# Construct a Multi-Lingual Speech Corpus in Taiwan with Extracting Phonetically Balanced Articles

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

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

## Abstract

In this paper, we describe an initial stage to construct **a multi-lingual speech corpus in Taiwan** with selecting phonetically balanced scripts. 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, constructing a multilingual phonetic alphabet, namely Formosa Phonetic Alphabet (ForPA), is the first step. In addition, the multilingual lexicons (Fomosa Lexicons) are also important parts for building the corpus. Recently, this corpus containing 2,300 speakers' speech database has been finished and is ready to be released. It contains about 200 hours of speech in 154,000 utterances.

## 1. Introduction

Multilingual speech recognition is one of the most popular research topics recently in speech signal processing. It is essential to collect a large-scale multilingual speech database for research, especially for designing a speaker independent and multilingual speech recognition system. People living in Taiwan (also called Formosa historically), 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 Sino-Tibetan language family. In the past several decades, most researchers of signal 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 [1] [2]. However, little has been done about the other two daily used languages. Therefore, we start to collect tri-lingual speech database by phonetically balanced word scripts.

Recently, it is decreasing severely number of speakers who could speak Taiwanese or Hakka. Therefore, it is hard to not only recruit the speakers but also invite experts to record prompt speech for speakers. In the other hand, this corpus is designed for speech recognition, which needs a variety of speech unit, so our database only need to covers enough speech units. In this paper, we describe the problem of phonetically balanced words (PBW) set [3]. We also modified the algorithm to select phonetically balanced scripts more efficiently to collect Formosa Speech Database (ForSDat), which is a large-scale multilingual speech corpus and cover those three languages used in daily life in Taiwan. Until now, we have collected about 2,100 speakers and hundreds of hours of speech.

## 2. The phonetic alphabet and the pronunciation lexicons

One of the preliminary tasks to construct speech corpus is to build up a pronunciation lexicon. We have set up a Formosa Lexicon (ForLex), which is a tri-lingual pronunciation lexicon containing Taiwanese, Hakka, Mandarin and Mandarin-Taiwanese bi-lingual lexicons. 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 paragraphs.

### 2.1. The Formosa Phonetic Alphabet (ForPA)

The Mandarin Phonetic Alphabet (MPA, also called Zhu-in-fu-hao) and Pinyin (Han-yu-pin-yin) is the most widely known phonetic symbol sets to transcribe Mandarin Chinese. They have been officially used in Taiwan and Mainland China respectively for a long time. However, both two systems are designed only for Mandarin. It's necessary to design a more suitable phoneme set to begin with multilingual speech data collection and labeling. A partial example of ForPA is listed in Table 1.

| ForPA | Syllable (字) | IPA | Pinyin | MPA |
|---|---|---|---|---|
| i | i(一) | i | y;i | ㄧ |
| u | u(吳) | u | w;u | ㄨ |
| yu | yuan(原) | y | yu;ü | ㄩ |
| ii | zii(資) | ɪ | i | |
| -nn | ann(鮟$^{TH}$) | ã | | |
| -p | ap(夾$^{TH}$) | -p | | |

*Table 1*: The partial example of 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. The Formosa Lexicon (ForLex)

A tri-lingual 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[4][5]. Some statistical information about

the lexicon was listed in Table 2. Besides, we employ the CKIP lexicon to construct a Mandarin-Taiwanese bi-lingual lexicon. There is at least a Taiwanese word corresponding to each Mandarin word in this bi-lingual lexicon. Each Taiwanese word has been transcribed into at least two pronunciations, containing literature (classic) and oral pronunciations with ForPA. With the bi-lingual lexicon, we can transcribe automatically abundant sentences or articles, which are written in Mandarin or Taiwanese from internet. The statistics of Mandarin-Taiwanese lexicon is shown in Table.3.

| | 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. (Syl: syllable)

| | LP-Taiwanese | OP-Taiwanese | Total |
|---|---|---|---|
| 1-Syl | 2319 | 8040 | 10359 |
| 2-Syl | 21337 | 49222 | 70559 |
| 3-Syl | 7163 | 11367 | 18530 |
| 4-Syl | 55 | 15525 | 15580 |
| 5-Syl | 1 | 711 | 712 |
| 6-Syl | 0 | 497 | 497 |
| 7-Syl | 0 | 478 | 478 |
| 8-Syl | 0 | 195 | 195 |
| 9-Syl | 0 | 3 | 3 |
| 10-Syl | 0 | 20 | 20 |
| Total | 30875 | 86060 | 116935 |

*Table 3*: The number of pronunciation of bi-lingual Lexicons, including literature pronunciation (LP) and oral pronunciation (OP) in Taiwanese (Syl: syllable)

## 3. The Process of phonetically balanced article sheets

In order to collect speech data with as much information about phonetic variations and keep the articles as small as possible, we have to choose article sheets to satisfy some criterions. It can be shown that the article 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. From the abundant articles transcribed in Mandarin-Taiwanese lexicon, we tried to extract the article sets, which include all possible distinct syllables and inter-syllabic bi-phones in those articles [6][7].

Based on the three pronunciation lexicons transcribed in ForPA, we extracted sets of distinct syllables and inter-syllabic bi-phones from the three languages. In order to collect speech data related to the co-articulation effect of continuous speech, we extracted phonetically abundant article 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, which is NP-hard. Here, we adopted a simple greedy heuristic approximate solution [8].

### 3.1. Problem definition of the phonetically balanced articles

Before we explain the algorithm, we define a set-covering optimization problem:

$A = \{A_i : 1 \le i \le M\}$ is the set of all articles, where M is the number of articles, $A_i$ is the *i-th* article.

$A_i = \{u_{i1}, u_{i2}, u_{i3} ..., u_{iN_i}\}$ is the set of phonetic units in $A_i$. and $N_i$ is the number of phonetic units in $A_i$, where $u_{im}$ may be the same as $u_{in}$ when m differ from n.

$U = A_1 \bigcup A_2 \bigcup ..... \bigcup A_M = \{u_1, u_2 ....., u_L\}$ is all distinct phonetic units in those articles, where $i \neq j \Rightarrow u_i \neq u_j$ and L is number of phonetic unit.

If we want to select a set of phonetically balanced articles, we should find the set $K^*, j_1^*, j_2^* ....., j_{K^*}^*$ such that

$$K^*, j_1^*, j_2^* ....., j_{K^*}^* = \underset{\substack{1 \le K \le M \\ 1 < j_1 < j_2 < ... < j_K < M \\ A_{j_1} \cup A_{j_2} \cup .... \cup A_{j_K} = S}}{\arg\min} (N_{j_1} + N_{j_1} + ..... + N_{j_K})$$

Where K is the number of selected articles, $A_{j_m}$ is one article and $j_m \neq j_n$ when m≠n

Therefore, if the number of articles is 1000, the search space is $2^{1000}$ candidates and it is an NP-complete problem.

### 3.2. The algorithm for selecting phonetically balanced articles set

However, a phonetically balanced words set algorithm has been proposed and the concept of the algorithm is that minimize the number of the chosen words to cover all distinct phonetic units [3]. Therefore, the same concept of algorithm can also be used to select article. Nevertheless, in order to minimize redundant words that exist in chosen articles, a slight improvement should be applied to [3]. The improvement method is that choose articles in which maximum ratio non-chosen phonetic units to the total distinct phonetic units. Besides, the multiple criterions are also used to select article in the new method, when the article cannot be chosen only by major criterion.

Before the method is described, some notations should be defined as followed:

$A = \{a_i : 1 \le i \le M\}$ is the set of all articles, where *N* is the number of articles, $a_i$ is the i[th] article.

$S^i$ and $P^i$ are the sets of total syllables and inter-syllabic bi-phones in the article $a_i$ respectively. In addition, $S_d^i$ and $P_d^i$ are the sets of all distinct syllables and inter-syllabic bi-phones in the article $a_i$

$A(t)$, $S(t)$ and $P(t)$ are the sets which has not been selected in all articles, all distinct syllables and all distinct inter-syllabic bi-phones at iteration t, respectively. $S_d^i(t)$ and $P_d^i(t)$ are the sets which has not been selected in all distinct syllables and inter-syllabic bi-phones at iteration t in the article $a_i$.

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

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

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

Based on the notations of the above, the new method should be satisfied as following equation: Choose the word $a_{i^*}$ for

$$i^* = \arg\max_{0 \le i \le N(A(t))-1} N(U_d^i(t)) / N(U^i)$$

Where $U_d^i(t)$ can be $S_d^i(t)$ or $P_d^i(t)$, and $U^i$ can be $S^i$ or $P^i$.

The ratio $N(S_d^i(t)) / N(S^i)$ is the first criterion. If two or more candidates should be selected by the first criterion, the algorithm will choose the maximum ratio $N(P_d^i(t)) / N(P^i)$ of the articles. Iterate the above equation with the residue articles, until all distinct syllables and inter-syllabic bi-phones has been covered in the chosen articles.

### 3.3. The process of producing phonetically balanced articles sheets

We set the requirements of the article 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-article data sheets. The process of producing phonetically balanced word sheets is shown in Fig 1 and some examples from the data sheets used to collect ForSDat are list in Table 4.

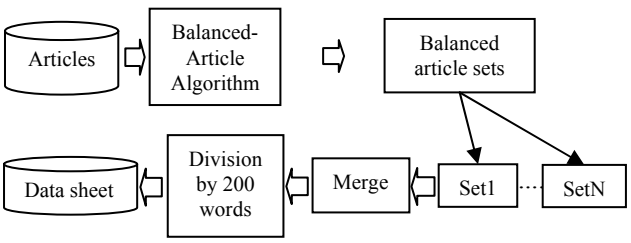


*Fig. 1*: The process of producing data sheets

| 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 |

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

## 4. Database Information

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. 2). 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 |

*Fig 2: A portion of a speaker's profile in the database.*

The database has been collected over both microphone and telephone channels, namely, ForSDat-TW01, ForSDat-MD01, ForSDat-TW02 and ForSDat-TW03, 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.3. All the statistics of the database are listed in Table 5.

| | 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 | 10.1 | 0.7 |
| | TW02-M1 | | Male | 277 | 11.66 | |
| | TW03-M0 | | Female | 409 | 74.61 | |
| | TW03-M1 | | Male | 264 | 45.56 | |
| | TW02-T0 | TEL | Female | 580 | 29.21 | 0.95 |
| | TW02-T1 | | Male | 412 | 19.37 | |

*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).

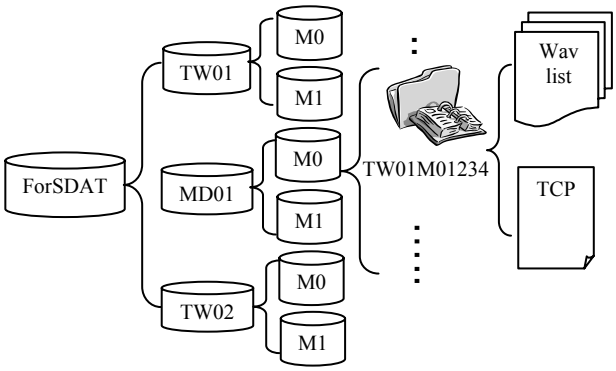


*Fig. 3:* The structure of database for Taiwanese and Mandarin. (TW01: Taiwanese database collected in 2001; M0: the microphone channel was used and the gender was female, T1: the telephone channel was used and the gender was male; and so on. There is a transcription file (TCP) for each unique speaker.)

## 5. 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, where the functions described in the following subsections.

### 5.1. 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 in turns in one work sheet, etc. These directories are also considered unusable.

### 5.2. 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 [9] to perform forced-alignment automatically on an utterance using all possible syllable combinations. We keep the highest scores for combinations to transcribe the speech.

### 5.3. Step 3: manual phonetic transcription

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 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.

## 6. Conclusions

Until now, the release of this corpus containing 600 speakers' speech of Taiwanese (Min-nan) and Mandarin Chinese has been finished and ready to release. We have collected speech of about 2,300 people, including about 200 hours and 154,000 utterances in Taiwanese and Mandarin. Because the project is on going, more speech data of Hakka and Mandarin will be collected.

## 7. References

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

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

[3] Liang, M. S., R. Y. Lyu and Y. C. Chiang, "*An efficient algorithm to select phonetically balanced scripts for constructing corpus*," Beijing, NLP-KE 2003.

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

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

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

[7] Shen, J. L., etc. "*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, 1999.

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

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