Computational Linguistics and Chinese Language Processing
Vol. 9, No. 2 , August 2004, pp. 1-12

Toward Constructing A Multilingual Speech Corpus for Taiwanese (Min-nan), Hakka, and Mandarin

Ren-yuan Lyu[*], Min-siong Liang[+], Yuang-chin Chiang[**]

Abstract

The **Formosa** speech database (ForSDat) is a multilingual speech corpus collected at Chang Gung University and sponsored by the National Science Council of Taiwan. It is expected that a multilingual speech corpus will be collected, covering the three most frequently used languages in Taiwan: Taiwanese (Min-nan), Hakka, and Mandarin. This 3-year project has the goal of collecting a phonetically abundant speech corpus of more than 1,800 speakers and hundreds of hours of speech. Recently, the first version of this corpus containing speech of 600 speakers of Taiwanese and Mandarin was finished and is ready to be released. It contains about 49 hours of speech and 247,000 utterances.

Keywords: Phonetic Alphabet, Pronunciation Lexicon, Phonetically Balanced Word, Speech Corpus

1. Introduction

To design a speaker independent speech recognition system, it is essential to collect a large-scale speech database. Taiwan (also called ***Formosa*** historically), which has become famous for its IT industry, is basically a multilingual society. People living in Taiwan usually speak at least two of the three major languages, including Taiwanese (also called Min-nan in the linguistics literature), Hakka and Mandarin, which are all members of the Chinese language family. In the past several decades, most of the researchers studying natural language processing, speech recognition and speech synthesis in Taiwan have devoted themselves to research on Mandarin speech. Several speech corpora of Mandarin speech have, thus, been collected and distributed [Wang *et al.*, 2000; Godfrey, 1994]. However, little has been done

[*] Dept. of Computer Science and Information Engineering, Chang Gung University, Taoyuan, Taiwan
Email: rylyu@mail.cgu.edu.tw Tel: 886-3-2118800ext5967, 5709
[+] Dept. of Electrical Engineering, ,Chang Gung University, Taoyuan, Taiwan
[**] Inst. of Statistics, National Tsing Hua University, Hsin-chu, Taiwan

on the other two languages used in daily life. In this paper, we describe a government-sponsored project which aims to collect a large-scale multilingual speech corpus, namely, the ***Formosa Speech Database*** (ForSDat), covering these three languages used in Taiwan. The construction of ForSDat is a 3-year project, the goal of which is to collect hundreds of hours of speech from up to 1,800 speakers. So far, we have finished about one-thrid of what the project is expected to achieve.

This paper is organized as follows. Section 1 is the introduction. Section 2 describes the ***Formosa Phonetic Alphabet*** (ForPA), which is being used to transcribe all the speech data and the pronunciation lexicons. Section 3 discusses the phonetically balanced word sheets used to record speech utterances. Section 4 reports the software tools used for corpus collection. Section 5 describes the information obtained about speakers. Section 6 provides information about the database information. Section 7 discusses data validation, and section 8 is a conclusion.

2. The Phonetic Alphabet and the Pronunciation Lexicon

One of the preliminary jobs involved in constructing a speech corpus is to build up a pronunciation lexicon. We have set up several pronunciation lexicons composed of more than 60,000 words for Taiwanese, more than 70,000 words for Mandarin and more than 20,000 words for Hakka. Each item in the lexicons contains a Chinese character string and a string of phonetic symbols encoded in the ***Formosa Phonetic Alphabet*** (ForPA), which will be described in the following paragraphs.

2.1. Formosa Phonetic Alphabet (ForPA)

Many symbolic systems have been developed for labeling the sounds of languages used throughout the world. One of the most popular systems is the International Phonetic Alphabet (IPA). Since many IPA symbols are not defined in the ASCII code set and are not easy to manipulate, many ASCII-coded IPA symbolic sets have been proposed in the literature. Two popular systems are SAMPA [Wells, 2003] and WorldBet [Hieronymus, 1994]. It has claimed that one can select parts of these phone sets for a specific language. However, both ASCII-coded phonetic systems have many symbols that are difficult to read, such as “@”or “&”. In addition, since these systems are designed for all the languages used around the world, they are too complex to be applied to some local languages, like those that will be addressed here.

The most widely known phonetic symbol sets used to transcribe Mandarin Chinese are the Mandarin Phonetic Alphabet (MPA, also called Zhu-in-fu-hao) and Pinyin (Han-yu-pin-yin), which have been officially used in Taiwan and Mainland China, respectively, for many

years. However, both systems are inadequate for application to the other members of the Chinese language family, like Taiwanese (Min-nan) and Hakka. Among the phonetic systems useful for Taiwanese and Hakka, there are Church Romanized Writing (CR, also call Peh-e-ji, 「白話字」) [Chiung 2001] for Taiwanese and the Taiwan Language Phonetic Alphabet (TLPA) [Ang 2002] for Taiwanese and Hakka. Because the same phonemes are represented using different symbols in Pinyin, CR and TLPA, it is confusing to learn these phonetic systems simultaneously. For example, the syllable "pa(ㄅㄚ)" in TLPA and "pa(趴)" in CR may be confused with each other because the phoneme /p/ is pronounced differently in the two systems.

Therefore, it is necessary to design a more suitable phoneme set for multilingual speech data collection and labeling [Zu, 2002][Lyu, 2000]. The whole phone set for the three major languages used in Taiwan is listed in Table 1 for four phonetic systems: MPA, Pinyin, IPA, and the newly proposed ForPA. Table 1 also lists examples of syllables and characters which contain the target phonemes.

It is known that phonemes can be defined in many different ways, depending on the level of detail desired. The labeling philosophy adopted in ForPA is that when faced with various choices, we prefer not to divide a phoneme into distinct allophones, except in cases where the sound is clearly different to the ear or the spectrogram is 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 any finer distinctions that might exist within any context. Generally speaking, ForPA might be considered as a subset of IPA, but it is more suitable for application to the languages used in Taiwan.

Table 1. The phone set for the three languages in Taiwan, represented as different phonetic systems. The Chinese character in parentheses followed by a syllable, is an example character used in Mandarin, e.g., "ba(ㄦ)" is pronounced in Mandarin as syllable "ba", without considering the tone. For phonemes not found in Mandarin Chinese, we use Chinese character pronounced as Taiwanese (^{T}) or Hakka (^{H}) to be example characters. For example, "bha(肉^{T})" meaning "肉" is pronounced "bha" in Taiwanese.

ForPA	b	p	m	f	d	t	n	l	g	k	h	zh	ch	sh
Syllable (字)	ba(八)	pa(趴)	ma(媽)	fa(發)	da(搭)	ta(他)	na(那)	la(拉)	ga(嘎)	ka(咖)	ha(哈)	zha(渣)	cha(差)	sha(殺)
Pinyin	b	p	m	f	d	t	n	l	g	k	h	zh	ch	sh
MPA	ㄅ	ㄆ	ㄇ	ㄈ	ㄉ	ㄊ	ㄋ	ㄌ	ㄍ	ㄎ	ㄏ	ㄓ	ㄔ	ㄕ
IPA	p	p'	m	f	t	t'	n	l	k	k'	x	tʂ	tʂ'	ʂ
SAMPA	p	p_h	m	f	t	t_h	n	l	k	k_h	x	ts'	ts_h'	s'
WorldBet	p	ph	m	f	t	th	n	l	k	kh	x	tsr	tsrh	sr

ForPA	rh	z	c	s	r	bh	gh	v	ng
Syllable (字)	rhan(然)	za(匝);zi(機)	ca(擦);ci(七)	sa(撒);si(西)	ru(如^{T});ri(字^{T})	bha(肉^{T})	ghua(我^{T})	voi(會^{H})	ang(氨);nga(雅);ng(黃)
Pinyin	r	z;j	c;q	s;x					-ng
MPA	ㄖ	ㄗ;ㄐ	ㄘ;ㄑ	ㄙ;ㄒ					ㄤ
IPA	z̩	ts;tɕ	ts';tɕ'	s;ɕ	z;ü	b	g	v	N
SAMPA	z'	ts';ts\	ts_h;ts_h	s;s\	z;z\	b	g	v	N
WorldBet	zr	ts;cC	tsh;chC	s;C	z;zr	b	g	v	N

ForPA	a	o	er	e	err	i	u	yu	ii	-nn	-p	-t	-k	-h
Syllable (字)	a(阿)	o(喔)	er(鵝)	ie(也)	err(而	i(一)	u(吳)	yuan(原)	zii(資)	ann(餡TH)	ap(壓TH)	at(握TH)	ak(沃TH)	ah(鴨TH)
Pinyin	a	o	e	ê	er	y;i	w;u	yu;ü	i					
MPA	ㄚ	ㄛ	ㄜ	ㄝ	ㄦ	ㄧ	ㄨ	ㄩ	ㄭ					
IPA	a	o	ɣ	ɛ	ɚ	i	u	y	ɨ	ã	-p	-t	-k	-h
SAMPA	A	o	7	E	@'	i	u	y	i'					
WorldBet	A	o	2	E	&r	i	u	y	4r	~				

2.2. Formosa Lexicon (ForLex): A Pronunciation lexicon composed of Taiwanese, Hakka and Mandarin

Before producing word sheets for speakers to utter, a complete pronunciation lexicon needs to be prepared. A lexicon has been collected in this project to meet the requirement. This lexicon, called the Formosa Lexicon (ForLex), was adapted from three other lexicons: the CKIP Mandarin lexicon , Gang's Taiwanese lexicon, and Syu's Hakka lexicon [CKIP 2003] [Syu 2001]. Some statistical information about the lexicon was listed in Table 2.

Table 2. The distribution of words in three lexicons: Gang's Taiwanese lexicon, Syu's Hakka lexicon, and the CKIP Mandarin lexicon.

	1-Syl	2-Syl	3-Syl	4-Syl	5-Syl	
Gang	8027	44846	12129	1823	161	
Syu	7322	9161	4948	2382	21	
CKIP	6863	39733	8277	9074	435	
	6-Syl	7-Syl	8-Syl	9-Syl	10-Syl	Total
Gang	0	0	0	0	0	66986
Syu	3	0	0	0	0	23837
CKIP	223	125	52	2	8	64792

3. The process of producing phonetically balanced word sheets

Based on the three pronunciation lexicons transcribed in ForPA, we extracted sets of distinct syllables and inter-syllabic bi-phones from the three languages. The statistics of the phonetic units considered here are listed in Table 3. In order to collect speech data related to the co-articulation effect of continuous speech, we extracted phonetically abundant word sets. Therefore, the chosen phonetic units were not only base-syllables, phones, and 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 [Shen *et al*., 1999], which is NP-hard. Here, we adopted a simple greedy heuristic approximate solution [Cormen, 2001].

First, we set the requirements of the word set as to cover the following phonetic units: Base-syllables and Inter-syllabic RCD phones. Accordingly, the selected word set could cover all the phones, Initial-Finals, RCD phones, RCD Initial-Finals, Base-syllables and Inter-syllabic RCD phones. In this way, we could obtain several sets of words for our balance-word data sheets. All the statistics of the phonetic units considered here are listed in Table 3. [Liang 2003].

Table 3. The numbers of distinct subwod units for each of the three languages and their unions, where T: Taiwanese; H: Miaulik-Hakka M: Mandarin; ∪: union.

Language	Base syllable	Phones	Within-syllabic bi-phones	Inter-syllabic bi-phones
T	832	53	410	716
H	683	53	327	696
M	429	45	208	234
T∪H	1134	70	583	1036
T∪M	1055	64	486	809
H∪M	939	71	435	797
T∪H∪M	1326	78	600	1105

3.1. Data sheets

The process of producing data sheets is depicted in Fig.2. Before we produced the data sheets, we defined the sheets' coverage rate. The coverage rate of the sheets was defined as the total number of base-syllables (or inter-syllabic phones) over the number of all possible distinct base-syllables (or inter-syllabic phones). The format of the data sheet is partially shown in Table 4.

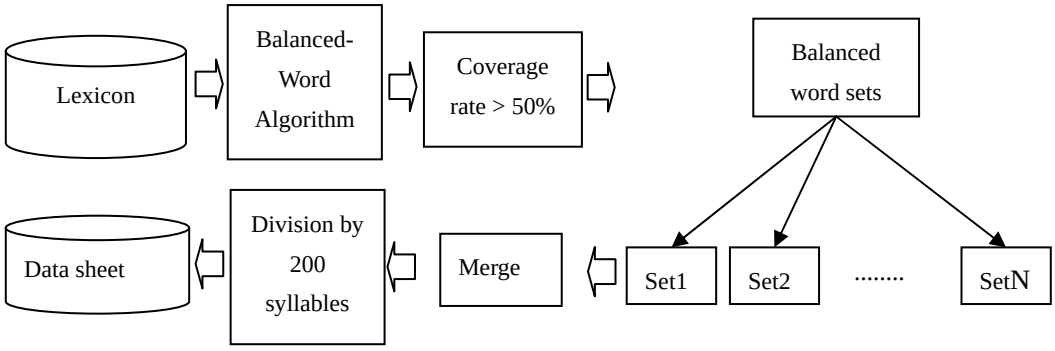

Figure 2. The process of producing data sheets.

Table 4. Some examples from the data sheets used to collect ForSDat.

Filename	Text	Transcription in ForPA
blwr00000	觀世音菩薩	guan1_se3_im1_po5_sat7
blwr00001	驚 ga 刺激著	giann1_ga2_ci3_gik7_diorh6
blwr00002	藥檢實驗室	iorh6_giam4_sit6_ghiam2_sik7
blwr00003	藝術工作者	ghe2_sut6_gang1_zok7_zia4

In terms of Taiwanese sheets, although we produced 364 balanced-word sets in total, we only used sets whose coverage rates exceeded 50%. Because the variation in the numbers of syllables or words in some sets was very high, we merged those sets and then re-segmented them to produce data sheets. Finally, each sheet contained about 200 syllables. The numbers of data sheets and total words were 446 and 37,275, respectively.

As for Miaulik-Hakka sheets, all the balanced-word sets were concatenated in sequence and then segmented into data sheets, each of which contained 70 words. Finally, we got 340 data sheets, which consisted of 23,837 words.

For Mandarin, one phonetically-rich set was segmented equally into ten sheets, and every sheet consisted of roughly 300 words. In addition, all the tonal-syllables were segmented into ten equal-size data sheets.

4. The software tools used for corpus collection

Two kinds of database collection systems are being used to create ForSDat. They are microphone and telephone systems, respectively.

4.1. The telephone recording system

The telephone system is set up in the Multi-media Signal Process Laboratory at Chang Gung University. The speakers dial into the laboratory using a handset telephone. Before recording, we give the speakers prompt sheets. The input signal is in format of 8K sampling rate with 8-bits μ-law compression. The speakers utter words while reading the prompt sheet, and supervised prompt speech is played to help the speakers follow the prompt speech to finish the recording. After recording, all speech data are saved in a unique directory. Figure 3 shows the recording process carried out using the telephone system.

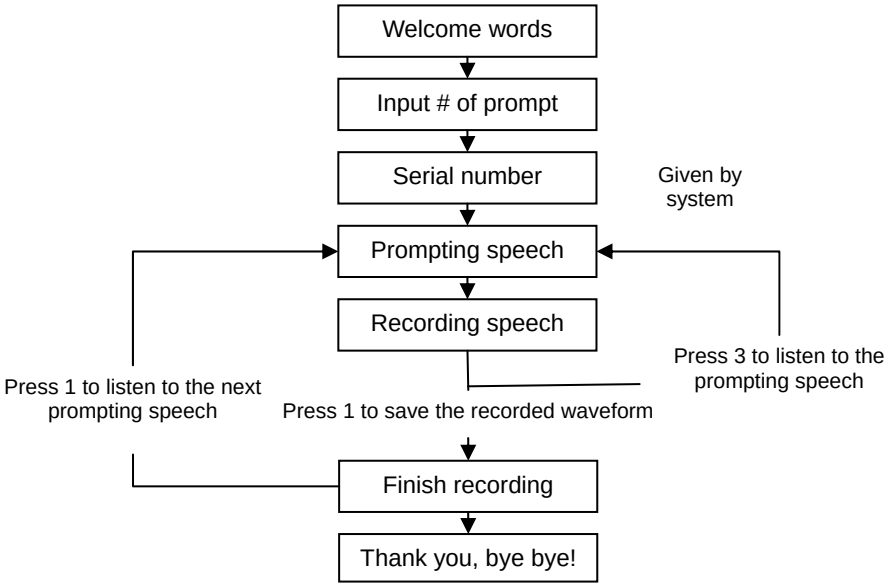

Figure 3. The telephone recording system.

4.2. The microphone recording system

When we record a waveform into a computer, it is not convenient to type the file name necessary for saving it. Therefore, we use a good tool (DQS3.1) [Chiang 2002] to record speech. If we create a script in a specific form for this software, we can record the waveform easily and get a labeled file, which contains information of transcription using ForPA. Then, we simply set up the system on a notebook computer and take it wherever we want to record speech.

5. Speaker recruiting

We employ several part-time assistants to recruit speakers around Taiwan. Each speaker is

asked to record one sheet and receives a remuneration after finishing recording. Each part-time assistant receives a remuneration when they recruits a speaker.

5.1 Profiles of speakers

After a recording is finished, we ask the speakers to provide us with their profiles. This is useful for arranging speech data later. The user can also design experiments according to these profiles (see Fig. 4). The profile of a speaker includes the following attributes:

i. the name and gender of the speaker;

ii. the age and birthplace of the speaker;

iii. the location of the speaker and time;

iv. the number of years of education of the speaker.

編號	unusable	人員編號	姓名	性別	年齡	錄音劇本	教育程度	語言能力
518	*	t007-g021	張筑涵	F	*	050	3	10000
796	*	t004-b013	蘇裕盛	M	*	026	3	11010
795	*	t004-b013	蘇裕盛	M	*	025	3	11010
517	*	t007-g021	張筑涵	F	*	049	3	10000
549	*	t007-g005	蕭雯華	F	*	013	3	10010
550	*	t007-g005	蕭雯華	F	*	014	3	10010
553	*	t007-g003	林怡萱	F	*	009	3	11010
554	*	t007-g003	林怡萱	F	*	010	3	11010

Figure 4. A portion of a speaker's profile in the database.

5.2. Speech data format

We save the utterance in a binary file. If the speech is recorded using a microphone, we save it as a 16KHz/16bits PCM file and a corresponding label file that contains the phonetic transcription for a word. Otherwise, we save the utterances as a 8KHz/8bits μ-law file if the speech data were obtained over a telephone.

6. Database information

The database has been collected over both microphone and telephone channels, namely, ForSDat-TW01, ForSDat-MD01 and ForSDat-TW02, respectively. The tag "TW01" means that a portion of the database was collected in 2001 in Taiwanese. In the other hand, the tag

"M0" means that the recording channel used was a microphone and gender was female, and so on. Every speaker has a unique serial number and speech data, which contain a transcription of waveforms made in the early stage and are stored in a unique folder named according to the serial number. The database structure is shown in Fig.5. All the statistics of the database are listed in Table 5.

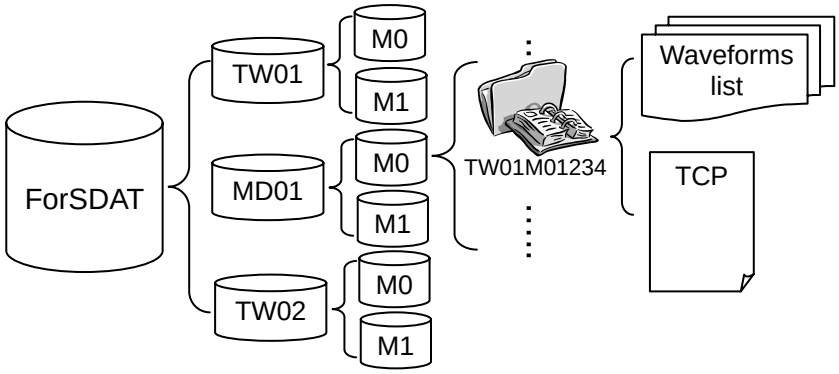

Figure 5. The structure of database for Taiwanese and Mandarin. (TW01: Taiwanese database collected in 2001; M0: the microphone channel was used and the gender was female, T1: the telephone channel was used and the gender was male; and so on. There is a transcription file for each unique speaker.)

Table 5. The statistics of utterances, speakers and data length for speech collected over microphone and telephone channels in Taiwanese and Mandarin (MIC: microphone; TEL: telephone).

	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	

7. Database validation

After the speakers have finished recording, the speech data need to be validated. This step can guarantee that the speech data will be useful for training the acoustic models of the speech recognizer. Although the data sheets are designed to be as readable as possible and we provide prompting speech for speakers, the utterances still are not compatible with the prompt. We thus validate the speech data using a specially designed software tool, which has the user interface shown in Fig.6 and the functions described in the following subsections.

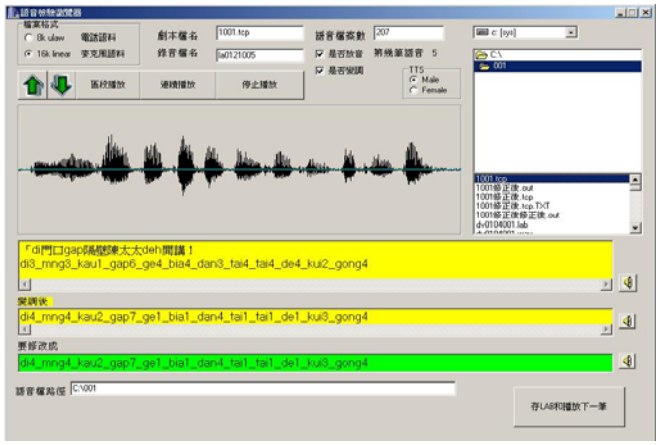

Figure 6. The software tool for validation.

7.1. Step 1: pre-processing

We browse all the waveforms using the validation tool and check whether the following problems occur:

1. the voice is cut off;i.e., the speakers pronounce too fast;

2. the voice file is empty;

3. there are other sounds mixed into the waveform, such as the voices of other people or the sounds of vehicles;

4. the speakers laughed when the waveform was being recorded.

If any one of the above problems are found, the speech file is considered unusable. If the total number of unusable files exceeds 10% of all the files in the directory, the directory is considered unusable. The speaker will then be asked to record the work sheet again.

Other problems may also occur. For example, two speakers may record speech data

inturns in one work sheet, etc. These directories are also considered unusable.

7.2. Step 2: phonetic transcription by means of forced alignment

After the speech data is pre-processed, we validate it to determine whether the labels that consist of phonetic transcriptions correspond to the speech data. We use two methods to achieve this goal. First, we use HTK [Steven, 2002] to perform forced-alignment automatically on an utterance using all possible syllable combinations. We keep the highest scores for combinations to transcribe the speech. Secondly, we use the TTS (text-to-speech) technique to synthesize all the labels that were transcribed using HTK and then we transcribe the speech manually using more appropriate phonetic symbols. Finally, we can construct a relational database using ACCESS to record all the profiles of the speakers (see Fig.3) and what they recorded. Therefore, we can query the speech database using the SQL language to find the waveforms transcribed using the specific phones or syllables or even query who recorded the specific-phone waveforms. This step is on-going and will be finished soon.

8. Conclusion

Version 1.0 of this corpus containing the speech of 600 speakers of Taiwanese (Min-nan) and Mandarin Chinese has been finished and is ready to be released. We have collected the speech of 1,773 people, including 49.47 hours of speech and 247,027 utterances. As work on this project continues, more Hakka and Mandarin speech data will be collected.

References

Wang, H. C., F. Seide, C.Y. Tseng and L.S. Lee, "Mat-2000 – design, collection, and validation of a mandarin 2,000-speaker telephone speech database," *International Conference on Spoken Language Processing 2000*, Beijing, China, 2000.

Godfrey, J., "Polyphone: Second anniversary report," *International Committee for Co-ordination and Standardisation of Speech Databases Workshop 94*, Yokohama, Japan, 1994.

Zu, Y., "A super phonetic system and multi-dialect Chinese speech corpus for speech recognition," *International Conference on Spoken Language Processing 2002*, Denver, USA, 2002.

Lyu, R. Y., "A bi-lingual Mandarin/Taiwanese (Min-nan), Large Vocabulary, Continuous speech recognition system based on the Tong-yong phonetic alphabet (TYPA)," *International Conference on Spoken Language Processing 2000*, Beijing, China, 2000.

Liang, M. S., R. Y. Lyu and Y. C. Chiang, "An efficient algorithm to select phonetically balanced scripts for constructing corpus," *IEEE International Conference on Natural Language Processing and Knowledge Engineering 2003*, Beijing, China, 2003.

Shen, J. L., H. M. Wang, R. Y. Lyu and L. S. Lee, "automatic selection of phonetically distributed sentence sets for speaker adaptation with application to large vocabulary mandarin speech recognition," *Computer speech and language*, vol. 13, no. 1,pp. 79-97, Jan. 1999.

Cormen, T. H. ect, "Chapter 37: Approximation Algorithm", *Introduction to Algorithm*, pp. 974-978, 2001.

Chiang, Y. C., and R. Y. Lyu, the speech recording system which are developed by Dr. Yuang-Chin Chiang at Nation Tsing Hua University, 2002.

Steven, Y., "The HTK book version 3.2", Cambridge University Engineering Department, 2002.

Wells, J., SAMPA (Speech Assessment Methods Phonetic Alphabet), http://www.phon.ucl.ac.uk/home/sampa/home.htm, April, 2003.

CKIP, Chinese Knowledge Information Processing, http://rocling.iis.sinica.edu.tw/CKIP/, 2003.

Syu, J. C., "Hakka dictionary of Taiwan", Nantian Bookstore published, 2001.

Hieronymus, J., "ASCII phonetic symbols for the world's languages: Worldbet," *AT&T Bell Laboratories, Technical Memo*, 1994.

Ang, U., Taiwan Language Phonetic Alphabet(TLPA), Taiwan Languages and Literature Society, http://www.tlls.org.tw/, 2002.

Chiung, W. V. T., "Romanization and Language Planning in Taiwan," *The Linguistic Association of Korea Journal 9(1)*, pp. 15-43, 2001.