Biphone-rich versus Tripohne-rich: A Comparison of Speech Corpora in Automatic Speech Recognition

Yong-Chang Yio[1], Min-Siong Liang[3], Yuang-Chin Chiang[1], Ren-Yuan Lyu[2]

[1]. Inst. of Statistics, National Tsing Hua University, Hsin-chu, Taiwan
[2].Dept. of Computer Science and Information Engineering, Chang Gung University,Taiwan
[3]. Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan
E-mail: chiang@stat.nthu.edu.tw, rylyu@mail.cgu.edu.tw, Tel: 886-3-5715131 ext 2645

ABSTRACT

In this paper, we compare the performance of a speech recognition system trained with two speech corpora. We select two set of words such that they covered all the cross-syllable bi-phones and tri-phones, and are called phonetically bi-phone-rich and tri-phone-rich respectively. It is required about 10 times more words than that of cross-syllable bi-phones to cover all the cross-syllable tri-phones. To facilitate fair comparison, the bi-phone-rich corpus is thus consisted of ten sets of words that each covers all the cross-syllable bi-phones. With those words as data sheets, a male Taiwanese speaker recorded all the words as microphone speech. The resulting speech corpora, about 100 minutes for each set, are used to train for the acoustic models. Although both perform quite well in tasks with recognition networks of linear net and free syllable net, the tri-phone-rich corpus does not show much advantages over the bi-phone-rich corpus.

I. INTRODUCTION

In Taiwan, There exist three languages in Taiwan, including Mandarin, Taiwanese and Hakka. Taiwanese is one of three major languages (Mandarin, Taiwanese and Hakka) and is widely used as the native tongue of more than 75% population in Taiwan. Besides Mandarin and Taiwanese, there are 10% people can use Hakka and Hakka and most people in Taiwan are bi-lingual, even trilingual. Unfortunately, due to lack of elementary education for Taiwanese and Hakka, most people can not read, write or even speak fluently in Taiwanese although they speak and listen to it every day. 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. In recent years, Taiwan government started to pay much more attention to mother-tongue language, and made more effort and budget for it.

In order to design a speech recognition system, it is essential to collect abundant phonetically-rich speech corpus. Several attempts have been made to extract automatically phonetically-rich sentences from large text corpus in European, Mandarin and Indian [1-2]. Recently, more relative works for Taiwanese is increasing [3-5]. In this paper, we select two set of words such that they covered all the cross-syllable bi-phones and tri-phones, called phonetically bi-phone-rich and tri-phone-rich respectively, to construct two corpora. A speech recognition system trained with two speech corpora in Taiwanese can be built to compare the performance between these corpora. For this purpose, one of the preliminary tasks to construct speech corpus is to build up a pronunciation lexicon. We have set up pronunciation lexicons of more than 70 thousand words for Taiwanese. 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.

This paper is organized as follows. In Section 2, we discuss the phonetic alphabet and the pronunciation lexicons. In Section 3, we then describe how we extract acoustic units from the corpus. In Section 4, several experiments are done for comparison on two corpora. Finally, we summarize our major findings and outline for our future work.

II. THE PHONETIC ALPHABET AND THE PRONUNCIATION LEXICONS

A The Formosa Phonetic Alphabet (ForPA)

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 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」) for Taiwanese and the Taiwan Language Phonetic Alphabet (TLPA) 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(ㄚ)" in TLPA and "pa(ㄞ)" 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, newly proposed ForPA, for multilingual speech data collection and labeling [6].

B Gang's Taiwanese lexicon: one of Formosa Lexicons (ForLex)

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 [7]. Some statistical information about the lexicon was listed in <Table 2>. With the Gang's Taiwanese pronunciation lexicon transcribed in ForPA, we can extract the sets of distinct syllables and cross-syllable bi-phones and tri-phones for Taiwanese. The statistics of phonetic units considered are listed in <Table.3>

	Taiwanese
1-Syllable	8153
2- Syllable	46587
3- Syllable	13241
4- Syllable	2106
5- Syllable	175
Total	70262

<Table 2>: The distribution of Gang's Taiwanese lexicon.

	Distinct Number	Total Number
Syllable	830	150349
cross-word bi-phone	866	220611
cross-word tri-phone	11760	300698

<Table 3>: The distribution of the number of phonetic units for syllable, cross-word bi-phone and cross-word tri-phone in Taiwanese.

III. THE PROCESS OF PHONETICALLY RICH CORPORA

In order to collect speech data with as much information about phonetic variations and keep the words as small as possible, we have to choose sheets to satisfy some criterions. Two sets of words are selected such that they covered all the cross-syllable bi-phones and tri-phones, and are called phonetically bi-phone rich and tri-phone rich word set respectively. In cross-syllable bi-phone words set, total distinct cross-syllable bi-phones are selected before total distinct syllables are selected. In tri-phone words set, total distinct cross-syllable tri-phones are selected before total distinct syllables are selected and the similar procedure are used to collect bi-phone rich set. The selection of algorithm in detail will be described in the following statement.

A The algorithm for selecting phonetically rich 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. $S(w_i)$ 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 or tri-phone set till time *t*. $W^c(t) = W - W(t)$ is non-chosen word set, $S^c(t) = S(W) - S(t)$ is non-chosen distinct syllable set and $U^c(t) = U(W) - U(t)$ is non-chosen distinct bi-phones set or tri-phone set till time *t*.

Based on the notations of the above, the algorithm could be described as 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 maximize the union of $S^c(t-1)$ and $S(w_i)$, 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 $S^c(t) = \phi$ continue next step

Step 4):

Choose the word w_{i^*} as C_t^* such that maximize 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^*

$U^c(t) = U^c(t-1) - U(C_t^*)$, $W^c(t) = W^c(t-1) - C_t^*$

$t = t + 1$

Step 5):

If $B^c(t) \neq \phi$ and $W^c(t) \neq \phi$, repeat the step 4

else if $B^c(t) = \phi$ or $W^c(t) = \phi$, exit the algorithm.

B The analysis of bi-phone rich and tri-phone rich word sets

The statistics of the word set selected by the algorithm for selecting phonetically rich word sets is shown in <Table 4>. In <Table 4>, the number of syllables in tri-phone rich word set is about 7.9 times bigger than bi-phone rich word set. In order to compensate bi-phone rich word set, the other 9 bi-phone rich word sets are selected with the phonetically rich algorithm and appended to a new word set. Ideally, all reasonable syllable, bi-phone and tri-phone units are 830, 1427 and 57227 in Taiwanese, but the collected units dose not reach 100% coverage rate which coverage equal distinct units is divided by all reasonable distinct units. Finally, the statistics of the word sets for recording is shown in <Table 5>.

	Word	DS	AS	DBP	AB	DTP	AT
BR	801	830	1810	866	2611	2339	3374
TrR	6050	830	14590	866	20640	11760	27847

<Table 4> The statistics of the number of the word set selected by the algorithm for selecting phonetically rich word sets, where BR: bi-phone rich set, TrR: tri-phone rich set, DS: distinct syllable units, AS: all syllable units, DBP: distinct bi-phone units, AB: all bi-phone units, DTP: distinct tri-phone units, AT: all tri-phone units

	Word	DS	SC	AS	DBP	BC	AB
TrR_S1	6050	830	100%	14590	866	60.69%	20640
BR_S10	6363	830	100%	14642	866	60.69%	21005
	DTP	TC	AT				
TrR_S1	11760	20.53%	27847				
BR_S10	7570	13.22%	28017				

<Table 5> The statistics of the number of the phonetically rich word sets for recording, where BR: bi-phone rich set, TrR: tri-phone rich set, S1: set1, S10: set1~set10, DS: distinct syllable units, SC: syllable coverage rate, AS: all syllable units, DBP: distinct bi-phone units, BC: bi-phone coverage rate, AB: all bi-phone units, DTP: distinct tri-phone units, TC: tri-phone coverage rate, AT: all tri-phone units

C The bi-phone rich and tri-phone rich corpora

The corpora, SDG-B and SDG-T, recorded by a person are speaker dependent by BR_S10 and TrR_S1 sheets. The speaker is male and about 50 years old. The recording data is through Creative HS-300 microphone, Creative Audigy2 NX external sound blaster card and sampled in 16K, 16bits waveform. In order to compare results with different amount of corpora, we combine SDG-B and SDG-T corpora into a new corpus SDG-C. The testing data is 1000 utterances about 18 minutes [4]. The distribution of these corpora is shown as <Table 6>.

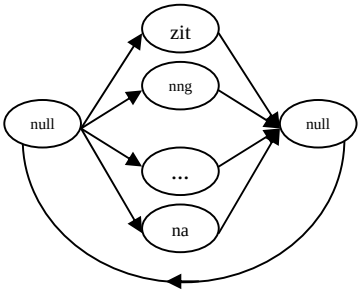
Sheet	NDS	NAS	Utterance	Time
SDG-B	830	14,642	6,363	106.9min
SDG-T	830	14,590	6,050	100.4min
SDG-C	830	23,392	10,019	167.4min
Testing data	463	2426	1,000	18.4min

<Table 6> The distribution of three corpora, where NDS: the number of distinct syllable units, NAS: the number of all syllable units.

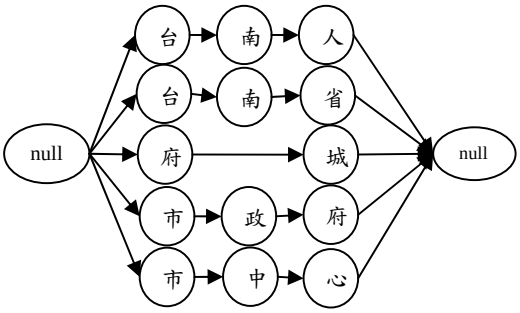
IV. EXPERIMENT

A HMM acoustic model and search net setup

We conducted our experiments with three corpora. 39 dimension MFCC features were computed for each 20 ms frame with 10 ms frame shift. Twelfth speaker independent models consisting of cross-syllable bi-phone, cross-syllable tri-phone, inside-syllable bi-phone and inside-syllable tri-phone models were built using a decision-tree state tying procedure for each three corpora. The recognition was carried out using 830-free-syllables net and 20K-word linear net shown in <Fig.1> and <Fig.2>. Each HMM model has three states except short pause model with one state. The number of Gaussian mixtures is dynamically tuned up depended on the amount of observation of each state in model.



<Fig1> The free-syllable search net with 830 nodes.



<Fig2> The linear search net with 20k-vocabulary size.

B Experiment results

1) Recognition rate in linear net

The phone, inside-syllable bi-phone model and inside-syllable tri-phone model trained with dynamic Gaussian mixtures and the maximum number of mixture is twelfth. The recognition rate is shown in <Table 7>. Maybe the training data is not enough that the recognition rate of inside-syllable tri-phone is lower than inside-syllable bi-phone and cross-syllable tri-phone is lower than cross-syllable bi-phone, but total recognition rate in <Table7> can reach above 90%.

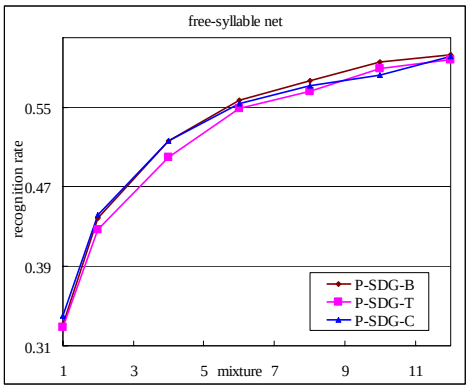
Model	database / Mixture	SDG-B	SDG-T	SDG-C
Phone	12	90.27%	89.04%	90.07%
IS Bi-phone	12	94.81%	95.34%	95.38%
IS Tri-phone	12	91.22%	90.67%	93.08%
CS Bi-phone	4	94.97%	95.09%	**95.75%**
CS Tri-phone	4	93.08%	93.12%	94.48%

<Table 7> The syllable recognition rate in three corpora for phone, inside-syllable bi-phone model and inside-syllable tri-phone model in 20K-vocabulary linear net.

2) Recognition rate in free-syllable net

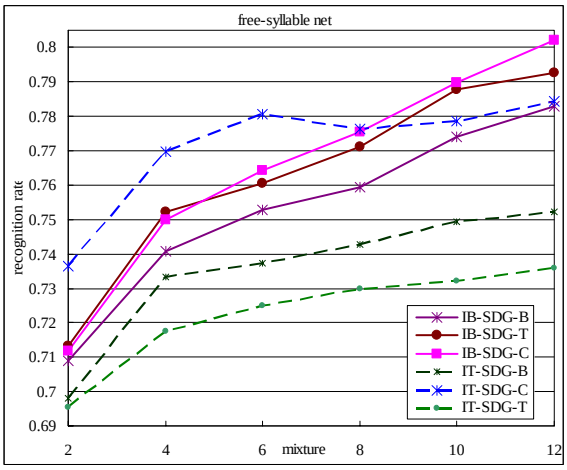
In <Fig 3>, the phone recognition rate is not significantly different in three corpora and the difference is less than 0.5% in the same Gaussian mixtures model. Even though the amount of

SDG-C is 2 times more than SDG-B, the model of SDG-C is not better than the model of SDG-B. It seems to mean that the amount of corpus could not influence recognition rate in a small model set, where the number of phone model is 50.



<Fig 3> Syllable recognition rate with phone model for SDG-B, SDG-T and SDG-C corpora

In comparison between inside-syllable bi-phone and inside-syllable tri-phone models, the number of inside-syllable tri-phone models is 2.8 times more than inside-syllable bi-phone models. But the recognition rate of inside-syllable tri-phone models is decaded 3.05%, 5.69% and 1.77% shown in <Fig 4> for SDG-B, SDG-T and SDG-C with 12 Gaussian mixture respectively. It seems to mean the recognition is lower in bigger number of models for lack of abundant speech data. On the other hand, the cross-syllable bi-phone and tri-phone model is better than inside-syllable bi-phone and tri-phone model in <Fig 5>.

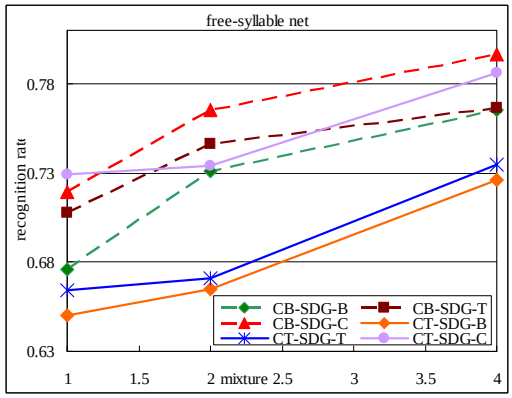


<Fig 4> Syllable recognition rate with inside-syllable bi-phone (IB) and inside-syllable tri-phone (IT) model for SDG-B, SDG-T and SDG-C corpora

V. CONCLUSION

In this paper, the corpora (SDG-B, SDG-C) are recorded by the cross-syllable bi-phone rich and cross-syllable tri-phone rich sheets. The corpus SDG-C is combined from SDG-B and SDG-C. With three corpora, we can research the difference in performance for these three corpora.

At first, we expect that the performance of tri-phone rich database is better than bi-phone rich database because the coverage rate of tri-phone rich database is bigger than bi-phone rich database in cross-syllable tri-phone units. In practice, the performance does not show significant difference between these two databases. It maybe has two reasons: first, the amount of corpus is not enough to train for cross-syllable tri-phone models. Second, the amount of corpora is adequate for cross-syllable bi-phone models. In the future, we will collect more data to prove whether the performance of tri-phone rich models is better than bi-phone rich models.



<Fig 5> Syllable recognition rate with cross-syllable bi-phone (CB) and cross-syllable tri-phone (CT) model for SDG-B, SDG-T and SDG-C corpora

VI. REFERENCES

[1] Hsin-min Wang, “Statistical Analysis of Mandarin Acoustic Units and Automatic Extraction of Phonetically Rich Sentences Based Upon a Very Large Chinese Text Corpus”, *ICLCLP vol.3 no.2, August 1998 pp. 93-144*

[2] K. Arora, S. Arora, K. Verma and S.S. Agrawal, “Automatic Extraction of Phonetically Rich Sentences from Large Text Corpus of Indian Languages”, *ICSLP 2004*, Jeju, Korean, 2004.

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

[4] Dau-Cheng Lyu, et al., "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Twaiwanse/Mandrain Continuous Speech Recognition," *In Proc. SST*, Melbourne, December 2002.

[5] Min-siong, Laing et al., “A Taiwanese Text-to-Speech System with Applications to Language Learning,” In Proc. *ICALT*, Joensuu, Finland, 2004.

[6] Ren-yaun 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] Liang, M. S., R. Y. Lyu and Y. C. Chiang, “An efficient algorithm to select phonetically balanced scripts for constructing corpus,” *IEEE NLP-KE*, Beijing, China, 2003.