A Taiwanese (Min-nan) Text-to-Speech (TTS) System Based on Automatically Generated Synthetic Units

Ren-yuan Lyu [1], Zhen-hong Fu [1], Yuang-chin Chiang [2], Hui-mei Liu[2]
[1.] Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan
[2.] Inst. of Statistics, National Tsing Hua University, Hsin-chu, Taiwan
Email: rylyu@mail.cgu.edu.tw, rylyu@ms1.hinet.net; Tel: 886-3-3283016ext5677

Abstract

A Taiwanese (Min-nan) Text-to-Speech (TTS) system has been constructed in this paper based on automatically generated synthetic units by considering several specific phonetic and linguistic characteristics of Taiwanese. Some basic facts about Taiwanese useful in a TTS system is summarized, including the issues of tone sandhi, the writen format and the others. Three functional modules, namely a text analysis module, a prosody module, and a waveform synthesis modules is described sequentially. The synthetic units in the waveform synthesis module come from 2 sources, i.e., (1) a set of isolated-uttered tonal syllables and (2) a set of designed continuous speech corpus. A HMM-based large vocabulary Taiwanese speech recognizer is used to do the forced alignment for the speech corpus. A 85.17% segmentation consistency rate within 20 ms can be achieved..

1. Introduction

In Taiwan, although many Mandarin text-to-speech (TTS) systems and dictation systems have been proposed during the past few years [1][2], little work has been done about Taiwanese (Min-nan), which is widely used as the native tongue of more than 75% population in Taiwan. In recent years, we have reported some results on speech recognition research about Taiwanese. [3][4] In this paper, we attempt to construct a Taiwanese TTS system, which should be able to read out any modern Taiwanese articles rather naturally. This TTS system is composed of 3 major functional modules, namely a text analysis module, a prosody module, and a waveform synthesis module. The system architecture can be shown as <fig.1>.

First of all, the input Taiwanese text is analyzed by the text analysis module, which is governed by a phonetic transcription rule and a digit sequence processing rule based on a Taiwanese lexicon with about 70 thousand words. The output of the text analysis module is a sequence of tonal syllables. Since Taiwanese is a tonal language with rather complex tone sandhi phenomenon, it's necessary to deal with such an issue to generate natural speech. In our system, there is a prosody module followed by the text analysis module to deal with the tone sandhi issue. The output of the prosody module is a sequence of tonal syllables with proper tones. The waveform synthesis module takes the tonal syllables as the input and generates the synthetic speech waveform as the final output of the system. It uses syllabic speech units extracted from the speech corpus of both a set of 4521 pre-recorded tonal syllables and a set of phonetically balanced sentences.

This paper is organized to describe all 3 major modules in detail sequentially in the following sections, and finally a discussion and conclusion is given.

2. Text Analysis Module

The text analysis is more difficult in Taiwanese than in Mandarin. One of the major difficulties is that Taiwanese has not been assigned as an official language historically and thus the written form is not consistent at all. It's usually found that modern Taiwanese texts consist of Chinese characters and English characters simultaneously. This is because many daily used Taiwanese words do not have commonly acceptable Chinese characters as their written forms, although most Taiwanese words can be written in the form of Chinese characters. These words are said originated from non-Chinese culture and are usually represented by a spelling system consisting of English (Roman) characters. Take the following sentence as an example: "**咱** e-dang **去佗位** cit-tor ? " (in English: "Where can we go to play ?"). In this sentence, "e-dang" (can) and "cit-tor" (play) can not find the acceptable Chinese written forms and thus are represented as spelling words based on a certain spelling system called Tong-yong Phonetic Alphabet (TYPA) [4] For this reason, the text analysis module should be able to deal with the Chinese-English mixed texts first of all.

Since there is no natural boundary between 2 successive words, we have to transcribe the text to phonetic representation by word segmentation first. We use a Taiwanese pronunciation dictionary with about 70 thousand words as the knowledge source and then adopt a word segmentation algorithm based on the sequentially maximal-length matching in the lexicon. For each segmented word, there may exist not only one pronunciation. To deal with the multi-pronunciation problem, a network with word frequencies as node information and word transitional frequencies as arc information has been constructed for each sentence and a Viterbi search for the best pronunciation is then conducted. A 77.5% correct rate of word segmentation and 87.2% correct rate of phonetic transcription can be achieved. The correct rate is obtained by comparing the output of the word segmentation module and the human transcription of 28 Taiwanese articles.

Another important issue for text analysis is the normalization of the digit sequences. In Taiwanese, each of the digits has 2 distinct pronunciations, i.e., the literature (classic) pronunciation and the oral pronunciation. Both pronunciations of digits can be listed in <table.1>. In fact, each of almost all Taiwanese single-syllabic words has 2 distinct manners of pronunciation: one for classic literature like poems, and the other for oral expression in daily lives. However, for digits,

these 2 manners of pronunciation exist in daily lives. Take the following phrase as an example. “1221 **公斤**” is pronounced as “1(*zit*) **千**(cing) 2(*nng*) **百**(bah) 2(*ri*) **拾**(zap) 1(*it*) **公斤** (gong-gin)”, where the first “1” and the last “1” are pronounced as “zit” (oral) and “it” (classic) respectively, and similarly, both “2”’s are pronounced as “nng” (oral) and “ri” (classic) respectively. The manner of pronunciation depends on the position of the digit in a sequence, which can be summarized in rules. On the other hands, if a digit sequence does not represent a quantity, it is pronounced digit by digit as the classic pronunciation. For examples, “**西元** 1221 **年**” is pronounced as “**西**(se) **元**(quan) 1(it) 2(ri) 2(ri) 1(it) **年**(ni) ”, where all “1” and “2” are pronounced as their classic pronunciations “it” and “ri”, respectively.

3. Prosody Module

As a member of Sino-Tibetan language family, Taiwanese is a tonal language. Traditionally speaking, it has 7 lexical tones, two of which are carried in syllables ended with stop vowels, such as /ap/, /at/, /ak/ and /ah/ (called entering-tone traditionally) and the other five are carried in those without stop-vowels (called non-entering tone traditionally). Let’s define the number 1 to 7 to encode the 7 Taiwanese tones as follows: “1” High-Level （like **東**）, “2” Mid-Level （like **洞**）, “3” Low-Falling （like **棟**）, “4” High-Falling （like **黨**）, “5” Mid-Rising（like **同**）, “6” High-Stop（like **獨**）, “7” Mid-Stop（like **督**）.An example of these 7 tones with one corresponding Chinese character for each tone is shown in <table.2>. Some phonetic/acoustic characteristics, including contour of fundamental frequency (*F0*), the description of relative frequency level (*RF*), and the proposed tone-to-digit (*TD*) mapping are also shown. In this table, one can also find 2 additional tones, namely “8” Low-Stop and “9” High-Rising, which are necessary for tone-sandhi issue discussed in next paragraph.

The tone sandhi issue is relatively complex in Taiwanese. Every Taiwanese syllable has 2 kinds of tones called the lexical-tone and the sandhi-tone depending on the position it appears in a word or a sentence. One of the most frequently referred sandhi rules says that , for most cases, if a syllable appears at the end of a sentence, or at the end of a word, then it is pronounced as its lexical tone, otherwise, it is pronounced as its sandhi tone.[3] The sandhi rules for each lexical tone is as follows:

(1) tone “1” will change to tone “2”;
(2) tone “2” will change to tone “3”;
(3) tone “3” will change to tone “4”;
(4) tone “4” will change back to tone “1”;
(5) tone “5” may change to tone “2” or tone “3” for two different major sub-dialects;
(6) tone “6” will change to tone “8”;
(7) tone “7” will change to tone “6”.

The above is summarized in <fig.2>, which is called the “tone sandhi sailboat”.

There is also a common oral usage which has a very special sandhi rule, i.e., the triple adjective, where the first character of 3 duplicative adjectives will carry a very different tone other than the traditional 7 lexical tones mentioned previously. We map such a “High-Rising” tone to digit “9”, and call it tone “9” in the following. However, not every first character of a triple-adjectives will carry tone “9”. In some cases, it just obey the rules of the “tone sandhi sailboat”. Take the following phrases as examples: “**紅紅紅**” with lexical tones “555” will change its tone pattern to “925”, where the last syllable is pronounced as its lexical tone, the second syllable obeys the “tone sandhi sailboat”, while the first syllable change its lexical tone to tone “9”. The tone sandhi rules for triple adjectives are summarized in <table.3>.

Furthermore, one of the most important issues in prosody of speech is the normalization of the duration and energy of each synthesis unit. It is observed that some optional short pauses between two successive syllables improve the naturalness of the synthesized speech. To add the short pause adequately, we find three types of syllable concatenation in naturally continuous speech waveform. They are overlap concatenation, tight concatenation and loose concatenation. A short pause between the syllables is added when these successive 2 syllables are determined to be loosely concatenated.

The concatenation type of two successive syllables depends on the ending phone of the first syllables and the beginning phone of the second syllable. All 3 types of concatenation are listed in <table.4> according to the ending phones of syllables. For an example, “**阿伯** a-beh” is loosely concatenated because the key phones are “a” and “b”. The statistics of such a table is obtained by an automatically segmented speech corpus described in the next section.

4. The Waveform Synthetic Module

In many modern TTS systems, the output speech was generated by concatenating possible basic synthetic units, e.g. words, phones, bi-phones for English and syllables, Initials/Finals for Mandarin. Here we use the tonal syllables as the basic synthetic units in our start-up system. There are 2 ways to prepare the sets of synthetic units. One way is to record all thousands of Taiwanese tonal syllables in an isolated syllabic mode. It avoids the possible segmentation of units but is less natural to concatenate such units to generate speech of a sentence. The other way is to record speech as sentences. It is more natural. But how to design a script to record and how to segment the speech of a sentence into desired synthetic units is uneasy. Here we describe the design and process of a phonetically balanced speech corpus and the procedure to obtain the desired synthetic units from the corpus automatically.

To design a phonetically balanced corpus for Taiwanese is more difficult than Mandarin or English since there is very little text material available. Here we use a book containing about 6 thousand sentences (about 60 thousand syllables) as the basic text material. To record all 6 thousand sentences is too lengthy and we try to extract a subset of the sentences which cover all possible tonal syllables and phonetic concatenation. A set of 846 sentences with 5813 syllables are obtained, which

contains a set of 1793 distinct tonal syllables. The problem to select such a sentence set is actually a set-covering problem and can be approximately solved by a greedy approximation algorithm. [4]

To segment and label the recorded speech corpus, we use a HMM based speech recognition system to do the force alignment. To estimate the syllable boundary accuracy we must have a reference answer to compare. For this purpose, we design a manual segmentation tool to be used. Two graduate students spent weeks to segment all the sentences at their best. These manually labeling boundaries are referred to compare with the automatically segmented boundaries obtained by the HMM recognizer. The consistency rate of the segmentation between 2 students is 80.28% when we ignore the inconsistency of two boundaries below 10 ms (<10ms), and a 92.02% consistency rate is achieved when we ignore the inconsistency of two boundaries below 20 ms (<20ms). Take this human inconsistency as a reference, we compare the HMM segmentation with human segmentation. a 71.58% and 85.17% consistency rates are achieved fro 10 ms and 20 ms respectively. The result of the segmentation experiments are listed in <table.5>, where we also see that the average inconsistency of human segmentation is 7.84 ms, while the inconsistency between HMM segmentation and human segmentation is about 13.02ms.

After the segmentation of the corpus, the statistics of separation of two successive syllables can be obtained and the information about the concatenation types of syllables described in the previous section is then obtained as shown in <table.4>

However, since the phonetically balanced sentences can contain all units in the basic text material, there are also many units missed. Thus we have to add the second source of speech corpus, which include all possible syllables with all possible tonal variations. This database contains 4521 isolated uttered syllables.Therefore, in the Waveform synthetic module, the synthetic units come from 2 sources: (1) the automatically generated synthetic units from continuous speech corpus, and (2) the isolated-uttered tonal syllables.

Finally, since the recording environment always cause the energy or pitch level of the prerecorded speech to be inconsistent. Therefore, an energy or pitch normalization is performed before the extracted waveform can be concatenated to a sentence..

The interface of our Taiwanese TTS system was shown in <fig.3>, where users can input any Taiwanese text or sentence to the system, and then the system will output the result by speech. It can select Qyuan-zhou tone or Zhang-zhou tone, 2 major sub-dialects of Taiwanese (Min-nan), male or female sound and synthetic units coming from isolated-uttered tonal syllables or automatic generation from assigned corpus.

5. **Conclusions**

We have successfully construct a Taiwanese TTS system and learned a lot about Taiwanese phonetics/linguistics. To make such a prototype system more intelligible and natural, we adopt modern corpus-based TTS technology. However, there are still a lot to do. In the future, we'll make emphasis on the following:

(1) Improving the lexicon by adding part of speech into it. This will improve our text analysis module, and then make better the correct rate of word segmentation and transcription..

(2) Improving the recognizer by developing a speaker independent speech recognition system such that the segmentation accuracy for multi-speakers will be improved and we can switch TTS sound between several different speakers.

(3) Using signal processing techniques to smooth the waveform to reduce the incontinuous concatenation burst.

References

[1] Lin-shan Lee, "Voice Dictation of Mandarin Chinese", IEEE Signal Processing Magzine, July, 1997,p.63~p.101

[2] Fu-chiang Chou, Chiu-yu Tseng, "Corpus-based Mandarin Speech Synthesis with Contextual Syllabic Units Based on Phonetic Properties", ICASSP98,

[3] Ren-yuan Lyu, Yuang-chin Chiang, Wen-ping Hsieh, Ren-zhou Fang "A Large-Vocabulary Speech Recognition System for Taiwanese (Min-nan)", *Journal of the Chinese Institute of Electrical Engineering*, Vol7, No.2, p.123~p136, May, 2000

[4] Ren-yuan Lyu [1], Chi-yu Chen [1], etc, "A Bi-lingual Mandarin/Taiwanese(Min-nan), Large Vocabulary, Continuous Speech Recognition System Based on the Tong-yong Phonetic Alphabet (TYPA)", ICSLP2000, Beijing, China

[5] H. Hon, etc，"Automatic Generation Of Synthesis Units For Trainable Text-To-Speech System"，ICASSP'98.

[6] Chang-yi Zhou, "A Taiwanese Min-nan Course", (In Chinese) , 1997, Anker Publication Company

Tables and Figures

<table.1> Two distinct pronunciation systems for Taiwanese digits, the number in the last of each syllable denotes the tone

numeral	classical	oral
1	it6	zit7
2	ri2	nng2
3	sam1	sann1
4	su3	si3
5	ngo4	qo2
6	liok6	lak6
7	cit7	cit7
8	bat7	beh8
9	giu4	gau4
0	kong5	ling5

<table.2> The Taiwanese tones

TYPA	dong1	dong2	dong3	dong4	dong5	dong9
Ch	東	洞	棟	黨	同	
F0						
RF	HL	ML	LF	HF	MR	HR
TD	1	2	3	4	5	9

TYPA	dok6	dok7	dok8
Ch	獨	督	
F0			
RF	HS	MS	LS
TD	6	7	8

TYPA: Tong-yong Phonetic Alphabet
Ch: an example Chinese Characer
F0: the fundamental frequency contour
RF: relative frequency level
H: High; M: Middle; L: Low
R:Rising; F: Falling; S: Stop

<table.3> The tone sandhi rules for triple adjectives

lexicical	sandhi-
1	9
2	9
3	4
4	1
5	9
6	9
7	6

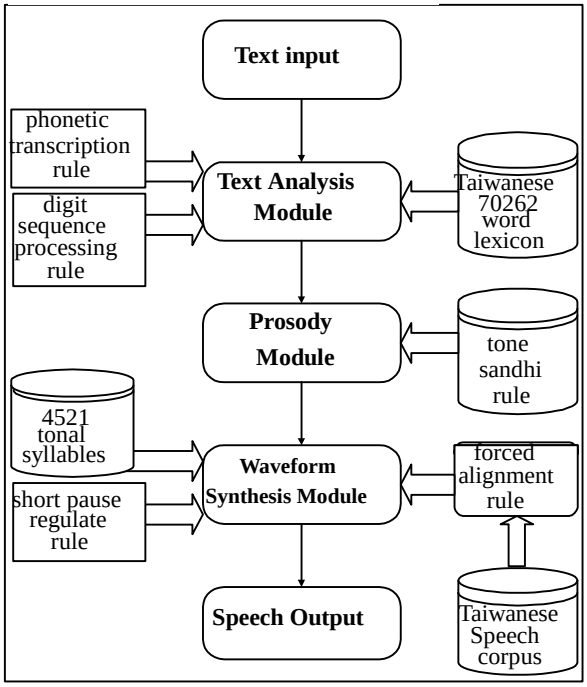
<table.4> all three types of concatenation of syllables

The beginning phone of 2nd syllable / the ending phone of the 1st syllable	b,f,g,k,p,t,z	ann,enn,inn,onn,unn, -ng,-m,ng,, a,ah,ak,ape,eh, i,im,it,ik,o,oh,ok r,rh,u,ut,c,h,l,m,n,q,s,v
a,e,I,o,r,u, ann,enn,inn,onn, unn,-m,-n,-ng	loose concatenation	overlap or tight concatenation
ah,ak,ap,at, e,eh,ih,ik,ip,it oh,ok,op,uh,ut rh,annh,ennh, innh,onnh	loose concatenation	loose concatenation

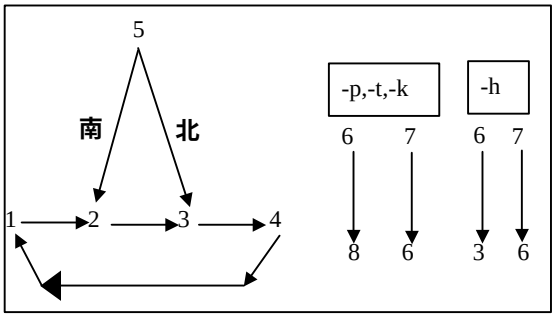
<table.5> The consistency rate for speech segmentation

	manual	computer
<10ms	80.28%	71.58%
<20ms	92.02	85.17
average length of inconsistency	7.84ms	13.02ms

<fig.1> The TTS system architecture



<fig.2> The Taiwanese tone sandhi rules



<fig.3> the interface of Taiwanese TTS system