baseline system." Note that the system configuration (a) was also based on exactly the same corpus developed by the corpus design principles mentioned in this paper, as well as unit selection/concatenation approaches developed in this paper, but other modular components of the system configuration (a) was conventional.

It should be pointed out here that it seems not easy for a subject to compare too many different versions of synthesized speech perceptually at the same time. This is why here we divided the tests into three separate experiments, in each of which only a few versions of synthesized speech were compared, plus some reference scores for the prior experiment given. This may be an approach to obtain incremental improvements for each modular component. However, with the three separate experiments, it may not make too much sense to compare the mean scores for the final system [3.67 for configuration (f)] and the baseline system [3.41 for configuration (a)] directly. In other words, if the two versions of synthesized speech for configurations (f) and (a) were compared in another experiment directly, the results may be quite different. But the point here is to show that with the incremental improvements provided by different modular components (though not very significant for each individual case), the final system was actually improved step-by-step. With these difficulties in speech quality assessment, the authors are planning to construct a Website demonstrating synthesized speech samples used in the experiments in Table VIII after the paper is published, so that the readers will be able to assess the synthesized speech quality perceptually by themselves, as long as they are able to listen to Mandarin speech.

VIII. Concluding Remarks

It has been well known that there are many structural features of Mandarin Chinese. The tonal aspect has been one usually considered first, and tone Sandhi was often taken as a core phenomenon for prosody studies in Mandarin Chinese. In this research, it was found from the experiences in manual labeling of speech corpora that the hierarchical structure may provide a better global picture of prosody in Mandarin Chinese, while the tone Sandhi reflects the local prosody behavior. By testing with the prototype system, it was easily verified that almost all tone Sandhi phenomena were automatically produced in the prosody structure generated by the proposed approach, although no special efforts or considerations for tone Sandhi were ever made. This verified that the technologies developed here provided a better overall picture of prosody in Mandarin Chinese, although there are still many questions unanswered. On the other hand, the monosyllabic structure and character/syllable mapping relation is another key structural feature of Mandarin Chinese, on which many prior text-to-speech systems have been based. In the approaches presented in this paper, however, the waveform units selected from the corpus very often have time spans across syllable boundaries, as can be easily observed in the prototype system. It was also found from the experiences in manual labeling of speech corpora that the hierarchical structure seems to play a much more important role than the individual syllable structure in the prosody of Mandarin Chinese, although the character/syllable mapping relation is always a key factor in TTS for Mandarin Chinese. It can be easily found throughout this paper that many of such well-known structural features of Mandarin

Chinese have in fact been carefully considered, from corpus design to automatic labeling, from prosodic structure generation to waveform unit selection, although a completely new set of technologies were actually developed here.

REFERENCES

[1] "Special issue on spoken language processing," *Proc. IEEE*, Aug. 2000.

[2] W. N. Campbell and A. W. Black, "Prosody and the selection of source units for concatenative synthesis," in *Progress in Speech Synthesis*. Berlin, Germany: Springer Verlag, 1996, pp. 279–282.

[3] A. W. Black and P. Taylor, "CHATR: A generic speech synthesis system," in *Proc. COLING-94*, 1994, pp. 983–986.

[4] W. N. Campbell, "CHATR: A high-definition speech re-sequencing system," in *Proc. 3rd ASA/ASJ Joint Meeting*, 1996, pp. 1223–1228.

[5] C. Shih and R. Sproat, "Issues in text-to-speech conversion for Mandarin," *Int. J. Computat. Linguis. Chin. Lang. Process.*, vol. 1, no. 1, pp. 37–86, 1996.

[6] L. S. Lee, C. Y. Tseng, and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 9, pp. 1309–1320, 1989.

[7] C. S. Liu, M. F. Tsai, Y. H. Hsyu, C. C. Lu, and S. M. Yu, "A Chinese text-to-speech system based on LPC synthesizer," *Telecommun. Lab. Tech. J.*, vol. 19, no. 3, pp. 269–285, 1989.

[8] J. Choi, H. W. Hon, J. L. Lebrun, S. P. Lee, G. Loudon, V. H. Phan, and S. Yogananthan, "Yanhui, a software based high performance Mandarin text-to-speech system," in *Proc. ROCLING VII*, 1994, pp. 35–50.

[9] B. Ao, C. Shih, and R. Sproat, "A corpus-based Mandarin text-to-speech synthesizer," in *Proc. Int. Conf. Spoken Language Processing*, 1994, pp. 1771–1774.

[10] L. Cai, H. Liu, and Q. Zhou, "Design and achievement of a Chinese text-to-speech system under windows," *Microcomput.*, vol. 3, 1995.

[11] S. H. Hwang, S. H. Chen, and Y. R. Wang, "A Mandarin text-to-speech system," in *Proc. Int. Conf. Spoken Language Processing*, 1996, pp. 1421–1424.

[12] R. Sproat, C. Shih, W. Gale, and N. Chang, "A stochastic finite-state word segmentation algorithm for Chinese," *Comput. Linguist.*, vol. 22, no. 3, 1996.

[13] L. S. Lee, C. Y. Tseng, and C. J. Hsieh, "Improved tone concatenation rules in a formant-based Chinese text-to-speech system," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 287–294, July 1993.

[14] Y. C. Chang, Y. F. Lee, B. E. Shia, and H. C. Wang, "Statistical models for the Chinese text-to-speech system," in *Proc. EUROSPEECH*, 1991, pp. 227–240.

[15] S. H. Hwang and S. H. Chen, "A prosodic model of Mandarin speech and its application to pitch level generation for text-to-speech," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1995, pp. 616–619.

[16] C. Shih, "The prosodic domain of tone Sandhi in Chinese," Ph.D. dissertation, Univ. California, Berkeley, 1986.

[17] M. C. Chang, "A prosodic account of tone, stress, and tone Sandhi in Chinese language," Ph.D. dissertation, Univ. Hawaii, Hilo, 1992.

[18] K. E. A. Silverman, M. Beckman, J. F. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, 1992, pp. 867–870.

[19] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of nonuniform synthesis units," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1988, pp. 679–682.

[20] K. Takeda, K. Abe, and Y. Sagisaka, "On the basic scheme and algorithms in nonuniform units speech synthesis," in *Talking Machines: Theories, Models and Designs*. Amsterdam, The Netherlands: North-Holland, 1992, pp. 93–106.

[21] S. King, T. Portele, and F. Hofer, "Speech synthesis using nonuniform units in the Verbmobil Project," in *Proc. EUROSPEECH*, 1997, pp. 569–572.

[22] M. Chu *et al.*, "Selecting nonuniform units from a very large corpus for concatenative speech synthesizer," in *Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, UT, 2001.

[23] C. S. Liu, G. H. Ju, W. J. Wang, H. C. Wang, and W. H. Lai, "A new speech synthesizer for text-to-speech system using multipulse excitation with pitch predictor," in *Int. Conf. Computer Processing of Chinese and Oriental Languages*, 1991, pp. 205–209.

[24] M. Chu and S. Lu, "High intelligibility and naturalness Chinese TTS system and prosodic rules," in *Proc. XIII Int. Congr. Phonetic Sciences*, 1995, pp. 334–337.

[25] [Online]. Available: http://www.sinica.edu.tw/~tibe/2-words/modern-words.

[26] L. S. Lee, "Voice dictation of Mandarin Chinese," *IEEE Signal Processing Mag.*, vol. 14, pp. 63–101, July 1997.

[27] L. S. Lee *et al.*, "Golden Mandarin (I)—A real-time Mandarin speech dictation machine for Chinese language with very large vocabulary," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 158–179, Apr. 1993.

[28] J. Shen *et al.*, "Automatic selection of phonetically distributed sentence sets for speaker adaptation with application to large vocabulary Mandarin speech recognition," *Comput. Speech Lang.*, vol. 13, no. 1, pp. 79–97, Jan. 1999.

[29] Y. R. Chao, "Tone and intonation in Chinese," in *Bulletin of the Institute of History and Philology*. New York: Academia Sinica, 1933, vol. 4, pp. 2121–2134.

[30] ——, *A Grammar of Spoken Chinese*. Berkeley, CA: Univ. of California Press, 1968.

[31] A. T. Ho, "Intonation variations in a Mandarin sentence for three expressions: Interrogative, exclamatory, and declarative," *Phonetica*, vol. 34, pp. 446–456, 1977.

[32] F. C. Chou and C. Y. Tseng, *The Design of Prosodically Oriented Mandarin Speech Database*: Int. Congr. Phonetic Science, 1999.

[33] C. N. Li and S. A. Thompson, *Mandarin Chinese: A Functional Reference Grammar*. Berkeley, CA: Univ. California Press, 1981, pp. 520–563.

[34] C. Wells, "Computer-coded phonemic notation of individual languages of the European community," *J. Int. Phonetic Assoc.*, vol. 19, pp. 32–54, 1989.

[35] C. Y. Tseng and F. C. Chou, "Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan," in *Proc. Oriental COCOSDA Workshop*, 1998, pp. 179–183.

[36] ——, *A Prosodic Labeling System for Mandarin Speech Database*: Int. Congr. Phonetic Science, 1999.

[37] F. C. Chou, C. Y. Tseng, and L. S. Lee, "Automatic segmental and prosodic labeling of Mandarin speech," in *Int. Conf. Spoken Language Processing*, 1998.

[38] J. P. H. van Santen and R. W. Sproat, "High accuracy automatic segmentation," in *Proc. Eurospeech*, 1999.

[39] HMM Tool Kits (HTK).. [Online]. Available: http://htk.eng.cam.ac.uk.

[40] C. W. Wightman, A. K. Syrdal, G. Stemmer, A. Conkie, and M. Beutnagel, "Perceptually based automatic prosodic labeling and prosodically enriched unit selection improve concatenative TTS synthesis," in *Proc. ICSLP*, Beijing, China, 2000.

[41] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans Speech Audio Processing*, vol. 2, pp. 469–481, 1994.

[42] A. K. Syrdal and J. McGory, "Inter-transcriber reliability of ToBI prosodic labeling," in *Int. Conf. Spoken Language Processing*, Beijing, China, 2000.

[43] A. K. Syrdal *et al.*, "Corpus-based techniques in the AT&T NEXTGEN synthesis system," in *Int. Conf. Acoustics, Speech, Signal Processing*, 2001.

[44] F. C. Chou, C. Y. Tseng, and L. S. Lee, "Automatic generation of prosodic structure for high quality Mandarin speech synthesis," in *Int. Conf. Spoken Language Processing*, 1996, pp. 1624–1627.

[45] F. C. Chou, C. Y. Tseng, K. J. Chen, and L. S. Lee, "A Chinese text-to-speech system based on part-of-speech analysis, prosodic modeling and nonuniform units," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1997, pp. 923–926.

[46] K. J. Chen and S. H. Liu, "Word identification for Mandarin Chinese sentences," in *Proc. COLING*, 1992, pp. 101–107.

[47] L. P. Chang and K. J. Chen, "The CKIP part-of-speech tagging system for modern Chinese texts," in *Proc. ICCPOL*, 1995, pp. 172–175.

[48] Chinese Knowledge Information Processing Group, "The analysis of Chinese parts-of-speech," Inst. Inform. Sci., Academia Sinica, Beijing, China, Tech. Rep. 93-06, 1993.

[49] M. Q. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Comput. Speech Lang.*, pp. 175–196, 1992.

[50] A. Black and P. Taylor, "Assigning phrase breaks from part-of-speech sequences," in *Proc. Eurospeech*, 1997, pp. 995–998.

[51] F. C. Chou, C. Y. Tseng, and L. S. Lee, "Selection of waveform units for corpus-based Mandarin speech synthesis based on decision trees and prosodic modification costs," in *Proc. Eurospeech*, 1999.

[52] F. C. Chou and C. Y. Tseng, "Corpus-based Mandarin speech synthesis with contextual syllabic units based on phonetic properties," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1998, pp. 893–896.

[53] A. Conkie, M. C. Beutnagel, A. K. Syrdal, and P. E. Brown, “Preselection of candidate units in a unit-selection-based TTS synthesis system,” in *Int. Conf. Spoken Language Processing*, Beijing, China, 2000.

[54] F. Yvon *et al.*, “Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French,” *Comput. Speech Lang.*, vol. 12, pp. 393–410, 1998.

[55] G. P. Sonntag and T. Portele, “PURR—A method for prosody evaluation and investigation,” *Comput. Speech Lang.*, vol. 12, pp. 437–451, 1998.

[56] ITU-T Recommendation P.800, *Telephone Transmission Quality Subjective Opinion Tests: Methods for Subjective Determination of Transmission Quality*. Geneva, Switzerland: ITU, 1996.

[57] D. Gibbon, R. Moore, and R. Winski, *HANDBOOK of Standards and Resources for Spoken Language Systems*. Berlin, Germany: Mouton de Gruyter, 1997, pp. 481–563.

[58] J. Van Santen. Multi-lingual text-to-speech synthesis evaluation. [Online]. Available: http://www.itl.atr.co.jp/cocosda/synthesis/evaltext.html.

[59] J. Zhang, S. Dong, and G. Yu, “Total quality evaluation of speech synthesis system,” in *Int. Conf. Spoken Language Processing*, 1998, pp. 60–63.

Fu-Chiang Chou received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1989 and 1999, respectively.

He was a Postdoctoral Fellow at the Institute of Linguistics, Academia Sinica, Taipei, in 1999. He was the Chief Technology Officer of Applied Speech Technologies, Taipei, from 1999 to 2001. Since 2001, he has been with Philips Research East Asia, Taipei, as a Senior Researcher. He is now a Project Manager of Philips Speech Processing, Voice Control. His research interests are in the area of digital speech processing with special interests on text-to-speech and voice recognition systems.

Chiu-Yu Tseng received the Ph.D. degree in linguistics from Brown University, Providence, RI.

She is a Research Fellow with Institute of Linguistics, Academia Sinica, Taipei, Taiwan, R.O.C. She has collaborated with engineers extensively and worked on a number of speech science related projects. Her research in speech science has focused on building phonetic and prosodic oriented speech database for Mandarin, developing cross-dialect labeling systems for Chinese, and constructing a working organization for speech prosody on the basis of production, perception, speech planning, physiology and breathing. Her other research interests also include psycholinguistic and neurolinguistics basis for speech.

Lin-Shan Lee (S’76–M’77–SM’88–F’93) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He has been a Professor of electrical engineering and computer science at the National Taiwan University, Taipei, R.O.C., since 1982 and holds a joint appointment as a Research Fellow of Academia Sinica, Taipei. His research interests include digital communications and spoken language processing. He developed several of the earliest versions of Chinese spoken language processing systems in the world, including text-to-speech system, natural language analyzer, and dictation systems.

Dr. Lee was Guest Editor of a Special Issue on Intelligent Signal Processing in Communications of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS in December 1994 and January 1995. He was the Vice President for International Affairs (1996–1997) and the Awards Committee Chair (1998–1999) of the IEEE Communications Society. He has been a member of Permanent Council of International Conference on Spoken Language Processing (ICSLP), and is currently a member of the Board of International Speech Communication Association (ISCA).