AN EFFICIENT ALGORITHM TO SELECT PHONETICALLY BALANCED SCRIPTS FOR CONSTRUCTING A SPEECH CORPUS

Min-siong Liang[2], Ren-yuan Lyu[1,2], Yuang-chin Chiang[3]

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

ABSTRACT

In this paper, we describe an efficient algorithm to select phonetically balanced scripts for collecting a large-scale multilingual speech corpus. It is expected to collect a multilingual speech corpus covering three most frequently used languages in Taiwan, including Taiwanese (Min-nan), Hakka, and Mandarin Chinese. To achieve the objective, the first step is to construct a multilingual phonetic alphabet, namely Formosa Phonetic Alphabet (**ForPA**). In addition, the multilingual lexicons (**Fomosa Lexicons**) are also important parts for building the corpus. Until now, this corpus containing 600 speakers' speech of Taiwanese (Min-nan) and Mandarin Chinese has been finished and ready to release. There contains about 40 hours of speech in 247 thousand utterances in this release.

Keywords: Phonetic alphabet, Pronunciation lexicon, Phonetically-balanced word, Speech corpus

1. INTRODUCTION

It is essential to collect a large-scale speech database for research and development of speech technology, especially for designing a speaker independent speech recognition system. Taiwan (also called ***Formosa*** historically), being famous for its Information Technology (IT) Industry in the past decades, is basically a multilingual society. People living in Taiwan usually speak at least two of three major languages, including Taiwanese (also called Min-nan in linguistic literatures), Hakka and Mandarin Chinese, which are all members of the Chinese language family. In the past several decades, most researchers of natural language processing, speech recognition and speech synthesis in Taiwan devoted themselves to the research for Mandarin speech. Several speech corpus of Mandarin speech has thus been collected and distributed [Wang et. al. 2000]. However, little has been done about the other two daily used languages. In this paper, we describe an efficient algorithm to select phonetically balanced scripts to collect a large-scale multilingual speech corpus, namely ***Formosa Speech Database*** (ForSDat), covering those three languages daily used by people in Taiwan. The construction of ForSDat is a 3-year project, the goal of which is to collect up to 1800 speakers and hundreds of hours of speech. Until now, we have finished about 1/3 of what the project is expected to achieve.

This paper is organized as follows: section 2

describes the Phonetic Alphabet, the ***Formosa Phonetic Alphabet*** (ForPA), which was used to transcribe all the speech data, and the pronunciation lexicons; section 3 is about the process of the phonetically balanced word sheets for speakers to utter; section 4 is about database information; and finally section 5 is a brief conclusion.

2. THE PHONETIC ALPHABET AND THE PRONUNCIATION LEXICON

One of the preliminary task to construct speech corpus is to build up a pronunciation lexicon. We have set up pronunciation lexicons of more than 60 thousand words for Taiwanese, 70 thousand words for Mandarin and 23 thousand words for Hakka. Each item in the lexicon contains a Chinese character string and a string of phonetic symbols encoded in ***Formosa Phonetic Alphabet*** (ForPA), which will be described in the following paragraph.

2.1 Formosa Phonetic Alphabet (ForPA)

The most widely known phonetic symbol sets to transcribe Mandarin Chinese is 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 a long time. However, both two systems are insufficient for the other members of the Chinese languages like Taiwanese (Min-nan) or Hakka. Therefore, it's necessary to design a more suitable phoneme set to begin with multilingual speech data collection and labeling [Zu 2002] [Lyu 2002]. The whole phone set for three major languages used in Taiwan are listed in <Table.1>, where four kinds of phonetic systems have been used, including MPA, Pinyin, IPA, and the newly proposed ForPA. In <Table.1>, for each phoneme, there is also an example of syllable (and Chinese character) which contains the target phoneme as part of its pronunciation.

ForPA	Syllable (字)	IPA	Pinyin	MPA
b	ba(八)	p	b	ㄅ
p	pa(趴)	p'	p	ㄆ
m	ma(媽)	m	m	ㄇ
f	fa(發)	f	f	ㄈ
d	da(搭)	t	d	ㄉ
t	ta(他)	t'	t	ㄊ
n	na(那)	n	n	ㄋ
l	la(拉)	l	l	ㄌ
g	ga(嘎)	k	g	ㄍ
k	ka(咖)	k'	k	ㄎ
h	ha(哈)	x	h	ㄏ
zh	zha(渣)	tʂ	zh	ㄓ
ch	cha(差)	tʂ'	ch	ㄔ
sh	sha(殺)	ʂ	sh	ㄕ
rh	rhan(然)	ʐ	r	ㄖ
z	za(匝);zi(機)	ts;tɕ	z;j	ㄗ;ㄐ
c	ca(擦);ci(七)	ts';tɕ'	c;q	ㄘ;ㄑ
s	sa(撒);si(西)	s;ɕ	s;x	ㄙ;ㄒ
r	ru(如^{T});ri(字^{T})	z;ɹ̩		
bh	bha(肉^{T})	b		
gh	ghua(我^{T})	g		
v	voi(會^{H})	v		
ng	ang(鞍);nga(雅);ng(黃)	ŋ	-ng	ㄤ
a	a(阿)	a	a	ㄚ
o	o(喔)	o	o	ㄛ
er	er(鵝)	ɤ	e	ㄜ
e	ie(也)	ɛ	ê	ㄝ
err	err(而)	ɚ	er	ㄦ
i	i(一)	i	y;i	ㄧ
u	u(吳)	u	w;u	ㄨ
yu	yuan(原)	y	yu;ü	ㄩ
ii	zii(資)	ɨ	i	
-nn	ann(鮟TH)	ã		
-p	ap(壓TH)	-p		
-t	at(握TH)	-t		
-k	ak(沃TH)	-k		
-h	ah(鴨TH)	-h		

<Table 1>: The phone set for languages in Taiwan, decoded in four different phonetic alphabet, including ForPA, IPA, MPA, and Pinyin. An example of syllable and Chinese character （字） are also shown in the second column.

2.2 Formosa Lexicon (ForLex): A Pronunciation lexicon of Taiwanese, Hakka and Mandarin for Speech Recognition

A tri-lingual lexicon has been collected to produce word sheets for speakers to utter. This lexicon, called Formosa Lexicon (ForLex), was adapted from three other lexicons, including the CKIP Mandarin lexicon, the Gang's Taiwanese lexicon, and the Syu's Hakka lexicon. Some statistical information about the lexicon was listed in <Table.2>.

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

<Table 2>: The numbers of words of 3 Lexicons including CKIP Mandarin , Gang's Taiwanese, Syu's Hakka lexicons

3. THE PROCESS OF PHONETICALLY BALANCED WORD SHEETS

From the tri-lingual pronunciation lexicon transcribed in ForPA, we tried to extract the sets of distinct syllables and inter-syllabic bi-phones for the three languages. The statistics of phonetic units considered are listed in <Table.3>. In order to collect speech data with as much information about phonetic variations and keep the word lists as small as possible, we have to choose word sheets to satisfy some criterions.

First of all, the word set should cover the following phonetic units: Base-syllables and Inter-syllabic bi-phones. It can be shown that the word set which covers base-syllables and Inter-syllabic bi-phones can also cover all phones, Initial-Finals, within-syllabic bi-phones, right context Initials, context independent Finals, and Inter-syllabic right-context-dependent phones. All of the above speech units are widely used in the speech recognition research literatures in the past decades.

Language	Base syllable	Phone	Within-syllabic bi-phone	Inter-syllabic bi-phone
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

<Table 3>: The statistics of phonetic units for the 3 languages, Taiwanese (T), Hakka (H) and Mandarin (M); ∪ represents the union of two or three languages.

The selection of such a word set is actually a set-covering optimization problem [Shen 1999], which is NP-hard. Here we adopt a simple greedy heuristic approximate solution [Lim 1995], which will be described in the following paragraph

3.1 The algorithm for selecting phonetically balanced word set

Before we explain the algorithm, we define several notations as follows:

$W = \{w_i : 1 \le i \le N\}$ is the set of all words in the lexicon, where N is the number of words, w_i is the i^{th} word.

S^i and P^i are the sets of all distinct syllables and inter-syllabic bi-phones in the word w_i respectively.

$W(t)$, $S(t)$ and $P(t)$ are the sets of all words, all distinct syllables and all distinct inter-syllabic bi-phones except for those which has been selected from beginning to iteration t respectively.

$S^i(t)$ and $P^i(t)$ are the sets of all distinct syllables in the word w_i except for those which has been selected from beginning to iteration t, respectively.

In addition, *N*(set) represents the number of elements in the set. It can be shown that

$$W(0) = W,\ S(0) = \bigcup_{i=0}^{N-1} S^i,\ P(0) = \bigcup_{i=0}^{N-1} P^i$$

and $S^i(t) = S^i \cap S(t), P^i(t) = P^i \cap P(t)$

Based on the notations of the above, the algorithm could be described as following steps, whose flow chart is also shown as in <Fig. 1>:

Step 1): Initially t=0 and we have

$W(t) = W(0), S(t) = S(0),\ P(t) = P(0)$

Step 2):

Choose the word w_{i^*} for

$$i^* = \underset{0 \le i \le N(W(t))-1}{\operatorname{argmax}} N(S^i(t))$$

If $N(S^{i^*}(t)) = 0$, go to step 4.

Step 3):

$S(t+1) = S(t) - S^{i^*}$, $P(t+1) = P(t) - P^{i^*}$ and $W(t+1) = W(t) - W^{i^*}$. If $N(S(t+1)) \neq 0$ and $N(W(t+1)) \neq 0$, go to step 2 with *t=t+1*. Otherwise, if $N(S(t+1)) = 0$, go to step 4 with *t=t+1*. If $N(W(t+1)) = 0$, quit the algorithm.

Step 4):

Choose the word w_{i^*} for

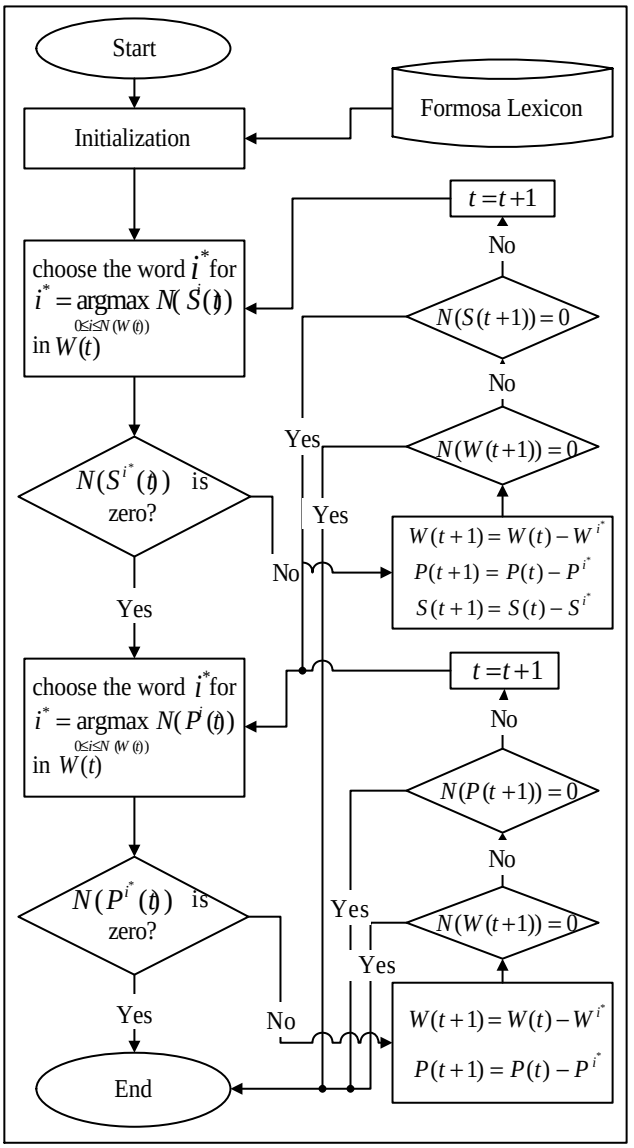
$$i^* = \underset{0 \le i \le N(W(t))-1}{\operatorname{argmax}} N(P^i(t))$$

If $N(P^{i^*}(t)) = 0$, quit the algorithm.

Step 5):

$P(t+1) = P(t) - P^{i^*}$, $W(t+1) = W(t) - W^{i^*}$.

If $N(P(t+1)) \neq 0$ and $N(W(t+1)) \neq 0$, return to step 4 with *t=t+1*. Otherwise, if $N(S(t+1)) = 0$ or $N(W(t+1)) = 0$, quit the algorithm.



<Fig 1>: The flow chart of the algorithm for selecting phonetically balanced word set

3.2 Data sheets

We define the coverage rate of the sheet to be the number of the speech units in the sheet divided by the number of all distinct speech units in the pronunciation lexicon. The word set selected by the algorithm for selecting phonetically balanced word set was further filtered using a criterion of coverage rate, as shown in <Fig.2>

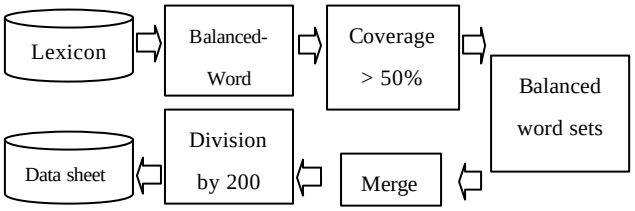


Fig 2: The process from lexicon to generate data sheets

In terms of Taiwanese sheets, although we produce 364 balanced-word sets totally, we only take sets whose coverage rate exceeds 50%. Because the number of syllables or words of some sets is highly different, we merge those sets in sequence and segment it equally to produce data sheets, which include 200 syllables roughly for each sheet. Therefore, the numbers of the data sheets and total words are 446 and 37275 respectively.

In terms of sheets of Hakka, all balanced-word sets are in sequence, and then segment it equally according to 70 words into the data sheets. Therefore, we get 340 data sheets which consist of 23837 words.

Otherwise, in Mandarin, one phonetically-rich set is segmented equally into ten sheets, and every sheet consists of 300 words roughly. In addition, all tonal-syllables are segmented equally into ten data sheets.

4. DATABASE INFORMATION

The database has been collected in both microphone and telephone channel, namely ForSDat-Mic600 and ForSDat-Tel1000 respectively. The statistical information is listed in <Table.4> and <Table.5>, where the number of utterances, the length of read speech of the current release of the database is shown.

	Length for MIC-channel	Length for TEL-channel
Read speech	24.3 hr	16 hr

Table 4: The statistics of utterances in microphone and telephone channel for Taiwanese (MIC: microphone, TEL: telephone)

	Speaker Number for MIC channel	Speaker Number for TEL channel
Male	328	626
Female	291	425
Total	619	1051

Table 5: The statistics of speakers in microphone and telephone channel for Taiwanese (MIC: microphone, TEL: telephone)

5. CONCLUSION

Until now, the version 1.0 of this corpus containing 600 speakers' speech of Taiwanese (Min-nan) and Mandarin Chinese has been finished and ready to release by adopting the balanced-word algorithm. We have collected speech of 1773 people, including 40.30 hours and 247,027 utterances. Because the project is on going, more speech data of Hakka and Mandarin will be collected.

References

[1] Wang, H.C. F. Seide, C.Y. Tseng, and L.S. Lee, 2000. Mat-2000 – design, collection, and validation of a mandarin 2000-speaker telephone speech database. ICSLP2000, Beijing.

[2] Zu, Y. 2002. A super phonetic system and multi-dialect Chinese speech corpus for speech recognition. ISCSLP 2002.

[3] Lyu, R. Y. 2002. A bi-lingual Mandarin/Taiwanese (Min-nan), Large Vocabulary, Continuous speech recognition system based on the Tong-yong phonetic alphabet (TYPA). ICSLP 2000.

[4] Shen, J. L., etc. 1999. 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, pages 79-97, Jan. 1999.

[5] Lim, Y. and Y. Lee, 1995. Implementation of the POW (phonetically optimized words) algorithm for speech database. ICASSP-95., pages: 89 -92 vol.1, 9-12 May.

[6] Steven, Y. 2002. The HTK book version 3.2. Cambridge University Engineering Department.