Speech Communication 7 (1988) 317–325
North-Holland

AN ACOUSTIC-PHONETIC ORIENTED SYSTEM FOR SYNTHESIZING CHINESE

Shun-an YANG and Yi XU
Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China

Received 20 August 1987
Revised 23 November 1987

Abstract. The study of synthesizing Chinese is faced with the pressing task of improving sound quality. This article presents the structure and features of our synthesis system for Standard Chinese. This system was built on the basis of an acoustic-phonetic analysis of Chinese syllables. Several original models and rules are employed in the system. All the 1268 syllables in Standard Chinese have been synthesized by this system, which produces a sound quality close to that of natural speech with respect to both intelligibility and naturalness.

Zusammenfassung. Die Verbesserung der Tonqualität ist das wichtigste Problem mit dem sich die Synthese der Chinesischen Sprache konfrontiert sieht. Dieser Beitrag erläutert die Struktur und die Merkmale unsers Synthesesystems für das Hochchinesische. Das System basiert auf einer akustisch-phonetischen Analyse von Chinesischen Silben. Einige originelle Modelle und Regeln sind in das System eingebaut. Alle 1268 Silben des Hochchinesischen sind synthetisch hergestellt worden mit einer Qualität welche, Deutlichkeit und Natürlichkeit betreffend, dem Original sehr nahe war.

Résumé. L'amélioration de la qualité du son synthétique est le problème le plus urgent auquel l'étude de la synthèse du Chinois est confrontée. Cet article présente la structure et les traits de notre système de synthèse pour le Chinois standard. Le système a été construit sur base d'une analyse acoustico-phonétique des syllabes du Chinois. Le système incorpore plusieurs règles et modèles originaux. Toutes les 1268 syllabes du Chinois standard ont été synthétisées avec une qualité proche de l'original en ce qui concerne l'intelligibilité et le naturel.

Keywords. Synthesis of Chinese, cascade formant synthesizer.

1. Introduction

In the field of speech science, the technics of speech synthesis are a very useful means for studying human speech. Nowadays, no one dares publish some grand theory of speech production without putting it to the test of synthesis (Coker, 1972). In the field of speech technology, speech synthesis is an important basis for developing a speech output system for computers.

In recent years, the study of synthesizing Chinese has made a good start (Li and Wolf, 1982; Huang et al., 1982; Lee et al., 1982; Zhang, 1986). However, due to a lack of proper linguistic and phonetic theory and a lack of systematic rules and accurate data, the sound quality of synthetic Chinese was not very satisfactory.

On the basis of the systematic acoustic-phonetic analysis of Standard Chinese, we have developed a software system for synthesizing Chinese, which uses a cascade formant synthesizer. With the original Tone Model, the Exponential Dynamic Model and the Initial-Final-Transaction Model employed in this system, we have successfully synthesized all the syllables (1268) in Standard Chinese; these came very close to natural ones both in intelligibility and in naturalness. The system has been used in the study of speech perception, and has laid a firm foundation for our further study of the building a text-to-speech system for Standard Chinese.

0167-6393/88/$3.50 © 1988, Elsevier Science Publishers B.V. (North-Holland)

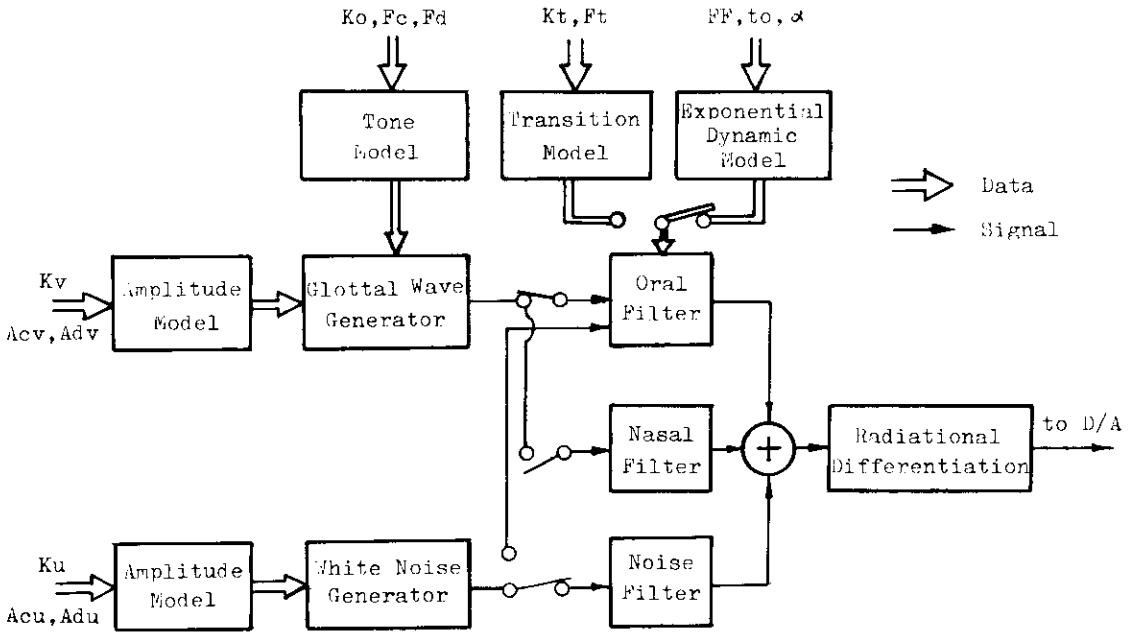

Fig. 1. Block diagram of the synthesis system for Standard Chinese.

2. Essential components of the synthesis system and their main characteristics

Formant parameters are widely used for specifying speech sounds, and formant synthesizers are therefore employed in our system. The formant synthesizer is based on the acoustic theory of speech production (Fant, 1960), according to which any speech sound is the product of source excitation, vocal tract transmission and oral and/or nasal radiation. There are two kinds of formant synthesizers: cascade and parallel, and there has been much debate as to the advantages and disadvantages of the two kinds of synthesizers (Holmes, 1983; Klatt, 1983; etc.). The kind of synthesizer we used in our system is a cascade synthesizer. We think that the basic configuration of the two kinds of synthesizers is well developed, and the key to fine quality of synthetic speech lies in having a good specific configuration for the synthesizer, a proper type of parameter input, accurate data and appropriate phonological and phonetic rules. Fig. 1 is a block diagram for this synthesis system, in which the white noise generator is a pseudo-random number generator, its amplitude distribution is Gaussian, and the glottal wave generator produces the kind of waveform tested by Rosenberg (1971) (see Fig. 2) that can be expressed as

$$g(t) = \begin{cases} 0.5[1 - \cos(t/T_p)], & 0 \leqslant t \leqslant T_p \\ \cos[(t - T_p)/2T_n], & T_p \leqslant t \leqslant T_p + T_n, \\ 0, & T_p + T_n \leqslant t \leqslant T_0. \end{cases} \quad (1)$$

The three vocal tract filters are all composed of several second-order cascade digital filters.

In Chinese each character corresponds exactly to a syllable, and the syllables are relatively stable in continuous speech (despite an inevitable coarticulation effect), therefore we chose the syllable

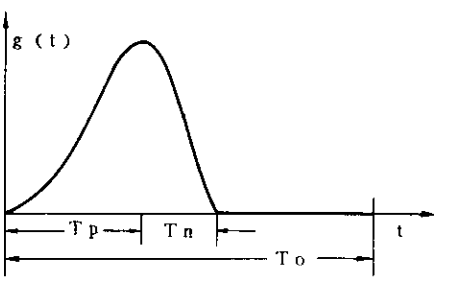

Fig. 2. The glottal pulse shape used in this synthesis system.

as the synthetic unit. On the basis of a great deal of acoustic-phonetic analysis (Wu, 1964; Lin, 1965, Wu and Lin, n.d.; Xu and Yang, n.d.), we proposed an acoustic-phonetic framework consisting of 7 sections for Standard Chinese syllables, as is shown in Fig. 3. According to this framework, any given SC syllable is a combination of certain (from one to all) of the 7 sections in the order shown in Fig. 3. After inputting proper data to a section, the parameters needed by the synthesizer are calculated frame by frame by means of relevant models or rules. The synthesizer then does the calculation one section after another, and the final output is the synthetic sound of the whole syllable.

The system is written in FORTRAN and is implemented on a microcomputer type BCM-3. The accuracy of the D/A converter is 12-bit, and its sampling frequency is 10 kHz. In the following sections we will describe briefly the main characteristics of this system and the synthesis method for several typical SC syllables.

2.1. Tone model

Chinese is a tone language, and the F_0 pattern of the synthetic syllables has a great influence upon the intelligibility as well as naturalness. In the light of the large amount of variability in the F_0 patterns of SC syllables, a normalized Tone Model (Yang, 1986) is used in this synthesis system for generating F_0 parameters. The formula for the model is as follows:

$$F_0(t) = \log^{-1}[F_c + F_d \cdot f_{ko}(t)]. \quad (2)$$

In the formula, t is normalized time; F_c is the cardinal pitch value representing the neutral pitch of the voice; F_d is pitch range, representing the sphere of frequency variation for a given tone; $f_{ko}(t)$ is an array designated by the symbol ko, defining a curve pattern for a given tone. This $f_{ko}(t)$ can be stored in the program as well as inputted by the user according to his needs. There are 11 elements in the array corresponding respectively to normalized time 0, 0.1, 0.2, ..., 1.

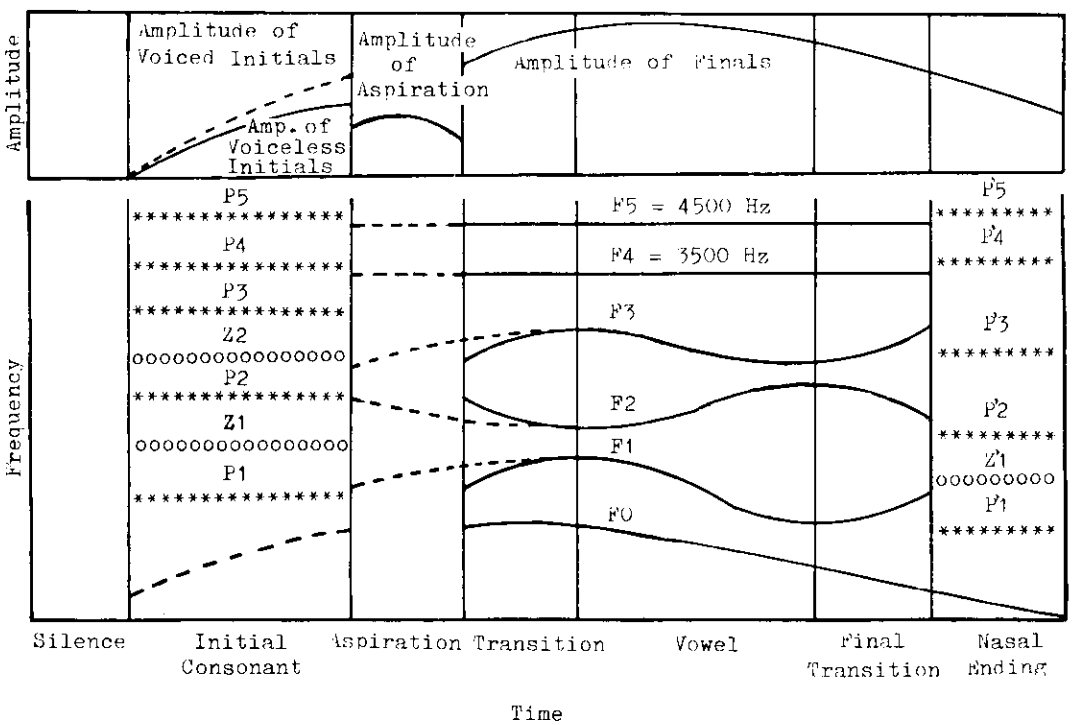

Fig. 3. The structural frame for synthetic syllables.

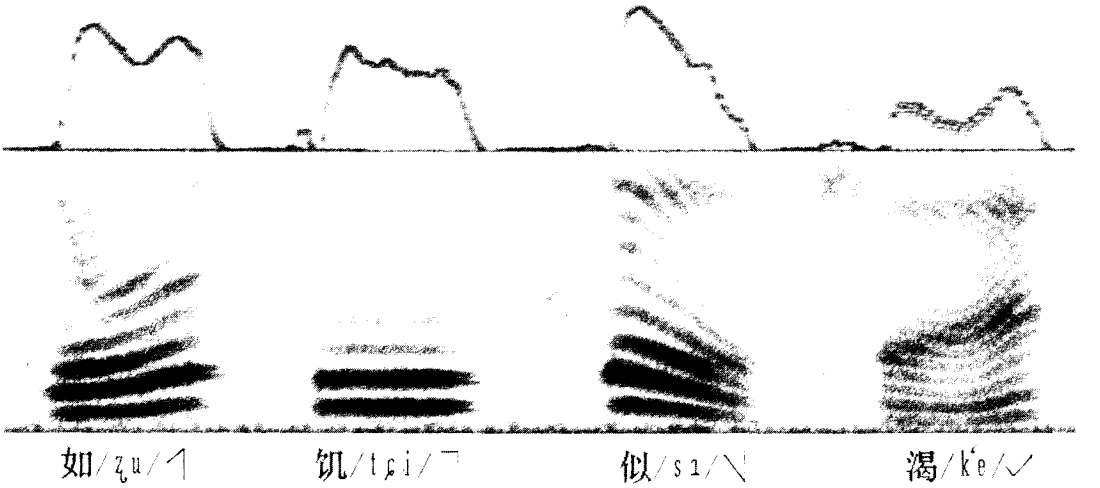

Fig. 4. Narrow-band spectrograms of the synthetic syllables in the four SC tones.

The $f_{ko}(t)$ values at any given time may be calculated by means of a linear interpolation formula. Thus, by inputting different F_c, F_d and ko, we can get the exact F_0-patterns we need. Fig. 4 shows the narrow-band spectrograms of the synthetic syllables with the four tones of Standard Chinese.

2.2. Amplitude model

The two time-varying parameters A_v and A_u, representing voice source amplitude and noise source amplitude respectively, are generated by the following formula:

$$A(t) = \log^{-1}[A_c + A_d \times f_{ka}(t)]/20. \quad (3)$$

In the formula, A_c is the maximum value (in dB) of the source amplitude; A_d is the range of amplitude variation (in dB); and $f_{ka}(t)$ is an array designated by the symbol ka, which defines the pattern of amplitude variation, the use of which is the same as that of $f_{ko}(t)$.

2.3. Transition model

In the transition section of a syllable, the frequency parameters of the formants are generated by the following formula:

$$F(t) = F_v + (F_t - F_v)f_{kt}(t), \quad (4)$$

F_v is the target value of a given formant frequency in the vowel section; F_t is the target value of the corresponding transition frequency of the consonant; $f_{kt}(t)$ is an array designated by the symbol kt, which defines the transition curve, the use of which is the same as that of $f_{ko}(t)$.

2.4. The synthesis of compound vowels

Half of the SC syllables contain compound vowels (diphthongs and triphthongs), and the formant frequencies of the compound vowels are highly variable. Therefore, finding a proper dynamic model is the key to obtaining good synthetic compound vowels. The dynamic model used in our system is an Exponential Dynamic Model (Yang, 1987). According to this model, for a given compound vowel that has n target values, the dynamic trajectory of the frequency of a given formant can be calculated by the following formula:

$$F(t) = \sum_{i=1}^{n} F_{i,i+1}(t) - \sum_{i=2}^{n} F_i \quad (n \geq 2),$$

$$F_{i,i+1}(t) = F_{ci} + 0.5S \cdot F_{di}\{1 + e^{[-\alpha_i(t - t_{0i})S]}\}, \quad (5)$$

$$F_{ci} = 0.5(F_i + F_{i+1}),$$
$$F_{di} = F_{i+1} - F_i,$$
$$S = \quad 1 \quad (t - t_{0i} \geq 0),$$
$$S = -1 \quad (t - t_{0i} \leq 0).$$

Figures 5(a) and 5(b) show the dynamic trajectories of a given formant in a certain diphthong or triphthong calculated by the above formula. It may be noted that the formant frequencies only approach the target values rather than actually reaching them at the two extremities (for both the diphthong and the triphthong) and in the central region (for the triphthong) of the trajectories.

With this feature of the model, we can synthesize the 13 compound vowels in Standard Chinese, which consist of $9 \times 2 + 4 \times 3 = 30$ phones with only 6 sets of target values for /i, e, a, o, u/ and /y/. The tracings for the formants are very smooth, and the synthetic vowels sound very natural. See Fig. 6 for the spectrograms of some of the SC syllables containing compound vowels.

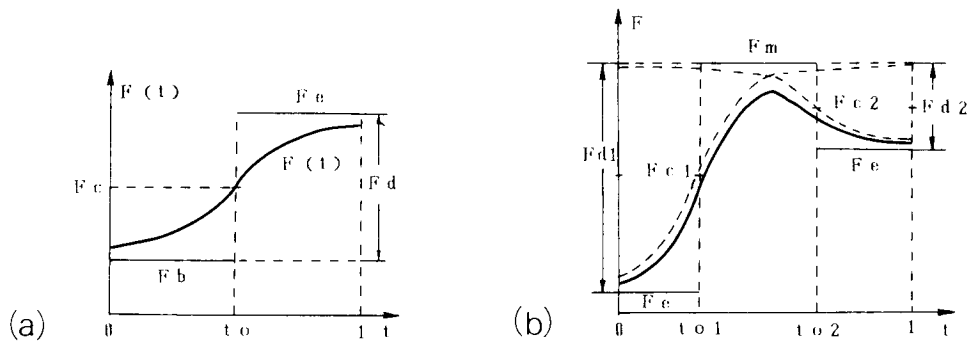

Fig. 5. The Exponential Dynamic Model: (a) for diphthongs, (b) for triphthongs.

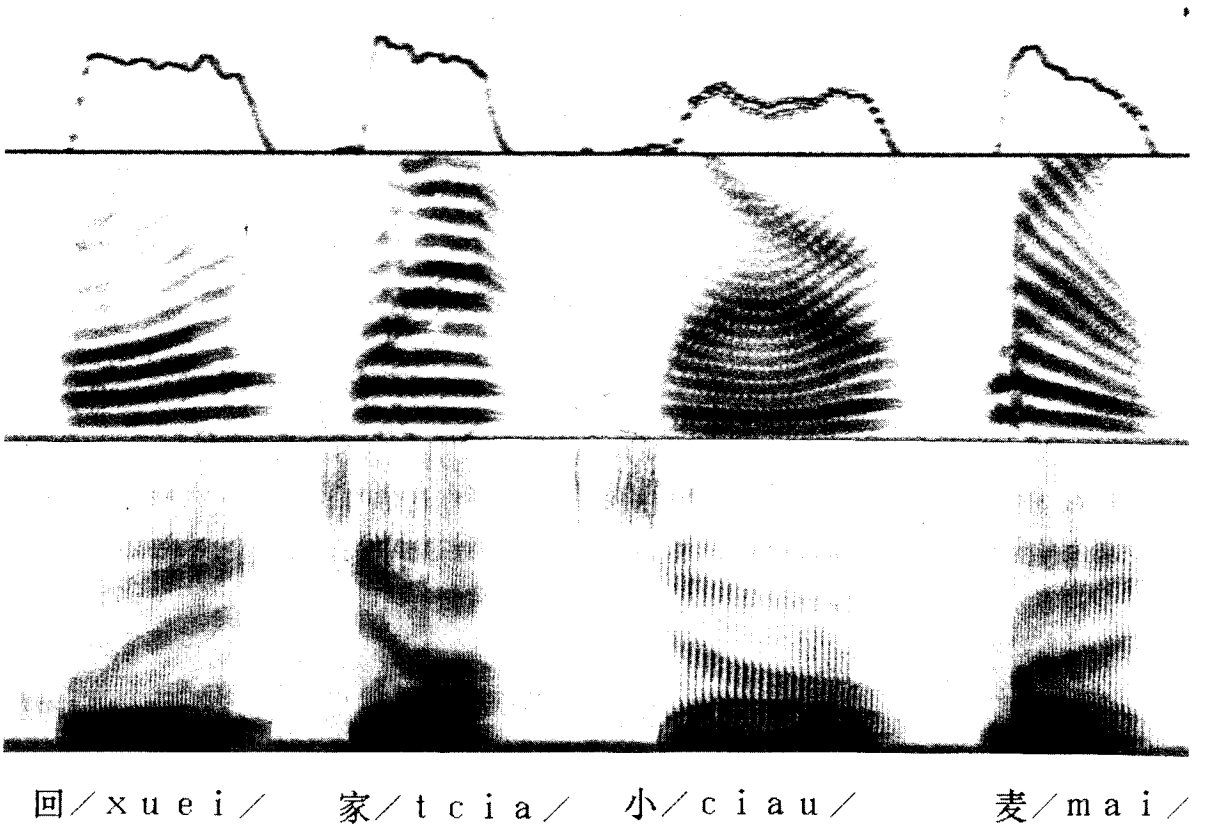

Fig. 6. Spectrograms for synthetic SC syllables containing compound vowels.

2.5. The synthesis of nasals

There are several key points in synthesizing syllables containing nasals with this cascade synthesizer. We will discuss the synthesis of /mei/ as an example. See Fig. 7 for the parameters needed.

Firstly, there is a jumping variation of voice source amplitude at the time of nasal release; secondly, the bandwidth of all the poles in the nasal section (nasal murmur) are all rather broad, and the main energy is concentrated in the low frequency region; thirdly, there is a fault transition of the formants at the nasal-vowel nexus; fourthly, the vowel following the nasal is nasalized, and the closer a vowel frame is to the nasal section, the more heavily it is nasalized.

Nasalization is realized by adding a pole-zero pair to the lower frequency region of the vowel, and by broadening the bandwidth of the inherent formants of the vowels. The pole and zero are separated just after the nasal release, and come closer to each other as time goes on, and finally they overlap. The tendency in the vowel formant bandwidth variation is the same: from broad to narrow. Figure 8 shows the spectrograms of synthetic and natural /mei/.

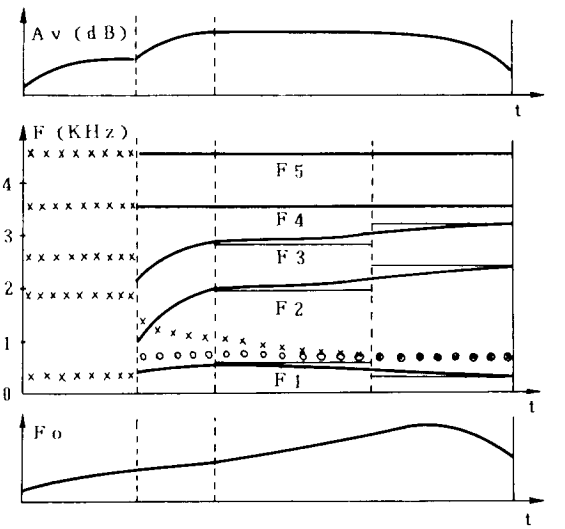

Fig. 7. Diagram of parameter arrangement for synthesizing /mei/.

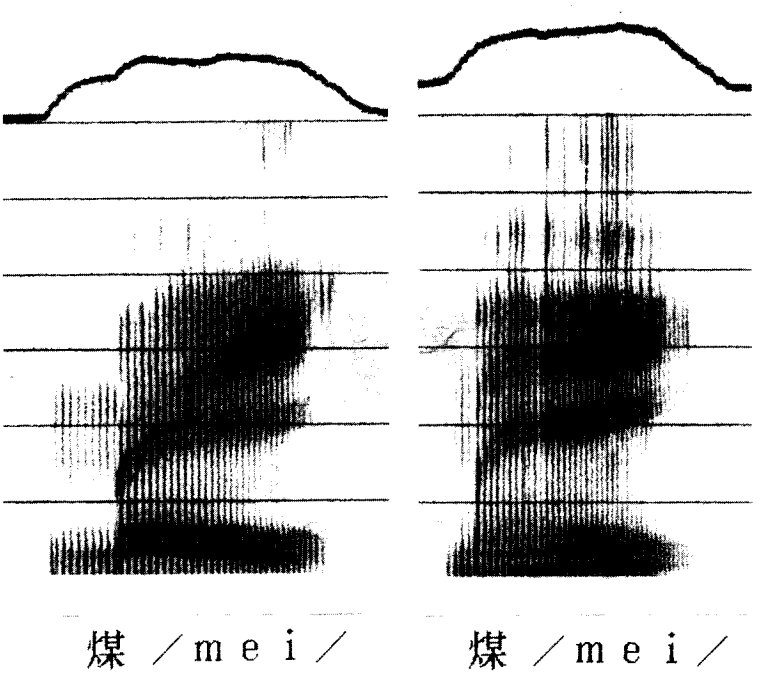

Fig. 8. Spectrograms of the syllable /mei/ (right: synthetic, left: natural).

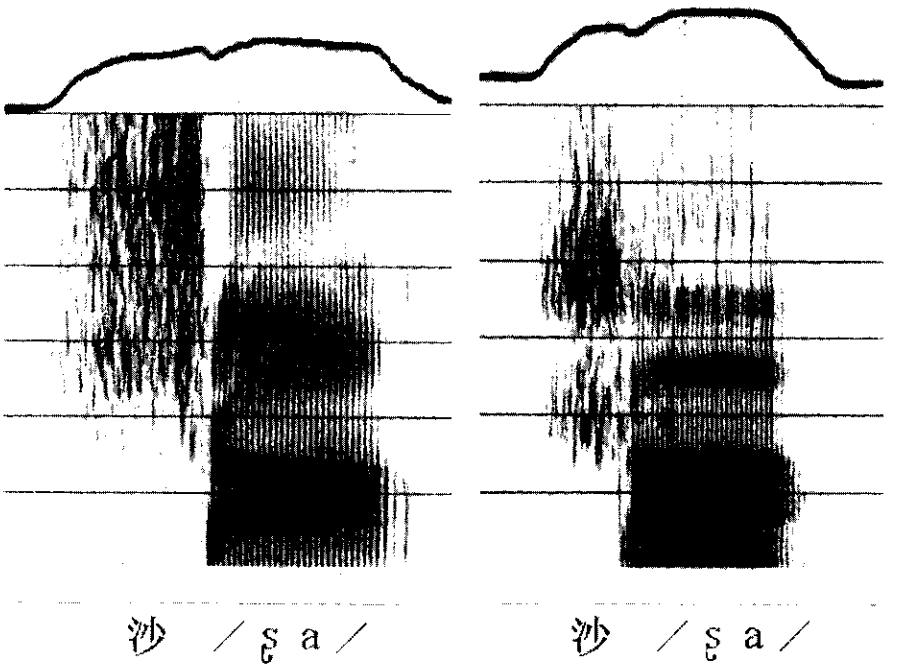

Fig. 9. Spectrograms of the syllable /ʂa/ (right: synthetic, left: natural).

2.6. *The synthesis of fricatives*

There are three key points in synthesizing fricatives with this cascade formant synthesizer:

(1) choosing such an amplitude variation curve so that the amplitude of the fricative section starts rather weak and becomes gradually stronger as time goes on;

(2) using appropriate pole and zero parameters so that the spectral patterns of the synthetic fricatives are similar to those of the natural ones;

(3) keeping a gradual weakening noise element in the voiced transition section, so that fricatives noise extends into the initial part of the vocalic section in the spectrogram, just like the fricatives in natural speech.

See Fig. 9 for the spectrogram of the synthetic and natural fricatives in the syllable /ʂa/.

2.7. *The synthesis of aspirated consonants*

Aspirated segments occur in aspirated stops and affricates in Standard Chinese. As an example, in synthesizing /tʂ'ai/ (see Fig. 10), the noise source is turned on in the aspiration section, and the vocal tract filter is the oral filter. The aspirated part is treated as the first half of the transition, i.e., as a voiceless transition. The formants

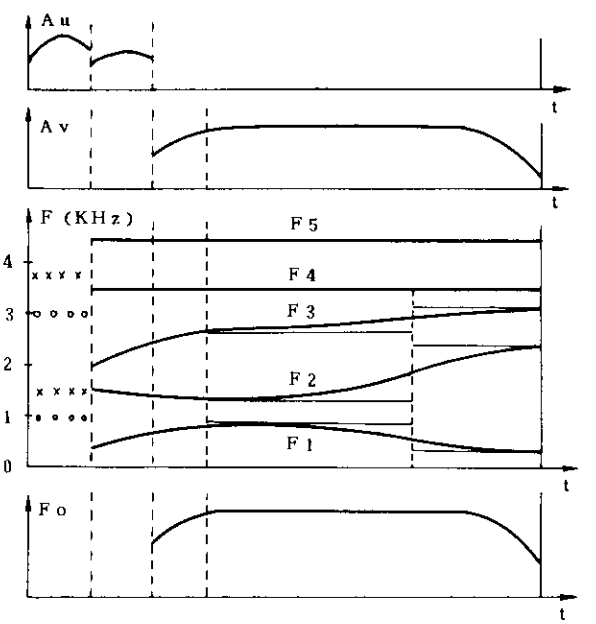

Fig. 10. Parameters for synthesizing /tʂ'ai/.

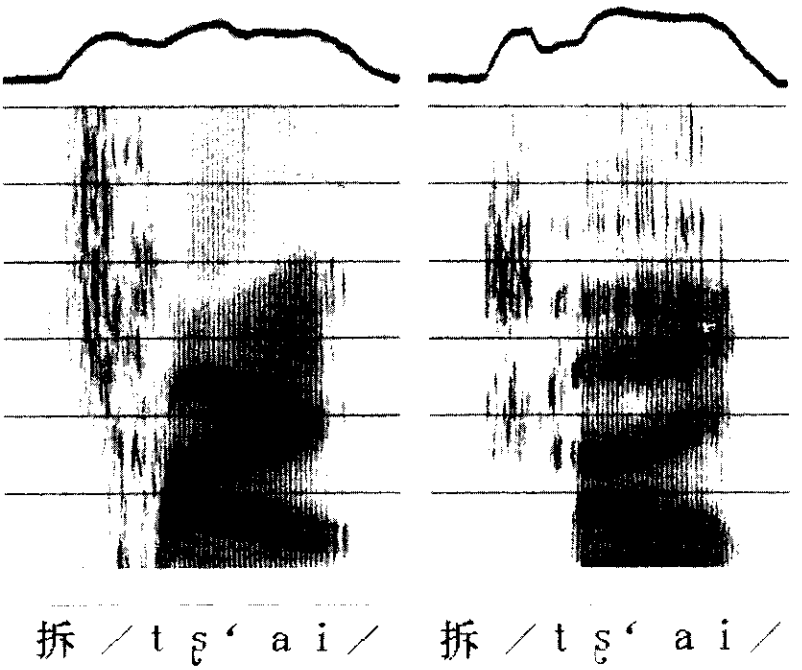

Fig. 11. Spectrograms of the syllable /tʂ'ai/ (right: synthetic, left: natural).

of the voiceless transition join smoothly with the formants of the following voiced transition. The aspirated consonants synthesized this way are very close to those in natural speech. See Fig. 11 for the spectrograms of the synthetic and natural syllable /tʂ'ai/.

2.8. The sound quality of the synthesized SC syllables

With the methods described above, we have synthesized all the 1268 syllables that can be produced in isolation in Standard Chinese. For checking the sound quality of the synthetic syllables, we held a formal listening test for the intelligibility and naturalness of the synthetic syllables. The syllable lists were taken from "*Manual of Acoustics*" (Ma and Shen, 1983). The result of the test is shown in Table 1.

Table 1.
Comparison in intelligibility (%) and naturalness (1–5 from worst to best) between synthesized and natural syllables in Standard Chinese

	Whole syllable	Initials	Finals	Tones	Naturalness
Synthesized	85	89	95	100	4.52
Natural	93	95	98	100	4.88

3. Applications of the synthesis system for Standard Chinese

The synthesis system for Standard Chinese has already been used in the study of speech.

With this system, Lin (1987) synthesized the syllables /ai/, /ʂi/ and /tuo/ with different tonal features by using different F_c, F_d in the Tone Model and with different syllable durations and amplitude patterns. In listening tests using these synthetic syllables as stimuli, he found that native speakers of Standard Chinese could not distinguish

the four tones by means of the amplitude pattern alone, but with the F_0 patterns, the accuracy of tone identification was as high as 95%. He thus concluded that F_0 patterns, were the necessary and sufficient cues for the identification of the four tones in Standard Chinese.

Using this synthetic system, Wu and Xu (1987) conducted an experiment into the properties of aspiration and found that when a vowel following a consonant was low (such as /a/), the aspiration was best synthesized by using the aspiration section in our system; however, when the vowel was high (such as in /ts'i/ and /tɕ'i/, the aspiration was best synthesized as the extension of the fricative section in the system. The conclusion was that the aspiration before different vowels was produced by different mechanisms.

After painstaking data analysis and meticulous work in debugging and parameter adjusting, the parameters and data for synthesizing all the SC syllables are now stored in the data files in the units for allo-initials and allo-finals (conditional varieties of initials and finals). This means that after the keying-in of the Pinyin (the Chinese phonetic alphabet) letters of the syllables to be synthesized, the program will automatically retrieve the parameters and data needed, and will then proceed to the calculation. Because of the coarticulation effects between initials and finals (i.e., between *C–V*), the parameters for a given initial are different when followed by different finals. And the same is true for finals. In order to simulate these coarticulation effects, an Initial-Final-Transaction Model was developed which resolved the problem satisfactorily (there will be a specific article for its discussion). And the size of parameter storage was greatly reduced: only 12 Kbytes are needed for storing all the parameters needed for synthesizing all the SC syllables. The programs and the data stored have thus laid a firm foundation for the future development of a speech output system for Standard Chinese.

References

C.H. Coker, (1972) "Facing the complexities of speech", in: *Papers in Interdisciplinary Speech Research*, ed. by J. Hirschberd (Akademia: Kisdo. Budapest) pp. 319–320.

G. Fant, (1960), *Acoustic Theory of Speech Production*, (Mouton, the Hague).

J.N. Holmes, (1983), "Research report: formant synthesizers: Cascade or parallel?", *Speech Communication*, Vol. 2, No. 4, pp. 251–273.

T.Y. Huang, et al. (1982), "A Chinese text-to-speech synthesis system based on an initial-final model", *Proc. ICASSP-82*, Paris, Vol. 3, pp. 1601–1603.

D.H. Klatt, (1980), "Software for a cascade/parallel formant synthesizer", *J. Acoust. Soc. Am.*, Vol. 67, No. 3, pp. 971–995.

S.C. Lee, et al. (1982), "Microcomputer generated Chinese speech", *Proc. 1982 Internat. Conf. Chinese-Language Computer Soc.*, pp. 157–168.

T.-Y. Li, and H.E. Wolf (1982), "A primary study of synthetic Chinese speech of unlimited vocabulary", *NTZ Archiv*, Bd.4, H.5. 121–125.

M.C. Lin, (1965), "The pitch indicator and the pitch characteristics of tone in Standard Chinese", *Acoustica Sinica*, Vol. 2, No. 1, pp. 15–18 (in Chinese).

M.-C. Lin, (1987), "The perceptual cues of tones in Standard Chinese", *Proc. 11th ICPHS*, Vol. 1, pp. 162-165.

D.-Y. Ma, and H. Shen, eds. (1983), *Manual of Acoustics* (Kexue Chubanshe, Beijing) pp. 417–423 (in Chinese).

A.E. Rosenberg, (1971), "Effect of glottal pulse shape on the quality of natural vowels", *J. Acoust. Soc. Am.*, Vol. 49, No. 2, pp. 583–590.

Z.-J. Wu, (1964), "The spectrographic analysis of the vowels and consonants in Standard Colloquial Chinese", *ACTS Acoustics Sinica*, Vol. 1, No. 1, pp. 33–40 (in Chinese).

Z.-J. Wu, and M.-C. Lin (n.d.) *A Course in Phonetics* (in Chinese).

Z.-J. Wu, (1987), "Aspirated vs. non-aspirated stops and affricates in Standard Chinese", *Proc. 11th ICPHS*, Vol. 5. pp. 209–212.

Y. Xu, and S.-A. Yang (n.d.), "The acoustic-phonetic structure of Standard Chinese syllables and their synthetic simulation" (in Chinese).

S.-A. Yang, (1986), "The effects of the dynamic characteristics of voiced source upon the quality of synthesized speech", *Zhongguo Yuwen*, No. 3, pp. 173–181 (in Chinese).

S.-A. Yang, (1987), "An articulatory dynamic model for diphthongs and triphthongs in Chinese", *Proc. 11th ICPHS*, Vol. 1, pp. 239–242.

J.-I. Zhang, (1986), "Acoustic parameters and phonological rules of a text-to-speech system for Chinese", *Proc. ICASSP-86*, pp. 2023–2026.