The Synthesis Rules in a Chinese Text-to-Speech System

LIN-SHAN LEE, SENIOR MEMBER, IEEE, CHIU-YU TSENG, AND MING OUH-YOUNG

Abstract—**An initial attempt to develop a preliminary Chinese text-to-speech system has been made recently. The design approaches are based on a syllable concatenation concept due to the special characteristics of the syllabically paced nature of Chinese language. This paper describes in some detail the synthesis rules developed for this system with special attention given to the lexical tones and other prosodic rules such as concatenation rules, sandhi rules, stress rules, intonation patterns, syllable duration rules, pause insertion rules, and energy modification rules. These rules were derived basically from the acoustic properties of Mandarin Chinese, and therefore are useful not only in designing other Chinese text-to-speech systems, but in understanding the characteristics of Mandarin sentences and processing Mandarin speech signals for other purposes such as segmentation or recognition.**

I. Introduction

With the expected and continued trend toward friendlier man–machine interfaces, the applications and demands for speech synthesis technology are growing rapidly [1]. Devices capable of synthesizing a limited number of sentences with fixed text, such as talking toys or clocks, have long been available commercially. The technology for these devices is basically independent of the target language, i.e., they can be trained to speak various languages but only for a limited number of sentences. On the other hand, text-to-speech systems capable of synthesizing an unlimited number of sentences from unrestricted text input [2]–[15] have been developed for different languages by many research groups. These systems are much more attractive because of their higher flexibility and potential in a wide range of applications. However, the technology for developing such systems is not only much more complicated and advanced, but generally language dependent. In other words, although the basic methodology and philosophy in developing text-to-speech systems may be quite similar regardless of the target language, the phonetic aspects of the synthesis rules have to be tailored specifically for different target languages. During the initial stage of developing our Chinese text-to-speech system, efforts were made to investigate different design approaches, strategies, and implementations of various text-to-speech systems [7]–[15], but always bearing in mind the phonetic properties of Mandarin Chinese. To our knowledge, very little work has been reported on Chinese text-to-speech systems [16]–[18]. As a result of our initial effort, a preliminary version of a Chinese text-to-speech system has been implemented at National Taiwan University [19], [20].

The design of the system to be reported in this paper depends heavily on the characteristics of Mandarin Chinese in the sense that it is a monosyllable based system. The rationale is that most Mandarin Chinese morphemes are monosyllabic with relatively simple syllabic structure. Another major consideration that requires substantial efforts has been the tones of Mandarin due to the fact that Mandarin Chinese is a tonal language. A brief description of the synthesis approaches used in this system is given in Section II. However, this paper will concentrate primarily on the phonetic rules used in the speech synthesis processes. The rules described here include the concatenation rules, especially sandhi rules, stress rules, intonation patterns, syllable duration rules, pause insertion rules, energy modification rules, etc. Because these rules are derived essentially from the phonetic characteristics as well as the acoustic properties of Mandarin Chinese, they are considered independent of the system design approaches and strategies. In other words, if some other Chinese text-to-speech systems are to be designed using different approaches and strategies, these rules will be equally helpful, although not necessarily directly applicable. Furthermore, because these rules demonstrate how a Mandarin Chinese sentence is synthesized, it is our hope that they can be used as research tools to understand the characteristics of Chinese sentences so they will also be helpful in processing Mandarin speech signals, such as continuous speech segmentation and recognition. Although there are some other Chinese text-to-speech system developments reported elsewhere [16]–[18], to our knowledge this is the first set of rules obtained which considers and includes a wider range of linguistic phenomena and acoustic properties of Mandarin Chinese, and this system is also the first one which has been very effectively implemented [20].

II. The Syllable Concatenation Concept for a Chinese Text-to-Speech System

From the viewpoint of Chinese orthography, there are at least some 13 000 commonly used Chinese characters

Manuscript received June 17, 1987; revised October 13, 1988.

L.-S. Lee is with the Department of Electrical Engineering and the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, Republic of China.

C.-Y. Tseng is with the Institute of History and Philology and the Institute of Information Science, Academia Sinica, Taipei, Taiwan, Republic of China.

M. Ouh-Young is with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, Republic of China.

IEEE Log Number 8929370.

(written symbols or ideographs), each corresponding to a monosyllable. However, there are at least 60 000 commonly used words in Chinese, each composed of one to several characters. Nevertheless, the total number of phonologically allowed syllables in Mandarin speech is only about 1300. That is, there are only 418 phonologically allowed syllables in Mandarin Chinese regardless of tones, and not all syllables have 4 tonally variant counterparts, which yields the total syllable number to approximately 1300 [21]. Almost all of these syllables are open syllable in structure, i.e., they always end with a vowel (with the exception of vowels plus nasals -n or -ng). As a result, some of the linguistic characteristics of Chinese include a relatively high number of homonyms, simple syllable structure, and the one-to-one correspondence between a morpheme/syllable to a character. Although polysyllabic words do exist in fairly large numbers resulting in a co-articulation across syllables, and intersyllabic as well as intrasyllabic co-articulation may also happen in running speech, relatively distinct syllables in spoken form appear to be a rather accepted mode in read, enunciated speech. Besides, since each syllable corresponds to a character in the orthography, it appears to be a natural choice for native speakers of Chinese as the smallest unit for written text. Based on the above observation on the special structure of Chinese language, the use of syllables as the basic units to synthesize Mandarin Chinese becomes a very natural choice. Speech waveforms for Chinese sentences can be synthesized directly by simply concatenating the syllables in the sentences and adjusting the parameters describing the acoustic properties of these syllables. The concept of syllable concatenation has thus become the basic idea of a Chinese text-to-speech system [19], [20].

Another very special important feature of Mandarin Chinese is the tonal aspects, since Chinese is a tonal language, i.e., the tones in Mandarin Chinese have lexical meaning. There are basically four lexical tones and one neutral tone in Mandarin Chinese. All tones described in this paper are in the forms of pitch period patterns, and are thus inverse in shape as compared to those patterns in terms of their corresponding fundamental frequencies. The four lexical tones consist of one level tone (hence Tone 1) and three contour tones (hence Tone 2, Tone 3, and Tone 4), with an additional neutral tone (hence Tone 5) whose phonetic manifestation is conditioned by its preceding lexical tone [22]. It has been shown [23] that the primary difference for the four tones is in the pitch contours, and in fact there exist standard patterns for the pitch contours which will produce the four tones. One example [24] is shown in Fig. 1, where the pitch contours for the four tones of three vowels and two diphthongs[1] [a, u, i, ai, au - 1, 2, 3, 4] for the same speaker are plotted. It can be seen that regardless of the different qualities of the

[1]The transliteration symbols used in this paper are in Mandarin Phonetic Symbols II (MPS II). The numerical numbers following each syllable denote the lexical tone of the syllable.

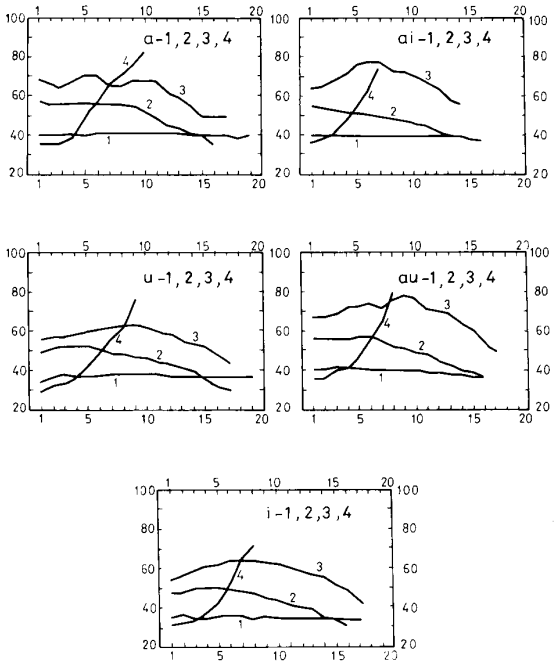


Fig. 1. The pitch contours of [a, u, i, ai, au -1, 2, 3, 4] for the same speaker, sampling period versus frame number. The horizontal axis is the time in units of frame number, the vertical axis is the pitch period in units of sampling period. The number on each curve indicates the tone.

vowels, the basic patterns for the pitch contours for the four tones remain essentially the same. A plot of many typical patterns of pitch contours for the same speaker [24] is shown in Fig. 2, where the pitch contours for different syllables such as ba, fu, li with the same tone are plotted together. It is clear that the basic patterns are almost identical across syllables when the syllables are produced in isolation. In other words, the pitch contours corresponding to different lexical tones appear to be relatively independent of the syllable structure that carries the pitch contour, provided that these syllables are produced as isolated tokens. As a result from such an observation, we were able to superimpose the pitch contour patterns related to isolated lexical tones onto a Mandarin syllable such as ba and yield four syllables with four distinct lexical tones such as ba-1, ba-2, ba-3, and ba-4 with relatively good quality [24]. As for the treatment of the neutral tone (Tone 5), see Section III phonetic rule 6, for more detailed discussion. As we stated earlier, only 418 possible syllables exist in Mandarin Chinese if we disregard the tonal information. Therefore, these 418 syllables with their tonally variant counterparts are sufficient in generating all possible tonally different syllables in Mandarin Chinese, which in turn further enable us to generate all possible syllable combinations (words) in Mandarin Chinese.

A block diagram of the Chinese text-to-speech system based on the above syllable concatenation concept is shown in Fig. 3. In the syllable database, the LPC coefficients for the 418 syllables of Tone 1 and the standard patterns for the pitch contours of the four lexical tones are

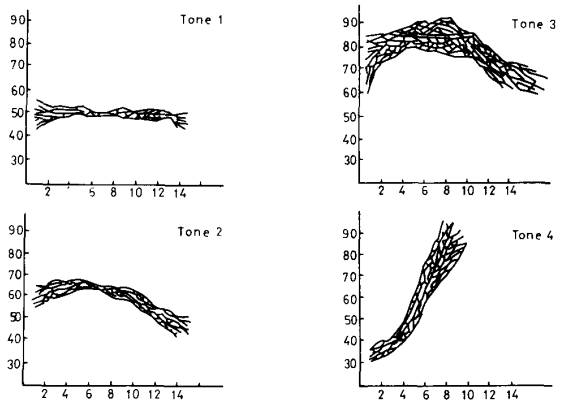


Fig. 2. The pitch contours of different syllables for the same speaker plotted for each individual tone, sampling period versus frame number.

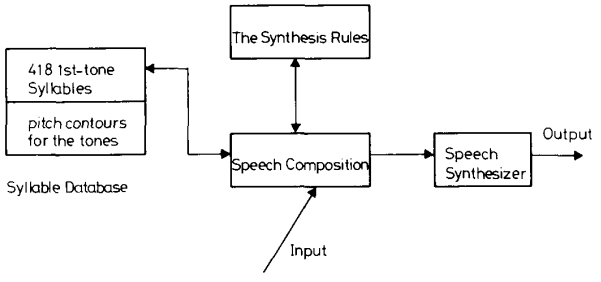


Fig. 3. The block diagram of the Chinese text-to-speech system based on the syllable concatenation concept.

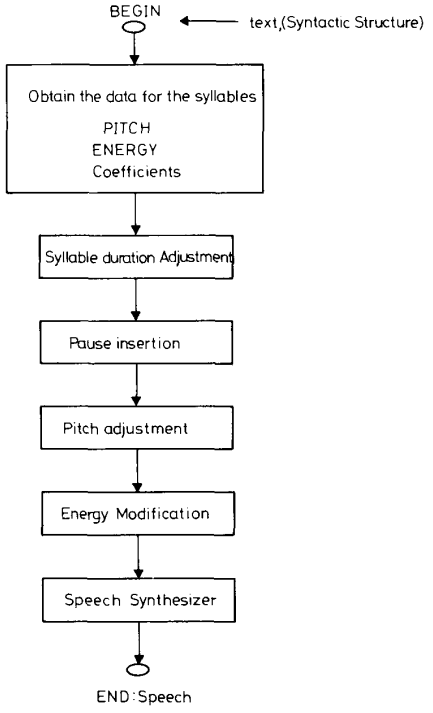


Fig. 4. The block diagram of the synthesis procedure.

stored. The synthesis rules are a set of general rules which determines how the parameters describing the acoustic properties of the syllables should be adjusted when the syllables are concatenated to form unrestricted sentences with arbitrary text. The speech composition is a set of software systems which tries to adjust the parameters obtained from the database according to the synthesis rules. A flow chart of the operations in the speech composition is shown in Fig. 4 which summarizes how the different synthesis rules are applied in the system. The system first extracts the parameters for the syllables from the database according to the input text. Syllable duration is then defined, followed by pauses being inserted, pitch periods adjusted, energy modified, and finally the speech synthesizer produces the speech output. All the relevant rules will be presented in detail in the following.

III. The Tone Concatenation Rules

It has been noted before that the so-called standard tone patterns for pitch contours are subject to various modifications in connected speech [22], [27]–[29]. In other words, considerable changes of pitch contour shapes occur in connected or running speech. After some initial analysis of the speech data in our sentence database, we have for the time being obtained six basic tone concatenation rules as an initial approximation by selecting the most important modifications and leaving the less significant ones for future improvement. These rules, although generally in agreement with the phonological rules of tone modification, more commonly known as the sandhi rules, are meant for synthesizing more natural sounding Mandarin speech output. A brief description of these rules is presented in this section.

In our sentence database, 92 Mandarin sentences selected from newspapers were produced by a male speaker in read form, band-limited to 4 kHz, and sampled at 10 kHz. The average speed of the sentences is approximately 3.4 syllables per second. The number of syllables per sentence ranges from 6 to 19 with the mean at about 13.8 syllables per sentence. Among the total of 1270 syllables (987 of them are distinct syllables), the distribution of Tones 1, 2, 3, 4, and 5 (the neutral tone) is around 28, 13, 18, 29, and 12 percent, respectively. As a result, each tone can be followed by any of the 5 tones, and a total of 25 possible tone combinations of adjacent syllables can be derived. Measurements of the variations of the pitch period contours for tone concatenation serve as the basis for the phonetic rules regarding syllable concatenation described in the following.

1) 3 → 2/__ 3: When a Tone 3 precedes another Tone 3 without any pause between them, the first Tone 3 is pronounced approximately as Tone 2.[2] An example is shown in Fig. 5(a), where the first two syllables ni-3 da-3 "you hit" in the sentence both are of Tone 3, but the first syllable ni-3 is pronounced almost exactly like Tone 2.

[2]See [22, p. 27].

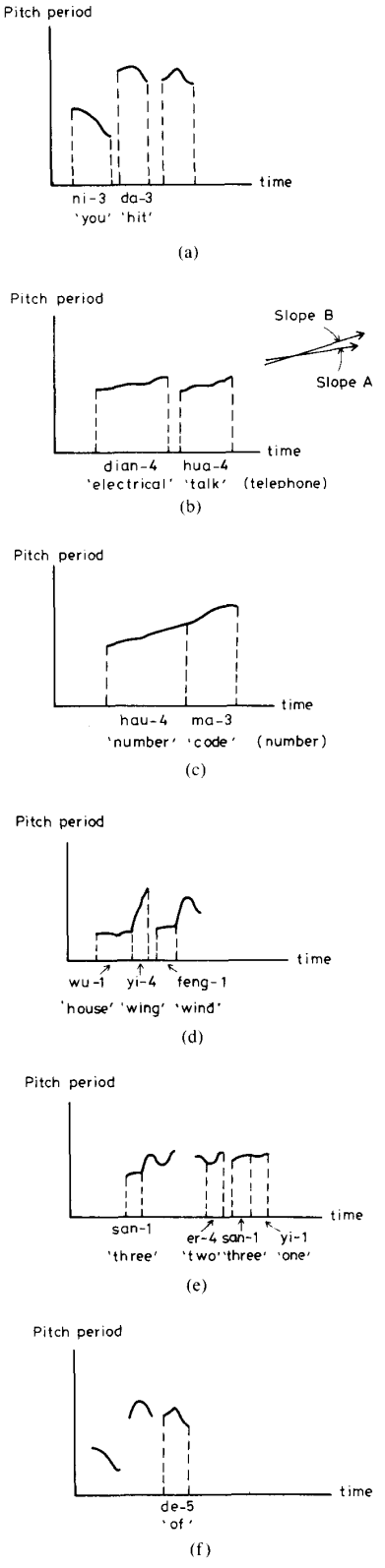


Fig. 5. (a)-(f) Examples of pitch contours demonstrating the tone concatenation rules 1)-6), respectively.

2) $4 \rightarrow 4'/\_\_\ 4$: When a Tone 4 precedes another Tone 4 without any pause between them, the first Tone 4 will be modified such that the slope of the pitch contour will be decreased by an order of about 20 percent.[3] An example is shown in Fig. 5(b), where the two syllables dian-4 hua-4 "telephone" both are of Tone 4, and the difference in the slopes in the two contours is quite clear.

3) $3 \rightarrow 3'/4\ \_\_$: When a Tone 3 follows a Tone 4, the Tone 3 will be modified such that the entire pitch contour should be slightly shifted up to make a continuous contour connecting the preceding syllable. An example is shown in Fig. 5(c), where the two syllables hau-4 ma-3 "number" have a continuous pitch contour which is caused by a shift of the pitch contour of the second syllable.

4) $1 \rightarrow 1'/\{3, 4\}\ \_\_$: When a Tone 1 follows a Tone 3 or a Tone 4, the pitch periods of the Tone 1 should be increased by an order of about 30 percent. An example is shown in Fig. 5(d) where the first and third syllables wu-1 "house" and feng-1 "wind" both are of Tone 1, but their pitch levels are different. The reason is that the third syllable feng-1 "wind" follows the second syllable yi-4 "wing" which is of Tone 4, therefore causing a slight increase in pitch level of the following Tone 1.

5) $1 \rightarrow 1'/1'\ \_\_$: When a Tone 1 follows another Tone 1, any modification made on the first syllable will be naturally repeated for the second. An example is shown in Fig. 5(e), where the last three syllables er-4 "two" san-1 "three" yi-1 "one" and the first syllable san-1 "three" demonstrate this phenomenon. The pitch level of the last two syllables (both of them are of Tone 1) is higher than that of the first syllable (also of Tone 1), because the last two syllables san-1 "three" yi-1 "one" follow er-4 "two" which is of Tone 4, therefore, san-1 is shifted according to the previous rule (4) and yi-1 is then modified accordingly.

6) $5 \rightarrow \underline{3}''$: Tone 5, the so-called "neutral tone," is phonologically related to weak stress, and is traditionally described as flattened to practically zero in tone range and reduced to relatively short in duration [25]. Also, the phonetic manifestation of a neutral tone is the result of the conditioning of the lexical tone of its preceding syllable. However, Tone 5 is treated in our system as a Tone 3 with reduced duration for the present form. The reasons are as follows. First, we intend to study stress patterns as well as intonation of Mandarin Chinese in detail in a later part of our project. Second, we have noticed some considerable change in spoken Mandarin particularly with respect to neutral tone in our data. Our preliminary observation showed less prominence regarding stress and neutral tone in current spoken Mandarin Chinese in Taiwan which seems to be of some distance from the description cited earlier. We therefore chose a simplified solution described in this rule, for the time being, as shown in Fig. 5(f), where the third syllable de-5 "of" (possessive or relative clause marker) is of the neutral tone, and the pitch

[3] See [22, p. 28].

contour possesses a more Tone 3 pattern with reduced duration.[4]

It was found that the above rules are most important in our Chinese text-to-speech system. Once they are implemented, the intelligibility of the synthesized speech is improved significantly.

IV. Other Tone Variation Rules

In addition to the above six general rules applied to the tones of the syllables, there are still some other tone variation rules which have to be taken into consideration. The most important cases include tone variation rules for Tone 3, and the rules for morphemes yi-1 "one," chi-1 "seven," ba-1 "eight," and bu-4, "negative marker" only.

1) Tone Variation Rules for Tone 3: The Tone 3 possesses the most complex pitch contours among the four lexical tones, as can be easily found in Figs. 1 or 2. However, such tone shape is produced fully only at sentence final position such as the third syllable in the sentence lau-3 shi-1 "teacher" tzau-3 "morning" ("good morning, Sir"). When it is followed by another Tone 3, only the second half will be pronounced and therefore is very close to a Tone 2 (see rule 1 in the previous section). Nevertheless, when it is followed by other tones, only the first half will be pronounced, for example, shou-3 du-1 "capital." The most difficult problem arises when more than two third tones are concatenated. For example, in sentences like wo-3 "I" you-3 "have" hau-3 ji-3 "several" ba-3 "quantifier" shiau-3 "small" yu-3 san-3 "umbrella" ("I've got several small umbrellas"), where all morphemes are of Tone 3, the above sandhi rule for connecting Tone 3's (rule 1, Section III) should not be applied recursively, otherwise, the phonetic output would be wo-2 you-2 hau-2 ji-2 ba-2 shiau-2 yu-2 san-3, which is a very unnatural way to say a sentence like this. In fact, syntactic boundaries within a sentence act like barriers, blocking the application of phonological/phonetic rules such as sandhi rules [26], with the exception (in Chinese) if the preceding word is monosyllabic. In other words, such sandhi rules are applied for morphemes within syntactic categories, unless the preceding syntactic category consist of only one monosyllabic word. Fig. 6 shows the syntactic structure of the above-mentioned sentence and a rough sketch of the tone shape of each syllable.

2) Sandhi Rules for Morphemes yi-1 "one," chi-1 "seven," ba-1 "eight," and bu-4 "negative marker" [29]: Morphemes yi-1 "one," chi-1 "seven," ba-1 "eight," and bu-4 "negative marker" form a special, although small, class whose behavior differs from other morphemes in terms of tonal variation by condition. Among them, the first three morphemes are of Tone 1, namely, yi-1 "one," chi-1 "seven," and ba-1 "eight," whereas bu-4 "negative marker" is of Tone 4. Their respective lexical tones remain unchanged under the following three conditions: a) when read in isolation; b) when they appear at phrase or sentence final position; and c) when they precede numerals. However, when these four morphemes precede morphemes other than the numerals, tone sandhi also occurs. The sandhi rules are summarized as follows.[5]

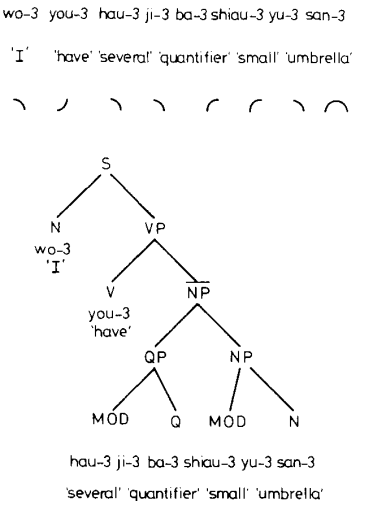


Fig. 6. Syntactic structure of the example sentence and the resulting tone shape of each syllable after the application of sandhi rules.

a) For morphemes yi-1 "one" and bu-4 "negative marker,"

$$\{1, 4\} \rightarrow 4/\_\_\_ \{1, 2, 3\}$$

$$\{1, 4\} \rightarrow 2/\_\_\_ \{4\}.$$

b) For morphemes chi-1 "seven" and ba-1 "eight,"

$$1 \rightarrow 2/\_\_\_ \{4\}.$$

Among these two rules, rule a) is obligatory, whereas rule b) is optional for some speakers. These rules, nevertheless, were established in our system.

V. Stress Rules and Intonation Patterns

The role of stress and intonation in Mandarin Chinese is probably similar to that in polysyllabic and nontonal languages. Generally speaking, for bisyllabic words, the stress usually falls on the second syllable, although the reverse does occur considerably frequently. For tri- or polysyllabic words, the primary stress usually falls on the last syllable, and the secondary stress falls on the first syllable; while those syllable(s) in between become unstressed. Fig. 7(a) is an example of a sentence where the energy level is plotted as a function of time. In the sentence, tzai-4 "in, at" juan-3 shuen-4 jian-1 "a moment" shiau-1 mie-4 "disappear" le-5 "aspect marker" tzueng-1 ying-3 "trace" ("all traces disappeared in a moment"), notice that in both the trisyllabic word juan-3

[4]For a more detailed discussion on stress related phenomena, see Tseng, 1988, forthcoming, "On some stress related acoustic features of disyllabic words in Mandarin Chinese."

[5]See [29, pp. 265–268].

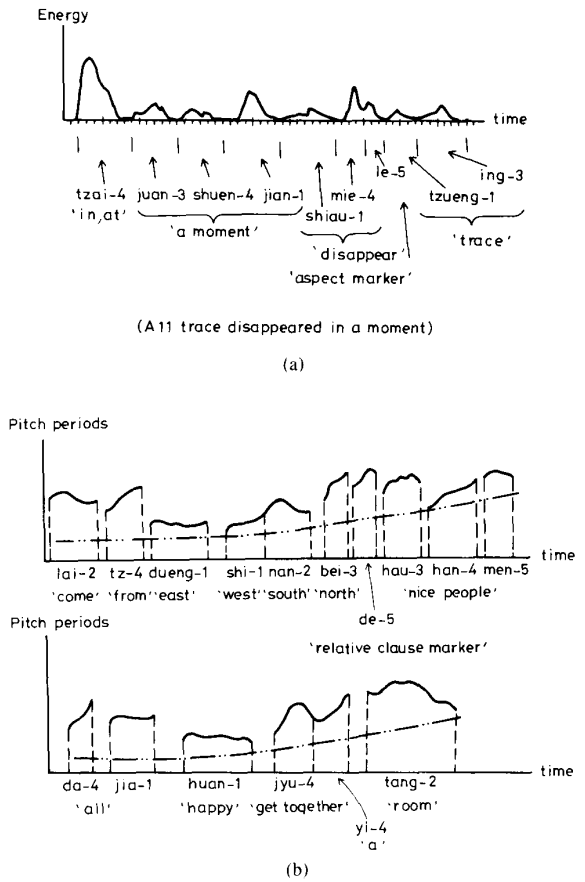


Fig. 7. Examples demonstrating (a) stressed syllables and (b) overall intonation patterns, plotted as a function of time.

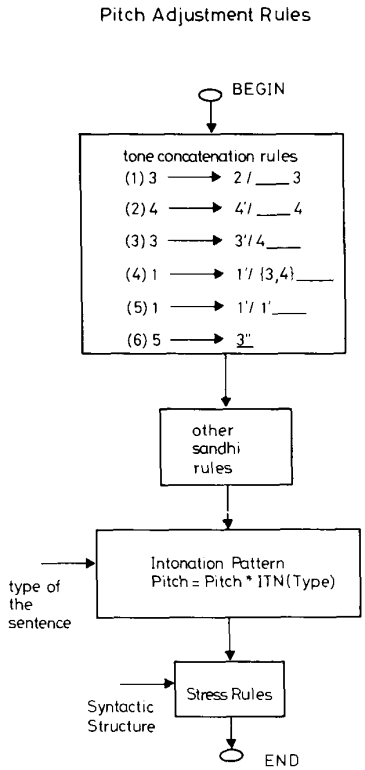


Fig. 8. The complete pitch modification rules.

shuen-4 jian-1 "a moment" and bisyllabic word shiau-1 mie-4 "disappear," the stress is on the last syllables jian-1 and mie-4, respectively. On the other hand, the intonation pattern for a declarative sentence is generally declining, although this is not true for an interrogative sentence. An example is shown in Fig. 7(b), where the pitch period contours of two parts of a declarative sentence, lai-2 "come" tz-4 "from" dueng-1 "east" shi-1 "west" nan-2 "south" bei-3 "north" de-5 "relative clause marker" hau-3 han-4 men-5 "nice people" da-4 jia-1 "all" huan-1 "happy" jyu-4 "get together" yi-4 "a" tang-2 "room" ("the nice people coming from many different places are all getting together happily in a room"), are plotted versus time. It can be seen that although the lexical tones of each syllable can still be traced through the corresponding pitch patterns, there is a tendency for the pitch period of the entire sentence to go upward, which means a declining intonation is interacting with the individual lexical tones. However, we have stated earlier that we will investigate the properties of stress and intonation in a later part of the project. Interestingly enough, it seems from our experiments that the intelligibility of the synthetic speech appears to be unaffected by the fact that the study of stress and intonation is incomplete.

The complete pitch modification rules for the current system are then shown in Fig. 8. The tone concatenation rules are applied first, additional sandhi rules are then used to make a further check and modification. The intonation pattern is then applied by making global pitch contour modifications depending on the sentence types, i.e., declarative or interrogative. The stress rule is finally applied by enhancing the dynamic range of the pitch variations for stressed syllables.

VI. The Syllable Duration Rules

In naturally spoken Mandarin Chinese, syllabic duration varies considerably depending on various linguistic and nonlinguistic factors. If a sentence is concatenated by syllables with equal durations, it sounds very artificial, although still relatively intelligible. A detailed analysis of the duration of the syllables is performed on the same sentence database of the 92 sentences as described above. The syllables in each sentence are first segmented manually from the waveform, and the durations are measured. Statistics obtained from duration measurements of syllables serve as the basis of the duration modification rules for the system presented in this paper. Table I(a) lists results of our duration measurements in terms of consonant type. Table I(b) lists results of our duration measurements in terms of lexical tone type. Our measurements reveal that both consonant type and tone have an effect on duration. Both are well-known facts that have been well researched, especially with regard to consonant effect. Since

TABLE I
STATISTICS ON SYLLABLE DURATIONS (a) WITH RESPECT TO SYLLABLE-INITIAL CONSONANTS, (b) WITH RESPECT TO THE TONES

Syllable-Initial Consonant Type	Syllable Duration Statistics (ms)	
	Mean	Standard Deviation
p, t, k, ts	314	36
h	319	41
ch, chi	332	49
s, f, sh, shi	348	46
r, l, m, n	293	33
without initial consonant	276	34
other initial consonants	302	36

(a)

Tones	Syllable Duration Statistics (ms)	
	Mean	Standard Deviation
1	288	59
2	302	66
3	327	50
4	272	45
5	210	33

(b)

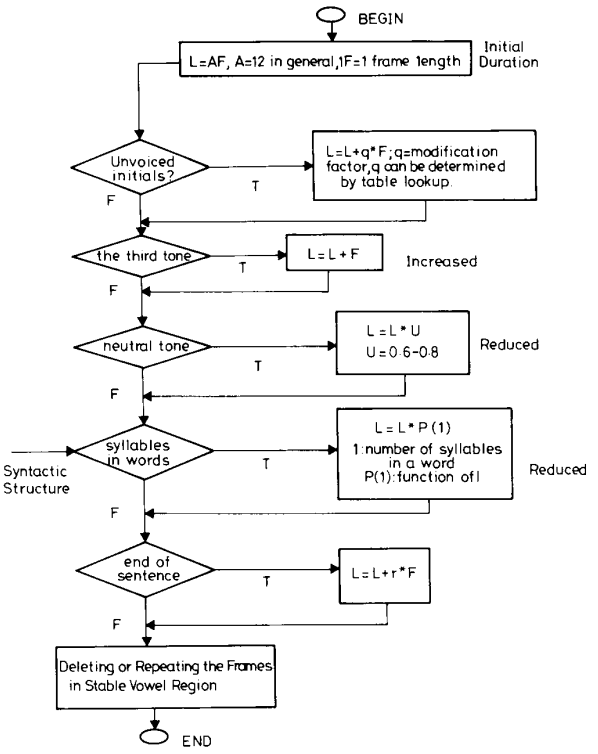


Fig. 9. Syllable duration rules developed for the current Chinese text-to-speech system.

our system is designed on the basis of one speaker only, we chose to base our duration rules on our measurements. A set of preliminary duration rules was then summarized in Fig. 9. Every syllable is first assigned an initial duration of 12 frames (each frame is 20 ms). If the syllable has an unvoiced initial, the duration will be increased by 2–3 frames depending on the initial found by table lookup. For example, the duration will be increased by 3 frames if the syllable has an initial fricative ch-, and by 2 frames if it has an initial velar fricative h-. The duration will be further increased if the syllable is of Tone 3, or it will be decreased if the syllable is of the neutral tone, because Tone 3 is usually longer and the neutral tone is shorter. The duration will then be reduced if the syllable is in a polysyllabic word, and the reduction factor varies from 0.62 to 0.93 depending on the number of syllables in the word. Because the current system developed at Taiwan University cannot analyze lexical structure, information regarding each lexical item has to be keyed in by the user to improve the synthesized speech quality. The duration of the syllable will be further increased if it is toward the end of the sentence. When the desired duration is finally calculated, the duration of the syllable obtained from the parameters in the syllable database will be adjusted to the desired value by deleting or repeating the frames in the specified steady-state portion of the vowel of each syllable. This steady-state portion of each syllable is in fact detected by hand to assure the syllable quality after significant modification of the duration. For example, if the duration of the syllable ban-1 is to be increased or decreased, all one has to do is to increase or decrease the duration of the vowel [a].

Two examples are shown in Fig. 10 to illustrate variation of duration. In Fig. 10(a), the sentence is: je-4 shie-1 "these" jr-3 "only" shr-4 "are" shr-2 yien-4 shr-4 "laboratory" jueng-1 "inside" de-5 "of" cheng-2 jiou-4 "accomplishments" ("these are only the accomplishments in the laboratory"). By examining the duration of the syllables, we note that the second syllable shie-1 in the bisyllabic word je-4 shie-1 "these" is relatively longer in duration partly because it is the first stressed syllable of the sentence, and partly because the initial consonant sh- consists of frication. For the trisyllabic word shr-2 yien-4 shr-4 "laboratory," note that the syllabic duration for each of the syllables is relatively shorter, partly because polysyllabic words tend to be more co-articulated, and partly because it happens to be in the sentence-middle position. This can also be seen in contrast with the last syllable of the sentence jiou-4 which is the second syllable of the bisyllabic word cheng-2 jiou-4 "accomplishments." When occurring in sentence-final position, and in this case, breath-group final as well, the syllabic duration is relatively longer for physiological reasons as well. This example can be seen as an illustration of the various effects on syllabic duration. The sentence in Fig. 10(b) is: ni-3 "you" ching-1 kuai-4 di-5 "briskly" tzou-3 guo-4 "walk through" je-4 "the" sen-1 lin-2 "forest" ("you briskly walked through the forest"). Note that besides the effect of consonant type which we discussed in the sentence in Fig. 10(a), the relatively short duration of

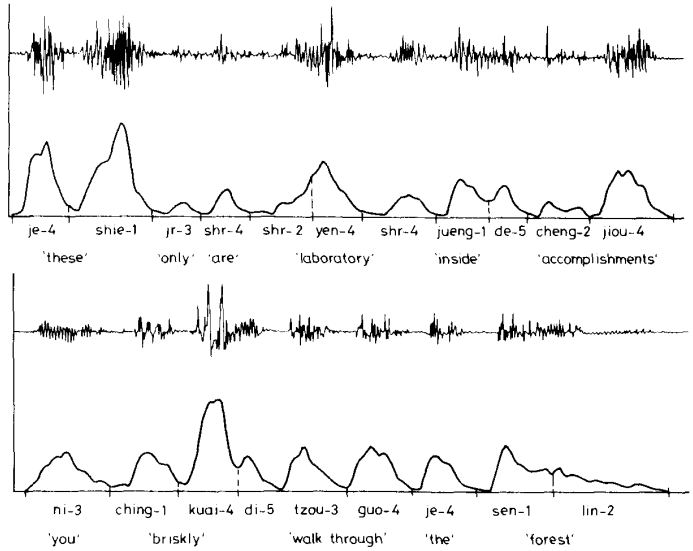


Fig. 10. Two example sentences for the syllable duration rules.

the neutral tone (Tone 5) in comparison to the other lexical tones is seen in the fourth syllable di-5 "adverbial particle." A similar effect can be found in the ninth syllable de-5 "of" of the sentence in Fig. 10(a).

VII. The Pause Insertion Rules and Energy Modification Rules

Pauses of different duration should also be inserted into the sentence in various positions between the syllables to make the sentences more natural. When the sentences in the sentence database as manually segmented into syllables as described above, the pauses between them are automatically obtained. After a similar statistical analysis on these pauses, the preliminary pause insertion rules are summarized in Fig. 11. Pauses for different punctuation marks are first assigned, pauses within a sentence are then assigned to major syntactic boundaries. The latter is a more complicated situation. In short, first, no pause should be inserted after pronouns (for example, wo-3 "I," ni-3 "you," and ta-1 "he") or before the relative clause marker de-5 "that." The major syntactic boundaries are then assigned pauses with different duration [30]. An example is shown in Fig. 12, where within the sentence: ni-3 "you" wo-3 "I" shiang-1 feng-2 "meet" tzai-4 "at" hei-1 "dark" ye-4 "night" de-5 "of" hai-3 shang-4 "sea" ("you and I met in a dark night at sea"), a longer pause is needed after shiang-1 feng-2 "met" because it is a major syntactic boundary. Fig. 12(a) shows the syntactic structure of the sentence, while Fig. 12(b) is the waveform and energy plot of the sentence. Note that the pause after bisyllabic word shiang-1 feng-2 "meet" is quite clear.

In naturally spoken Mandarin sentences, the energy level of the syllables also varies. After carefully examining the statistics of the energy levels of the syllables in

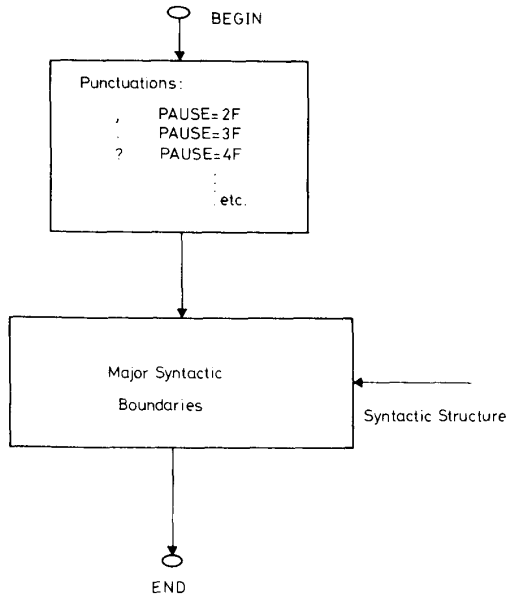


Fig. 11. Pause insertion rules.

the above sentence database just as was done for the durations, the preliminary energy modification rules obtained here are summarized in Fig. 13. Every syllable has its initial energy profile as obtained from the database. The energy will be reduced by a given factor if the syllable is of the neutral tone. It will then be increased if the syllable is stressed in a multisyllabic word as mentioned above, depending on the number of syllables in the word. Some intrinsic loudness adjustment is also made on the vowels, for example, the vowel [a] is always louder than the vowel [i]. Similar to the method employed for rules of duration adjustment, statistical analysis of measure-

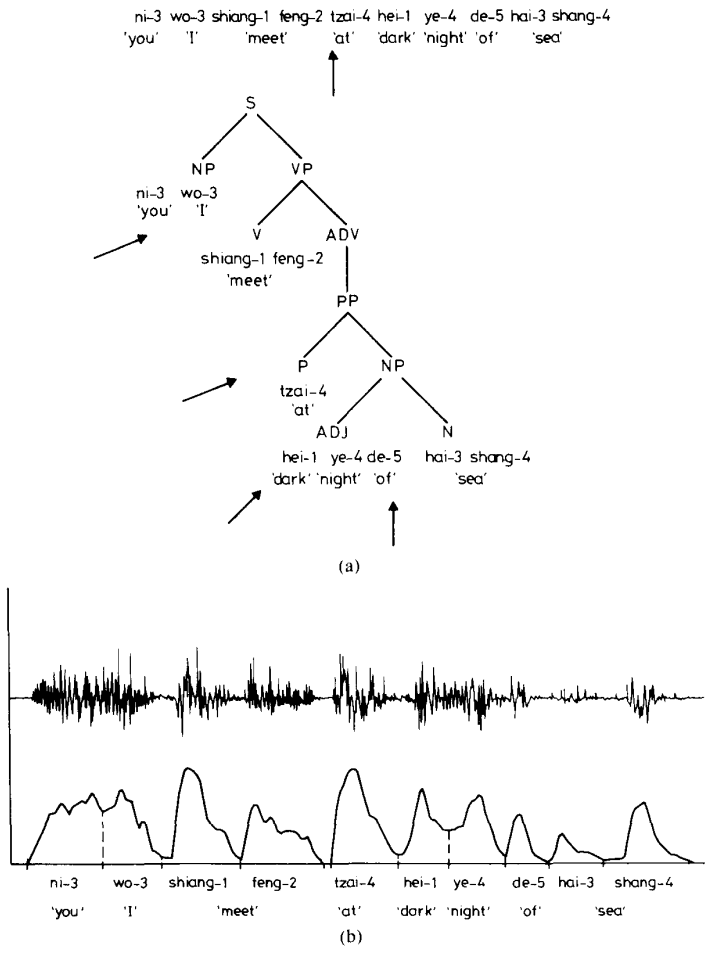


Fig. 12. Pause insertion into the syntactic boundaries (a) the syntactic structure of the sentence and (b) the sentence waveform and energy plot.

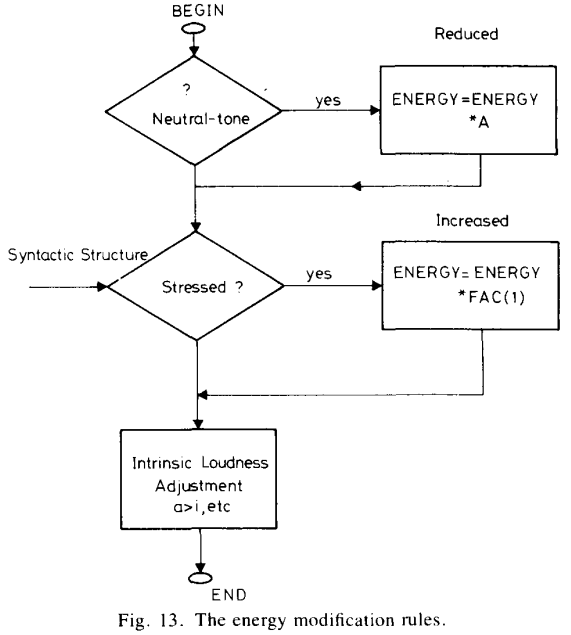


Fig. 13. The energy modification rules.

ments of energy levels on the vowels of the sentence database is obtained and serves as the basis of our rules for adjustment of loudness. The energy levels of the vowels in a sentence are first normalized with respect to the highest energy within the sentence to remove possible variation of energy levels across sentences of the database. As a result, statistical analysis of relative energy levels for different vowels can be derived. Two examples are shown in Fig. 14 to illustrate this point. The sentence in Fig. 14(a) is: yi-2 ge-5 "a + classifier" shen-2 jing-1 "neural" shi-4 bau-1 "cell" shiang-1 dang-1 yu-2 "correspond to" yi-2 ge-5 "a + classifier" da-4 shing-2 "large scale" ji-1 ti-3 dian-4 lu-4 "integrated circuit" ("a neural cell corresponds to a large scale integrated circuit"). Note that in this sentence, all of the following syllables shiang-1, dang-1 (in the trisyllabic word shiang-1 dang-1 yu-2 "correspond to"), da-4 (in the bisyllabic word da-4 shing-2 "large scale"), dian-4 (in the polysyllabic word ji-1 ti-3 dian-4 lu-4 "integrated circuit") consist of the vowel [a]. Even the diphthong of the syllable bau-1 (in the bisyllabic word shi-4 bau-1 "cell") consists of an [a] onglide. All of these syllables corre-

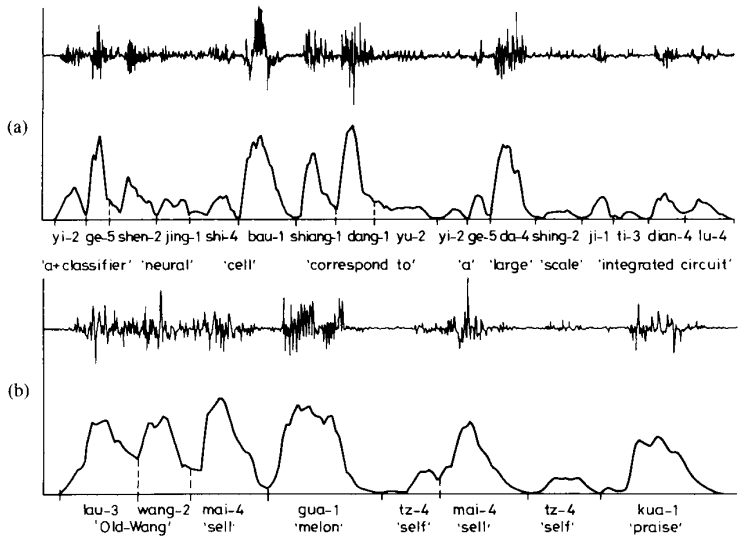


Fig. 14. Two example sentences for the syllable energy rules.

spond to a relatively high level of energy. On the other hand, those syllables consisting of the vowel [i], namely, jing-1 (in the bisyllabic word shen-2 jing-1 "neural"), shi-4 (in the bisyllabic word shi-4 bau-1 "cell"), yi-2 (in the bisyllabic words yi-2 ge-5 "a + classifier"), shing-2 (in the bisyllabic word da-4 shing-2 "large scale"), ji-1 and ti-3 (in the polysyllabic word ji-1 ti-3 dian-4 lu-4 "integrated circuit"), correspond to a relatively low level of energy. It might be possible to argue that this observation does not take into consideration the possible effect of lexical tones. However, the sentence in Fig. 14(b) is: lau-3 wang-2 "Old Wang" mai-4 "sell" gua-1 "melon" tz-4 "self" mai-4 "sell" tz-4 "self" kua-1 "praise" ("Old Wang who sells melons praises his own ware highly"). All of the syllables, except the two occurrences of syllable tz-4 "self" that ends in a back unrounded vowel [ɯ] twice in the sentence, consist of the vowel [a] or [a]-onglide in Tones 1, 2, 3, and 4. It is very obvious that vowel [a] is higher in energy levels regardless of possible effects on energy from tones.

VIII. The Implementation Experiences and Preliminary Performance Evaluation

The completed Chinese text-to-speech system is implemented on an IBM personal computer with an extra Digital Signal Processor board. The speech synthesizer is an LPC lattice structure synthesizer with order 10, implemented on a Digital Signal Processor integrated circuit. The sampling rate is 10 kHz. Each frame of speech signal contains 20 ms of speech waveform and is synthesized from 50 bits of data specifying the parameters. The entire system is in fact table driven, with an attribute table being the kernel. The attribute table is like a blackboard in which all relevant phonetic and phonological information for the desired sentences can be written down, and the speech synthesizer can finally synthesize the output speech according to the information here, such as the complete input text, the parameters for the syllables, and the modifications to be made on each syllable. The system works effectively in real time, with the output speech at a rate on the order of 3.3 syllables per second.

Some preliminary performance evaluation was conducted on this system to see how these rules jointly provide the synthesized speech quality. In a regular classroom setting without using earphones, recorded speech from the system was played to a total pool of 120 subjects, all of them are undergraduate university students. The subjects were asked to rate what they heard using the following criteria.

1) Intelligibility: In this test, the subjects' tasks were to listen to the synthetic speech output of several articles without prior knowledge of the content of the speech they heard. After the initial listening, the subjects were then provided with written text of the speech output when they were asked to listen to the same speech output again and mark all tokens that were not intelligible to them during the first trial.

2) Comprehensibility: Speech of several articles with various topics and lengths were played to the subjects first, then a pencil-and-paper test consisting of questions relating to the contents of the articles was given to the subjects. Two tests of both natural and synthesized speech with identical text and questions were performed on two different groups of subjects, and the relative ratio between their average scores was obtained.

The results were as follows. For intelligibility, the average score was 96 percent with a standard deviation of 3.6 percent with respect to the syllables. For comprehensibility, the average raw score was 89.3 percent with a

standard deviation of 8.1 percent for synthesized speech and 93.2 and 6.2 percent, respectively, for the natural speech, and the relative ratio was 95.9 percent. Although these tests are quite preliminary and more rigorous evaluation is needed, the results here serve as a rough sketch of the system performance.

IX. Conclusion

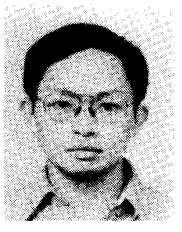
In this paper, the synthesis rules developed for a successfully implemented Chinese text-to-speech system are described in detail. The system design approaches encompass some of the most important features of Chinese language such as the monosyllabic aspect. Special attention was also given to the tonal aspect of Mandarin Chinese. All the phonological and phonetic rules were derived from linguistic properties, and thus are independent of the choice of design approaches. These rules can also be useful in understanding the characteristics of continuous Mandarin speech. It is our belief that studies of this kind can go beyond the synthesis of speech. Some of our major findings may also apply to continuous Mandarin speech segmentation and recognition in the future.

Acknowledgment

The authors would like to thank R. G. Chen, C. L. Huang, and C. K. Cheng for their help in establishing the sentence database and performing the statistical analysis of the syllable parameters.

References

[1] Special Issue on Man-Machine Communication by Speech, *Proc. IEEE*, vol. 73, Nov. 1985.
[2] J. Allen, "Synthesis of speech from unrestricted text," *Proc. IEEE*, vol. 64, pp. 422-433, 1976.
[3] J. Allen, S. Hunnicutt, R. Carlson, and B. Granstrom, "MITalk-79: The 1979 MIT text-to-speech system," in *ASA-50 Speech Communication Papers*, J. J. Wolf and D. H. Klatt, Eds. New York: Acoustical Society of America, 1979, pp. 507-510.
[4] J. Holmes, I. Mattingly, and J. Shearme, "Speech synthesis by rule," *Language and Speech*, vol. 7, pp. 127-143, 1964.
[5] D. H. Klatt, "The KLATTalk text-to-speech conversion system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1982, pp. 1589-1592.
[6] N. Umeda, "Linguistic rules for text-to-speech synthesis," *Proc. IEEE*, vol. 64, Apr., 1976.
[7] S. R. Hertz, J. Kadin, and K. J. Karplus, "The delta rule development system for speech synthesis from text," *Proc. IEEE*, vol. 73, pp. 1589-1601, Nov. 1985.
[8] S. R. Hertz, "From text to speech with SRS," *J. Acoust. Soc. Amer.*, vol. 72, no. 4, pp. 1155-1170, 1982.
[9] H. S. Elovitz, R. Johnson, A. McHugh, and J. E. Shore, "Letter-to-sound rules for automatic translation of English text to phonetics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 446-473, Dec. 1976.
[10] O. Fujimura, M. J. Macchi, and J. B. Lovins, "Demisyllables and affixes for speech synthesis" (Abstract), in *Contributed Papers, vol. 1, 9th Int. Congr. Acoust.*, Madrid, Spain, July 4-9. Madrid: Spanish Acoustics Society, 1977, p. 515.
[11] H. Dettweiler, "An approach to demisyllable speech synthesis of German words," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1981, pp. 110-113.
[12] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 67, no. 3, Mar. 1980.
[13] H. Sato, "Japanese text-to-speech conversion system," *Review Elec. Commun. Lab., Nippon Telegraph and Telephone Corp.*, vol. 32, no. 2, 1984.
[14] Y. Sagisaka *et al.*, "Accentation rules for Japanese text-to-speech conversion," *Review Elec. Commun. Lab.*, vol. 32, no. 2, 1984.
[15] Many other papers on text-to-speech systems in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, in recent years.
[16] T.-Y. Huang, C.-f. Wang, and Y.-h. Pao, "A Chinese text-to-speech synthesis system based on an initial-final model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, France, Apr. 1982, pp. 1601-1603.
[17] J. Zhang, "Acoustic parameters and phonological rules of a text-to-speech system for Chinese," presented at the IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo, Japan, Apr. 1986, Paper 38.7.
[18] K. C. Zhou and T. Cole, "A chip designed for Chinese text-to-speech synthesis," *J. Elect. Electron. Eng., Australia*, vol. 4, no. 4, pp. 314-318, Dec. 1984.
[19] M. Ouh-Young, C.-Y. Tseng, and L.-S. Lee, "Design considerations and preliminary results for a Chinese text-to-speech system," in *Proc. 1984 Int. Comput. Symp.*, Tamkang Univ., Taipei, Taiwan, Republic of China, Dec. 1984, pp. 1331-1341.
[20] ——, "A Chinese text-to-speech system based on a syllable concatenation model," in *Proc. 1986 Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan, Apr. 1986, pp. 2439-2442.
[21] *Guoyurbao Tzdian (Mandarin Chinese Daily Dictionary)*, R. He, Ed. Taipei, Taiwan, Republic of China: Guoyurbao, 1976.
[22] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkeley, CA: University of California, Berkeley Press, 1968.
[23] J. M. Havie, *Acoustical Studies of Mandarin Vowels and Tones*. Cambridge, U.K.: Cambridge University Press, 1976.
[24] S.-M. Lei and L.-S. Lee, "Digital synthesis of Mandarin speech using its special characteristics," *J. Chinese Inst. Eng.*, vol. 6, no. 2, pp. 107-115, Mar. 1983.
[25] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkeley, CA: University of California, Berkeley Press, 1968, pp. 35-39.
[26] W. E. Cooper, S. G. Laponite, and J. M. Paccia, "Syntactic blocking of phonological rules in speech production," *J. Acoust. Soc. Amer.*, vol. 61, pp. 1314-1320, 1977.
[27] C.-y. Tseng, "An acoustic phonetic study on tones in Mandarin Chinese," Ph.D. dissertation, Brown Univ., June 1981.
[28] V. A. Fromkin, *Tone—A Linguistic Survey*. New York: Academic, 1978.
[29] *Mandarin Phonetics*. Taiwan, R.O.C.: National Taiwan Normal University Press, 1982.
[30] M. Ouh-Young, "A Chinese text-to-speech system," Master thesis, Dep. Elec. Eng., Nat. Taiwan Univ., June 1985.



Lin-Shan Lee (S'76-M'77-SM'88) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, Republic of China, in 1974, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1975 and 1977, respectively.

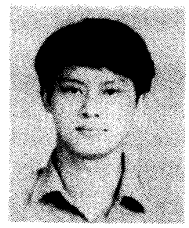
He was involved in research of communication systems and satellite systems while a graduate student at Stanford University, and was with EDUTEL Communications and Development, Inc., Palo Alto, CA, from 1977 to 1980 with research interests in various aspects of communication systems, technologies and developments, especially in satellite communications. He became an Associate Professor at the Department of Electrical Engineering, National Taiwan University in September 1979, and a Professor in August 1982. He also became the Acting Chairman of the Department of Computer Science and Information Engineering of the university in September 1982, and the Chairman from August 1983 to July 1987. He currently teaches courses in communication technologies and signal processing and does research in the areas of digital transmissions, satellite communications, and digital speech processing, especially concentrating on the problem of computer input/output techniques using Mandarin Chinese. He has authored about 100 technical papers in these areas in the last 10 years, among which about 25 are published in IEEE Transactions, and about 40 in IEEE-sponsored International Conferences.

Dr. Lee was the recipient of the Outstanding Young Engineer Award sponsored by the Institute of Chinese Engineers in 1983, the Ten Outstand-

ing Young Men Award of the Republic of China in 1984, the Outstanding Young Scientist Fellowship sponsored by URSI (Union of Radio Science International) in 1984, the Distinguished Research Award sponsored by the National Science Council of the Republic of China in 1985 and 1987 (to which only the top 5 percent of the researchers of the country are entitled for every two years), the Outstanding Youth Medal of the Republic of China in 1986, and the Distinguished Teaching Award sponsored by the Ministry of Education of the Republic of China in 1988 (to which only the top 2 percent of the professors in the country are entitled).

Chiu-Yu Tseng received the B.A. degree in English from National Taiwan Normal University in 1972, and the Ph.D. degree in linguistics from Brown University, Providence, RI, in 1982.

She has been an Associate Research Fellow at the Institute of History & Philology, Academia Sinica, Taipei, since 1982, and has been holding a joint position at the Institute of Information Science, Academia Sinica, and an Adjunct Associate Professorship at the Department of Computer Science and Information Engineering, National Taiwan University, since 1985. Her research has focused in the area of the synthesis and recognition of Mandarin Chinese as well as the acquisition of Mandarin by Chinese children.



Ming Ouh-Young received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981 and 1985, respectively.

He is working toward the Ph.D. degree in computer science, with an emphasis on man-machine interface, at the University of North Carolina, Chapel Hill. His interests include interactive computer graphics, and man-machine interface by kinesthesia.