Computational Linguistics and Chinese Language Processing
Vol. 12, No. 4, December 2007, pp. 391-410 


# Some Studies on Min-Nan Speech Processing

**Wei-Chih Kuo*, Chen-Chung Ho*, Xiang-Rui Zhong*,**

**Zhen-Feng Liang*, Hsiu-Min Yu+, Yih-Ru Wang*, and Sin-Horng Chen***

**Abstract**

In this paper, three studies of Min-Nan speech processing are presented. The first study concerns the implementation of a high-performance Min-Nan TTS system. On the basis of the waveform templates of 877 base-syllables used as basic synthesis units and through the application of the RNN-based prosody generation method and the PSOLA algorithm for prosody modification, this Min-Nan TTS system can convert texts, represented in both Han-Luo (漢羅) and Chinese logographic writing systems, into natural Min-Nan speech. An informal, subjective listening test confirms that the system performs well and the synthetic speech sounds natural for well-tokenized Min-Nan texts and for automatically tokenized Chinese logographic texts. The second investigation concerns the realization of a Min-Nan speech recognizer. It adopts the *initial-final*-based HMM approach with a simple base-syllable bigram language model. A base-syllable recognition rate of 65.1% has been achieved. Finally, a model-based tone labeling method is presented. This method adopts a statistical model to eliminate the affections of all factors other than tone on the syllable pitch contour for automatic tone labeling. Experimental results confirm that this method outperforms the conventional VQ-based approach.

**Keywords:** Min-Nan Text-to-Speech System, Speech Recognition, Model-Based Tone Labeling

## 1. Introduction

Min-Nan is one of the subcategories of the Min dialect, which is one of the seven Chinese dialect families [Yuan *et al.* 1989]. Aside from some pockets of speakers scattered over

* Department of Communication Engineering, Chiao Tung University
Tel: +886-3-5731844, Fax: +886-3-5710116
E-mail: yrwang@mail.nctu.edu.tw
+ Department of Foreign Languages and Literature, Chung Hua University

[Received November 8, 2006; Revised June 29, 2007; Accepted July 5, 2007]

Southeast Asia, varieties of Min-Nan are spoken in southern Fujian, eastern and southeastern Guangdong, and are spread over much of the islands of Hainan and Taiwan, where it is spoken by approximately 73.3 percent of the inhabitants [Huang 1995]; hence, it is often called Taiwanese. In recent years, even though Min-Nan has captured much attention in Taiwan's academic community, research related to its speech processing still remains small due to (1) non-unified writing standards, (2) the various accents of Min-Nan used in Taiwan, and (3) lack of non-public Min-Nan speech and text corpora. These multiple factors may lead to hindering progress in Min-Nan speech processing technology.

However, in spite of the aforementioned deficiencies, which add a degree of difficulty to the automatic processing of this language, three achievements in the technology of Min-Nan speech processing have been made in our study, including the implementation of a high-performance Min-Nan TTS system, the realization of a Min-Nan speech recognizer, and a model-based tone labeling method.

The paper is organized as follows. Section 2 gives a brief introduction to the background of Min-Nan. Section 3 presents the proposed Min-Nan TTS system. Section 4 discusses the realization of a Min-Nan speech recognizer. Section 5 describes a new model-based tone labeling method for Min-Nan speech. Some conclusions are given in the last section.

## 2. A Brief Description of Min-Nan

Like Mandarin and most other Chinese dialects, Min-Nan is monosyllabic in nature, which means that, basically, every syllable is a free morpheme with a meaning value, and that syllable is the unit for pronunciation and every character in text reading is assigned one, but not the only, syllabic sound. The syllabic structures of both Min-Nan and Mandarin can be described in terms of traditional Chinese philology, where syllable is conventionally viewed to be formed by two constituents: the "initial", a consonantal onset, and the "final", made up from a prenucleus onglide, the nucleus – the only obligatory syllabic element, and a coda. Compared with Mandarin, which has 21 initials, 37 finals, and 408 base-syllables, which are legitimate syllables formed by rule-governed combinations of initials and finals, Min-Nan has 18 initials, 82 finals, and 877 base-syllables. In addition to the differences in the numbers of the above-mentioned syllabic constituents and base-syllables, Min-Nan and Mandarin also show differences in the types of syllables, which are often classified by Chinese linguists into "checked" or "entering" syllables, namely syllables ending in a plosive coda (-p,t,k, and a glottal stop), and "smooth" or "slack" syllables, namely syllables ending in a non-plosive. Of the two dialects, only Min-Nan has checked/entering syllables, which leads to different prosodic features associated with syllable types from those of Mandarin.

Min-Nan is a tonal language, where every syllable has an inherent tone, and tones of different pitch values function to distinguish different lexical meanings. [Yang 1999] Min-Nan

has 8 tones, including 7 lexical tones and one degenerated tone, each of which displays a distinct pitch contour. Moreover, based on the type of syllable, tones inherent in entering/checked syllables are termed entering/checked tones accordingly, and those in smooth/slack syllables are called non-entering/non-checked tones. If syllabic tones are under consideration, Min-Nan has approximately 2000 syllables. It is also worth a mention in passing that, despite the fact that in Min-Nan mono-syllable is held to be the basic pronunciation unit, in actual speech mono-syllabic morphemes are not uttered independently; instead, two or more mono-syllabic morphemes, to convey meaning relationship, are concatenated to form meaningful or syntactic poly-syllabic units, which generates changes in the inherent pitch contours of the concatenated syllables. This tonal variation is called "tone sandhi," a very well-known term used to describe the tonal changes depending on the tonal environment in which poly-syllabic words occur.

As for the writing system, although no consistent written forms have been standardized for Min-Nan, two sets of writing systems have been more widely accepted in Taiwan, namely "Romanization" or "Luo Ma Pin Yin" (羅馬拼音) and "Han Luo" system (漢羅系統). In the former, Roman letters are used to spell or transcribe Min-Nan speech, and numbers to specify its tones. This writing system has been widely used among churches to transcribe the Bible that has been translated into Min-Nan. With limited letters and numbers, Romanization provides an easy way to learn the pronunciation of Min-Nan. Therefore, it is not uncommon to see many functionally illiterate Min-Nan elderly churchgoers who cannot read Chinese characters but can recite in Min-Nan scriptures in the Bible written in Romanization. However, since most of the Min-Nan native speakers are literate, and possible ambiguity may be caused by homophones when Chinese characters are not shown, the other writing system, namely Han-Luo system (a hybrid from Chinese characters and Romanization) is used more often in written texts. Unfortunately, the problem still exits in the inconsistency of the Chinese characters selected to represent Min-Nan words or expressions. Except for some popular words, people often choose by preference a string of Chinese characters with similar pronunciations to Min-Nan to represent a Min-Nan word. This increases the degree of difficulty of text analysis for Min-Nan speech processing.

Another linguistic phenomenon worth noting is that, for many Min-Nan syllables, two pronunciation styles co-exist. The first one is called Bai Hua (白話) – the vernacular reading – which is widely used in daily conversation. The other, referred to as Wen Yan (文言) – literary reading – is restrictedly used in reading poetry, some numbers, or in terms used for naming people, buildings, festivals, and so forth.

## 3. An Implementation of Min-Nan TTS System

In this section, the implementation of a high-performance Min-Nan TTS system is presented. Figure 1 shows a block diagram of the proposed Min-Nan TTS system. It is worth noting that such an approach has been successfully applied to developing a high-performance Mandarin TTS system [Chen *et al*. 1998] [Chen *et al*. 2000] [Ho *et al*. 2000]. The system consists of four main functional blocks: a text analyzer, a recurrent neural network (RNN)-based prosody generator, an acoustic inventory, and a PSOLA speech synthesizer. Input text is first tokenized into word/syllable sequence by the text analyzer. The waveform sequence corresponding to the syllable sequence is then formed by the acoustic inventory. Meanwhile, some linguistic features are extracted from the syllable sequence and used in the RNN-based prosody generator to generate necessary prosodic parameters. Afterwards, the PSOLA speech synthesizer uses these prosodic parameters to modify the prosody of the waveform sequence and generate the output synthetic speech. In the following subsections, we will discuss these four main functional blocks in detail.

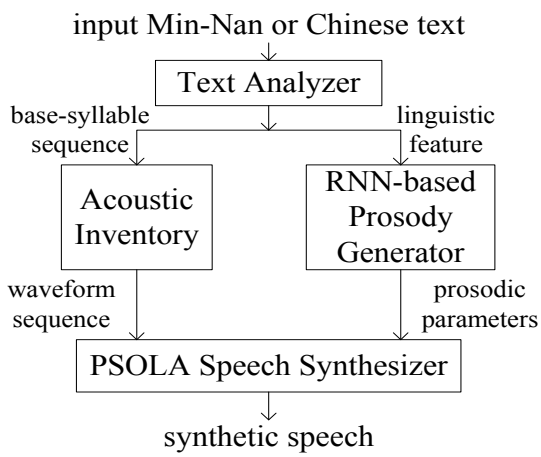


***Figure 1. A schematic diagram of the proposed Min-Nan TTS system.***

### 3.1 The Text Analyzer

The function of the text analyzer is first to tokenize the input text into word sequence and then extract relevant linguistic features from the sequence. Two kinds of input texts are processed. One kind is Min-Nan text represented in the hybrid written form of Han-Luo. Another kind of text is represented in Chinese characters only. Figure 2 displays the block diagram of the text analyzer. It first converts an input text into a Unicode sequence in preprocessing. Here, a look-up table is used to find all syllables represented in Romanized form. It then uses two lexica and a long-word-first criterion to convert the Unicode sequence into a word sequence.

The first lexicon is a Min-Nan lexicon. It contains about 120,000 entries represented in the Han-Luo system. Each entry is a word with a length in the range of 1-6 syllables. The second lexicon, with 110,000 entries, is a Chinese-to-Min-Nan lexicon. It is an extended version of our Chinese lexicon in which Chinese words are transferred to Min-Nan syllable sequence character by character. The use of the Chinese-to-Min-Nan lexicon helps us solve the out-of-vocabulary problem encountered in the text analysis. This also makes the system possess the capability of processing input Chinese text.

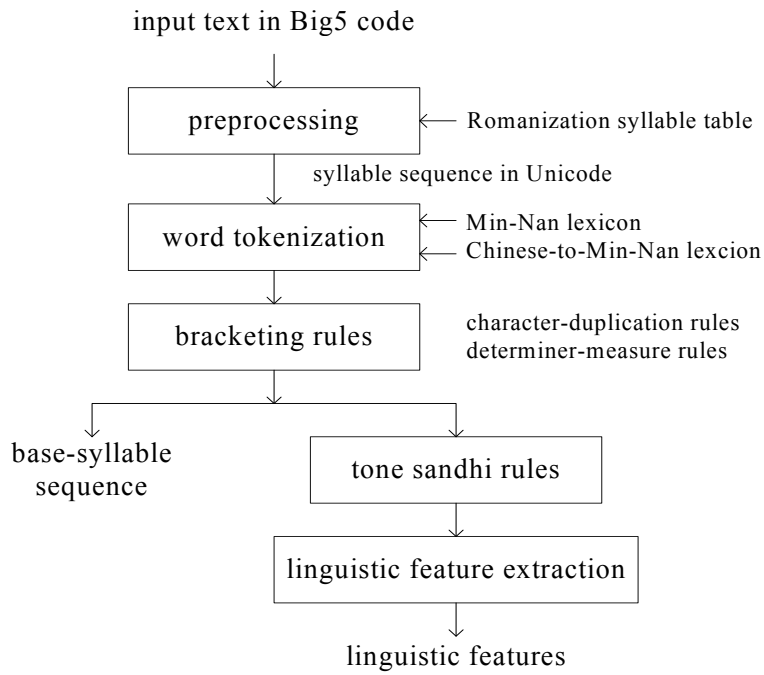


***Figure 2. A functional block diagram of the text analyzer.***

We then use two bracketing rules to construct two types of compound words which are not contained in the lexicon [Huang 2001]. One is for character-duplicated compound words and the other is for determiner-measured compound words. Here, we also decide whether to pronounce the number of a determiner-measured compound word in the style of vernacular reading or in literary reading. For instance, "1998" should be pronounced in the second style as "it kiu2 kiu2 bat", while "兩萬一千八百" (twenty one thousand eight hundred) is pronounced in the first style as "lng7 ban7 chit chheng peh pah".

After obtaining the word sequence, a set of tone *sandhi* rules is then explicitly applied to change the lexical tones of all syllables into the ones to be pronounced [Huang 2001]. Basically, all syllables except the final one of a word chunk (or pronunciation group) have to change their tones. These rules [Cheng 1993] are listed below:

$$
\begin{aligned}
&1 \rightarrow 7\\
&7 \rightarrow 3\\
&3 \rightarrow 2\\
&2 \rightarrow 1\\
&5 \rightarrow \begin{cases} 7 & \text{south} \\ 3 & \text{north} \end{cases}\\
&4\ (p, t, k) \leftrightarrow 8\ (p, t, k)\\
&4h \rightarrow 2\\
&8h \rightarrow 3
\end{aligned}
\tag{1}
$$

Here, an arrow indicates the way a tone changes, *e.g*., Tone 2 will change to Tone 1; "north" and "south" mean the northern and southern parts of Taiwan; and "*p*", "*t*", "*k*", and "*h*" represents entering tones. Besides, four additional rules [Cheng 1993] are used for special cases where a syllable preceding the special character "仔, a function word" (/a/) has been changed to Tone 2 or 3:

$$
\begin{aligned}
&7 \rightarrow 3 \rightarrow 7\\
&8h \rightarrow 3 \rightarrow 7\\
&3 \rightarrow 2 \rightarrow 1\\
&4h \rightarrow 2 \rightarrow 1
\end{aligned}
\tag{2}
$$

For instances, 鋸(ki3→ki1)仔(saw) and 葉(hioh8→hioh7)仔(leaf). An advantage of the approach of using explicit tone *sandhi* rules is that it results in obtaining an RNN-based prosody generator with high efficiency on learning phonological rules of human's prosody generation.

Two sets of linguistic features are then extracted from the word sequence. One is the syllable sequence, which is extracted directly from the word sequence by referring to the lexicon. This will be used in the acoustic inventory to form the basic waveform template sequence. Another consists of two subsets of syllable-level and word-level linguistic features and is used in the RNN-based prosody generator to synthesize proper prosodic parameters. The subset of syllable-level linguistic features contains four parameters: the *initial* type, *final* type, and tone of the current syllable, and the position of the current syllable in the current word. The subset of word-level linguistic features includes two sequences of word length and PM.

## 3.2 The RNN-Based Prosody Generator

The RNN-based prosody generator uses four RNNs to separately generate four types of prosodic parameters for the current syllable: 4 pitch-contour parameters [Chen *et al*. 1990],

*initial* and *final* durations, log-energy level, and the following pause duration. All four RNNs have the same architecture shown in Figure 3. Each RNN is a four-layer network with outputs of the two hidden layers and the output layer being fed back to their own inputs. An RNN of this architecture has been proven in previous studies to be effective in exploring the contextual information of the input linguistic features for the generation of proper output prosodic information [Chen *et al*. 1998]. Table 1 shows the input linguistic features used in these four RNNs.

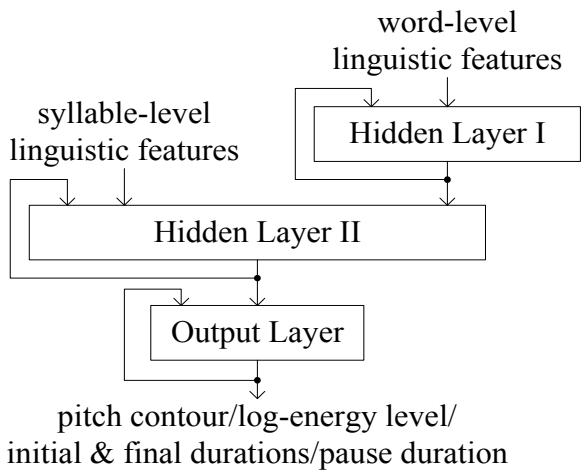


***Figure 3. The architecture of the RNN used in the TTS system.***

***Table 1. The input linguistic features used in the four RNNs for generating syllable pitch contour, initial and final durations, syllable energy level, and pause duration. Here "common" means features commonly used for all four RNNs.***

| | | |
|---|---|---|
| syllable-level linguistic features | common | 1. tone of current syllable<br>2. position of current syllable in a word |
| | Pitch contour | 1. tone of next syllable<br>2. *initial* types of current and next syllables |
| | *Initial* and *final* durations | 1. *initial* and *final* types of current syllable<br>2. light pronunciation of current syllable |
| | energy level | 1. *initial* and *final* types of current syllable<br>2. light pronunciation of current syllable |
| | pause duration | 1. *initial* and *final* types of current syllable<br>2. light pronunciation of current and next syllables<br>3. tone of next syllable<br>4. existence a break following a long word? |
| word-level linguistic features (common) | 1. lengths of current and next words<br>2. existence of special PM following the next word whose length equals to 1?<br>3. PM type following the current word<br>4. POSs of the current and next words | |

These four RNNs can be trained using a large, single-speaker speech database following the back-propagation through time (BPTT) algorithm [Haykin 1994]. The BPTT algorithm is a supervised training algorithm used to learn the mapping from input linguistic features extracted from the input text to output prosodic parameters extracted from the associated utterance. For preparing those inputs and outputs, all texts of the database are manually processed to obtain the word and POS sequences, and the associated utterances are also manually segmented. A further processing of the database is also done for extracting some additional features to improve the efficiency of RNN training. The further processing includes: (1) all minor and major breaks occurring at inter-syllable locations without PMs are manually detected and labeled with special marks; (2) some special characters (referred to as "虛詞, function word") which are consistently pronounced lightly and short are marked, *e.g.*, "甲" in "互氣甲, be angered" and "仔" in "囝仔, child"; (3) all 5-syllable and 6-syllable words are classified respectively into {2-3, 3-2} and {2-2-2, 3-3} pronunciation patterns; and (4) pitch contours of all short syllables are manually refined. Finally, we modify the learning process of the RNN for inter-syllabic pause duration $d$. Instead of letting the RNN learn the real pause duration, we first classify the pause duration into four classes: short ($d \le 75$ ms), medium ($75 \text{ ms} \le d \le 175$ ms), long ($175 \text{ ms} \le d \le 475$ ms), and very long ($475 \text{ ms} \le d$). The pause duration of the "short" class was further normalized with respect to the mean and standard deviation of the final types (2 types: with and without entering tone) of the processing syllable and the initial types (4 types) of the preceding syllable. We then let the RNN learn (1) the class of the pause duration and (2) the pause duration when it belongs to the "short" class. This change can let the RNN take care of both the detail of short pause duration and rough classification of long pause duration.

## 3.3 The Acoustic Inventory

The function of the acoustic inventory is to generate a waveform template sequence for each base-syllable sequence given by the text analyzer. It is a look-up table containing waveforms templates of all 877 base-syllables which are the basic synthesis units used in our system. All of the waveform templates are obtained from isolated-syllable utterances pronounced clearly by a male speaker. All of the speech signals are directly recorded digitally, using a PC with a sound card. The sampling rate is 20 kHz. Each utterance is manually pre-processed to detect ending-points and to label pitch marks.

## 3.4 The PSOLA Speech Synthesizer

The function of the PSOLA speech synthesizer is to generate the output synthetic speech by modifying the waveform template sequence of the base-syllable sequence given by the acoustic inventory using the prosodic parameters given by the RNN-based prosody generator.

Modifications include changing the pitch contour for each syllable, adjusting *initial* and *final* durations for each syllable, scaling the energy level for each syllable, and setting the pause duration for each inter-syllable location. Finally, the output synthetic speech is generated by a 16-bit Sound Blaster card.

## 3.5 Experimental Results

Performance of the proposed Min-Nan TTS system was examined by simulation, using a male speaker database. The database contains 255 utterances including 130 sentential utterances with lengths in the range of 5-30 syllables and 125 paragraphic utterances with lengths in the range of 85-320 syllables. The total number of syllables is 23,633. In addition, a set of 877 isolated base-syllable utterances was recorded for constructing the acoustic inventory. Most of these 877 utterances are syllables with Tone 1. All speech signals were digitally recorded at a 20 kHz rate. All utterances and associated texts were manually pre-processed in order to extract the acoustic features and the linguistic features required to train and test the system.

We first examined the performance of the RNN-based prosody synthesizer. Table 2 lists the root mean square errors (RMSEs) of the synthesized prosodic parameters. RMSEs of 10.2 (12.4) ms, 26.2 (32.4) ms, 15.1 (21) ms, 0.79 (0.80) ms/frame and 2.28 (3.12) dB were achieved in the inside (outside) test for initial duration, final duration, pause duration, pitch contour and log-energy level, respectively. Here, in the calculation of RMSE for pause duration, we set the target pause duration of the three classes of "medium", "long" and "very long" to be 75ms. The classification errors for the four pause duration classes were 12.1% and 13.8% for the inside and outside tests, respectively. Actually, over 80% of classification errors were associated with Class 2. Figure 4 shows a typical example of these synthesized prosodic parameters. It can be seen from the figure that the synthesized prosodic parameters of most syllables matched well with their original counterparts.

***Table 2. The experimental results of RNN prosody generation.***

| | inside | outside |
|---|---|---|
| *initial* duration (ms) | 10.2 | 12.4 |
| *final* duration (ms) | 26.2 | 32.4 |
| pause duration (ms) | 15.1 | 21.0 |
| pitch contour (ms/Frame) | 0.79 | 0.80 |
| energy level (dB) | 2.28 | 3.12 |

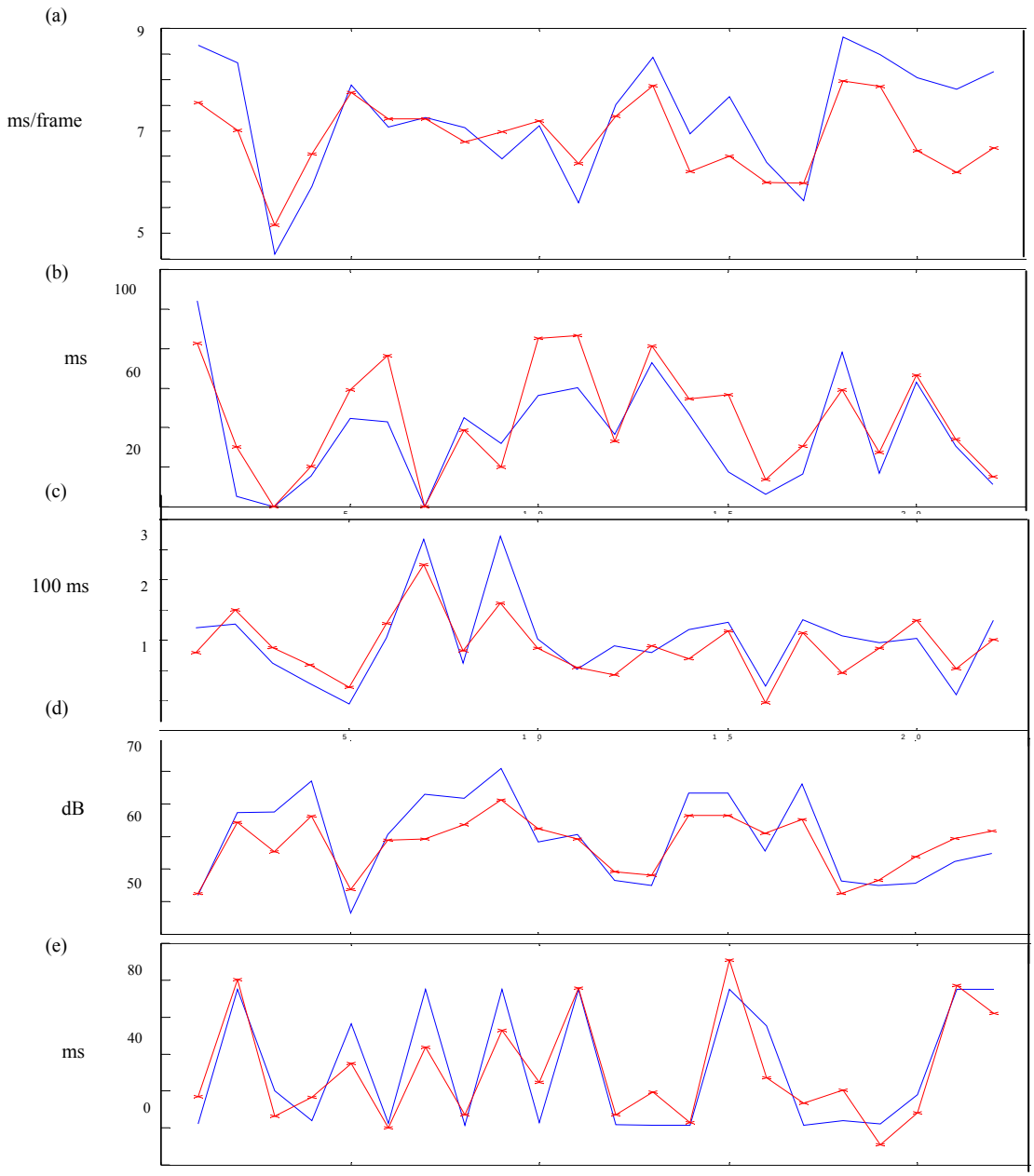


***Figure 4. A typical example of synthesized prosodic parameters: (a) pitch mean, (b) initial duration, (c) final duration, and (d) log-energy level of syllable, and (e) inter-syllable pause duration. The text is "生活應該是鮮豔、開朗、充實，自信滿滿 the 享受人生才著。 seng-oah8 eng2-kai-si7 sian-iam7、khai-long2、chhiong-sit8，chu7-sin3-moa2-moa2 the hiang2-siu7 lin5-seng chai5 tioh8。".***

The whole system has been implemented in software on a PC with a 16-bit Sound Blaster card. An informal subjective listening test using various texts not covered in the database was finally derived to examine the performance of the system. Many participants confirmed that all of the synthesized speeches sounded natural for well-tokenized Min-Nan (Han-Luo) texts and for automatically tokenized Chinese texts. However, the sound quality was only fair for automatically tokenized Min-Nan texts because of the lack of a standardized written form.

## 4. A Min-Nan Speech Recognizer

As can be expected, complicated linguistic properties would affect the performance of speech recognition. Compared with Mandarin, Min-Nan has an inventory of base-syllables double that of Mandarin, and contains syllables ending in a plosive coda, which are not found in Mandarin. These linguistic properties lead to a syllable recognition rate for Min-Nan significantly lower than for that of Mandarin. In [Lyu *et al*. 2000] [Lyu *et al*. 2003] , 825 basic syllables were used for Min-Nan speech recognition system, and a 58% syllable recognition rate was achieved when tri-phone HMM models were used. A Min-Nan speech recognizer is also implemented in this paper. Following the idea of using syllable *initial* and *final* as basic recognition units in Mandarin automatic speech recognition (ASR), the Min-Nan speech recognizer adopts 101 right-*final*-dependent *initial*s and 84 context-independent *final*s as basic acoustic modeling units. Each *initial* is modeled by a 3-state HMM and each *final* is modeled by a 5-state HMM. All 877 base-syllables, including 28 base-syllables with entering tone, can be represented by using these 185 sub-syllable units. Additionally, a 3-state silence model and a one-state short pause model are used to represent the background long silences and inter-syllabic short pauses, respectively. The recognition features consist of 12 MFCCs, 12 delta-MFCCs, 12 delta-delta-MFCCs, delta-log-energy and delta-delta-log-energy. They are extracted for each 30-ms frame with 10-ms frame shift. The cepstrum mean normalization (CMN) technique is also applied to remove the speaker bias.

We first examined the performance of the baseline recognizer (Scheme 1) using only acoustic models by simulation on a large Min-Nan speech database. The database was recorded in 16-kHz sampling rate. It consisted of many sentential and paragraphic utterances generated by 197 speakers, including 91 males and 106 females. We divided the database into two parts, one for training and the other for testing. The training set contained 105,687 syllables while the test set contained 12,211 syllables. The number of syllables in the database is only one-third of the TCC database, which is the most commonly used database for Mandarin ASR in Taiwan. The experimental result is displayed in the 2$^{nd}$ row of Table 3. A base-syllable recognition rate of 46.1% was achieved. The recognition result is relatively low as compared with a typical Mandarin base-syllable recognizer whose base-syllable recognition rate is usually over 60%. This could result, in part, from the fact that the number of

base-syllables in Min-Nan speech is almost twice as many as found in Mandarin speech, and in part from the high confusion of base-syllables of entering tone and their non-entering-tone counterparts. An error analysis showed that base-syllables of the same phonemic constituent with and without entering-tone are highly confusing pairs.

***Table 3. The base-syllable recognition rates of the Min-Nan speech recognizer.***

| | Inclusion rate | deletion error | insertion error | recognition rate |
|---|---|---|---|---|
| Scheme 1 | 48.87% | 2.90% | 2.80% | 46.1% |
| Scheme 2 | 52.73% | 2.91% | 2.56% | 50.2% |
| Scheme 3 | 66.50% | 3.14% | 1.36% | 65.1% |

We then improved the baseline Min-Nan speech recognizer by considering the effect of tone *sandhi* rules. As shown in Equations (1) and (2), base-syllables with */h/* entering tone may change to their counterparts of non-entering tone. This tone *sandhi* will cause serious errors in both HMM model training and recognition test. The total number of *final*s with */h/* entering tone is 17 (out of 28 *final*s with entering tone). We, therefore, relabeled all syllables with */h/* entering tone in both the training and test data sets. Except when located before a long pause, 10-frame silence in our study, all syllables with */h/* entering tone were changed to their non-entering-tone counterparts. The performance of the modified recognizer (Scheme 2) is displayed in the 3rd row of Table 3. A base-syllable recognition rate of 50.2%, or a 4.1% improvement, was obtained.

The recognizer was further improved by incorporating it with a language model (LM). Due to the fact that it is very difficult to collect a large text database with proper tagging or parsing, we considered a simple base-syllable bigram LM instead of the conventional word bigram LM. A text database containing 325,267 syllables was used to train the LM. Texts in the database are news, articles, and stories. The performance of the improved recognizer (Scheme 3) is displayed in the 4th row of Table 3. A base-syllable recognition rate of 65.1%, which corresponded to a 30% error reduction rate, was achieved.

Last, the Min-Nan speech recognizer was applied to a domain-specific task, an in-car speaking assistant prototype for an intelligent transportation system (ITS). An in-car speaking assistant is a user-friendly spoken dialogue human-machine interface acting as an agent to allow the driver to easily control a variety of in-car equipment while keeping his hands and eyes on the road. To add the new module to the existing Mandarin-based in-car speaking assistant system, a Min-Nan grammar for ITS dialog management was needed. In this study, we simply implemented it by directly translated the Chinese grammar into a Min-Nan version. With some simple modifications, the Min-Nan speech recognizer with the ITS grammar was invoked in the ATK [Young 2007] as a real-time Min-Nan ASR module. It successfully expanded the function of the in-car speaking assistant to process Min-Nan input speech. Some

examples of on-line recognition results of the system are shown in Table 4.

***Table 4. Some examples of on-line recognition results for Min-Nan input speec***

| User input | Recognition result (in Mandarin) |
|---|---|
| 系統你好 | 系統你好 |
| 即馬我欲挃交大, 該按怎走? | 然後我過去交大, 那會怎麼走 |
| 等下我要走科技路還是寶山路? | 等一下我要走科技路還是寶山路 |
| 繼續直直走對不對? | 繼續在馬路之後哪一個到 |

## 5. A Model-based Tone Labeling Method for Min-Nan Speech

The task of tone labeling is to determine the tone sequence pronounced in each utterance of a speech database [Li 2002] [Kuo *et al*. 2004]. A database with proper tone labeling should be good to be used in either TTS or ASR. Several approaches can be employed to tackle the task. First, a direct approach is to do the job manually by listening to and/or observing the pitch contour. However, as mentioned above, this approach will suffer from the difficulties of inconsistency and heavy workload. Another approach is to determine the tone sequence by applying the above tone *sandhi* rules to the associated text. As shown in [Liang *et al*. 2004], the tone sandhi rules have been applied to all syllables except for the ones word/sentence final. The results indicated that the tone labeling accuracy for the tonal variations was about 62-65%. The main problem of this approach is that it is not known exactly how to automatically form word chunks from the word sequence. Besides, determining tones only from texts may suffer from errors. The third approach is to regard it as a classification problem by classifying the pitch contours of all syllables with the same lexical tone using an unsupervised clustering technique such as vector quantization (VQ). A drawback of the third approach is that errors may occur because the pitch contour of a syllable in a continuous speech is influenced by many factors other than just the tone itself. The fourth approach is to tackle the task by an efficient pitch contour model which can separate all major affecting factors that control the variation of the pitch contour.

### 5.1 The Proposed Tone Labeling Method

In this study, we adopt the last approach by using a statistical pitch contour model [Wang *et al*. 2000] [Chen *et al*. 2005] [Yang 1999]. We first represent the pitch contour of each syllable by using a 3-rd order orthogonal polynomial expansion [Chen *et al*. 1990]. The basis polynomials used are normalized, in length, to [0,1] and can be expressed as:

$$\phi_0(\tfrac{i}{M}) = 1$$

$$\phi_1(\tfrac{i}{M}) = [\frac{12 \cdot M}{M+2}]^{1/2} \cdot [\frac{i}{M} - \frac{1}{2}]$$

$$\phi_2(\tfrac{i}{M}) = [\frac{180 \cdot M^3}{(M-1)(M+2)(M+3)}]^{1/2} \cdot [(\frac{i}{M})^2 - \frac{i}{M} + \frac{M-1}{6 \cdot M}] \tag{3}$$

$$\phi_3(\tfrac{i}{M}) = [\frac{2800 \cdot M^5}{(M-1)(M-2)(M+2)(M+3)(M+4)}]^{1/2}$$
$$\cdot [(\frac{i}{M})^3 - \frac{3}{2}(\frac{i}{M})^2 + \frac{6M^2 - 3M + 2}{10 \cdot M^2}(\frac{i}{M}) - \frac{(M-1)(M-2)}{20 \cdot M^2}]$$

for $0 \le i \le M$, where $M$+1 is the length of the current syllable log-pitch contour and $M \ge 3$. They are, in fact, discrete Legendre polynomials. A syllable pitch contour $f(\frac{i}{M})$ can then be approximated by:

$$\hat{f}(\tfrac{i}{M}) = \sum_{j=0}^{3} \alpha_j \cdot \phi_j(\tfrac{i}{M}) \qquad 0 \le i \le M\,, \tag{4}$$

where

$$\alpha_j = \tfrac{1}{M+1} \sum_{i=0}^{M} f(\tfrac{i}{M}) \cdot \phi_j(\tfrac{i}{M}) \tag{5}$$

The four coefficients are then divided into two parts: $\alpha_0$ representing the mean and $[\alpha_1 \quad \alpha_2 \quad \alpha_3]$ representing the shape. They are separately modeled. The pitch mean model used can be expressed by:

$$Y_n = F_n + \beta_{pt_n} + \beta_{t_n} + \beta_{ft_n} + \beta_{p_n} \tag{6}$$

where $Y_n$ is the observed pitch mean $\alpha_0$ of the nth syllable; $F_n$ is the normalized pitch mean and is modeled as a normal distribution with mean $\mu$ and variance $\nu$; $\beta_r$ is the compressing-expanding factor (CF) for affecting factor $r$; $t_n$, $pt_n$ and $ft_n$ represent, respectively, the lexical tones of the current, previous and following syllables; and $p_n$ represents the prosodic state of the current syllable. Here, prosodic state roughly represents the state of the syllable in a prosodic phrase and is treated as hidden. Note that $t_n$ ranges from 1 to 22 including 7 standard patterns of lexical tones and all their *sandhi* tones, while both $pt_n$ and $ft_n$ ranges from 0 to 22 with 0 denoting the cases of major punctuation marks {，, 。, ！, ；, ？, 、, ：} or the non-existence of the previous or following syllable. The CFs for $pt_n = 0$ and $ft_n = 0$ are set to zero because we do not want to count the effect of tone across a punctuation mark.

The pitch shape model used can be expressed by:

$$\mathbf{Z}_n = \mathbf{X}_n + \mathbf{b}_{pt_n} + \mathbf{b}_{t_n} + \mathbf{b}_{ft_n} + \mathbf{b}_{p_n} \tag{7}$$

where $\mathbf{Z}_n$ is the observed shape vector $\begin{bmatrix}\alpha_1 & \alpha_2 & \alpha_3\end{bmatrix}^T$ of the nth syllable's pitch contour; $\mathbf{X}_n$ is the normalized pitch shape vector and is modeled as a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix R.

To estimate the parameters of these two models, an expectation-maximization (EM) algorithm is adopted. The EM algorithm is derived based on the maximum likelihood (ML) estimation from incomplete data with prosodic state and pronounced tone pattern being treated as hidden or unknown. To illustrate the EM algorithm, an auxiliary function is firstly defined in the expectation step (E-step) as:

$$Q(\bar{\lambda}, \lambda) = Q_1(\bar{\lambda}_1, \lambda_1) + Q_2(\bar{\lambda}_2, \lambda_2) \tag{8}$$

where

$$Q_1(\bar{\lambda}_1, \lambda_1) = \sum_{n=1}^{N} \sum_{p_n=1}^{P} \sum_{t_n} p(p_n, t_n \mid Y_n, \bar{\lambda}_1) \log p(Y_n, p_n, t_n \mid \lambda_1), \tag{9}$$

$$Q_2(\bar{\lambda}_2, \lambda_2) = \sum_{n=1}^{N} \sum_{t_n} p(p_n, t_n \mid \mathbf{Z}_n, \bar{\lambda}_2) \log p(\mathbf{Z}_n, p_n, t_n \mid \lambda_2), \tag{10}$$

N is the total number of training syllables, P is the total number of prosodic states, $p(p_n, t_n \mid Y_n, \bar{\lambda}_1)$, $p(Y_n, p_n, t_n \mid \lambda_1)$, $p(p_n, t_n \mid \mathbf{Z}_n, \bar{\lambda}_2)$ and $p(\mathbf{Z}_n, p_n, t_n \mid \lambda_2)$ are conditional probabilities, $\lambda = \lambda_1 \cup \lambda_2$, $\lambda_1 = \{\mu, \nu, \beta_t, \beta_{pt}, \beta_{ft}, \beta_p\}$ and $\lambda_2 = \{\boldsymbol{\mu}, \mathbf{R}, \mathbf{b}_{pt}, \mathbf{b}_t, \mathbf{b}_{ft}, \mathbf{b}_p\}$ are the sets of parameters to be estimated, and $\lambda$ and $\bar{\lambda}$ are respectively the new and old parameter sets. Based on the assumption that the normalized pitch mean $F_n$ and shape $\mathbf{X}_n$ are both normally distributed, $p(Y_n, p_n, t_n \mid \lambda_1)$ and $p(\mathbf{Z}_n, p_n, t_n \mid \lambda_2)$ can be derived from the assumed model given in Eqs.(6) and (7) and expressed by:

$$p(Y_n, p_n, t_n \mid \lambda_1) = N(Y_n; \mu + \beta_{pt_n} + \beta_{t_n} + \beta_{ft_n} + \beta_{p_n}, \nu), \tag{11}$$

and

$$p(\mathbf{Z}_n, p_n, t_n \mid \lambda_2) = N(\mathbf{Z}_n; \boldsymbol{\mu} + \mathbf{b}_{pt} + \mathbf{b}_t + \mathbf{b}_{ft} + \mathbf{b}_p, \mathbf{R}) \tag{12}$$

Similarly, $p(p_n, t_n \mid Y_n, \bar{\lambda}_1)$ and $p(p_n, t_n \mid \mathbf{Z}_n, \bar{\lambda}_2)$ can be expressed by:

$$p(p_n, t_n \mid Y_n, \bar{\lambda}_1) = \frac{p(Y_n, p_n, t_n \mid \bar{\lambda}_1)}{\sum_{p'_n=1}^{P} \sum_{t'_n} p(Y_n, p'_n, t'_n \mid \bar{\lambda}_1)}, \tag{13}$$

and

$$p(p_n, t_n \mid \mathbf{Z}_n, \bar{\lambda}_2) = \frac{p(\mathbf{Z}_n, p_n, t_n \mid \bar{\lambda}_2)}{\sum_{p'_n=1}^{P} \sum_{t'_n} p(\mathbf{Z}_n, p'_n, t'_n \mid \bar{\lambda}_2)} \tag{14}$$

Then, sequential optimizations of these parameters can be performed in the maximization step (M-step). At the end of each iteration, the pronounced tone pattern for each syllable is re-assigned to one of its possible patterns by:

$$t_n^* = \arg\max_{t_n} p(t_n \mid Y_n, \lambda_1)\, p(t_n \mid \mathbf{Z}_n, \lambda_2) \tag{15}$$

To execute the EM algorithm, initialization of the parameter set $\bar{\lambda}$ is needed. This can be done by estimating each individual parameter independently. Specifically, the initial CF for a specific value of an affecting factor is assigned to be the difference of the mean (mean vector) of $Y_n(\mathbf{Z}_n)$ with the affecting factor equaling the value of the mean of all $Y_n(\mathbf{Z}_n)$. Notice that, in the initialization of CFs for prosodic states, each syllable is pre-assigned a prosodic state by vector quantization. After initialization, all parameters are sequentially updated in each iteration. The iterative procedure is continued until a convergence is reached.

## 5.2 Experimental Results

Performance of the proposed model-based Min-Nan tone labeling method was examined by simulation on the same single-male speaker database used in the Min-Nan TTS system to train and test the RNN prosody generator. Four tone labeling methods were then realized and compared. The first one was the manual approach, which determined the tone sequence to be pronounced by examining the text. Although the results might contain some errors, we still took them as the reference target because of the lack of anything superior. It is referred to as MANUAL. Another two systems were the VQ-based methods which used 4 (mean + shape) and 3 (shape) orthogonal expansion coefficients of syllable pitch contour as classification features, respectively. They are referred to as VQ-4 and VQ-3. The last one was the proposed model-based method and referred to as MODEL. The RMSEs of the reconstructed pitch contour are 0.815 and 0.286 ms/frame for VQ-4 and MODEL, respectively. The superior results of MODEL show the effectiveness of the pitch mean and shape models. Table 5 shows the correct rates of tone labeling for the latter three methods by taking the results of MANUAL as reference target. Correct rates of 50.9, 52.4, and 61.9% were obtained by VQ-4, VQ-3, and MODEL, respectively. Obviously, MODEL outperformed both VQ-4 and VQ-3. It can also be found in Table 5 that Tone 1 and Tone 2, which share a single *sandhi* tone pattern, have better labeling results.

***Table 5. The correct rates of the three tone labeling methods of VQ-4, VQ-3, and MODEL. (unit: %)***

| Tone (*sandhi* tones) | 1 (7) | 2 (1) | 3 (2,1) | 4 (2,1,8) | 5 (7,3,7) | 7 (3,7) | 8 (3,7,4) | Ave. |
|---|---|---|---|---|---|---|---|---|
| VQ-4 | 61.9 | 82.9 | 55.4 | 40.9 | 28.1 | 34.0 | 33.9 | 50.9 |
| VQ-3 | 58.7 | 84.8 | 44.1 | 28.7 | 43.7 | 47.2 | 35.8 | 52.4 |
| MODEL | 72.4 | 89.3 | 51.7 | 55.7 | 50.6 | 51.1 | 41.9 | 61.9 |

By examining all 22 tone patterns obtained in the pitch mean and shape models, we found that most *sandhi* tone patterns matched with those tone patterns suggested by the above-mentioned *sandhi* rules. Figure 5 displays the standard and *sandhi* tone patterns for lexical Tone 1 and Tone 2. Can be seen from Fig. 5(a) (Fig. 5(b)) that the shape of the *sandhi* tone pattern of Tone 1 (2) resembles the standard pattern of Tone 7 (1). Figure 6 displays pitch contour patterns of standard and *sandhi* tones for Tone 3 and Tone 2. It can be seen from Fig. 6(a) (Fig. 6(b)) that all three (two) *sandhi* Tone 3 (2) patterns resemble to the standard Tone 3 (2) pattern.

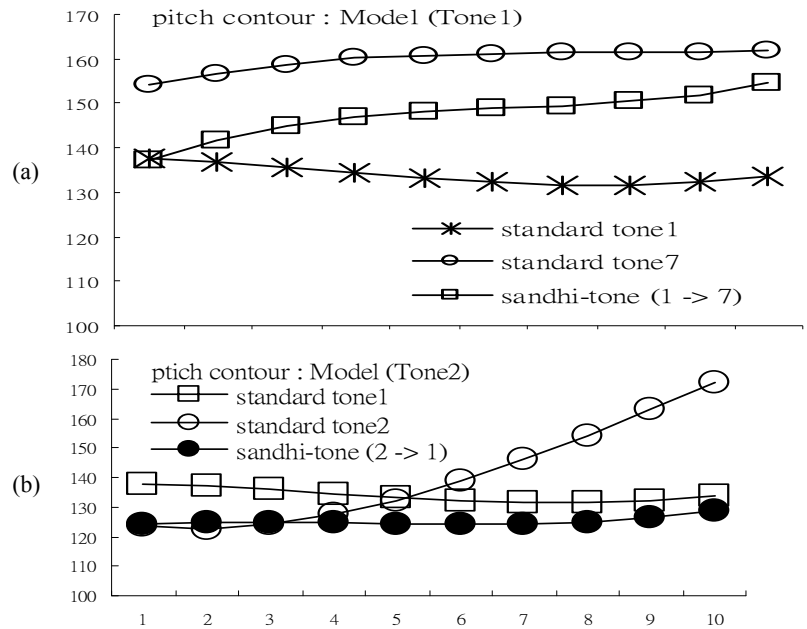


***Figure 5. Comparison of standard and sandhi tone patterns for lexical (a) Tone 1 and (b) Tone2.***

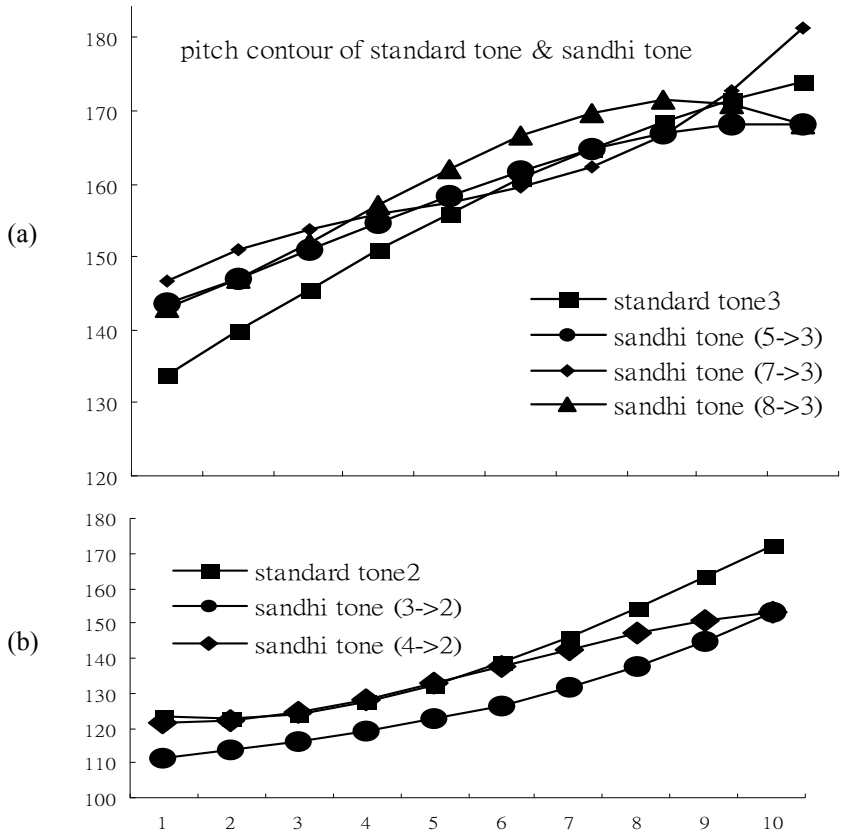


***Figure 6. Comparison of pitch contour patterns of standard tone & sandhi tones for (a) Tone 3 and (b) Tone 2.***

## 6. Conclusions

In this paper, three studies of Min-Nan speech processing have been discussed. They included the implementation of a high-performance Min-Nan TTS system, the realization of a Min-Nan speech recognizer, and a model-based tone labeling method. Experimental results confirmed that all proposed methods are promising.

From these studies, we find that the most important factor to affect the research results is the database. Basically, a large, phonetically-rich, high-quality speech database with text being properly annotated is needed. The two databases used in current three studies are still not perfect on their size and text annotation. To improve the quality of these two databases for achieving a good progress on our future Min-Nan speech processing studies are therefore worth doing.

## ACKNOWLEDGEMENT

This work was supported in part by MOE under contract EX-94-E-FA06-4-4. The authors thank Prof. R. L. Cheng and Prof. Y. C. Chiang for supplying the lexicon and the text corpus.

## REFERENCES

Chen, S. H., and C. C. Ho, "An Implementation of Taiwanese Text-to-Speech System," In *Proceedings of ISCSLP'2000*, 2000, Beijing, vol.1, pp. 613-616.

Chen, S. H., S. H. Hwang, and Y. R. Wang, "An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to-Speech," *IEEE Trans. Speech and Audio Processing*, 6(3), 1998, pp. 226-239.

Chen, S. H., W.-H. Lai, and Y.-R. Wang, "A statistics-based pitch contour model for Mandarin speech," *J. Acoust. Soc. Am.*, 117(2), 2005, pp. 908-925.

Chen, S. H., and Y. R. Wang, "Vector Quantization of Pitch Information in Mandarin Speech," *IEEE Tarns. Communications*, 38(9), 1990, pp. 1317-1320.

Cheng, R. L., *Taiwanese pronunciation and Romanization – with rules and examples for teachers and students*, Wang Wen Publishing Company, Taipei, 1993.

Haykin, S., *Neural networks – A comprehensive foundation*, Macmillan College Publishing Company, 1994.

Ho, C. C., and S. H. Chen, "A Hybrid Statistical/RNN Approach to Prosody synthesis for Taiwanese TTS," In *Proceedings of ICSLP'2000*, 2000, Beijing.

Ho, C. C., and S. H. Chen, "A Maximum Likelihood Estimation of Duration Models for Taiwanese Speech," In *Proceedings of ISAS-SCI 2000*, Orlando, USA, vol. VI, pp. 395-399.

Huang, S.-F., *Language, Society and Ethnicit*y, 2nd ed.. Crane, Taipei, 1995

Huang, J. Y., "Implementation of Tone *Sandhi* Rules and Tagger for Taiwanese TTS," Master Thesis, Communication Eng. Dept., National Chiao Tung University, 2001.

Kuo, W.-C., Y.-R. Wang, and S.-H. Chen, "A Model-Based Tone Labeling Method for Min-Nan/Taiwanese Speech," In *Proceedings of ICASSP2004*, 2004, Montreal, Canada, Vol. 1, pp. 505-508.

Kuo, W.-C., X.-R. Zhong, Y.-R. Wang, and S.-H. Chen, "A High-Performance Min-Nan/Taiwanese TTS System," In *Proceedings of ICASSP2003*, 2003, Hong Kong, Vol. 1, pp. 512-515.

Li, A., "Chinese Prosody and Prosodic Labeling of Spontaneous Speech," In *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, 2002.

Liang, M.-S., R.-C. Yang, Y.-C. Chiang, D.-C. Lyu, and R.-Y. Lyu, **"**A Taiwanese text-to-speech system with applications to language learning," In *Proceedings of 2004 IEEE International Conference on Advanced Learning Technologies*, 2004, pp. 91-95.

Lyu, R.-Y., Y.-C. Chiang, W.-P. Hsieh, and R.-Z. Fang, "A Large-Vocabulary Speech Recognition System for Taiwanese (Min-nan)," *Journal of the Chinese Institute of Electrical Engineering*, 7(2), 2000, pp. 123-136.

Lyu , D.-C., B.-H. Yang, M.-S. Liang, R.-Y. Lyu, and C.-N. Hsu, "Speaker independent acoustic modeling for large vocabulary bi-lingual aiwanese/Mandarin continuous speech recognition," In *Proceedings of the ninth Australian international conference on Speech science and technology*, 2002, Melbourne, pp. 28-33.

Wang, W. J., Y. F. Liao, and S. H. Chen, "RNN-based Prosodic Modeling for Mandarin Speech and Its Application to Speech-to-Text Conversion," *Speech Communication,* 36, 2002, pp. 247-265.

Yang, Y. C., "An Implementation of Taiwanese Text-to-Speech System," Master Thesis, Communication Eng. Dept., National Chiao Tung University, Hsinchu, 1999.

Young, S., ATK: A Application Tool for HTK, http://mi.eng.cam.ac.uk/~sjy/software.htm, 2007.

Yuan, J. H., Hanyu Fangyan Gaiyao, *Outline of Chinese Dialects*, 2nd ed., Wenzi Gaige Chubanshe, Beijing, 1989.