A CHINESE TEXT-TO-SPEECH SYSTEM

BO SHI
Department of Phonetics and Linguistics
University College London
Gower Street, London WC1E 6BT, UK

ABSTRACT

This paper presents a Chinese text-to-speech system which aims to produce highly intelligible standard Chinese speech with natural sounding intonation from Pinyin text, using the LSI parallel formant speech synthesizer. Two specific features of the system are: the use of a demisyllable dictionary which permits an effective and reliable way of carrying out text to phonetic element conversion and the implementation of a flexible intonation framework which enables a very wide variety of natural sounding fundamental frequency (F0) contours to be generated automatically. Some evaluation data obtained from a syllable intelligibility test are provided.

1. INTRODUCTION

The techniques used for unlimited vocabulary speech synthesis can be divided into two groups: concatenation of coded human speech (e.g. LPC) and synthesis-by-rule. The former approach offers a relatively easy way to generate intelligible synthetic speech. But the latter approach is much more flexible. It allows greater control over style of speech, speaker variation, and prosody. These advantages make the rule based approach particularly attractive for synthesizing Chinese speech. First, because the intrinsically simple syllabic structure of Chinese is specially suitable to synthesis-by-rule techniques. Second, because it is a language with syllabic tone, Chinese offers a challenging richness of rule governed pitch variation. The possibility of dynamic control of prosody seems even more important for a Chinese speech synthesis system to enhance naturalness and acceptability of synthetic output than for other non-tonal languages.

The work of present study is to develop a complete Chinese text-to-speech system. The phonetic synthesis part is adapted from an existing English synthesis-by-rule system (SYNCON/LSI) and the language-dependent text-to-phoneme components are added to it.

SYNCON is a phonetic synthesis-by-rule software package written for the BBC microcomputer to interactively control of a parallel formant synthesizer manufactured by Loughborough Sound Images Ltd. (LSI). This phonetic synthesis system not only is capable of producing high quality sounds but also provides versatile control of F0 and timing. The input to the system is a sequence of subphonetic units called “phonetic elements” aligned with two prosodic control parameters: duration and F0 value. Synthesizer control parameters are generated by applying Holmes-Mattingly-Shearman (HMS) acoustic-phonetic rules driven by a set of phonetic tables [1].

The task of present study falls into two parts: (1) creating a rule system for Chinese text to phonetic element conversion with special emphasis on generating prosodic control parameters for natural sounding intonation; (2) building up a set of Chinese phonetic element tables. A flow diagram of the current system is shown in Figure 1. The components previously in existence in the SYNCON/LSI synthesis system are marked as shaded blocks.

This Chinese synthesis system has initially been aimed at generating male speech. All the data presented in this paper are intended to match the observations from a single native male speaker, PCT.

![Figure 1. Flow diagram of the Chinese text-to-speech system](image-url)
2. OVERVIEW OF THE TEXT TO PHONETIC ELEMENT CONVERSION SYSTEM

As shown in Figure 1, the rule system for text to phonetic element conversion consists of three modules: a phonological/phonetic module; a demisyllable dictionary; and a prosodic module. These modules will be described in detail below.

2.1 Text input representation

A Pinyin\(^1\) representation (with tonal marks) of each input sentence is required by the system. In addition to the Roman alphabet, the symbols recognized by the system are the space character and punctuation which identify syntactic boundaries as shown:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Syntactic Boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single space</td>
<td>word boundary</td>
</tr>
<tr>
<td>Double space</td>
<td>phrase boundary</td>
</tr>
<tr>
<td>; ; ; ;</td>
<td>clause boundary</td>
</tr>
<tr>
<td>,</td>
<td>end of statement sentence</td>
</tr>
<tr>
<td>?</td>
<td>end of question sentence</td>
</tr>
<tr>
<td>!</td>
<td>end of exclamatory sentence</td>
</tr>
</tbody>
</table>

2.2 Demisyllables

A relatively simple syllabic structure and the occurrence of lexical tone are characteristics of standard Chinese. Following traditional Chinese phonological approach, the Chinese syllable may be analyzed into three constituents:

1. The initial: maximally a single consonant;
2. The rhyme: a vocalized component with or without a nasal ending;
3. The tone: a prosodic pattern carried by the voiced part of the syllable.

A specific feature of this approach is that the highly coarticulated vocalized syllable internal clusters are treated as units "rhymes". For convenience, the term "demisyllable" applies to initials and rhymes at a phonetic level in the current system.

A demisyllable dictionary has been used in the system for carrying out text to phonetic element conversion. There are several attractive features for this approach: (1) Demisyllables can be used effectively as a basis for determining syllable pronunciation. This is because the demisyllables are units with a relatively fixed internal phonetic structure. Co-articulation within the syllable then can be dealt with at initial/rhyme boundary only. (2) The inventory of demisyllables is small. The number of Chinese segmental syllables is approximately 400, whereas the total number of demisyllables, as used in the current system, is only 68. In practical, a demisyllable based system is simple in construction and fast to operate. (3) The rules for the decomposition of input syllables into demisyllables are easy to write. (4) By careful design, a demisyllable dictionary can supply intrinsic temporal information for each demisyllable which is essential to determine the distribution of the F0 contour over the voiced portion of the syllable.

2.3 Text to demisyllable conversion

The function of the phonological/phonetic module is to convert Pinyin text into demisyllable strings. The conversion is carried out in the following steps:

1. A group of orthographic/phonological rules are applied to each Pinyin syllable to decompose it into its three components: initial, rhyme, and tone;
2. Phonetic and allophonic substitution rules take effect to convert Pinyin initials and rhymes to appropriate demisyllables;
3. Syntactic information concerning word and phrase structure, sentence type is derived from the text and diverted directly to the prosodic module.

2.4 Demisyllable to phonetic element conversion

Demisyllable to phonetic element conversion is based on a dictionary look-up strategy. A demisyllable dictionary supplies the names of the phonetic elements and the associated temporal information for each demisyllable.

Because the prosodic properties of the rhymes are more subject to contextual variation than those of the initials, in the current system, prosodic control (duration adjustment and F0 modification) is exclusively focussed on the rhymes. The durations of the elements which constitute an initial are specified by their inherent absolute durations in centiseconds. Temporal information concerning rhymes, however, is provided by assigning to each phonetic element a value proportional to the total length of the rhyme. The absolute value of the rhyme element duration is then obtained by multiplying its proportional value by the total duration of the rhyme. This arrangement allows a rule-based duration assignment scheme to operate more effectively, because durational variations can be modelled merely by the rules which control the total duration of rhymes, without the need to adjust the duration for individual elements.

2.5 Prosodic parameters assignment

The prosodic module accepts syntactic and phonetic information from the preceding modules (see Figure 1). The function of this prosodic module is to apply prosodic rules at different syntactic levels and to assign the duration and F0 values to each phonetic element in the current element string as required by SYNCON.

2.5.1 The intonation model

The basic principle of the intonation model implemented in the prosodic module is that intonation and syllabic tone are intrinsically capable of separate, notional, definition. In real utterances the F0 contours are the product of syllabic-tone configurations modulated by intonation. Intonation, as a prosodic feature applying to units larger than the syllable, can be regarded abstractly as a sequence of "envelopes". Each
envelope defines the overall variation range of F0 and duration of a syllable at particular locations in an utterance. The effects of semantic, syntactic, pragmatic, and expressive factors on intonation can be represented by a series of rules which control the sizes and positions of the envelopes.

The implementation of the intonation model is as follows. For manipulatory convenience, the four syllabic tonal configurations are generated as a group of cosine curves defined by equation (4) (see 2.5.2 below for details). Here the tonal configurations refer to the F0 contours over the rhyme demisyllable only, since it is the rhyme which carries most of the differences in F0 between different realizations of tonal configurations. These cosine curves are constrained within a rule-controlled envelope (Figure 2). The durational variation range is conditioned by the length of the envelope, “d”, and F0 variation range (tonal register) is defined by the height of the envelope, “r”, and F0 value at baseline, “l”. The parameters d, r, and l are functions of speaker-specific prosodic properties, such as speaker’s F0 range and speech rate, and calculated by a set of rules presented in detail below. Each rule modifies the values of d, r, and l, and leads to deformations of the basic tonal contours. The model is summarized by the formulas:

\[d = D \left( \text{PRCEN}/100 - 0.1 \right)\]  
\[r = R \left( 1 - 0.15 \text{P} - 0.2 k/n \right)\]  
\[l = L \left( 1 + 0.8 \text{P} - 0.1 k/n \right)\]

Where D is the “citation value” of duration and R and L of tonal register. Citation value means the average value to be expected in isolated syllables spoken at normal speed. PRCEN is the percentage shortening/lengthening of duration determined by applying the durational rules listed below. P is the level value of prominence. k is the ordinal number of the syllable in the utterance and n is the total number of the syllables in the utterance.

![Figure 2. Four tonal contours are constrained within a rule-controlled envelope](image)

1.Durational rules
The durational rules implemented so far are pause insertion rule, null consonant lengthening rule, and sentence-final lengthening rule. Where the last two rules modify the percent value "PRCEN".

2. Prominence rules
Prominence rules have been used to deal with tonal reduction phenomena conditioned by word and phrase structures. Such as a word-final or phrase-final syllable tends to have a bigger F0 variation and longer duration than the ones in the middle of the word or phrase. The term "prominence" is intended to apply not only to F0 but also to duration. The appearance of tonal reduction shows up as shortening of duration and narrowing of the tonal register.

The degree of prominence may be divided into five levels. In the present system, the highest level is represented by level-1 and the lowest by level-5. The level of prominence determines the size of the envelope. The higher the prominence level, the bigger the envelope.

The default prominence patterns are:

- monosyllabic word 2
- disyllabic word 42
- trisyllabic word 352
- phrase-cluster-final 1

The exceptions can be specified manually when desired. By inserting prominence value, P, in the equations (1), (2), and (3), the factors d, r, and l will be modified according to the different weight factors of P.

It appears that most of the syntactic effects on contours at word/phrase level can be fairly accurately modelled in this five-level approach.

3. Declining sentence-intonation rules
Declination effects show up both in terms of an overall downward trend in F0 values, and with regard to the relative decrease in the F0 range towards the end of a contour. The final term in equations (2) and (3) specifies the downturn of r and l in steps with slopes determined by 0.2/n and 0.1/n respectively, as k increases.

An illustration of the intonation model is given in Figure 3.

![Figure 3. An illustration of F0 contours generated by the intonation model](image)
Where subscript $i$ is the tone number which varies from one to four. $FO(t)$ is the logarithmic $FO$ value of tone $i$ at time $t$. $Ai$ and $Bi$ are variables of the $FO$ domain. While parameters $Cl$, $Di$, $Ei$, and $Ti$ are variables of the time domain. $Ti$ is the duration from the start point of the $FO$ curve to the turning point (where the $FO$ curve changes direction). $Ei$ is the duration from the turning point to the offset point (where the effective end of the $FO$ curve occurs). $Di$ is the whole duration of the rhyme in tone $i$. All the variables in the equation (4) are functions of $r$, $d$, and $i$.

2.5.3 Duration and F0 value assignment

The requirement of the SYNCON system is that each phonetic element must be accompanied by two prosodic control parameters: phonetic element duration in centiseconds and F0 value. In order to allow a pitch change within the confines of a single phonetic element, an upper arrow symbol "\u2193" is used to specify the additional pitch point(s) within the range specified by the phonetic element duration.

However because most initials are voiceless consonants in Chinese, i.e. the rhymes are the main carriers of tone, over initials, the F0 contour can be approximated as an interpolation between the F0 contours of the adjacent rhymes. Only the control parameter of duration, which can be obtained directly from the demisyllable dictionary, is needed for the elements of initials. What follows has therefore been limited to a consideration of the rhymes. The procedure to assign the two prosodic control parameters for a rhyme with tone $i$ are:

- apply prosodic rules (see section 2.5.1) to assign appropriate values for $d$, $r$, and $i$;
- apply tone sandhi rule if there are two or more adjacent tone-3's: $T3 \rightarrow T2 /-T3$;
- compute the whole duration of the rhyme, $Di$, by equations:
  \[ D_1 = 1.1d, \quad D_2 = 1.29d, \quad D_3 = 1.7d, \quad D_4 = d \]
- calculate the duration of element $j$, $D_{ij}$, by equation: $D_{ij} = D_i \ast PROP$ value;
- calculate F0 value at the start point and at the boundaries of elements by equation (4);
- if the turning point and the offset point are not at the boundary of the element, add a "\u2193" to the appropriate element with the associated duration and F0 values.

3. CHINESE PHONETIC ELEMENT TABLES

Finally, the synthesizer control parameters are calculated by applying the HMS-rules to the element text string generated by the text to phonetic element conversion system. The acoustic-phonetic rules to be applied for each phonetic element are defined by a set of phonetic tables containing 51-numbers for each phonetic element. By suitable choice of table entries a very wide variety of transitions, appropriate for formant frequencies, formant amplitudes and degree of voicing, can be constructed. The dominance system, determined by the ranks, is capable of providing many of the co-articulation effects [2]. The Chinese phonetic element tables were created by a trial-and-error procedure. The database used for guidance in generating the phonetic tables includes all 1265 possible syllables in Chinese uttered by the speaker PCT. Formant synthesis parameters were produced for every individual syllable using the JSRU formant tracking package [3]. Initial table entries were derived from these formant-analyzed data, guided by phonetic knowledge. Further modifications of the table values have been carried out by means of perceptual experiments and spectrographic comparisons between the original natural speech and the synthetic output. There are total 104 phonetic element tables in the current version of the system.

4. PERCEPTUAL EVALUATION

An intelligibility test of isolated syllables has been carried out in order to evaluate the quality of the synthetic speech produced by the current system. An open response test was conducted. The average intelligibility scores of 300 presentations across ten listeners are: 89.1% for initials, 96.1% for finals, and 100% for tones.

5. CONCLUSIONS

Although the present system is still under development, its overall performance appears quite satisfactory on the basis of monosyllabic word recognition. The performance on isolated tone recognition was excellent. It can be predicted that the error rate will be reduced substantially in continuous speech recognition due to contextual constraint.

A flexible intonation model has been implemented in the system. The most important feature of this model is that intonation and syllabic tone have been dealt with as two separate but related prosodic systems. The prosodic rules operate on an abstract representation basis (envelopes). This approach not only allows us to dispense with complex absolute timing-related equations in modelling pitch variations and to generate the F0 contours in a single pass, but also gives the model flexibility and more generative power. The prosodic rules implemented so far are only those which appear to be the most essential ones for generating natural sounding intonation. The work concerning the implementation of more detailed prosodic rules covering both segmental and suprasegmental effects is in progress.

6. REFERENCE