PRINCIPLES OF IMPLEMENTING A TEXT-TO-SPEECH SYSTEM FOR TONAL CHINESE (MANDARIN)

K. C. Zhou

School of Electrical Engineering
University of Sydney

ABSTRACT - Unlike other rule synthesis systems for non-tonal languages, a CTS system (Chinese Text-to-Speech or Chinese Type-Speaker) gives particular attention to the tone variation for each syllable. The principles and the implementation of CTS both in software and hardware, are presented in this paper.

PRINCIPLES OF A MANDARIN-CHINESE TEXT-TO-SPEECH SYSTEM

In general, a text-to-speech system which forms speech from non-voice information is language dependent. It is different from the language-independent speech encoding-transmitting-retrieving systems (vocoders). In a text-to-speech synthesis system, the language features are reflected in the rules at three message formation levels for mapping non-voice information from the written text domain onto the acoustic sound domain. The lowest level is the phonetic unit. For a vocabulary-unlimited system, smaller phonetic units below the word level must be chosen for storing acoustic information. Rules are set to adjust the amplitude, duration and pitch variation of these units. The second level is to form a meaningful utterance (a word usually) from the retrieved phonetic units. Rules involved at this level mainly deal with the transitions between these units. The top level is the sentence level where the prosodic rules dominate. The operations at any level re-organize the written information into a format that the system can follow in order to manipulate the acoustic information.

The encoding techniques are involved heavily at the bottom level and partly at the middle level. If the filter model is employed for speech generation, the encoding technique would decide the structure of the filter and the type of parameter set. Not much attention has been paid to pitch variation rules at these two lower levels for generating a voiced sound, since pitch variation is only a prosodic feature, and does not affect the meaning for non-tonal languages.

Chinese has thousands of scriptures known as characters in its written form. Each character is monosyllabic. Spoken Chinese (Mandarin) can be regarded as a string of pronounced syllables. Traditionally, a Chinese syllable can always be split into the initial (a consonantal begining part or a blank called zero-initial) and the voiced final part with a V or VN structure (here V is a vowel or vowel string, N represents a nasal -n or -ng). The pitch variation, called the tone of the syllable, is applied over the voiced section of the syllable. In Mandarin there are a total of 22 initials (4 voiced, 17 unvoiced and the zero-initial) and 38 finals to be used as the phonetic units from which all the possible sounds can be spelt. As a tonal language, each Mandarin syllable has 4 basic tones (Chao, 1968) to represent different meanings and the tone sandhi changes further complicate the pitch variations in multi-syllabic words. Because tone is tied up with meaning, it can be regarded as one of the three elements of a Chinese syllable, as the initial or the final (Wu, 1980). The system

is special in that tone is dealt with at the bottom level or phonetic unit level in retrieving the voiced phonetic units.

The Chinese Text-to-Speech (CTS) system has been planned to be simple and effective, and the choices have been made as follows:

(1)an unlimited vocabulary so that the system can produce all the possible Mandarin sounds;
(2)an initial-final syllable model which requires only 59 phonetic units (excluding zero-initial);
(3)fixed lengths for phonetic units as well as spelt syllables (at a sampling rate of 8 kHz, 1024 points for an initial and 4096 points for a final or a spelt syllable).
(4)waveform encoding (original waveforms for unvoiced sounds and impulse responses for voiced sounds);
(5)a revised version of "pinyin" used as input through a standard keyboard;
(6)artificial tone pitch trains (4 basic tones and 3 sandhied tones);
(7)no transitions between initials and finals.

IMPLEMENTATION OF THE SYSTEM

The software implementation (Zhou, 1984) is actually an algorithm simulation. A Log Magnitude Approximation Filter (Zhou, 1985; Zhou, 1986b) using cepstral parameters was initially tried to reduce memory requirements. However, the filter complexity and the cost has forced the author to use waveform coding in the hardware realization also (Zhou, 1986a). A LSI custom chip has been designed for generating tone pulse trains (Zhou and Cole, 1984). Testing of the fabricated Tone Generator of Chinese (TGC) chip has shown the design to be successful (Zhou, 1986c). The author has also solved the problem of how to evaluate such a tonal language system (Zhou, 1986d).

The general description is shown in the block diagram in Figure 1.

The system consists of three parts. The first part, named "data preanalysis", prepares data for each phonetic unit and stores the data into one of the sound libraries for use in synthesis. The second part is a tone generator which produces tone pulse trains for four basic and three sandhied tones for the synthesis. The third part is the synthesis system itself, from a "pinyin" input to the synthetic speech output.

The first two parts prepare the data stored for the third part, the CTS, so that it can re-generate all the possible Chinese sounds at a phonetic unit level.

The pre-analysis section is a signal processing block for each phonetic unit. It includes sampling(8 kHz) and digitizing (12 bits) natural utterances; cutting phonetic units from them; for a voiced unit, segmenting to get short-time analysis frames (512 samples each frame, a quarter frame shift and the use of Hamming Window) (Rabiner and Schafer, 1978); and homomorphic filtering to obtain the vocal tract filter's impulse response for each frame (Oppenheim, 1975). The properly cut waveform for unvoiced sounds and the impulse responses obtained for voiced sounds are stored in two sound libraries respectively.

The tone generation block makes artificial tone pulse trains as ex-

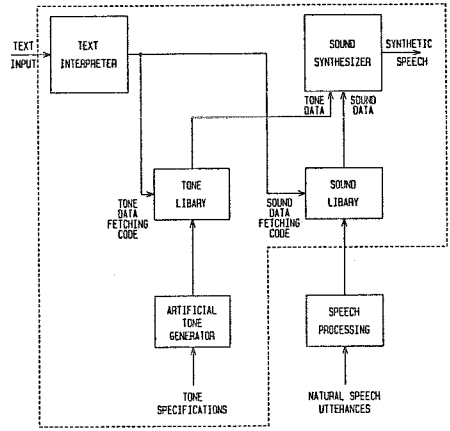


Figure 1 General Description of CTS

citation for the vocal tract filter to produce the sounds in four basic and three sandhied tones. In traditional Chinese phonology, Five Level Descriptions (FLD) (Chao, 1968) have defined the levels of the begining, turning (for tone 3) and the ending points for the Mandarin tones (Figure 2(a)). For setting up the transit trace, the author has added a rule to FLD, saying that within a monotonic section of a syllable the pitch period is linearly changed with time rather than the pitch frequency, as shown in Figure 2(b). Those contours could be considered as the rules of tone generation for Mandarin.

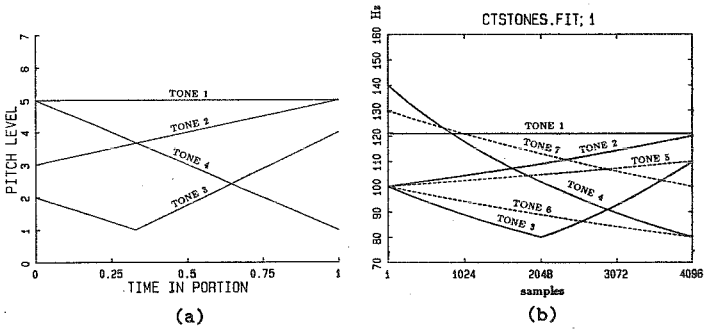
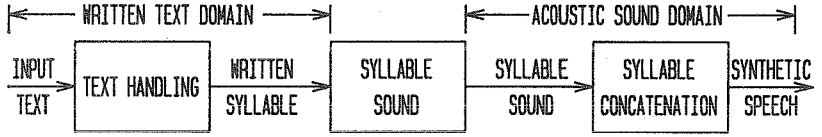
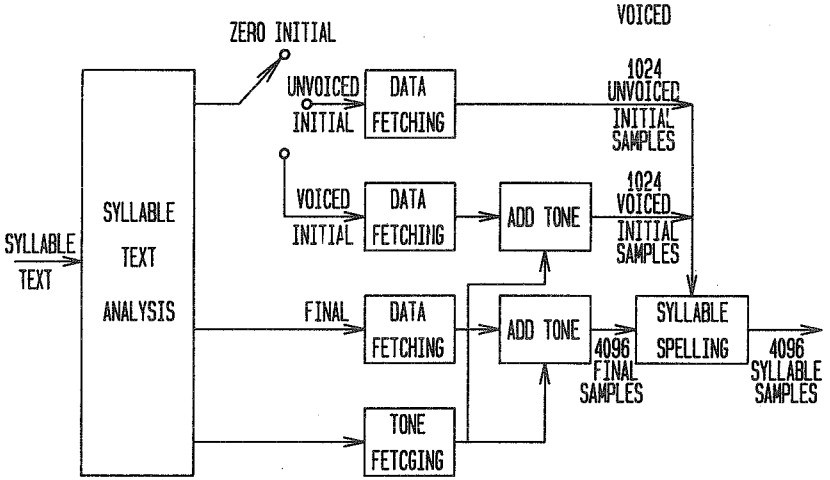


Figure 2 Pitch contours (a) in FLD and (b) in CTS

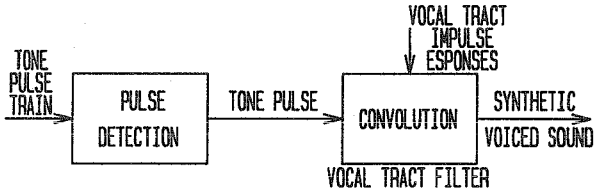
The CTS is the main body of this work. It has the three level structure shown in Figure 3.



Level 1 (Text-to-Spech conversion)



Level 2 (Syllable-to-sound synthesis)



Level 3 (Voiced sound retrieving)

Figure 3 Three levels of synthesis.

The top level is text-speech level. At this level, the text is a string of spelling of Chinese characters, as well as spaces and punctuation marks, and the speech is a string of sounds of these characters and pauses. For example, WO3MEN2 LAI2LEO (We came). The work principle at the text end is to apply spelling rules, tone sandhi rules and other rules (pause lengths for space and punctuation marks) to the text to decompose it into separate syllables. The work principle at the speech end is to concatenate the resulting sound, syllable by syllable, into a joined sound sequence file for later pronunciation.

A lower level is the syllable level. Based on the initial-final model a syllable synthesis is carried out at this level. The input at this level is a "pinyin" syllable with its tone number. At the entry end the principle, is to decompose the syllable into initial, final and tone number and further to interpret these three items into their location addresses in the memory. At the exit end of this level the first operation is to distinguish whether the initial is voiced. If it is voiced, the second step is to replace the first 8 frames of the final data with the initial data. The third step is to get the tone data and to apply it to the vocal tract data to obtain a sound time sequence. For a zero initial case the second step is to apply the tone data to the final data to form the final sound time sequence. If the initial is unvoiced, then after forming the final's time sequence, the first 1024 points of the sequence will be replaced by the initial sequence.

A filter model for the vocal tract and the principle of retrieving time waveforms for short-time analysis frames are involved in the bottom level, the sound unit level, for the voiced sounds. Adding the tone to a voiced sound is actually a convolution process. The vocal tract filter is described by its impulse responses measured in short time frames. A tone pulse will simply be a trigger signal to fetch the impulse response of the corresponding frame and this will be overlap-added to form an output time sequence. A vocoder retrieves an utterance from extracted articulation and pitch information. Here the only difference is that in synthesis the original pitch information is thrown out and artificial tone pitch trains are used instead.

DISCUSSIONS AND CONCLUSIONS

The results obtained from listening tests (Zhou, 1986c) have shown that the principles work quite well. Very high scores have been achieved in testing the tone rules. The software version of CTS on the VAX11-780 gives very intelligible continuous speech (e.g. poetry). The hardware version (M6800 as controller with a specially designed syllable synthesis board) (Zhou, 1986a) is less intelligible because of data truncation from 12 bits to 8 bits. These versions of CTS have shown the following principles:

(1) The Initial-Final model of Mandarin syllable structure is also a very good and practical model for Mandarin spelling in a computer-speaking system. It requires only 22 initials and 37 finals as phonetic units.

(2) The linear pitch period model of tone variation is an alternative to the linear fundamental frequency variation tone model. Artificial tones have also given quite satisfactory results.

(3) The sandhi rules are necessary for obtaining good quality synthetic speech. There are a number of tone sandhi rules in Mandarin: Three

of the sandhied tones described in this system can be regarded as new tone patterns.

(4) The homomorphic method produces good synthetic speech. The artificial tone patterns can be incorporated easily. It is a good alternative to Linear Predictive Coding (LPC).

(5) A fully automatic system has been achieved, in software and in hardware that will respond to typed-in text with synthetic speech.

REFERENCES

CHAO, Y. R. (1968) "A grammar of spoken Chinese", Chapter 1, (University of California Press, Berkeley).

OPPENHEIM A. V. (1975) "Digital signal processing", (Englewood Cliffs, NJ. Prentice Hall).

Rabiner, L. R., Schafer, R. W. (1978) "Digital processing of speech signal", (Bell Laboratories, Incorporated).

Wu, Z. (1980) "The distinctive features and their correlations in phonology of 'Putonghua' ", ZHONGGUO YUWEN (Chinese Language), No.5, 321-327, (in Chinese).

ZHOU, K. C. (1984) "Make a computer talk Chinese", Asian Studies Association of Australia, 5th National Conference, Adelaide, Australia, May.

ZHOU, K. C., Cole, T. W. (1984) "A chip designed for Chinese text-to-speech synthesis", Journal of Electrical and Electronics Engineering, Australia, Vol.4, No.4, pp. 314-318, December.

ZHOU, K. C. (1985) "The logarithmic magnitude filter and its application to text-to-speech synthesis of Mandarin", Proceedings of IREECON, Melbourne, Australia, pp. 1022-1025, September.

ZHOU, K. C. (1986a) "Hardware realization of Chinese text-to-speech synthesis", Proceedings of The Second Australian Computer Conference, pp. 211-220, August.

ZHOU, K. C. (1986b) "A Chinese Text-to-Speech Synthesis System Using the Logarithmic Magnitude Filter", Journal of Electrical and Electronics Engineering, Australia, (accepted).

ZHOU, K. C. (1986c) "Implementation of LSI tone generation chip for Chinese", Proceedings of the First Australian Speech Science and Technology Conference, November.

ZHOU, K. C. (1986d) "Preliminary evaluation of a Chinese text-to-speech system", Proceedings of First Australian Speech Science and Technology Conference, November.