CHAPTER 17

TAIWANESE MIN-NAN SPEECH RECOGNITION AND SYNTHESIS

Ren-Yuan Lyu[†], Min-Siong Liang[‡], Dau-Cheng Lyu[‡], and Yuan-Chin Chiang[§]

[†]*Department of Computer Science & Information Engineering,*
[‡]*Department of Electrical Engineering, Chang Gung University, Tao-yuan*
[§]*Institute of Statistics, National Tsing-hua University, Hsin-chu*
Email: {renyuan.lyu@gmail.com, minsiong@gmail.com}

In this chapter, we review research efforts in automatic speech recognition (ASR), text-to-speech (TTS) and speech corpus design for Taiwanese, or Min-nan – a major native language spoken in Taiwan. Following an introduction of the orthography and phonetic structure of Taiwanese, we describe the various databases used for these tasks, including the Formosa Lexicon (ForLex) – a phonetically transcribed database using Formosa Alphabet (ForPA), an alphabet system designed with Taiwan's multi-lingual applications in mind – and the Formosa Speech Database (ForSDat) – a speech corpus made up of microphone and telephone speech. For ASR, we propose a unified scheme that includes Mandarin/Taiwanese bilingual acoustic models, incorporate variations in pronunciation into pronunciation modeling, and create a character-based tree-structured searching network. This scheme is especially suitable for handling multiple character-based languages, such as members of the CJKV (Chinese, Japanese, Korean, and Vietnam) family. For speech synthesis, through the use of the bilingual lexicon information, the Taiwanese TTS system is made up of three functional modules: a text analysis module, a prosody module, and a waveform synthesis module. An experiment conducted to evaluate the text analysis and tone sandhi modules reveals about 90% labeling and 65% tone sandhi accuracies. Multiple-level unit selection for a limited domain application of TTS is also proposed to improve the naturalness of synthesized speech.

1. Introduction: The Languages and People of Taiwan

Taiwan's inhabitants are multi-lingual, while Taiwanese (also called *Min-nan* or Hokkienese) is the mother tongue of more than 70% of the island's population.[1] Although a majority of the people in Taiwan use Taiwanese as a native language, this language has been very much marginalized in the period after World War II.

Along with the democratic and economic achievements in Taiwan in recent years, there is a renewed confidence and interest to use Taiwanese as the main language of communication.

Linguistically, Taiwanese is a branch of the *Han* (Chinese) language family possessing many Chinese characteristics such as being a syllabic and tonal language, using *hànzì* (Chinese characters) as the major orthography in its writing system, and having a unique and systematic way to pronounce these Chinese characters. However, Taiwanese does not have a strong written tradition. Up to the 19th century, Taiwanese speakers wrote using a form of literary Chinese (文言文), which would be mostly unintelligible nowadays. A writing system made up of entirely roman characters was developed for colloquial Taiwanese in the 19th century by Western missionaries to facilitate translations of the Bible. This system is commonly called Church Romanization, or "peh-oe-ji" (POJ) in Taiwanese. A new orthographic system called *Hanlor*, proposed by Dr. IokDik Ong in the 1960's that uses both *hànzì* and roman characters, started to gain popularity. Since the 1990's, *Hanlor* has become the main mode of writing for Taiwanese and has been frequently used by major newspapers. Just like the increasing usage of vernacular Chinese (白話文) in Mandarin, the use of the more literary writing form becomes increasingly rare in Taiwanese.

Looking at Taiwanese phonetically, a majority of the Chinese characters used have multiple pronunciations. A character can be pronounced in the classic, literary way (文讀音, *Wen-du-in*) or in the "everyday" way (白讀音, *Bai-du-in*). It has been observed that if a word comes from classical literature, then *Wen-du-in* is used to pronounce that word. But if the origin of the word is vague, the pronunciation of the word tends to vary.

Taiwanese is a member of the Han language family, and not a dialect of Mandarin. Taiwanese, however, does have its own dialects. These varieties of Taiwanese can be classified as northern Taiwanese and southern Taiwanese, roughly corresponding to their origins from mainland China. The dialectal differences between them appear small and often insignificant to native speakers. As a result of well-developed transportation and communication systems, few pure dialectal tongues exist.

1.1. *Phonetic Structure of Taiwanese Syllables*

Phonetically, Taiwanese is a syllabic and tonal language with extensive tone sandhi rules. Similar to Mandarin, a syllable in Taiwanese can be defined by its three components: *initial* consonant, rhyme and tone; rhyme is also called *final* in speech community. There are 18 initials, and 47 finals[2] which are made up of the

28 phonemes listed in Tables 1 and 2. In the table, phonemes are expressed in the International Phonetic Alphabet (IPA) as well as in *TongYong Pinin*, a phonetic spelling system used in Taiwan since 2000. Two features are worth noting. (1) Nasalized vowels:. In Mandarin, a vowel is actually nasalized if preceded with nasal consonants /m/ and /n/ such as the /a/ vowel in "ma" and /au/ in "nau". This applies to Taiwanese as well, but Taiwanese has a rich set of nasalized vowels/rhymes even without preceding nasal consonants. Note that Taiwanese also has a third initial nasal consonant /ng/ not found in Mandarin. (2) Rhymed consonants: Among the 47 rhymes, two of them are consonants, namely /m/ and /ng/. That is, these two consonants can be rhymes and preceded by other consonants.

Table 1**.** The 17 Taiwanese consonants in IPA and in TongYong Pinin (in parentheses).

	Voiced	Unvoiced Unaspirated	Unvoiced Aspirated	Nasal
Alveolar		[s] (s)		
Palatoalveolar	[ʤ] (r)	[ʦ] (z)	[ʦʰ] (c)	
Bilabial	[b] (bh)	[p] (b)	[pʰ] (p)	[m] (m)
Dental	[l] (l)	[t] (d)	[tʰ] (t)	[n] (n)
Velar	[g] (gh)	[k] (g)	[kʰ] (k)	[ŋ] (ng)
Glottal			[h] (h)	

Table 2. The 11 Taiwanese vowel phonemes in IPA and in TongYong Pinin (in parentheses).

vowel phoneme	a (a)	ɛ (e)	i (i)	o (o)	ɤ (or)	u (u)
nasalized vowel	ã (aⁿ)	ɛ̃ (eⁿ)	ĩ (iⁿ)	õ (oⁿ)		ũ (uⁿ)

Traditional tone studies specify that there are seven tone classes in Taiwanese when a syllable is pronounced individually. However, if syllables are articulated consecutively in connected speech, a tone sandhi (or tone change) usually sets in, requiring at least two more tone classes for speech synthesis purposes. These nine tones are listed in Table 3. In contrast to the traditional order of tones, the order of tone classes in Table 3 is adopted from the *TongYong Pinin* system, mainly because of its ease in teaching as well as its simplicity in specifying tone sandhi rules.

Note that tone 6 and 7 (tone 4 and tone 8 in traditional tonal numbering) are the so-called *entering* tones and their phonetic transcriptions differ from other

tones. Syllables with these tones are shorter in duration, and are traditionally treated as tonal variations. However, in speech recognition, they are handled as different syllables.

Taiwanese is known to be rich in tone sandhi, and in our on-going T3 Taiwanese treebank – a bracketed corpus of more than 180,000 words[3] – creation, we recently started to annotate the corpus with tone sandhi marks. Based on about 7,000 phrases/sentences, the tone sandhi rate is more than 80% in syllable count.

There are two questions relevant to tone sandhi: when does the tone of a syllable change, and where does it change to. At the word level, in multi-syllabic words, most syllables would undergo tone changes except the final one. However, at the sentence level, the tone sandhi may appear even at the word boundary. This phenomenon seems closely related to the syntactic roles of words in a sentence, and is being studied in an on-going research.

As for the where-to problem of the Taiwanese tone sandhi rules, the "Taiwanese boat"[2] in Figure 1 illustrates the simplest way to recap the rules. Figure 1 also shows the tone sandhi rules of entering tones.

Table 3. Tones in Taiwanese. Note that tone 8 and 9 exist only in tone sandhi.

Tone	1	2	3	4	5	6	7	8	9
Example	東 dong[1]	洞 dōng[2]	擋 dǫng[3]	黨 dòng[4]	同 dǒng[5]	獨 dok[6]	督 dōk[7]	(獨) dǫk[8]	(黨) dóng[9]
Descrip-tion	high-level	mid-level	low-falling	high-falling	falling-rising	high-short	mid-short	low-short	rising
Traditional classes	1(陰平) InPing	7(陽去) IangChyu	3(陰去) InChyu	2(陰上) InSang	5(陽平) IangPing	8(陰入) InRu	4(陽入) IangRu		

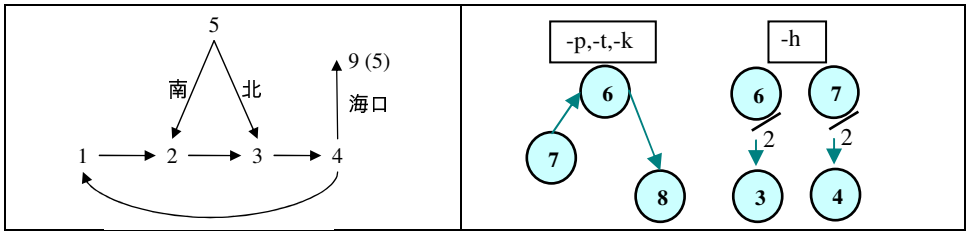

Fig. 1. The major tone sandhi rules of Taiwanese. On the left is the *Tone sandhi Boat* which captures the rules neatly. On the right are the rules for *entering* tones: syllables ending with –p, –t, –k, or –h.

1.2. *ForPA: Formosa Phonetic Alphabet*

With multi-lingual applications in mind, we design a Formosa Phonetic Alphabet (ForPA) for the three major languages in Taiwan: Taiwanese, Hakka, and Mandarin. Other systems, such as SAMPA[4] and WorldBet,[5] have been developed, but have not been adopted here for reasons of simplicity. For the phonetic transcription of Taiwanese, the symbols in ForPA are very similar to those in TongYong Pinin. It is known that phonemes can be defined in several different ways, depending on the level of detail desired. The philosophy driving the labeling process in ForPA is that when faced with choices, we prefer not to divide a phoneme into distinct allophones, except in cases where their sounds are clearly distinct (to the ear), or when their spectrograms look clearly different (to the eye). Since labeling is often performed by engineering students and researchers (as opposed to professional phoneticians), it is generally safer to keep the number of units as small as possible, assuming that the recognizer will be able to learn the finer distinctions that might exist within any context. Generally speaking, ForPA can be considered a subset of IPA, but it has been suited for applications relating to languages in Taiwan.[6]

13. *Multi-lingual ForLex: Formosa Pronunciation Lexicon*

Three lexicons have been collected to be used for corpus collection, speech recognition and speech synthesis. The first is the Formosa Lexicon, which contains about 123,000 words in Taiwanese Chinese text with their Taiwanese Mandarin pronunciations. It is a combination of two lexicons: Formosa Mandarin-Taiwanese Bilingual lexicon and Gang's Taiwanese lexicon.[7] The former is derived from a Mandarin lexicon, and thus many commonly used Taiwanese terms are missing due to the fundamental difference between these two languages. The latter lexicon contains more widely-used Taiwanese expressions from samples of radio talk shows. Some statistics of the Formosa Lexicon are summarized in Table 4, where out of a total of 123,438 pronunciation entries, 65,007 entries are with Wen-du-in pronunciations, while 58,431 entries are with Bai-du-in pronunciations. For all entries with Wen-du-in pronunciations, there are 6,890 mono-syllabic word entries, 39,840 entries of words with two syllables, and so on.

For the other two lexicons, the CKIP Mandarin lexicon and Syu's Hakka lexicon, the distribution of words according to the number of syllables are listed in Table 5.

Table 4. The number of pronunciation of Formosa bilingual Lexicons, including classic literary pronunciation (Wen-du-in) and everyday pronunciation (Bai-du-in).

	Taiwanese Wen-du-in pronunciation	Taiwanese Bai-du-in pronunciation	Total
1-Syllable	6890	2377	9267
2-Syllable	39840	36176	76016
3-Syllable	8308	15214	23522
4-Syllable	9119	4117	13236
5-Syllable	438	399	837
6-Syllable	225	94	319
7-Syllable	125	28	153
8-Syllable	52	22	74
9-Syllable	2	2	4
10-Syllable	8	2	10
Total	65007	58431	123438

Table 5. The distribution of words in two lexicons: Syu's Hakka lexicon, and the CKIP.

	1-Syl	*2-Syl*	*3-Syl*	*4-Syl*	*5-Syl*	
Syu	7322	9161	4948	2382	21	
CKIP	6863	39733	8277	9074	435	
	6-Syl	*7-Syl*	*8-Syl*	*9-Syl*	*10-Syl*	Total
Syu	3	0	0	0	0	23837
CKIP	223	125	52	2	8	64792

2. Speech Corpus

To implement a speaker-independent automatic speech recognition system, it is essential to collect a large-scale speech database. However, the years of marginalization of Taiwanese makes this task difficult in at least in two ways. Firstly, such a mammoth undertaking requires a huge amount of funding, which is difficult to obtain as funding is typically limited. Secondly, due to the limited level of education, only a small number of speakers have the capability to write Taiwanese. This low literacy level makes Taiwanese text collection difficult, which in turn makes the collection of speech data difficult – be it collecting read speech from existing texts, or phonetically transcribing existing speech data.

To overcome this problem, a moderate-sized speech database using only the lexicon is developed. In brief, we: (1) design sheets of phonetically-balanced words from the lexicon; (2) record the microphone and telephone speech of those words; and then (3) validate this speech database. The result is the ForSDat speech corpus which is detailed in the following subsections.

2.1. *Producing Phonetically-Balanced Word Sheets*

Given a lexicon, we can extract phonetically-abundant word sets such that the chosen phonetic units are not only base-syllables, phones, and right context dependent (RCD) phones, but also initial-finals, RCD initial-finals and inter-syllabic RCD phones. The process of selecting such a word set is actually a set-covering optimization problem, which is NP-hard. Here, we adopt a simple greedy heuristic approximate algorithm.[8]

First, some notation definitions. Let $W = \{w_i : 1 \le i \le N\}$ be the set of all words in the lexicon, where N is the number of words, w_i is the i-th word. $S(w_i)$ are the sets of all distinct syllables and $U(w_i)$ are denoted as cross-syllable bi-phones or cross-syllable tri-phones in the word w_i respectively. C_t^* is selected word in time t, $W(t) = \{C_1^*, \cdots, C_t^*\}$ is selected word set, and $S(t) = \{S(C_1^*), \cdots, S(C_t^*)\}$ is selected distinct syllables set and $U(t) = \{U(C_1^*), \cdots, U(C_t^*)\}$ is selected distinct bi-phones set till time t. $W^c(t) = W - W(t)$ is a non-chosen word set, $S^c(t) = S(W) - S(t)$ is a non-chosen distinct syllable set and $U^c(t) = U(W) - U(t)$ is non-chosen distinct bi-phones set or tri-phones set till time t. The algorithm of extracting phonetically-abundant word sets can then be described in the following steps:

Step 1: Initially t=0 and we have $W(0) = W$, $S(0) = S(W)$, $P(0) = P(W)$

Step 2: Choose the word w_{i^*} as C_t^* such that the union of $S^c(t-1)$ and $S(w_i)$ is maximized, i.e. $w_i = \arg\max_{w_i \in W^c(t-1)} \#(S^c(t-1) \cap S(w_i))$ -- (1) then $C_t^* = w_i$; if w_{i^*} is not unique in (1), choose $w_i = \arg\max_{w_i \in W^c(t-1)} \#S(w_i)$ -- (2) as C_t^*; if w_{i^*} is not unique in (1) and (2), choose the preceding index word as C_t^*, $S^c(t) = S^c(t-1) - S(C_t^*)$, $W^c(t) = W^c(t-1) - C_t^*$, $t = t+1$

Step 3: If $S^c(t) \neq \phi$ and $W^c(t) \neq \phi$, repeat Step 2
else if $W^c(t) = \phi$, exit the algorithm.
else if $S^c(t) = \phi$, proceed to the next step

Step 4: Choose the word w_{i^*} as C_t^* that maximizes the union of $U^c(t-1)$ and $U(w_i)$, i.e. $w_i = \arg\max_{w_i \in W^c(t-1)} \#(U^c(t-1) \cap U(w_i))$ -- (3) then $C_t^* = w_i$; if w_{i^*} is not unique in (3), choose $w_i = \arg\max_{w_i \in W^c(t-1)} \#S(w_i)$ -- (2) as C_t^*; if w_{i^*} is not unique in (3) and (2), choose the preceding index word as C_t^*, then $U^c(t) = U^c(t-1) - U(C_t^*)$, $W^c(t) = W^c(t-1) - C_t^*$ and $t = t+1$.

Step 5: If $B^c(t) \neq \phi$ and $W^c(t) \neq \phi$, repeat Step 4
else if $B^c(t) = \phi$ or $W^c(t) = \phi$, exit the algorithm.

Applying the algorithm to the three lexicons mentioned above, we identified a number of balanced-word sets. For Taiwanese, 446 balanced-word sets were generated, each sheet containing 200 syllables making up a total of 37,275 words.

2.2. *Speaker Recruitment and Recording System for Corpus Collection*

Several part-time assistants were employed to recruit speakers from around Taiwan. Each speaker was asked to record readings from one sheet, and the speaker, along with the assistant, received remuneration. We also noted the following information relating to the speaker:

(i) the name and gender of the speaker;
(ii) the age and birthplace of the speaker;
(iii) the location and time of the recording;
(iv) the number of years of education the speaker has completed.

This speaker profile information would be useful later on for organizing the collected speech data. The user can also design experiments according to these profiles.

Two systems were designed for collecting microphone and telephone speech for the ForSDat database. For the telephone system, the speakers dialed into the laboratory using a handset telephone. The input signal is in the format of 8K sampling rate with 8-bits μ-law compression. A speaker was given a prompt sheet before recording, and every word on the sheet was first played to the speaker before he/she recorded that word. The data gathered from different speakers were saved in different directories.

Table 6. The statistics of utterances, speakers and data length for speech collected over microphone and telephone channels in Taiwanese and Mandarin. (*MIC*: microphone; *TEL*: telephone; also denoted in Name by sub-tag after dash).

	Name	**Channel**	**Gender**	**Quantity**	**Train(hr)**	**Test (hr)**
ForSDAT	TW01-M0	*MIC*	Female	50	5.92	0.29
	TW01-M1		Male	50	5.44	
	MD01-M0		Female	50	5.65	0.27
	MD01-M1		Male	50	5.42	
	TW02-M0		Female	233	10.10	0.70
	TW02-M1		Male	277	11.66	
	TW02-T0	*TEL*	Female	580	29.21	0.95
	TW02-T1		Male	412	19.37	

For the microphone system, we used a speech recording tool that simplifies the processes of text prompting, speech prompting and saving the recorded speech and its associated phonetic label file in ForPA. This system was simply set up on a notebook computer and taken to wherever the recording needed to be done. Table 6 shows some statistics of the ForSDat speech database.

2.3. *Speech Data Verification*

To verify the consistency of the speech data and its corresponding transcriptions, an automatic as well as a manual checking are carried out after the recordings. Automatic verification involves superficially checking if the speech file is empty, or if the speech is too short, and so on. If more than 10% of such errors are found in one set, the whole dataset is considered unusable.

Manual checking is done by employing a simple concatenated TTS system to read out the prompt word, and that word's corresponding recorded speech is also played out. If the pair does not sound alright, the utterance is potentially in error, and further attention is given to it. In cases where a recorded speech does not correspond to its prompt, the prompt is changed, particularly its phonetic transcription, to match the speech. This approach seems quite effective since verification is such a tedious process, and it seems easier to detect errors by listening out for unnaturalness, compared to looking out for them visually.

Finally, a relational database using ACCESS is created to record the profiles of all speakers and their recorded items. This relational database can be queried using the SQL language to locate the waveforms transcribed using the specific phones or syllables, or even to locate the speaker who recorded the specific-phone waveforms.

3. Recognition

3.1. *Large Vocabulary Word Recognizer*

Figure 2 illustrates a series of four components, including a feature extractor, a unified acoustic model, a bilingual pronunciation dictionary and a language model. The feature extractor receives the waveform as its input and transforms it into the frame-based feature vectors O. The final outputs, the Chinese characters C, are closely dependent on the other pre-processors, and each of these pre-processors influences the results of our proposed bilingual speech recognizer.

Given the acoustic information O and the tonal syllable-based pronunciation S, the most likely Chinese character sequences C are found using the following expression:

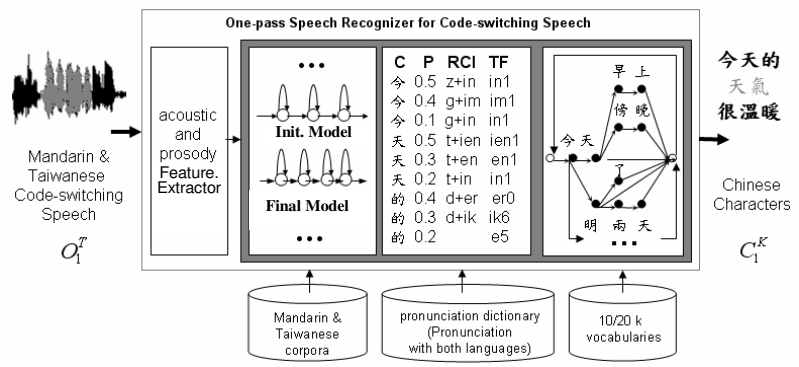

Fig. 2. The diagram of the one-pass speech recognizer.

$$\arg\max P(C \mid O, S) \tag{1}$$

Using standard probability theory, this expression can be equivalently written as

$$\arg\max P(O \mid S, C)P(S \mid C)P(C) \tag{2}$$

The three probability expressions in (2) are organized in such a way that acoustics of pronunciation and language information are contained in separate terms. In modeling, these terms are known as

1. $P(O \mid S, C)$ Tonal syllable acoustic model
2. $P(S \mid C)$ Pronunciation model
3. $P(C)$ The language model

In this framework,[9] Chinese character based decoding can be implemented by searching in a three-layer network composed of an acoustic model layer, a lexical layer, and a grammar layer. There are at least 2 critical differences between our framework and the conventional one. 1) In the lexicon layer, character-to-pronunciation mapping can easily incorporate multiple pronunciations in multiple languages, including Japanese, Korean, and even Vietnamese which also use Chinese characters. 2) In the grammar layer, characters instead of syllables are used as nodes in the searching network. Under this ASR structure, it does not matter which language the user speaks. Whether it is Taiwanese, Mandarin or a mixture of them even in one sentence, the ASR outputs only the Chinese characters, making the framework language/dialect independent. In another work,[10] we also used the framework to recognize Taiwanese-Mandarin code-switching speech.

3.2. *Feature Extraction with Tone Information*

Many researchers have included tone features in their tonal language recognizers, such as for Mandarin, Cantonese[11] and Taiwanese. They report that recognition accuracy rates increase as tonal features are incorporated into their recognizers. Our prior work[12] also confirms this point. Features carrying only acoustic information in one-pass recognizers will result in confusion and increase the number of searching nodes during the decoding process. Thus, tonal feature is necessary.

In this system, we adopted an algorithm based on auto-correlation and harmonic-to-noise ratio to estimate pitch. Pitch can only be correctly estimated in the voiced region of a waveform. In unvoiced portions, pitch is usually assigned zero in traditional pitch analysis programs. However, for speech recognition purposes, this zero padding strategy may not be appropriate because it will lead to problems of zero variances and undefined derivatives at voiced/unvoiced transitions. We fill the pitch gap with exponential decay/grow functions to connect the pitch contours of two voiced regions, and we call this *pitch smoothing by exponential functions*.

3.3. *Bilingual Acoustic Model*

For acoustic modeling, a unified approach for a hidden Markov model (HMM) based multi-lingual acoustic model is adopted. In this approach, a knowledge-based phone mapping method is applied to map phones across languages and reduce the effective number of phones in the multi-lingual acoustic model. When combining acoustic models of two different languages, we need to identify which phones are acoustically similar between the 2 languages, while knowing that other phones still need the use of language-dependent models. It is well known that language-dependent systems perform better than language-independent ones. Driven by this, using ForPA, we group all the phones in the different languages into phonetically and acoustically similar clusters.

Furthermore, in order to more efficiently merge the similar parts of the sounds from both languages, we use a tying algorithm based on a decision tree to cluster the HMM models by using the maximum likelihood criterion.[13] This approach has the advantage that every possible context acoustic model state can be classified by the tree, so that any back-off models can be avoided. In practice, this approach significantly reduced the syllable recognition error rate and the overall system parameters for unseen context acoustic model in previous LVCSR experiments.

3.4. *Tree-structured Language Searching Net*

A tree structure is a natural choice of representation for a large vocabulary lexicon, as many phonemes can be shared eliminating redundant acoustic evaluations. The advantage of using a lexicon tree representation is obvious: It can effectively reduce the state-search space of the trellis. Ney *et al.*[14] reported that a lexical tree has a saving factor of 2.5 over the linear lexicon. Besides, the efficiency of using a lexical tree is substantial, not only because it results in considerable saving of memory for representing state-search space, but it also saves a significant amount of time by searching in far fewer potential paths.

Figure 3 shows examples of a linear searching net and a tree-structured searching net. The perplexity of the linear searching net was found to be 5 while the tree-based one has a smaller perplexity of 4.89.

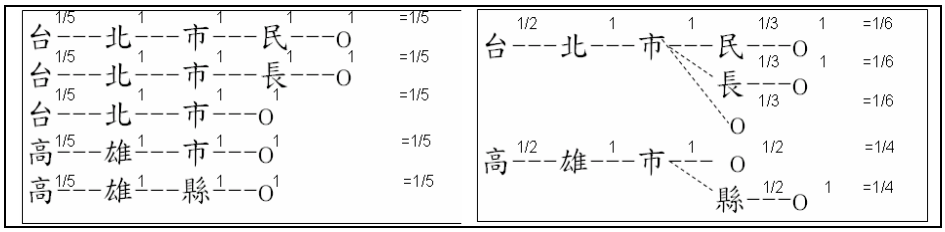

Fig. 3. The examples of isolated linear (left) and tree-structured (right) searching net with their probability values.

3.5. *Pronunciation Modeling Using Pronunciation Variation*

The pronunciation model plays an important role in our proposed one-pass Chinese character based ASR engine.[15] It not only provides more choices during decoding when the speaker exhibits variations in pronunciation, but also handles various speaking styles in different languages. As mentioned above, one Chinese character has more than two pronunciations in the combined phonetic inventory of Mandarin and Taiwanese. Accent and regional migration are also factors that influence the pronunciation or speaking style of speakers. In the following sub-sections, we propose two different methods, knowledge-based and data-driven methods, for obtaining rules of pronunciation variation.

3.5.1. *Knowledge-Based Method*

As shown by Strik,[16] information about pronunciation can be derived from knowledge sources, such as pronunciation dictionaries handcrafted by linguistic

experts, or from pronunciation descriptions and rules extracted from the literature. In this approach, a pronunciation variation rule is simply the multiple pronunciations that appear in the lexicon for the same character. Associated probabilities can be calculated as follows. 1) The character-pronunciation pairs are derived; 2) the frequencies of the pairs are counted, and the relative frequency with respect to the total frequency of the same Chinese character is calculated, and; 3) the pairs with high relative frequencies are kept as multiple pronunciation rules.

3.5.2. *Data-Driven Approach*

Although regular pronunciation variations can be obtained from existing linguistic and phonological information, such as from a dictionary, this knowledge base is not exhaustive. Many language variation phenomena in real speech have not yet been described or captured. Therefore, another way to derive pronunciation variations from acoustic clues is the data-driven method.[17] The algorithms from this method can then be used to derive formalizations. The information about pronunciation variation can be represented in terms of rewrite rules,[18] decision trees,[19] or neural networks.[17] Several other measures, for example confusability measures,[20] have been used to select rules or variants. In this section, we used a forced recognition approach to align the variation between transcripts of acoustic signals and the transcriptions of single tonal syllables in the lexicon. These variations then become the pronunciation rules added to the dictionary if the frequency measure of a particular variation falls within the selection criteria.

In practice, we combine both the knowledge-based and data-driven approaches for our pronunciation model. The reasons are twofold: 1) to be able to handle multiple pronunciations for one Chinese character in both languages by using knowledge-based extraction of pronunciation rules, and 2) to be able to handle pronunciation variations resulting from the speaker's speaking style or personal articulation by using the data-driven method of obtaining rules from acoustic signals. It is difficult to select an "optimum" number of pronunciations to be represented in the pronunciation model. Therefore, the weighting in these approaches are the same, and the determining of the final number of pronunciations is done by a different task.

4. Taiwanese Text-to-Speech

The TTS system proposed is composed of 3 major functional modules, namely a text analysis module, a prosody module, and a waveform synthesis module. The system architecture is shown in Figure 4. Since Taiwanese is a tonal language, we will describe the process of the tone sandhi rules for text analysis module. In the waveform synthesis module, the system adopts TD-PSOLA to modify waveforms by adjusting the prosody parameters of selected units so that the synthesized speech sounds more natural. Finally, the new TTS architecture is implemented in a multiple-level unit selection for a limited domain application.

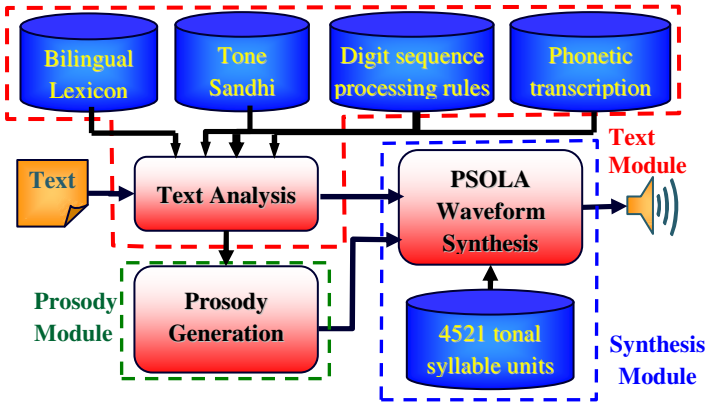

Fig. 4. The TTS system architecture is composed of 3 major functional modules: a text analysis module, a prosody module, and a waveform synthesis module.

4.1. *Word Segmentation and Mandarin-Taiwanese Translation (Sentence-to- Morpheme)*

Although the Hanlor orthography is the most common writing style of Taiwanese in contemporary Taiwan, all three types of written texts (see Section 1) can be analyzed by our text analysis module. Since there are no natural boundaries between two successive words, we must segment a Mandarin text into its word sequence first. The bilingual pronunciation dictionary is used as a basis for our word segmentation algorithm based on the sequentially maximal-length matching, which segments Mandarin sentences into maximal-length word combinations.[21]

Finally, we directly translate Mandarin into Taiwanese word-for-word, and transcribe the Taiwanese words phonetically into ForPA. This segmentation, translation and transcription process is exemplified in Figure 5, where the input Mandarin sentence is “他今天心情很好” (“he is very happy today”).

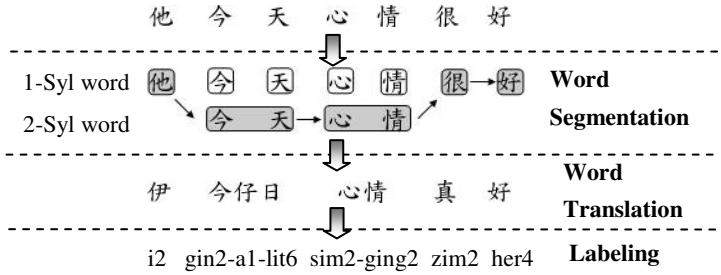

Fig. 5. The text analysis process, which combines word segmentation, translation and labeling (see next section), where the input Mandarin sentence is "他今天心情很好", "1-Syl Word" denotes one-syllable words, and so on.

4.2. *Labeling* (*Morpheme-to-Phoneme*) *and Normalization of Digit Sequences*

For each segmented word, there are more than one Taiwanese pronunciation. This multiple-pronunciation problem is tackled by a two-stage strategy. The first stage: Choose the everyday, or Bai-du-in, pronunciation first for initial phonetic labeling. If the everyday pronunciation does not exist, the classic literary pronunciation is considered. The second stage: Build a searching network with pronunciation frequencies as node information and pronunciation transitional frequencies as arc information for each sentence. Best pronunciation selection is then conducted by a Viterbi search.

Another important issue for text analysis is the normalization of digit sequences. However, for digits, the 2 manners of pronunciation are commonly heard. The choice of pronunciation depends on the position of the digit in a sequence. Another regularity is, if a digit sequence does not represent a quantity, it is pronounced in the classic literary way. These digit pronunciation rules are summarized in Table 7.

4.3. *Application of Tone Sandhi Rules*

One of the most frequently referred to sandhi rules states 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 with its lexical tone. Otherwise, it is pronounced with its sandhi tone. The Taiwanese sandhi rules for each lexical tone have been shown in Figure 1.

To produce more natural synthesis, finer aspects like the triple adjectives – where the first character of 3 duplicated adjectives will carry a very different tone

Table 7. The rules for normalization of digit sequences.

Position	Pronunciation
Ten Million	Read 0-9 as EP
Million	Read 0-9 as EP
Hundred Thousand	No sound for 1, 2 as LP and others as EP.
Ten Thousand	If the digit is 0 in hundred thousand position, read 0-9 as EP. If the digit is not 0 in hundred thousand position, 1, 2 read as LP, and others as EP.
Thousand	Read 0-9 as EP
Hundred	Read 0-9 as EP
Ten	No sound for 1, 2 as LP, and others as EP.
Unit digit	If the digit is 0 in hundred thousand position, read 0-9 as EP. If the digit is not 0 in hundred thousand position, 1,2 read as LP, and others as EP.

LP: classic, literary pronunciation, EP: everyday pronunciation.

Table. 8. Tone sandhi rules for triple adjectives.

lexical	1	2	3	4	5	6	7
sandhi-tone	9	9	4	1	9	9	6

other than the traditional 7 lexical tones mentioned previously – are handled. We map this "High-Rising" tone to digit 9, and call it tone 9. The tone sandhi rules for triple adjectives are summarized in Table 8.

4.4. *Evaluation of Text Analysis and Prosody Modules*

Following text analysis and prosody generation, an evaluative experiment is conducted. Its main target is to assess the accuracy rate of automatic transcription, which is produced by the text analysis and prosody modules, in comparison with manual transcription. A large amount of news reports are collected from the internet. The selection of these is random, without emphasis on any particular news category. Of these, a set of 200 sentences to cover all distinct Chinese characters are chosen. The comparative performance of manual and automatic transcription is shown in Table 9, with three sets of results: word segmentation, labeling and tone sandhi accuracy rates.[22]

From Table 9, we infer that the system can segment and translate most articles accurately into Taiwanese words with over 97% accuracy. If we do not consider tone sandhi, the system can translate an article into its correct

pronunciation close to the 88% rate, and most errors apply to names and out-of-vocabulary words. Because Taiwanese does not have uniform tone sandhi rules, it is acceptable that the accuracy rate of tone sandhi is lower.

Table 9. The statistics of performance in parts of word segmentation and transfer, labeling and tone sandhi.

	Expert1 (automatic)	**Expert2 (manual)**
Word Segmentation & Transfer	97.80%	98.76%
Labeling	89.96%	88.27%
Tone sandhi	65.43%	62.43%

4.5. *Waveform Synthesis Module*

There is a variety of synthesis methods, the most popular being the TD-PSOLA. We adopted this to modify the prosodic features of selected units. Synthesis components are used to not only raise or lower pitch, but also to extend or reduce duration. After the analysis of tonal syllables, we can gather duration and short pause information for each syllable. Based on the above information, the following instances were applied to the speech synthesis process:

Case 1: If the syllable consists of an unvoiced consonant (/p/-, /t/-, /g/-, /k/-, /z/-, /s/-, /c/-, /h/-), the system modifies the duration of the unvoiced initial, and modifies the duration and pitch of final.

Case 2: The system modifies duration and pitch both on the initial and final if an unvoiced initial is not present.

Case 3: The system replaces the short pause with a zero-value section.

4.6. *Multiple-Level Unit Selection for a Limited-Domain Application*

In the past, the production of audio books is a demanding process, effort-wise and time-wise. The application of TTS to a limited-domain audio book can save a considerably large amount of time, and the synthesized sentences are likely to turn out comparable to sentences recorded in natural speech.

The Taiwanese Bible seems to be a good example to prove this, being a high-quality, reliably translated document which has been validated by many high-ranking church officials and Christian devotees in general (domain experts). An audio book for this Bible would be a great help for speakers unable to read, making the Bible more accessible to the masses.[23]

Table 10. The textual statistics of the Taiwanese New Testament Bible.

Number of Chinese characters	278,633
Number of chapters	27
Number of sentences	39,171
Number of distinct words	7,189

From the statistics in Table 10, we find that the usage of duplicate words is highly frequent. If these much-duplicated syllables, words or sentences could serve as a gauge, the production cost of an audio book could be greatly reduced. But sounding natural is proportionally related to the number of synthesis units. Synthesis systems with fewer units tend to generate less natural-sounding speech. For instance, logically, any arbitrary sentence can be synthesized by all 4,609 tonal syllables in Taiwanese, but the prosody in this synthesized sentence would rank low in naturalness. When words are used as synthesized units, a total of 7,189 distinct words are required. However, the result of this would only be a slight improvement from the syllable-synthesis method. This is because, most Taiwanese words are monosyllabic anyway. Therefore, finding compromise units between the word and the sentence becomes a very important issue.

From our observation, poor quality (least natural sounding) synthesis occurs often in the concatenation of monosyllabic words. In fact, the naturalness of multi-words (or multiple words) unit combined by multi-syllabic words could result in only a slight improvement. It is therefore necessary to divide the multi-words units into two categories: concatenation of monosyllabic words and multi-syllabic words. Preference is given to concatenating monosyllabic rather than multi-syllabic words as synthesis units. To find high-frequency and longer-length synthesis units, we adopted an evolution method of maximum maximal-length words matching, called *maximal-multi-words matching*, which is described as follows:

Step 1. Let W_i be the i-th input sentence in the text corpus, where W_i is composed of N_i words and each word is separated by the equal sign denoted as $W_i = \left\{ w_i^1 = w_i^2 = w_i^3 = ... = w_i^{N_i} \right\}$. The length of the matching pattern is set to n, i.e. the pattern is $W_i^n = \left\{ w_i^1 = w_i^2 = w_i^3 = ... = w_i^n \right\}$, where n is smaller than or equal to N_i.

Step 2. Initially, let $n = N_i$, i.e. the matching pattern is W_i. Let $N(w_i^n)$ denote the count of the pattern of the multi-words W_i^n of the text corpus.

Step 3. If $N(w_i^n) \neq 0$ and $n = N_i$, repeat Step 1 with the next sentence. Else, if $N(w_i^n) = 0$, repeat Step 2 with $n = N_i$-1.

Finally, the number of multi-words synthesis units is 9,992, extracted from the 280,000-syllable Taiwanese New Testament Bible. There are 22,949 Chinese characters in the set, including 7,189 distinct words in this Bible. The total amount of recording counted in syllables is one of tenth of the whole Bible.

5. Conclusion

To collect a Taiwanese speech corpus, we designed a new phonetic alphabet, called the ForPA, to transcribe recorded speech read out from the phonetically-abundant word sheets. These sheets were generated by the application of a greedy heuristic algorithm and contained several kinds of context-dependent phone units. To date, the validated multi-lingual speech database has reached 92.77 hours of validated speech from 1,700 speakers.

For Taiwanese speech recognition, a new framework based on a unified approach with a one-stage searching strategy was implemented using a multiple pronunciation lexicon and a large vocabulary searching network with Chinese characters as its nodes. The lexicon was generated by both data-driven and knowledge-based statistical approaches. This framework shows its validity and efficiency to deal with the Mandarin/Taiwanese bilingual speech recognition issue from a unified angle.

For the TTS system, we significantly exploited knowledge on Taiwanese phonetics/linguistics to improve the natural quality of the synthesized speech. In addition, the Mandarin-to-Taiwanese bilingual TTS system has been successfully constructed based on the information from our bilingual lexicon. Most Mandarin articles can thus be translated into Taiwanese and automatically converted to a speech signal in Taiwanese. In a limited-domain application, a new multi-level unit selection algorithm was used to produce an audio book of the Taiwanese Bible. This technology has good potential to be used in language-learning tools and applications.

The collected Taiwanese corpus (in ForPA) is significant and can be used for further research of Taiwanese ASR and TTS. In future, a similar corpus will be created for yet another language in Taiwan, namely Hakka. It must be noted that the prediction of tone sandhi should be improved to achieve more accurate/natural synthesis. The use of signal processing techniques is also imperative to smoothen out waveforms and reduce discontinuity in most TTS systems, including ours. Finally, there is indeed a big need to propose more novel frameworks and approaches in speech recognition to significantly improve its performance for character-based languages.

References

1. Wikipedia. Available from http://en.wikipedia.org/wiki/Demographics_of_Taiwan (2006).
2. YuangChin Chiang, *A Course in Taiwanese Pinim*, (in Taiwanese) AnKor Publishing, PingTong (2005).
3. S.-Y. Zhou, *T3 Taiwanese Treebank and Brill Part-of-Speech Tagger*, (in Chinese), Master thesis of National TsingHua University, HsinChu, Taiwan (2006).
4. J. Wells, SAMPA (Speech Assessment Methods Phonetic Alphabet), http://www.phon.ucl.ac.uk/home/sampa/home.htm, April, (2003).
5. J. L. Hieronymus, "ASCII Phonetic Symbols for the World's Languages: Worldbet," *Technical Report AT&T Bell Labs*, (1994).
6. R.-Y. Lyu et al. "Toward Constructing A Multilingual Speech Corpus for Taiwanese (Min-nan), Hakka, and Mandarin," *ICLCLP Vol. 9, No. 2*, (August 2004), pp. 1-12
7. YuangChin Chiang, *An input method editor (IME) in Taiwanese Pinim*, (2005).
8. M.-S. Liang et al. "An Efficient Algorithm to Select Phonetically Balanced Scripts for Constructing Corpus," *NLP-KE*, (2003).
9. R.-Y. Lyu et al. "A Unified Framework for Large Vocabulary Speech Recognition of Mutually Unintelligible Chinese *Regionalects*," *In Proc. of ICSLP 2004*, (2004).
10. D.-C. Lyu et al.. "Speech Recognition on Code-Switching Among the Chinese Dialects," *In Proceedings of IEEE ICASSP'06*, (2006).
11. P. F. Wong and M. H. Siu, "Integration of Tone-related Feature for Chinese Speech Recognition," *in Proceedings on ICMI*, (2002) pp. 476-479.
12. D.-C. Lyu et al. "Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling," *In Proc. of Eurospeech*, (2003).
13. D.-C. Lyu et al. "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Taiwanese/Mandarin Continuous Speech Recognition," *In Proceedings of the 9th SST*, Melbourne (2002).
14. H. Ney et al. "Improvements in Beam Speech for 1000-Word Continuous Speech Recognition," *In Proc. of the ICASSP'92*, California, (1992), pp. 9-12.
15. D.-C. Lyu et al.. "Modeling Pronunciation Variation for Bi-Lingual Mandarin/Taiwanese Speech Recognition" *IJCLCLP, Vol. 10. no. 3*. (2005), pp. 363-380.
16. H. Strik and C. Cucchiarini, "Modeling Pronunciation Variation for ASR: Overview and Comparison of Method," *Speech Communication, Vol. 29*, (1999) pp. 225-246.
17. T. Fukada, et al. "Automatic Generation of Multiple Pronunciations Based on Neural Networks and Language Statistics," *In Proceedings of ESCA*, (1998), pp. 103-108.
18. N. Cremelie, and J. P. Martens, "In Search of Better Pronunciation Models for Speech Recognition," *Speech Communication 29*, Vol. 4 (2), (1999), pp. 115-136,
19. J. J Humphries et al. "Using Accent-Specific Pronunciation Modelling for Robust Speech Recognition," *In Proc. ICSLP-96*, (1996), pp. 2324-2327.
20. M. Wester and E. Fosler-Lussier, "A Comparison of Data-Derived and Knowledge-Based Modeling of Pronunciation Variation," *In: Proc. ICSLP 2000, Vol. 4,* (2000), pp. 270-273.
21. M.-S. Liang et al. "A Bi-lingual Mandarin-To-Taiwanese Text-to-Speech System," In *Proceedings Int. Conf. on Spoken Language Processing (ICSLP)*, (2005).
22. M.-S. Liang et al. "A Taiwanese Text-to-Speech System with Applications to Language Learning," *In Proc. ICALT 2004*, Joensuu, Finland, (2004).
23. K.-C. Chuang, *Phrase-based Synthesis Units and Study of Phrase Tone-Sandhi for Taiwanese Text-To-Speech and Application*, Master Thesis, University of Chang Gung, Taiwan, (2005).
24. S. C. Kumar et al. "Multilingual Speech Recognition: A Unified Approach," *In Proc. of Eurospeech*, Portugal, (2005).