International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI Core, Linguistics Abstracts, and ACL Anthology.

易繫辭曰上古結繩而
治後世聖人易之以書
契百官以治萬民以察
說文敘曰蓋文字者經
藝之本宣教明化之始
前人所以垂後後人所
以識古故曰本立而道
生知天下之至賾而不
可亂也教化既萌文心
雕龍則謂人之立言因
字而生句積句而成章
積章而成篇篇之彪炳

Vol.17 No.1 March 2012 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 17, no. 1, March 2012

Contents

Papers

The Association for Computational Linguistics and Chinese Language Processing

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 1, March 2012, pp. 1-26

Development and Testing of Transcription Software for a Southern Min Spoken Corpus

Jia-Cing Ruan[*], Chiung-Wen Hsu[*], James Myers[*], and Jane S. Tsay[*]

Abstract

The usual challenges of transcribing spoken language are compounded for Southern Min (Taiwanese) because it lacks a generally accepted orthography. This study reports the development and testing of software tools for assisting such transcription. Three tools are compared, each representing a different type of interface with our corpus-based Southern Min lexicon (Tsay, 2007): our original Chinese character-based tool (Segmentor), the first version of a romanization-based lexicon entry tool called Adult-Corpus Romanization Input Program (ACRIP 1.0), and a revised version of ACRIP that accepts both character and romanization inputs and integrates them with sound files (ACRIP 2.0). In two experiments, naive native speakers of Southern Min were asked to transcribe passages from our corpus of adult spoken Southern Min (Tsay and Myers, in progress), using one or more of these tools. Experiment 1 showed no disadvantage for romanization-based compared with character-based transcription even for untrained transcribers. Experiment 2 showed significant advantages of the new mixed-system tool (ACRIP 2.0) over both Segmentor and ACRIP 1.0, in both speed and accuracy of transcription. Experiment 2 also showed that only minimal additional training brought dramatic improvements in both speed and accuracy. These results suggest that the transcription of non-Mandarin Sinitic languages benefits from flexible, integrated software tools.

Keywords: Speech Transcription, Southern Min, Taiwanese, Romanization, Key-in Systems.

[*] Graduate Institute of Linguistics, National Chung Cheng University, Minshiung, Chiayi 62102, Taiwan
Telephone: (05) 272-0411 ext. 21510; Fax: (05) 272-1654
E-mail: Lngmyers@ccu.edu.tw
The author for correspondence is James Myers.

1. Introduction

1.1 Constructing a Southern Min Speech Corpus

As with any language, corpora of spoken Southern Min (Taiwanese) have many uses, both scientific and practical. Corpora of written Southern Min exist (e.g., Iunn, 2003a,b, 2005, based on novels, prose, dramas, and poems; the Southern Min Archives of Academia Sinica, 2002; Ministry of Education, 2010, with word frequency statistics), but Southern Min, unlike Mandarin, is virtually never written at all. For this reason, there has been increasing interest in corpora of spoken Southern Min, including the NCCU corpus of spoken Chinese (Chui, 2009), which includes everyday conversation in Southern Min, and ForSDat (Formosa Speech Database) of Lyu, Liang, & Chiang (2004), which is a multilingual speech corpus for Southern Min, Hakka and Mandarin.

One area where a spoken corpus is essential is in the study of first language acquisition. This consideration motivated the construction of the Taiwanese Child Language Corpus (TAICORP; Tsay, 2007), which contains about two million morphemes in half a million utterances, based on about 330 hours of recordings of spontaneous conversations between children and their caretakers. Speech corpora are also essential for understanding the use of language in adult conversation, motivating our corpus of adult spoken Southern Min (Tsay & Myers, in progress), based on spontaneous conversations from radio broadcasts in Chiayi county. Except for the coastal towns, the majority of the population (including the hosts and guests in the radio programs recorded) in this area speak a variety of Southern Min historically derived from that spoken in Zhangzhou in Southern Fujian, although due to language contact over the years this variety has been mixed with the other variety historically derived from Quanzhou Southern Min. As of December 2011, the completely double-checked and confirmed portion of this corpus has almost 800,000 word tokens (詞), based on about 3,800 minutes of recordings.

Both TAICORP and the Taiwanese Spoken Corpus are transcribed in cognate Chinese characters (本字) wherever applicable, and otherwise in the romanization system of the Ministry of Education (MOE), Taiwan (Ministry of Education, 2008). The most important features of the MOE transcription notation for the present discussion are the marking of coda glottal stop with “h” (e.g., 肉 <bah4> ‘meat’), the marking of vowel nasality with “nn” (e.g., 甜 <tinn1> ‘sweet’), and the marking of tone categories with digits (e.g., 詩 <si1> ‘poem’ vs. 時 <si5> ‘time’).

These two corpora have been used to generate a lexical bank, which as of December 2011, has approximately 20,000 entries. Each entry contains four elements (see Table 1): (1) the word written in Chinese characters (or romanization if no corresponding characters exist), with homographs distinguished with numerals; (2) the pronunciations in romanization (including possible alternative pronunciations, typically due to borrowings from the Quanzhou variety of Southern Min); (3) near-synonyms or an explanatory definition in Mandarin; and (4) an example. Elements (3) and (4) are used to disambiguate homographic or homophonic entries.

Table 1. Sample entries in Southern Min lexicon.

Characters	Pronunciation	Explanation	Example
愛1	ai3	喜歡、愛	你 有 愛1 食 糖仔 oo02。
愛2	ai3	需要(加單賓)	這1 愛2 兩1 支 la0。

1.2 Challenges in Transcribing Southern Min

The usual challenges of transcribing a spoken language are compounded for Southern Min because it lacks a conventionalized orthography. With sufficient training in any adequate orthography, character-based or romanization-based, it should be possible for a native transcriber to write Southern Min as easily as Mandarin. Thus it is essential for Southern Min transcription to be assisted by some sort of automated orthography checker, to confirm that transcribers are consistent and to give hints when they get stuck.

The Southern Min lexicon we have been developing plays a key role in this orthography checking. Any entry can be accessed either via Chinese characters (if available) or via romanization, and once it is accessed, the explanation can confirm to the transcriber that the intended entry has been found. If an entry is not found, this either means that the transcriber has misspelled the word, or that the word has not previously appeared in the corpus.

For several years, transcribers for the Taiwanese Spoken Corpus have relied on a set of independent software tools developed for TAICORP (designed by James Myers and Jane Tsay, and written by Ming-Chung Chang and Charles Jie): a lexical access tool, a transcription tool, and a segmentation tool. For convenience we will call this package of tools Segmentor. As described in Tsay (2007), Segmentor requires the user first to transcribe speech into Chinese characters (wherever possible), and then run a program to segment the character strings into words defined by the lexicon, resulting in segmented text as shown in Appendix C, where each word is represented both in characters and in romanization within < > brackets. If any mistake is found at this point (i.e., if the program cannot find a word in the lexicon), the transcriber performs the above process again. Initial transcription is in Chinese characters, rather than Southern Min romanization, because we assumed that our student transcribers have many years of experience using Mandarin key-in systems and no experience with a systematic Southern Min key-in system.

However, transcribing Southern Min using Chinese characters has a number of shortcomings. First, transcribers must choose the correct Chinese characters (本字), which may be low-frequency characters in Mandarin, even for high-frequency Southern Min morphemes (e.g., 囥 <khng3>, glossed as "放", "to put/place/lay"). Second, most transcribers use phonetic key-in systems for Chinese characters, so they must mentally activate the Mandarin pronunciation, not the Southern Min pronunciation, to key in a character. Third, even if the characters are familiar from Mandarin, the Southern Min compound may not be, so they cannot rely on word auto-completion tools (e.g., 鐵齒 <thih4khi2>, glossed as "不聽勸/不信邪", "stubborn" is a compound in Southern Min but not in Mandarin). Fourth, there are many common words in Southern Min that have no Chinese character form at all (e.g., chit4tho5 "to play").

Segmentor also has limitations of its own. First, although the segmented text shows the romanization, this can only help transcribers uniquely identify words if they clearly recall which tone digit goes with which tone category, but we have found that native speakers have great trouble doing this. Second, because Segmentor only supports ANSI format text files, while the lexicon file is in UTF-8 format, it does not support Southern Min morphemes that must be written with Chinese characters outside of the traditional Mandarin set. Although this problem can be solved by incorporating Unicode BuWanJiHua (http://uao.cpatch.org/), the resulting transcription still cannot be properly handled by the segmentation tool, since its server settings support only Big5, not UTF-8. Finally, the source code of the segmentation program is no longer available for updating.

The purpose of this study, then, was to develop a new tool for transcribing Southern Min. Our intuition was that transcription might be more efficient if the student assistants could transcribe text word by word, rather than relying on a segmentation program, and directly in Southern Min romanization, rather than indirectly via Mandarin. Because new assistants have no prior experience writing a standardized Southern Min romanization system, a new software tool must provide considerable assistance. In particular, the tool cannot require users to enter tone digits, which are very hard to remember, and should use auto-completion so that users need only enter part of a compound word for it to be accessed from the lexicon.

In 2010, during the period of our study, the Ministry of Education released an input system for transforming Southern Min romanization into cognate Chinese characters (本字, or 漢字 in their terms); see Ministry of Education (2012) for the latest version of this system. The MOE is to be applauded for producing a very useful and flexible writing tool. However, it does not suffice for the transcribers of spoken corpora, who would benefit from being able to interact directly and simultaneously with sound files, the written corpus, and full lexical entries (including both character and romanized transcriptions, as well as other information for distinguishing among homonyms). In the remainder of this paper, we describe the development of just such a system (ACRIP), and demonstrate its effectiveness in experiments on naive participants learning to transcribe with it.

2. Adult-Corpus Romanization Input Program (ACRIP)

The key weakness of romanization input is that it requires student transcribers to be very familiar with the MOE Southern Min romanization system, and to be consciously aware of phonemic contrasts that do not exist in Mandarin, and hence are not associated with writing in their usual experience (despite their fluency with perceiving and producing Southern Min aurally and orally). The Adult-Corpus Romanization Input Program (ACRIP) helps transcribers in a number of ways when using the romanization system, by exploiting our large and growing corpus-based dictionary of Southern Min. The program was written by the first author in Microsoft Visual Basic 6.0, running in Microsoft Windows.

2.1 ACRIP Architecture

The architecture of ACRIP is presented in Figure 1.

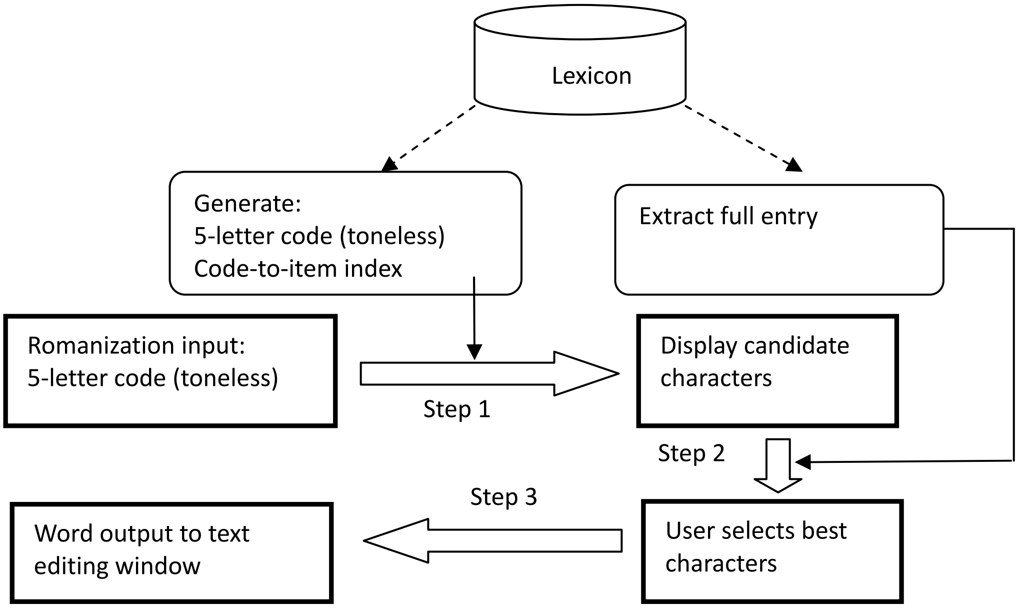


Figure 1. ACRIP architecture diagram.

The original corpus-based lexicon was edited to add a code of up to five letters for each entry, and a code-to-item index was established to link codes to candidate character-based entries, which were then linked to the other three elements of the entry (details are described in section 2.3). Each code is simply the first letters (up to five) of the romanization of a word, thus permitting a form of auto-completion: users only need to enter short strings of letters, without tone digits, to access full Southern Min words. More precisely, by entering a code, users get a list of candidate items, and then select the best item as the output according to the other elements in the entry (including explanation and example). When new entries are added to the lexicon, the coding can be updated automatically using an Excel macro.

2.2 The Main Interface for ACRIP 1.0

ACRIP integrates many functions for the transcription of Southern Min. The first version of this program, ACRIP 1.0, has the main interface shown in Figure 2 (ACRIP 2.0 retains the same functions, but adds others).

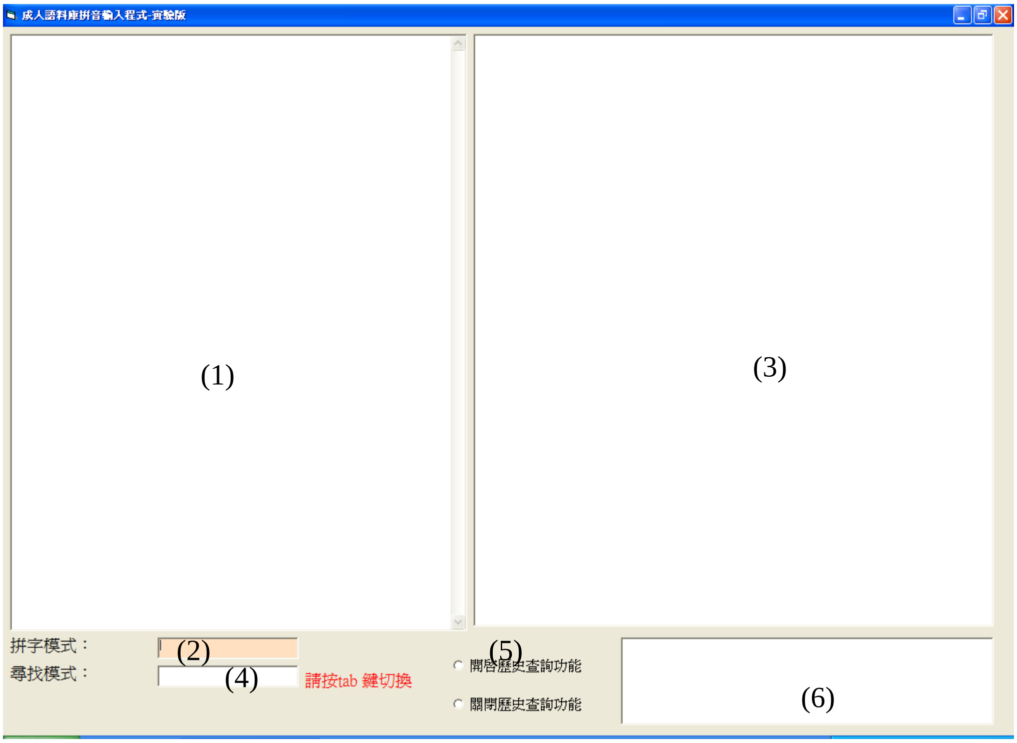
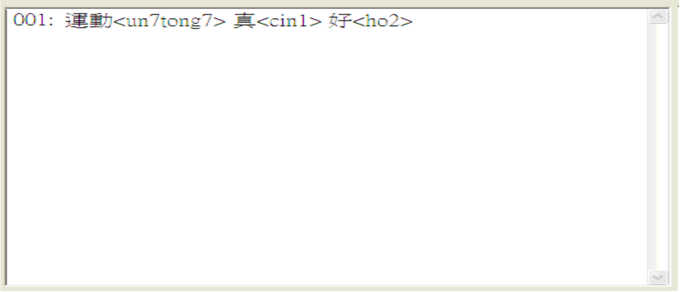


Figure 2. The main interface of ACRIP 1.0.

In contrast to the Segmentor tools, ACRIP integrates the three processes of accessing the lexicon, writing the transcription, and segmenting transcribed utterances into words, into a single interface. The corpus is transcribed by entering and checking one word (詞) at a time. The components of the ACRIP interface are as follows (identified by the numbers shown in Figure 2).

(1) Text editing window

This is the output window for segmented transcribed utterances (see Figure 3). The other components of ACRIP are designed to help the user fill this window with completed transcriptions. After transcriptions are complete, users can manually edit the contents of this window, or select the contents to copy or cut them to other editing programs.



001: 運動<un7tong7> 真<cin1> 好<ho2>

Figure 3. Window for text editing.

(2) Romanization search box

Transcribers enter up to five letters, without tone digits, to represent the word they hear in the spoken corpus. The words in the lexicon matching the first five letters will show up in the word candidate window. The example in Figure 4 shows the entry "unton", which is associated with the entry 運動<un7tong7>.

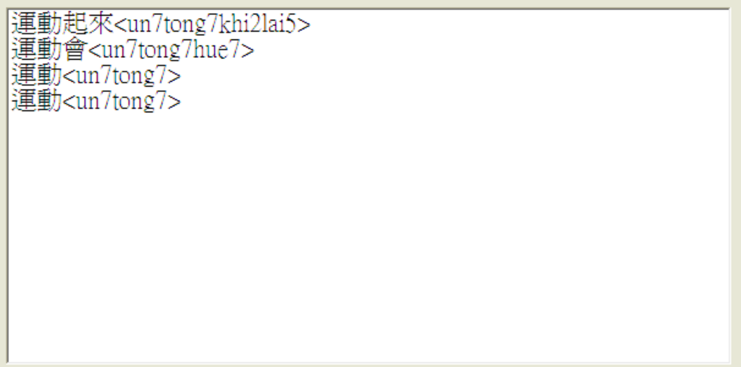
拼字模式：

unton

Figure 4. Text box for romanization input.

(3) Word candidate window

After entering a romanization code, all candidates in the lexicon with this code are shown in this window (see Figure 5). Users can then select the best candidate item to paste into the transcription being completed in the text editing window.

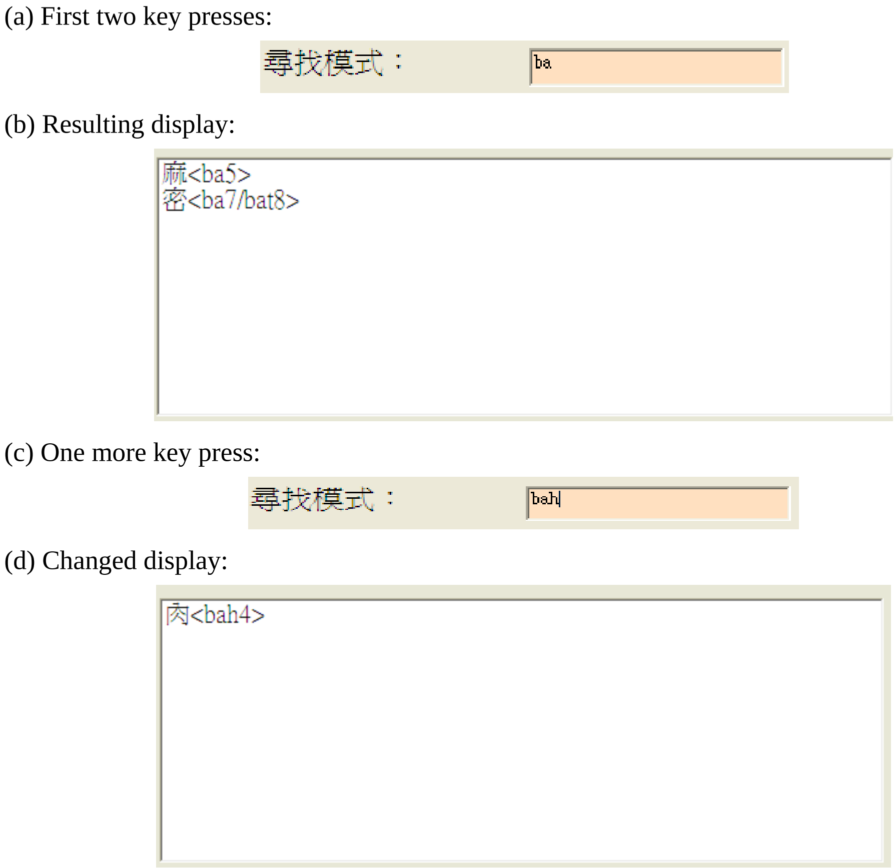


運動起來<un7tong7khi2lai5>
運動會<un7tong7hue7>
運動<un7tong7>
運動<un7tong7>

Figure 5. Window for candidate items

(4) Incremental romanization search box

This provides letter-by-letter search of romanization code for beginning users. This tool is helpful because pilot studies showed that the most difficult segments to perceive were the voiced onset obstruents (e.g., /b/ for 賣<be2> "sell", /g/ for 牛<gu5> "cow") and voiceless coda stops (e.g., /p/ for 汁<ciap4> “juice”, /t/ for 結<kat4> “knots”, /k/ for 角<kak4> “chunk”, glottal stop for 肉<bah4> “meat”). For example, transcribers often have trouble hearing glottal stop codas, as in the word 肉 (correctly transcribed in the MOE system as “bah4”). As shown in Figure 6, entering just the letters “ba” (a) only brings up the choices “ba5” (麻) and “ba7” (密) (b), immediately showing the transcriber that a coda is needed. Adding “h” (c) will then immediately change the list to the intended “bah4” (肉) (d).


(a) First two key presses:

尋找模式： ba

(b) Resulting display:

麻<ba5>
密<ba7/bat8>

(c) One more key press:

尋找模式： bah

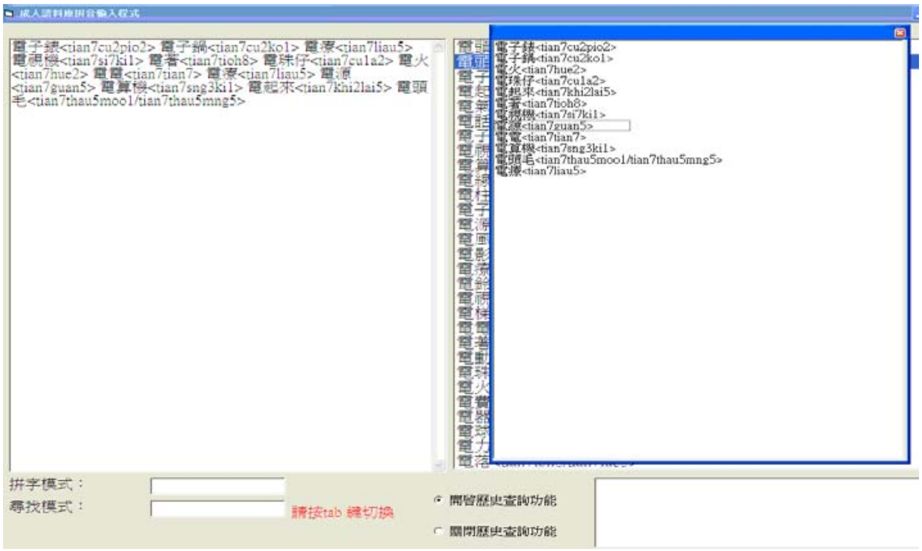
(d) Changed display:

肉<bah4>

Figure 6. Incremental romanization search.

(5) Toggle to save/erase work history

By turning on this function, users can avoid having to type the same code repeatedly for frequently occurring words. Instead, users can double-click strings in the work history to make them appear in the word candidate window. In the example shown in Figure 7, a user accessed the item 電腦<tian7nau2> by entering the code “tiann”. If the user needs to enter this item again, the user does not need to re-type the code, but can simply double click the string listed in the historical record. Users can also toggle this function off, erasing the work history.



電子錶<tian7cu2pio2> 電子鍋<tian7cu2ko1> 電療<tian7liau5> 電視機<tian7si7ki1> 電著<tian7tioh8> 電珠仔<tian7cu1a2> 電火<tian7hue2> 電電<tian7tian7> 電療<tian7liau5> 電源<tian7guan5> 電算機<tian7sng3ki1> 電起來<tian7khi2lai5> 電頭毛<tian7thau5moo1/tian7thau5mng5>

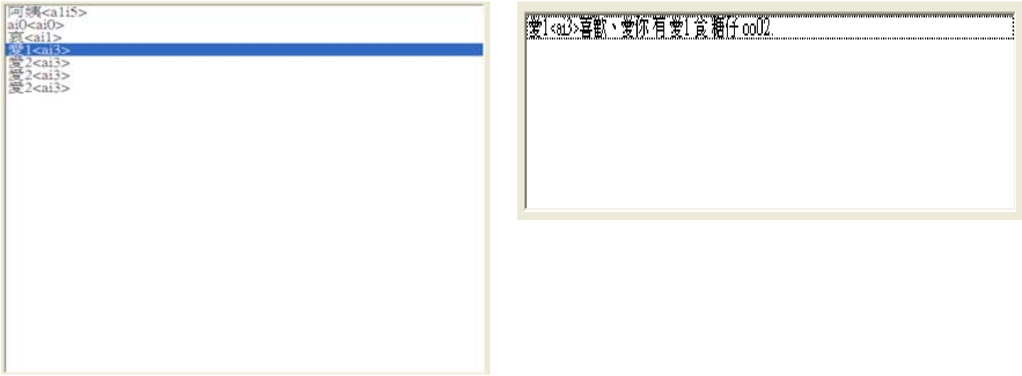
電子錶<tian7cu2pio2>
電子鍋<tian7cu2ko1>
電火<tian7hue2>
電珠仔<tian7cu1a2>
電起來<tian7khi2lai5>
電著<tian7tioh8>
電視機<tian7si7ki1>
電源<tian7guan5>
電電<tian7tian7>
電算機<tian7sng3ki1>
電頭毛<tian7thau5moo1/tian7thau5mng5>
電療<tian7liau5>

拼字模式：
尋找模式：
請按tab 鍵切換
● 開啟歷史查詢功能
○ 關閉歷史查詢功能

Figure 7. Using the history window.

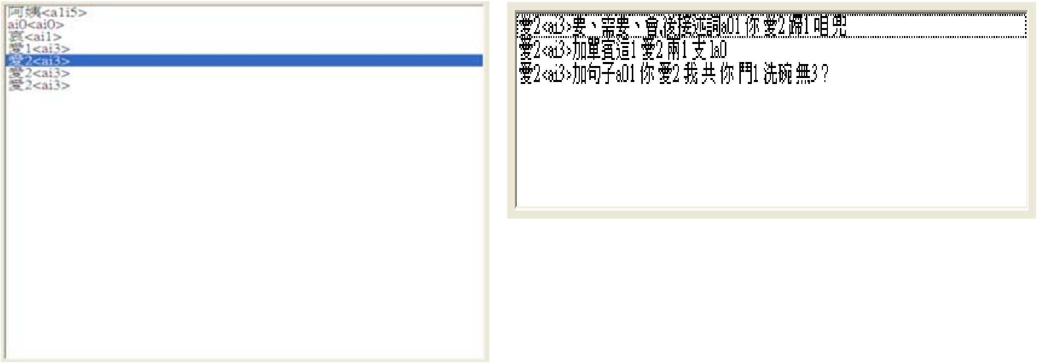
(6) Pop-up lexical entry display window

After the list of candidate words has appeared in the candidate word window, there may be homonyms, as for example 愛1 and 愛2 shown earlier in Table 1. Prior to the development of ACRIP, transcribers would need to memorize the difference or to shift to a separate lexicon program to look them up. ACRIP's built-in lexical entry display window appears as a pop-up when users choose any item in the word candidate window and press the space bar. This tool helps disambiguate the intended word and saves time by not requiring users to change to a separate program or to retype items for lexical look-up (see Figures 8 and 9).


阿姨<a1i5>
ai0<ai0>
哀<ai1>
愛1<ai3>
愛2<ai3>
愛2<ai3>
愛2<ai3>

愛1<ai3>喜歡、愛你有愛1食糖仔oo02.

Figure 8. Looking up 愛1



```
阿姨<a1i5>
ai0<ai0>
賓<ai1>
愛1<ai3>
愛2<ai3>
愛2<ai3>
愛2<ai3>
```

```
愛2<ai3>要、需要、會,後接述詞a01 你愛2歸1咱兜
愛2<ai3>加單賓這1 愛2 兩1 支 la0
愛2<ai3>加句子a01 你愛2我共你鬥1洗碗無3？
```

Figure 9. Looking up 愛2

2.3 Generation of the Romanization Input Codes and Code-to-item Index

In the development of ACRIP, the input romanization codes were generated from our original corpus-based lexicon by first deleting the tone digits and then extracting the first letters (up to five) as input code. This recoding was precompiled to speed up actual use of ACRIP (i.e., codes are stored in the lexicon rather than generated online).

One challenge faced when generating the input code was that the lexicon has many items that have alternative pronunciations, and therefore different romanizations, as shown in Table 2.

Table 2. Alternative pronunciations in a lexical entry.

Characters	Pronunciation	Explanation	Example
密密	ba7ba7/bat8bat8	滿滿	指緊密無縫

In this case, ‘baba’ and ‘batba’ are both codes for the entry ‘密密’. This problem was handled by editing the character and pronunciation elements of the lexical entries (using global replace in Microsoft Word and a macro in Microsoft Excel) to generate separate lexical entries for alternative pronunciations, so that each could be accessed separately.

After generating the romanization input code for each entry, we then incorporated them into the lexicon file using another macro in Microsoft Excel. The result was a file in which each lexical entry had a fifth element, representing the input code, as illustrated in Table 3.

Table 3. Revised lexical entries including romanization input code.

Input code	Characters	Pronunciation	Explanation	Example
baba	密密	ba7ba7	滿滿	指緊密無縫
batba	密密	bat8bat8	滿滿	指緊密無縫

3. Experiment 1: ACRIP 1.0 vs. Segmentor

In order to test whether ACRIP 1.0 improved the speed and accuracy of transcription of Southern Min using word-by-word romanization entry, we ran an experiment to compare it with the original Segmentor package for Chinese character transcription with post hoc segmentation. Naive native speakers of Southern Min transcribed short passages using both systems, and we examined the speed and accuracy of their transcriptions.

3.1 Methods

3.1.1 Participants

Twenty college students at National Chung Cheng University, who acquired Southern Min before kindergarten and without prior linguistic training, took part in the experiment. They were paid for their participation.

3.1.2 Design and Materials

The experiment had three phases: romanization training, romanization practice, and transcription testing. The romanization training phase used 30 nonlexical syllables that conformed to the phonotactic constraints of Southern Min (i.e., they were accidental gaps); see Appendix A. The romanization practice phase used 50 high-frequency Southern Min lexical items that together contain all of the segments and tone categories available in the phonological system of Southern Min (see Appendix B).

For transcription testing, two auditory passages were selected from the corpus of adult spoken Southern Min, Passages A and B; see Appendix C. Each passage was about 35 seconds long; based on piloting, we estimated that each would take less than an hour to transcribe. The two passages, which had already been transcribed and checked by our assistants, had roughly the same number of word tokens (Passage A: 129; Passage B: 122). The words were also matched in token frequency (based on our entire corpus), so we expected them to be approximately equal in transcription difficulty.

The transcription phase of the experiment used a Latin square design, balancing the presentation order of the two passages and the order of the two transcription systems across four groups of participants (five participants per group). Thus there was no confound among passage, order, or transcription method.

3.1.3 Procedure

In the romanization training phase, which lasted about an hour, the 30 nonlexical syllables were presented auditorily using Windows Media Player, and participant responses were made by pen and paper. Feedback on correctness was immediately given by the experimenter (second author). The purpose of this phase was to familiarize participants with the contrasting onsets, vowels, codas, and tones of Southern Min, with special focus on codas (e.g., distinguishing glottal stop from /k/).

In the romanization practice phase, which also lasted about an hour, the 50 Southern Min words were presented in random order, both auditorily and visually, using E-Prime 2.0 (Schneider, Eschman & Zuccolotto, 2002). Participants were asked to transcribe the lexical items by typing romanization. Before they made their response, participants were allowed to play the word up to ten times. When they typed their response, subjects received feedback on the correctness of their transcription.

In the transcription testing phase, participants transcribed the two corpus passages, in their assigned order (see 3.1.2). Segmentor was used to transcribe using Chinese characters, with post-hoc segmentation, while ACRIP was used to transcribe word-by-word using romanization. All participants were given no more than one hour to transcribe each passage. Thus the entire experiment took approximately four hours for each participant.

3.2 Results

Separate by-participant analyses were conducted on transcription speed and accuracy. In both analyses, the independent variables were Passage (A vs. B) and Transcription System (Segmentor/characters vs. ACRIP/romanization). Our focus was on the effect of transcription system, with Passage included in the analysis merely to test for possible confounds.

The mean number of transcribed words (transcription speed) and percentage of mistranscribed words (error rate) are shown in Table 4.

Table 4. Mean number of transcribed words and percentage of mistranscribed words for the two transcription systems.

System	Transcribed words	Mistranscription rate (%)
Segmentor	92.15	36.94
ACRIP 1.0	83.85	38.11

Both measures formed normal distributions, so a parametric test was used. We chose linear mixed-effects regression modeling because it is more flexible than analysis of variance (Baayen, 2008). Passage and Transcription System (both within-participant) were coded as effect variables (i.e., their values were coded as -1 vs. 1), and their interaction was included in the analyses. As is standard with this test, we computed *p* values from Markov chain Monte Carlo samples (using the pvals.fnc function of the languageR package; Baayen, 2008) in R (R Development Core Team, 2011).

As shown in Table 4, the use of ACRIP 1.0 was associated with slightly fewer transcribed words than Segmentor and a slightly higher error rate, but neither difference was statistically significant (*p*s > .1). The only significant effect was a main effect for Passage on the number of transcribed words (*B* = 12.4, *p* = .0001), but this was merely because Passage A had more words (129) than Passage B (122). There were no other main effects and no interactions for either measure.

3.3 Discussion

The results showed no significant effects of transcription method on the number of transcribed words or transcription accuracy. Putting these null results in a positive light, we found no evidence that romanization-based transcription of Southern Min is inherently less efficient or error-prone than character-based transcription. Of course, these null results may also relate to a floor effect for both transcription methods: two hours of training, and one hour of transcription per passage, may not be enough for a naive transcriber to develop adequate competence, regardless of which system is used.

Each software tool has its own problems. As we mentioned earlier, Segmentor requires users to translate the heard Southern Min into Mandarin so that they can enter Chinese characters, and they also get feedback only as the segmentation tool is run, not word by word. Moreover, even after typing a word in Chinese characters, they may have to choose among a list of candidate Southern Min words distinguished partly by Southern Min romanization. Using Segmentor also requires users to enter the etymologically correct characters (本字), which are often unfamiliar to naive users (assuming any character form exists at all), so that it is not uncommon for them to type a semantically or phonologically related character instead of the correct one.

Nevertheless, ACRIP 1.0 has its limitations too. Although romanization entry solves the above problems in principle, naive transcribers are far more familiar with Chinese characters than with Southern Min romanization. Opinions on whether learning this romanization system is worthwhile seemed to be divided across the participants. After the experiment, a survey was emailed to participants to ask for their opinion about the two transcription tools. Of the five participants who replied, three acknowledged the efficiency of the romanization system and agreed that if they had had more practice with it, they would have been able to do the transcription more quickly with it than with Chinese character entry. However, the other two thought that using Chinese characters as input was more intuitive to them and saved time compared with correcting mistakes in their romanized entries.

4. ACRIP 2.0

Based on the results of Experiment 1, some novice transcribers still seem to need an option for Chinese character word entry. Therefore, we modified the input program to combine ACRIP 1.0 with the advantages of Segmentor, calling the new version ACRIP 2.0 (also written in Microsoft Visual Basic 6.0 by the first author). The main interface of ACRIP 2.0 is shown in Figure 10.

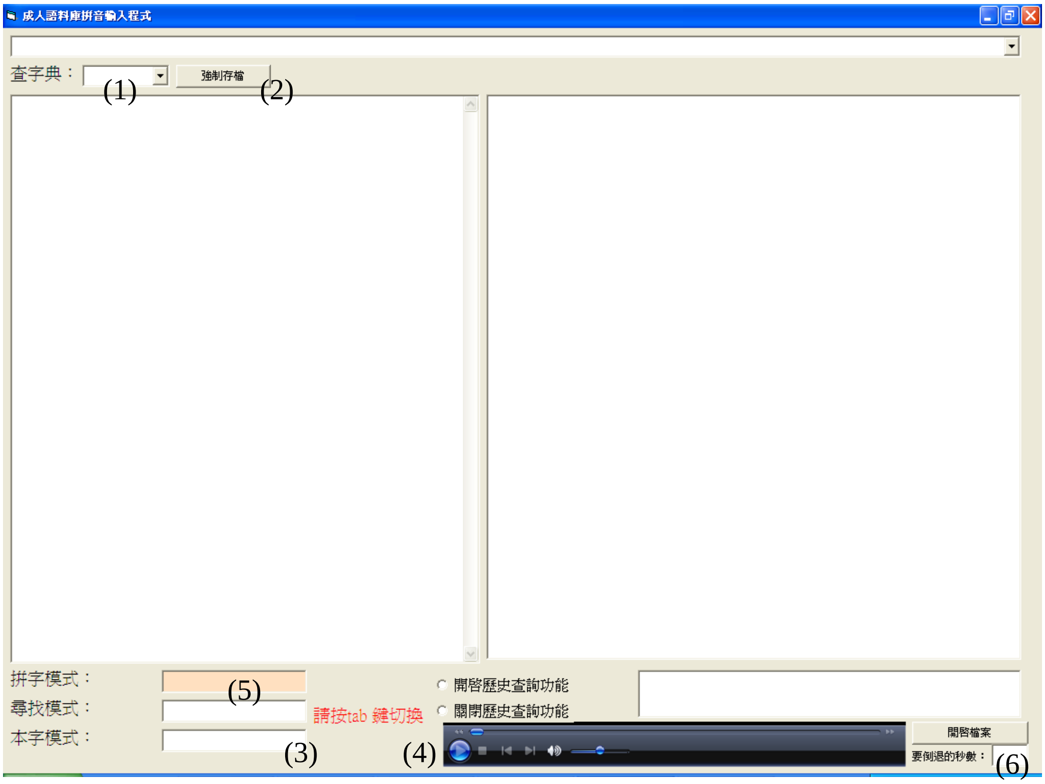
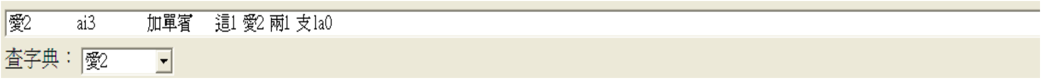


Figure 10. The main interface of ACRIP 2.0.

ACRIP 2.0 maintains all of the components of ACRIP 1.0, but adds the following new ones (see number labels in Figure 10).

(1) Integrated lexicon search box

Users can use this function to look up an item in the Southern Min lexicon by entering any of the four elements of an entry: Chinese characters, Southern Min romanization, Mandarin near-synonyms, or the explanatory example or definition (see Figure 11).



愛2 ai3 加單賓 這1 愛2 兩1 支la0

查字典：愛2

Figure 11. Looking up 愛2 in the integrated lexicon interface

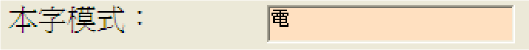
(2) Auto-save into the editing area

For safety, this new function allows users to save data in the text editing window at any time. In addition, an automatic function operates invisibly to save data in the text editing window whenever any changes are made in this window.

(3) Incremental Chinese character search box

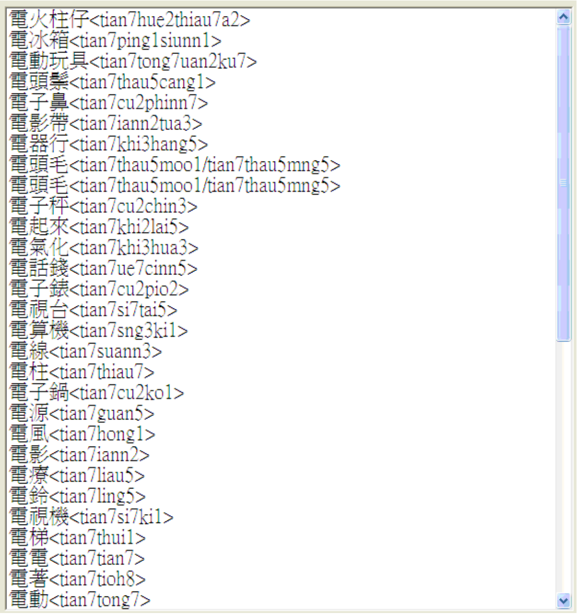
This provides a fuzzy search for lexical entries via the first character of the Chinese character element. For example, as shown in Figure 12, if a user enters “電” (a), the output list will be all items in the lexicon with Chinese character elements beginning with “電” (b).

(a) Character insertion:



本字模式： 電

(b) Resulting display in candidate item window:



```
電火柱仔<tian7hue2thiau7a2>
電冰箱<tian7ping1siunn1>
電動玩具<tian7tong7uan2ku7>
電頭鬃<tian7thau5cang1>
電子鼻<tian7cu2phinn7>
電影帶<tian7iann2tua3>
電器行<tian7khi3hang5>
電頭毛<tian7thau5moo1/tian7thau5mng5>
電頭毛<tian7thau5moo1/tian7thau5mng5>
電子秤<tian7cu2chin3>
電起來<tian7khi2lai5>
電氣化<tian7khi3hua3>
電話錢<tian7ue7cinn5>
電子錶<tian7cu2pio2>
電視台<tian7si7tai5>
電算機<tian7sng3ki1>
電線<tian7suann3>
電柱<tian7thiau7>
電子鍋<tian7cu2ko1>
電源<tian7guan5>
電風<tian7hong1>
電影<tian7iann2>
電療<tian7liau5>
電鈴<tian7ling5>
電視機<tian7si7ki1>
電梯<tian7thui1>
電電<tian7tian7>
電著<tian7tioh8>
電動<tian7tong7>
```

Figure 12. Incremental Chinese character search.

(4) Integrated Microsoft Windows Media Player

ACRIP 2.0 interfaces directly with Microsoft Windows Media Player so that users can play the portion of the audio file that they are currently transcribing.

(5) Play/stop the sound file

This function is attached to the romanization search box, and permits readily accessible keyboard control. When users press ESC, Microsoft Windows Media Player will play the sound file, and when they press ESC again, Microsoft Windows Player will stop playing.

(6) Automatic rewind timer

This function provides an automatic rewind operation which saves users the trouble of having to rewind sound files manually while replaying speech files during transcription. For example, if the timer is set to 3 seconds, when the sound file is off and users press ESC, Microsoft Windows Media Player will automatically rewind 3 seconds before replaying the speech file.

ACRIP 2.0 is intended to create a unified environment for the transcription of speech files. We observed that when using ACRIP 1.0, naive transcribers frequently needed to shift from this program to Microsoft Windows Media Player (in order to press the play/stop button and locate the time point they would like to replay in a speech file), and to the dictionary files (to look up items in Chinese characters when they did not know the Southern Min romanization). ACRIP 2.0 is designed to minimize the time needed to switch between these tasks: users first set up a default rewind time in the timer (6), and operate (4) and (5) via the ESC key (thus saving even more time by avoiding the need to use the mouse).

By permitting Chinese character search, including fuzzy search, and integrating Microsoft Windows Media Player for playing back speech files, users have more flexibility in entry options, have more powerful help tools, and can save time by not having to shift to other programs.

5. Experiment 2: ACRIP 2.0 vs. Segmentor and ACRIP 1.0

We hoped that the added features of ACRIP 2.0 would make it a much more efficient tool than either ACRIP 1.0 or Segmentor. To test this, we asked a new set of naive native speakers of Southern Min use ACRIP 2.0 to transcribe the same passages tested in Experiment 1. We also tested whether additional training brought any further improvements in speed and/or accuracy with using ACRIP 2.0.

5.1 Methods

5.1.1 Participants

Twenty college students at National Chung Cheng University, who acquired Southern Min before kindergarten and without prior linguistic training, took part in the experiment. None of the participants in Experiment 2 took part in Experiment 1. All participants were paid for first-session training and testing, and the half who received second-session training and testing were paid an additional fee.

5.1.2 Design and Materials

Experiment 2 had the same three phases as Experiment 1. The romanization training, romanization practice, and first-session transcription phases used the same materials as in Experiment 1. For the second-session transcription, two new passages, Passages C and D, were selected from the corpus of adult spoken Southern Min; see Appendix C. Both passages are about 39 seconds long, approximately the same length as Passages A and B, and had already been transcribed and checked. As with these earlier passages, we expected that the two new passages should take less than an hour to transcribe. The two passages have roughly an equal number of word tokens as the two passages in the Experiment 1 (Passage A: 129; Passage B: 122; Passage C: 123; Passage D: 121), and the words were matched in token frequency.

In the first session, half (10) of the participants transcribed Passage A before Passage B, while the other half transcribed the passages in the reverse order. To test the effect of additional training, half (10) of these participants were invited to join the second session, where half of these (5) transcribed Passage C before Passage D, while the other half transcribed the passages in the reverse order.

5.1.3 Procedure

The procedure for both sessions of Experiment 2 was identical to the procedure in Experiment 1, except that ACRIP 2.0 was the only transcription tool used. In both the first and second sessions, there was a romanization training phase, a romanization practice phase, and a transcription testing phase, each taking about an hour. Thus each experimental session lasted approximately three hours.

5.2 Results

We first compared the results for ACRIP 2.0 (the first phase of Experiment 2) with those for Segmentor and ACRIP 1.0 (Experiment 1), performing separate between-group by-participant analyses on transcription speed and accuracy. In all analyses, the independent variables were Passage (A vs. B) and Transcription System (ACRIP 2.0 vs. Segmentor, and ACRIP 2.0 vs. ACRIP 1.0). Our focus was on the effect of software tool: the mixed-system ACRIP 2.0 as compared with the Chinese character system Segmentor and with the romanization system ACRIP 1.0.

Table 5 shows the mean number of transcribed words (transcription speed) and percentage of mistranscribed words (error rate) for Experiment 1 (repeated from Table 4) and for the twenty participants in the first session of Experiment 2.

Table 5. Mean number of transcribed words and percentage of mistranscribed words for the three transcription systems.

System	Transcribed words	Mistranscription rate (%)
Segmentor	92.15	36.94
ACRIP 1.0	83.85	38.11
ACRIP 2.0	104.9	23.27

As can be seen in Table 5, ACRIP 2.0 yielded both a greater number of transcribed words and a lower mistranscription rate than either of the other two transcription tools. In two separate analyses, we compared ACRIP 2.0 with Segmentor and with ACRIP 1.0. Because the comparisons were being across different groups of participants, we used ordinary linear regression (equivalent to ANOVA, but chosen to facilitate comparison with the analyses used for Experiment 1). For each analysis, Passage and Transcription System were coded as effect variables, and their interaction was included in the analyses.

Both measures showed a statistically significant benefit of ACRIP 2.0 over Segmentor (number of transcribed words: $B = 6.375$, $p = .02$; mistranscription rate: $B = -6.83375$, $p = .002$). Similar positive results were found in the comparison of ACRIP 2.0 with ACRIP 1.0 (number of transcribed words: $B = 10.53$, $p = .0004$; mistranscription rate: $B = -7.42$, $p = .004$). significant main effect of Passage ($B = 14.175$, $p < .00001$). In addition, for the number of transcribed words, there were significant main effects of Passage (comparison with Segmentor: $B = 14.175$, $p < .00001$; comparison with ACRIP 1.0: $B = 12.23$, $p < .0001$), but again this was merely because Passage A had a few more words than Passage B. There were no other main effects and no interactions.

We then examined the effect of additional training with ACRIP 2.0 for the ten participants who received a second session of training and testing. The mean number of transcribed words (transcription speed) and percentage of mistranscribed words (error rate) for these ten participants are shown in Table 6.

Table 6. Mean number of transcribed words and percentage of mistranscribed words as a function of training on ACRIP 2.0.

Training	Transcribed words	Mistranscription rate (%)
First session	104.9	23.27
Second session	118.4	14.12

As shown in Table 6, additional training both increased the number of transcribed words and reduced the mistranscription rate. We analyzed both measures with Experience (-1 = first session, 1 = second session) as the only independent variable (Passage was confounded with session, since the first session used only Passages A and B and the second session used only

Passages C and D). Because Experience was a within-participant factor, we again used linear mixed-effects modeling with *p* values computed using Markov chain Monte Carlo samples. The results showed that the improvement in mistranscription rate was statistically significant (*B* = -5.38, *p* = .002) and the improvement in the number of transcribed words was marginally so (*B* = 6.46, *p* = .08).

5.3 Discussion

The results showed that transcription errors were significantly reduced when participants used the multi-functional, mixed-entry tool ACRIP 2.0, compared either with the character-based Segmentor or the romanization-based ACRIP 1.0. The number of transcribed words completed within the hour-long session also increased with the new tool.

Moreover, with additional training, transcriptions improved still further, with slightly more completed words and an even lower mistranscription rate. Projecting linearly, the drop in mistranscription rate from 23% to 14% from the first three-hour session to the second predicts that near-perfect accuracy could be attained with merely one further three-hour session. More realistically, of course, errors can never be expected to be eliminated entirely, so as is standard practice in the transcription of spoken corpora, the work of one transcriber must always be checked by another.

6. Conclusions

In this paper, we compared three software tools for assisting the transcription of the Taiwanese Spoken Corpus by interfacing with our Southern Min lexical bank. Segmentor requires users to transcribe passages as a string of Chinese characters, with segmentation performed later. The first version of Adult-Corpus Romanization Input Program (ACRIP 1.0) requires users to transcribe word by word, using romanization. The revised version, ACRIP 2.0, requires users to transcribe word by word, but permits them to input words either with Chinese characters or with romanization. In both versions of ACRIP, romanization input can be made without tone digits, and can use a form of auto-completion so that even longer words can be accessed with up to five letters. ACRIP 2.0 adds more flexibility to the input methods and also interfaces directly with Microsoft Windows Media Player so that audio files can be played and replayed from the same interface as word entry.

Our experiments found no significant disadvantage in using romanization entry compared with Chinese character entry, despite the native transcribers being much more familiar with the latter orthographic system. More importantly, ACRIP 2.0 was shown to permit significantly faster and more accurate transcriptions than either Segmentor or ACRIP 1.0. Efficiency and accuracy increased even more with only three additional hours of training. Since conducting this study, our trained graduate assistants use only ACRIP 2.0 as they continue to transcribe sound files for the Taiwanese Spoken Corpus.

Of all of the innovations of ACRIP, the most surprising for compilers of Chinese speech corpora may be its use of word-based and romanization-based input. Chinese text is traditionally entered into a computer character by character, supplemented by auto-completion for multi-character words where relevant. Yet as our results suggest, this may not be the most efficient method for transcribing fluent speech in Southern Min, a language with a distinct lexicon and phonology from Mandarin.

Nevertheless, given the great increase in performance of ACRIP 2.0 over ACRIP 1.0, it seems that a major strength of the tool lies more in its transcription-specific interface rather than in the type of transcription notation. That is, accuracy and speed were improved in large part because ACRIP 2.0 makes it possible for transcribers to have direct and simultaneous access to sound files, written corpus fragments, and full lexical entries. It is conceivable that additional benefits may result by integrating ACRIP 2.0 more fully into the MOE Southern Min writing tool (Ministry of Education, 2012), but this has yet to be tested.

Given this success, it seems reasonable to ask whether an ACRIP-like corpus transcription tool would applicable to other languages like Hakka or Formosa languages. For the most part, the new functions in ACRIP 2.0 are designed for facilitating the mechanics of transcription regardless of language. The only feature that may be less universally applicable is the ‘Incremental Chinese character search box’ function, which is not relevant for languages without cognate characters.

We hope that our findings will encourage compilers of other non-Mandarin Sinitic spoken corpora to explore the greater efficiency of input systems beyond the traditional Chinese character-based systems.

References

Academia Sinica. (2002). Southern Min archives: A database of historical change and language distribution. *National Digital Archives Program*. (Retrieved 2010/10/25) http://southernmin.sinica.edu.tw/

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, UK: Cambridge University Press.

Chui, K, & Lai, H. L. (2009). The NCCU corpus of spoken Chinese: Mandarin, Hakka, and Southern Min. *Taiwan Journal of Linguistics*, 6(2),119-144.

Iunn, U. G. (2003a). *Online Taiwanese syllable dictionary.* (Retrieved 2010/10/25) http://iug.csie.dahan.edu.tw/TG/jitian/.

Iunn, U. G. (2003b). *Online Taiwanese concordancer system.* (Retrieved 2010/10/25) http://iug.csie.dahan.edu.tw/TG/concordance/.

Iunn, U. G. (2005). *Taiwanese corpus collection and corpus based syllable / word frequency counts for written Taiwanese*. (Retrieved 2010/10/25) http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/guliau-supin.asp.

Lyu, R. Y., Liang, M. S., & Chaing, Y. C. (2004). Toward constructing a multilingual speech corpus for Taiwanese (Min-nan), Hakka, and Mandarin. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2), 1-12.

Ministry of Education. (2008). 臺灣閩南語羅馬字拼音方案使用手冊. (Retrieved 2011/04/11) http://www.edu.tw/files/bulletin/M0001/tshiutsheh.pdf

Ministry of Education. (2010). 教育部臺灣閩南語字詞頻統計. (Retrieved 2010/10/25) http://203.64.42.97/bang-cham/thau-iah.php

Ministry of Education. (2012). *Taiwan Southern Min Hanzi Input*, version 2.1 (Retrieved 2012/2/9) http://www.edu.tw/mandr/download.aspx?download_sn=3015&pages=0&site_content_sn=3364

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime reference guide.* Pittsburgh: Psychology Software Tools Inc.

Tsay, J. (2007). Construction and automatization of a Minnan child speech corpus with some research findings. *International Journal of Computational Linguistics and Chinese Language Processing*, 12(4), 411-442.

Tsay, J., & Myers, J. (in progress) *Taiwanese Spoken Corpus*. National Chung Cheng University, Chia-Yi, Taiwan.

Appendix A: Fake syllables for romanization training.

bai3	counn7	giem5	hounn3	jeunn7
khoop8	luek8	neunn3	nuiunn5	phoong5
pou7	teinn5	thoinn2	suat8	tot8
bam2	liak4	cei3	ngut8	jen3
pim3	ken5	hoi3	ngang3	coong5
nuoong2	bom2	gooi2	kiai1	sion1

Appendix B: Real syllables for romanization practice.

剝	pak4	合	hap8	芳	phang1	慢	ban7
莫	mai3	針	ciam1	姆	m2	焦	ta1
踢	that4	讀	thak8	雷	lui5	軟	nng2
零	lan5	鹹	kiam5	國	kok4	囥	khng3
牙	ge5	夾	ngeh4	摸	bong1	黃	ng5
蝦	he5	割	kuah4	走	cau2	食	ciah8
手	chiu2	深	chim1	衫	sann1	仙	sian1
爪	jiau3	南	lam5	路	loo7	無	bo5
煎	cuann1	病	penn7	傷	siong1	歹	phainn2
原	guan5	廟	bio7	唱	chiunn3	橫	huainn5
市	chi7	鮮	chinn1	膽	tann2		
枕頭	cim2thau5	田嬰	chan5enn1	硬拗	nge7au2	泡茶	phau3te5
踅街	seh8ke1	庄跤	cng1kha1	避雨	phiah4hoo3		

Appendix C: Passages from the Southern Min Spoken Corpus. The passages here have been modified by hand to remove alternative pronunciations listed in the lexical bank but not used by the speakers in these passages.

Passage A (Duration: 36sec)

Participants: 001 (hostess 1), 002 (hostess 2)

Filename: RC002

002: 阿媽<a1ma2> e0<e0> 話<ue7>。

001: 分享著<hun1hiang2tioh8> 老祖先<lau7coo2sian1> 所<soo2> 流傳<liu5thuan5> e0<e0> 智慧<ti3hui7> e0<e0> 話<ue7>，m0<m0>，這 1<ce1> 咱<lan2> e0<e0> [m 開場白]。來<lai5>，啥人<siann2lang5> 先<sing1> 講<kong2>？

002: 啥人<siann2lang5> 先<sing1> 講<kong2> ne0<ne0>？

001: m0<m0>，我<gua2> 先<sing1> 來<lai5> 講<kong2> 好<ho2> a02<a0>。

002: 好<ho2> 好<ho2> 好<ho2>，你<li2> 先<sing1> 講<kong2>。

001: henn0<henn0> 我<gua2> 欲<beh4> 講<kong2> 這 2<cit4> 句<ku3> hoonn0<hoonn0>，伊<i1> 講<kong2>，食<ciah8> 人 1<lang5> 一 1<cit8> 斤<kin1>，嘛<ma7> 就<to7> 還<hing5> 人 1<lang5> 四<si3> 兩 2<niu2>。

002: oo0<oo0>，食<ciah8> 人 1<lang5> 一 1<cit8> 斤<kin1>，嘛<ma7> 就<to7> 還<hing5> 人 1<lang5> 四<si3> 兩 2<niu2>。

001: henn0<henn0> a02<a0> hoonn0<hoonn0>。

002: hm0<hm0> hm0<hm0>。

001: 這 2le0<cit4e0> 這 2<cit4> 句<ku3> 話<ue7> 所<soo2> 講<kong2> e0<e0>，就是<to7si7> 講<kong2> hoonn0<hoonn0>，咱<lan2> 做人<co3lang5>，這 1<ce1> 人 1<lang5> 佮<kah4> 人 1<lang5> 咧<teh4> 交際<kau1ce3> hoonn0<hoonn0>，咧<teh4> 交往<kau1ong2> e0<e0> 這 2<cit4> 个<e5> 過程<kue3ting5> le02<le0>，總是<cong2si7> hoonn0<hoonn0>，愛 2<ai3> [m 禮尚往來] la0<la0>。就是<to7si7> 講<kong2> hoonn0<hoonn0>，愛 2<ai3> 有來有去<u7lai5u7khi3> la0<la0>。譬論<phi3lun7> 講<kong2>，

002: 未當<be7tang3> 單仔 1<kan1na7> 食<ciah8> 人 1<lang5> e0<e0>，嘛<ma7> 愛 2<ai3> 分 2<pun1> 人 1<lang5> 食<ciah8> la0<la0> hoonn0<hoonn0>。

001: 著 1<tioh8> 著 1<tioh8> 著 1<tioh8> 著 1<tioh8> 著 1<tioh8> hoonn0<hoonn0>。

Passage B (Duration: 35sec)

Participants: 001 (hostess 1), 002 (hostess 2),

Filename: RC002

001: 這 1<ce1> 受<siu7> 人 1<lang5> e0<e0> 恩惠<un1hui7> ne0<ne0>，就<to7> 愛 2<ai3> 知影<cai1iann2> 回報<hue5po3> hoonn0<hoonn0>。

002: henn0<henn0> a02<a0>。

001: e01<e0> 當然<tong1jian5>，

002: 未當<be7tang3> 講<kong2> hoonn0<hoonn0>，受<siu7> 人 1<lang5> e0<e0> 恩惠<un1hui7>，猶閣<a2koh4> 開始<khai1si2> 佇<ti7> 後壁<au7piah4> 共人<kang9> 創空<chong3khang1> 按呢<an3ne1>。

001: 敢 1<kam2> 會 1<e7> 創空<chong3khang1>？未<be7> la0<la0>，可能<kho2ling5> 是<si7> hoonn0<hoonn0>， [若是<na7si7> 上界<siong7kai3> i] 比較<pi2kau3> 較<khah4> 人 1<lang5> 無 1<bo5> 法度<huat4too7> 接受<ciap4siu7> 就是<to7si7> 講<kong2> hoonn0<hoonn0>，a01<a0>，onn0<onn0>，算講<sng3kong2>，受<siu7> 人 1<lang5> e0<e0> 恩惠<un1hui7>，a01<a0> 伊<i1> 閣<koh4> 毋<m7> 知影<cai1iann2> 講<kong2> hoonn0<hoonn0>，欲<beh4> 來<lai5> [m 知恩圖報] la0<la0> hoonn0<hoonn0>。

002: m0hm0<m0hm0>。

001: henn0<henn0>。e01<e0> 當然<tong1jian5> 今仔 2<cim2a2> 現代<hian7tai7> hoonn0<hoonn0>，就是<to7si7> 講<kong2>，社會<sia7hue7> 上<siong7>，真<cin1> 濟<ce7> 人 1<lang5> 就是<to7si7> 講<kong2>，咧<teh4> 幫助<pang1coo7> 別人<pat8lang5> hoonn0<hoonn0>，in1<in1> 感覺<kam2kak4> 講<kong2>，a0<a0>，咱<lan2> 就是<to7si7> [m 日行一善] ，hoonn01<hoonn0> 咱<lan2> 本底<pun2te2> ne0<ne0>，就是<to7si7> 欲<beh4> 來<lai5> 幫助<pang1coo7> 別人<pat8lang5> e0<e0> hoonn0<hoonn0>。所以<soo2i2> 講<kong2>，伊<i1> 是<si7> xxx<xxx> 真<cin1> 好意<ho2i3>，真<cin1> 善心<sian7sim1>，a01<a0> 伊<i1> 嘛<ma7> 無 1<bo5> 求<kiu5> 對方<tui3hong1> 來<lai5> 回報<hue5po3>。

Passage C (Duration: 40sec)

Participants: 001 (hostess 1)

Filename: RK006

001: 做陣<co3tin7> 收聽<siu1thiann1> 幸福<hing7hok4> 萬事通<ban7su7thong1>。

001: 我<gua2> 是<si7> [m 幸福] [m 妹妹] e0<e0> 淑芬<siok4hun1>。

001: 來<lai5> 今仔日<kin1a2jit8> 幸福<hing7hok4> 銀行<gin5hang5> 咱<lan2> 來<lai5> 儉 1<khiam7>，o0<o0> 兩 1<nng7> 个<e5> 朋友<ping5iu2> e0<e0> 故事<koo3su7>。

001: 咱<lan2> 講<kong2> a02<a0>，人生<jin5sing1> 旅途<lu2too5> oo02<oo0>，有<u7> 朋友<ping5iu2> hoonn0<hoonn0>，m0<m0> 咱<lan2> 會1<e7> 感覺<kam2kak4> 誠<ciann5> 幸福<hing7hok4>。

001: 因爲<in1ui7> 朋友<ping5iu2> e0<e0> 好處<ho2chu3> 就是<to7si7> 講<kong2> 會當<e7tang3> 佮<kah4> 你<li2> 分擔<hun1tam1> 你<li2> o0<o0> 心內<sim1lai7>，你<li2> 歡喜<huann1hi2> e0<e0> 事志<tai7ci3>，o0<o0> 你<li2> 感覺<kam2kak4> m0<m0> 悲傷<pi1siong1> e0<e0> 事志<tai7ci3> 攏<long2> 會當<e7tang3> 佮<kah4> 對方<tui3hong1> 講<kong2> la0<la0> hoonn0<hoonn0>。

001: a01<a0> 咱<lan2> 今仔2<cim2a2> 講著<kong2tioh8> 這2<cit4> 兩1<nng7> 个<e5> 朋友<ping5iu2> a02<a0>，in1<in1> 就是<to7si7> 相招<sio1cio1> 去<khi3> chit4tho5<chit4tho5>，hoonn01<hoonn0>。

001: a01<a0> in1<in1> 去<khi3> chit4tho5<chit4tho5> 這2<cit4> 个<e5> 所在<soo2cai7> hoonn0<hoonn0>，ai0ioo0<ai0ioo0> 去<khi3> [m 沙漠] [m 旅行] ne0<ne0>，hoonn01<hoonn0>。

001: 但是<tan7si7> 咱<lan2> 講<kong2> a02<a0>，閣<koh4> 較1<khah4> 好<ho2> e0<e0> 人1<lang5> hoonn0<hoonn0> 嘛<ma7> 有<u7> 可能<kho2ling5> 會1<e7> 冤家<uan1ke1> hoonn0<hoonn0>，e01<e0> 翁仔某<ang1a2boo2> 較1<khah4> 好<ho2> 嘛<ma7> 會1<e7> 相觸<sio1tak4> le02<le0> hoonn0<hoonn0>。

<u>Passage D (Duration: 39sec)</u>

Participants: 001 (hostess 1)

Filename: RK007

001: 來<lai5> 共<ka7> 聽眾<thiann1ciong3> 朋友<ping5iu2> 講<kong2> 一1<cit8> 个<e5> 鳥仔<ciau2a2> e0<e0> 故事<koo3su7> hoonn0<hoonn0>，onn0<onn0> 有<u7> 一1<cit8> 个<e5> 拍獵<phah4lah8> e0<e0> 人1<lang5> a02<a0> hoonn0<hoonn0>，a01<a0> 伊<i1> 掠著<liah8tioh8> 一1<cit8> 隻<ciah4> 鳥仔<ciau2a2>，hoonn01<hoonn0>，這2<cit4> 隻<ciah4> 鳥仔<ciau2a2> 足<ciok4> 水2<sui2> 足<ciok4> 水2<sui2> e02<e0> hoonn0<hoonn0>，[是<si7> 一1<cit8> 隻<ciah4> 真<cin1> i] ，e01<e0> 恰若<kah4na2> 彩色<chai2sik4> e0<e0> 就<to7> 著1<tioh8>，好親像<ho2chin1chiunn7> 咱<lan2> 彼1<he1> 南部<lam5poo7> e0<e0>，onn0<onn0> 彼2le0<hit4le0> 彩色<chai2sik4> 鳥<ciau2> 共款<kang7khuan2>，m0<m0> [m 確實] [m 很] [m 美] ，

001: 但是<tan7si7> 這2<cit4> 隻<ciah4> 鳥仔<ciau2a2> ne0<ne0>，e01<e0> 予1<hoo7> 伊<i1> 掠著<liah8tioh8> 了後<liau2au7> a02<a0>，這2<cit4> 隻<ciah4> 鳥仔<ciau2a2> 講<kong2> 會1<e7> 講話<kong2ue7> la0<la0> hoonn0<hoonn0>，

001: a01<a0> 這1<ce1> 講話<kong2ue7> 是講<si7kong2> 啥物<siann2mih8> 話<ue7> ne0<ne0>？伊<i1> 就<to7> 共<ka7> 這1<ce1> 个<e5> 拍獵<phah4lah8> e0<e0> 人1<lang5> 講<kong2> a02<a0>，enn0<enn0> 爲著<ui7tioh8> 欲<beh4> 感謝<kam2sia7> 你<li2> 會1<e7> 共<ka7> 我<gua2> 放開<pang3khui1>，所以<soo2i2> 我<gua2> 送<sang3> 你<li2> 三1<sann1> 項<hang7> 寶<po2>，這1<ce1> 三1<sann1> 項<hang7> 寶<po2> ne0<ne0>，是<si7> 三1<sann1> 句<ku3> 話<ue7>。

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 1, March 2012, pp. 27-42

可變速中文文字轉語音系統

Variable Speech Rate Mandarin Chinese Text-to-Speech System

江振宇[*, #]、黃啓全[*]、王逸如[*]、余秀敏[+]、陳信宏[*]

Chen-Yu Chiang, Qi-Quan Huang, Yih-Ru Wang, Hsiu-Min Yu, and Sin-Horng Chen

摘要

本論文描述以隱藏式馬可夫模型爲基礎發展之「可變速中文文字轉語音系統」，訓練語料爲三種不同語速之平行語料，分別對三種語速訓練文脈相關隱藏式馬可夫模型，並利用給予不同語速模型權重值來內插調整語速。另外，從語料庫觀察發現到慢速語音之靜音停頓較多而快速語音較少，傳統以標點符號位置決定靜音停頓的簡單方法，在用於可變速語音合成是不適當的，因此本研究加入預估靜音停頓之機制，對於不同語速分別訓練靜音停頓預估決策樹，再利用調整權重值內插不同語速停頓決策樹機率的方法，達到不同語速下靜音停頓的預估。爲了評估本系統之效能，我們對系統進行客觀測試及主觀測試，在客觀測試中，評量靜音停頓預估之效能及量測合成語音和目標語音的誤差值；在主觀測試中，特別針對隱藏式馬可夫模型權重、靜音停頓決策樹權重以上兩組權重值的組合比較合成語音自然度，實驗結果顯示兩組權重值必須匹配才可合成出較自然的語音。期望以本論文提出方法建構之系統，較傳統單一語速之文字轉語音系統，更適合用於人機互動之中。

[*] 國立交通大學電機工程學系, National Chiao Tung University, Hsinchu, Taiwan
E-mail: gene.cm91g@nctu.edu.tw; cchangwo@yahoo.com.tw; yrwang@cc.nctu.edu.tw; schen@mail.nctu.edu.tw

[+] 中華大學語言中心, Language Center, Chung Hua University, Hsinchu, Taiwan
E-mail: Kuo@chu.edu.tw

[#] 國立台北大學通訊工程學系, National Taipei University, New Taipei City, Taiwan
E-mail: cychiang@mail.ntpu.edu.tw

關鍵詞：文字轉語音系統、中文韻律、語速、停頓預估

Abstract

This paper presents an Hidden Markov Model (HMM)-based variable speech rate Mandarin Chinese text-to-speech (TTS) system. In this system, parameters of spectrum, fundametal frequency and state duration are generated by a context dependent HMM (CDHMM) whose model parameters are linear-interpolated from those of three CDHMMs trained by corpora in three different speech rates (SRs), i.e. fast, medium and slow. In addition, three decision tree (DT)-based pause break predictors trained by using the three SR corpora are used to interpolate the probabilities for inserting pause breaks. The performance of the proposed TTS system were evaluated by several objective and subjective tests. Experimental results suggested that coherence between interpolation weights for CDHMMs and DT-based pasue predictors is crutial for naturalness of the synthesis speech in variable SR. We believe that the proposed variable speech rate Mandarin Chinese TTS system is more suitable than conventional fixed SR TTS systems for applications of human-machine interaction.

Keywords: Text-to-Speech System, Mandarin Prosody, Speech Rate, Break Prediction

1. 緒論

1.1 研究背景、動機

文字轉語音技術在人機界面裡扮演著重要的角色，隨著大型語料庫(corpus-based)以及隱藏式馬可夫模型(HMM-based)為基礎的文字轉語音技術興起，語音合成的品質較以往進步許多，在許多人機介面應用中已有不錯的表現，然而在不同的應用上會有不同說話速度語音的需求，以達到更有效的溝通。以電話語音訂票系統為例，對本國籍的互動者來說，一般速度的合成語音可能過慢而浪費時間，這時快速語音就很適合此使用情形，但對於老人或外國人士來說，提供比一般速度稍慢的電話語音，才能讓他們有足夠時間反應聽懂內容，因此，在一些特定的人機互動情境下，傳統只做單一語速之語音合成系統便顯得不夠實用，因此開發不同語速的語音合成系統是一個值得深入探討的議題。

1.2 相關研究

1.2.1 語音合成方法

近期語音合成系統廣為使用的合成方式主要有兩種，分別是大型語料庫(corpus-based) (Chou *et al.*, 2002) 及隱藏式馬可夫模型(HMM-based approach) (Tokuda *et al.*, 2000) 的語

音合成方法；大型語料庫合成法由錄製好的語料庫中，挑選適當的語音信號片段串接合成，因此可原音重現，有極佳的合成音質，但是如果要合成出不同特性的語音，如不同講話速度及多種情緒等應用，則須錄製大量的語料作爲挑選單元的基礎，然而欲收集不同特性之語料並不容易，因此，對於合成不同特性語音的應用，單元選取並不是一個適合的方法。

基於隱藏式馬可夫模型語音合成器是一種統計式參數語音合方法，是目前最爲廣泛採用的合成方法，它以文脈相關隱藏式馬可夫模型(Context-dependent HMMs, CDHMMs)來模擬不同語言參數或韻律架構下的聲學信號，從語料庫訓練得到頻譜模型(spectral parameter model)、基頻模型(F0 parameter model)及音長模型(duration model)。欲合成語音時，利用上述訓練好的三種模型，依據輸入文本的語言參數或預估之韻律標記找到適當 CDHMM 模型並串接之，再以特殊的演算法由串接之 CDHMM 參數產生 frame spectrum 及 frame F0 參數，最後將 spectrum 和 f0 參數輸入 MLSA 濾波器(Mel Log Spectrum Approximation filter) (Imai, 1983) 輸出合成出語音訊號。

當想要以現有模型去合成出不同特性的語音訊號，則可利用調整參數的方式達到目的，如內插(interpolation methods) (Yoshimura *et al*., 2000)、調適(adaptation methods) (Tamura *et al*., 2001)。跟單元挑選相反的，使用隱藏式馬可夫模型合成器，不需要大量目標的語料，只需要足夠的語料就能利用現有隱藏式馬可夫模型去合成出不同特性的語音信號。

1.2.2 不同語速韻律之研究

研究語音韻律的文獻雖然很多，但是討論相異語速的文獻卻很少，在 (Yu *et al*., 2007) 著作中，作者一開始先利用對話語音的說話速度較快，以及音高軌跡範圍較朗讀式語音爲窄的特性，將朗讀式語音利用 linear regression 的方式轉換成對話語音，另外，對話語音由於說話速度較快，音節的音高軌跡可能會因爲發音不完全導致軌跡不完整，相較於在朗讀式語音完整發音如呈現拋物線的音高軌跡，對話語音變成近似直線，利用相對於朗讀式語音音高軌跡不完整的特性，將朗讀式語音韻律轉換爲對話式語音。

在 (Li & Zu, 2008) 中，作者採用階層式韻律架構的觀念，採用三種不同語速之平行語料庫做分析，實驗對於語速的測量分爲兩類，一爲 speech rate (SR)，定義爲每秒鐘包含停頓時長(pause duration)的發音的音節個數；另一爲 articulation rate (AR)，定義爲每秒鐘的音節個數(不包含 pause duration)。實驗語料庫爲四個漢語文字段落，音節數分別爲 134、123、151 和 34，實驗語料有快、中、慢速的區別，分析了在不同語速下，不同韻律單元的 AR、SR 變化，發現改變說話速度對各韻律階層邊界的 silent pause 是非線性的，語速的快慢會影響基頻軌跡(F0)的平均，發現的現象是快速語料的音高比較高而慢速語料的音高比較低，且其音高軌跡的 dynamic range 比慢速語料小。此篇提出一些不錯的觀點，但是語料庫的資料量不夠大，導致其分析結果不夠一般性是比較可惜的地方。

在 (Tseng, 2008) 中，根據其所提出的階層式多短語韻律句群架構，使用線性回歸統計中的逐步回歸技術(step-wise regression technique)來估算語料，分析出三種不同中文語速之韻律詞、韻律短語和呼吸組層次的時長和音強 pattern，解析出不同語速下，各個層次韻律單元於時長和音強的貢獻，此實驗中平行語料庫之快速語料爲一位台灣男性播音員所發音，中速語料是由一位台灣女性播音員發音，而慢速語料則由北京女性播音員發音。此篇研究提出了不少新的發現，但因其語料庫不是由同一人發音，會導致有些影響實驗結果的因素沒考慮到。

上述這些文獻雖有探討到相異語速的韻律變化，但仍有幾項需要克服的因素，(Yu *et al*., 2007) 的方法提供了 bottom-up 的方式分析，僅從音節層次討論音高軌跡會忽略到韻律結構上層的影響；至於 (Li & Zu, 2008) 和 (Tseng, 2008) 的階層式韻律架構則提供一個 top-down 的分析方式，對於底層之音節層次分析較缺乏，此外，傳統韻律階層的研究都需要人工事先標記韻律邊界，因此，在文獻 (Chiang *et al*., 2009) 同時提供 bottom-up 和 top-down 的分析方式，盡可能從各個不同面向討論相異語速語音之韻律變化，採用的自動標記分析方法可以省時省力，同時還可兼顧採用大量語料做研究，分別對不同語速語料的訓練得到韻律模型，藉由分析不同語速之韻律模型參數，探討了不同語速的韻律特性，包含：（1）不同語速音節基頻軌跡之比較、（2）不同語速之 prosodic phrasing、（3）上層韻律單元的 patterns、以及（4）break 和語言參數的關係。此研究是近期對於不同語速韻律較大規模的研究，對於建構可變速語音合成提供了許多實用的資訊。

1.3 系統概述及研究方向

本研究是以 HMM-based 語音合成器爲基礎之「可變速中文文字轉語音系統」，系統架構如圖 1 所示，訓練語料爲一位女專業播音員所錄製的快、中、慢三種語速之平行語料庫，其文本爲中研院 Treebank 3.0 (Huang *et al*., 2000) 選出之 348 篇短文。本研究先以這三種語速的語音資料庫，各自訓練出不同語速的 HMM-based 語音合成器(包含頻譜及音高 CDHMM 模型及 state duration 模型)，另外，爲了由輸入的文字或語言參數決定音節之間靜音停頓的存在與否，我們分別對三種語速的語音，以決策樹的方法由語言參數預估靜音停頓的插入。爲了達到可變速的語音合成，本研究以調整不同語速之靜音停頓決策樹模型以及 HMM-based 語音合成器參數之權重，可內插出不同語速之合成語音，探討不同語速之靜音停頓決策樹權重和 HMM-based 語音合成器權重關係，找到影響合成可變速語音品質的重要因素。

1.4 漢語多語速語料簡介

本研究所採用的實驗語料庫，是由一位專業的女性播音員讀稿之快速、中速及慢速之文本平行語料庫，此平行語料庫含有 348 個音檔，共有 48035 個音節，其語速及音高統計資訊如表 1 所示，其中 AR 與 SR 的定義同 1.2.2 節。語料庫的錄製順序是在第一梯次先錄中速語速，接下來才將其他兩種速度錄製完成，音檔均爲 20kHz 的取樣頻率及 16-bit 之 PCM 格式，語料庫的錄製文字爲 Sinica Treebank 語料庫中選出的短篇文章，主要內

容大多摘錄自新聞、網路文章、國小教科書等，由數個句子所組成的段落。所有音節的切割標記和基頻軌跡(F0)的偵測均先自動由 Hidden Markov Model Tool Kit(HTK) (Young *et al.*, 2006) 和 WaveSurfer (Sjlander & Beskow, 2000) 完成，明顯的參數錯誤再以人工修正，平均每個語句(utterance)音節數爲 138，每個句子 10.37 個字，最短及最長分別爲 80 與 272 個音節。

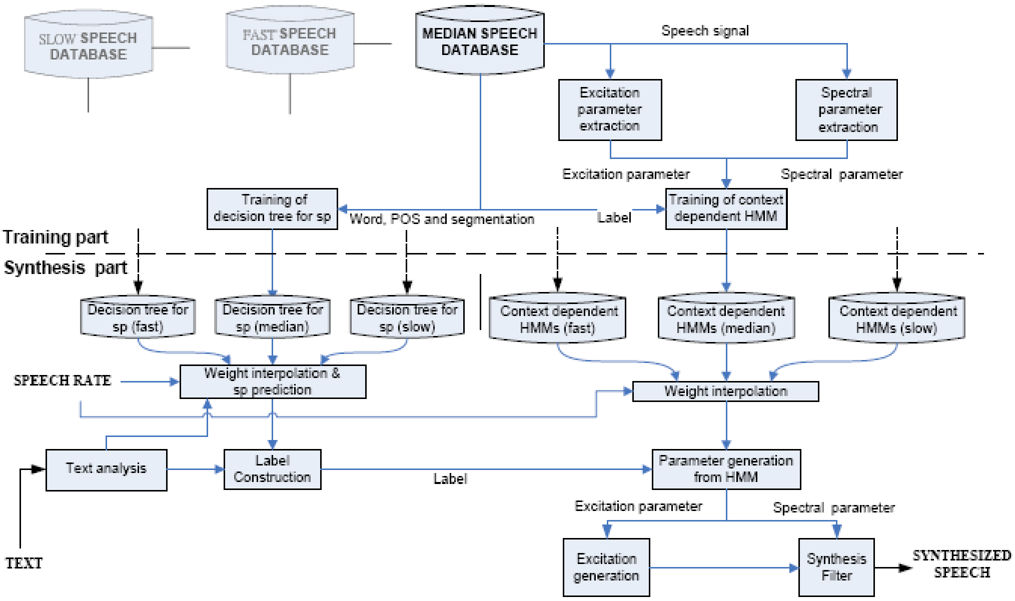

SLOW SPEECH DATABASE
FAST SPEECH DATABASE
MEDIAN SPEECH DATABASE
Speech signal
Excitation parameter extraction
Spectral parameter extraction
Excitation parameter
Spectral parameter
Training of decision tree for sp
Word, POS and segmentation
Label
Training of context dependent HMM
Training part
Synthesis part
Decision tree for sp (fast)
Decision tree for sp (median)
Decision tree for sp (slow)
Context dependent HMMs (fast)
Context dependent HMMs (median)
Context dependent HMMs (slow)
SPEECH RATE
Weight interpolation & sp prediction
Weight interpolation
Text analysis
Label Construction
Label
Parameter generation from HMM
Excitation parameter
Spectral parameter
TEXT
Excitation generation
Synthesis Filter
SYNTHESIZED SPEECH

圖 1. 多語速文字轉語音系統之訓練及合成部份

表1. 平行語料庫的平均音長、SRs和ARs

語料庫類型	Fast	Median	Slow
每字平均音長(秒)	0.183	0.241	0.267
SR(syllables/sec)	4.48	3.01	2.47
SR 的變異數	0.082	0.040	0.044
AR(syllables/sec)	5.56	4.19	3.79
AR 的變異數	0.144	0.070	0.065
F0 的平均值(Hz)	201.38	195.88	195.594
F0 的變異數	2489.27	2559.20	2773.37

2. 文字轉語音系統之訓練

本系統是由三種語速的 CDHMM 和靜音停頓決策樹共同加權合成出可變速之語音，我們分別對三種不同語速各自訓練出其 CDHMM 和靜音停頓決策樹，詳細方法如下。

2.1 基於隱藏馬可夫模型之語音合成 (HMM-based Speech Synthesis)

我們將中文聲母、韻母、長靜音（SIL）以及短靜音(SP)模擬成五個狀態的 HMM 模型，也就是將他們模擬成最小的 HMM 訓練單元，對於每個最小單元給予文本標示紀錄其文脈相關資訊，利用由語料求取好的語音聲學參數和文本標示，訓練出文脈相關的頻譜及音高 CDHMM 模型及 state duration 模型。

2.1.1 聲學參數 (Spectral and excitation parameter extraction)

本研究中 CDHMM 模擬的聲學參數為廣義梅爾倒頻譜係數(Mel-generalized cepstrum, MGC) (Tokuda *et al.*, 1994) 及基頻 (F0)。廣義梅爾倒頻譜係數可藉由調整其 γ 參數，將語音信號頻譜以 all pole (γ=-1)、Cepstrum (γ=0) 或是以廣義的 pole 和 zeros 一起表示 (γ≠-1,0)，亦可調整 α 參數以代表不同的 frequency wrapping，以方便考量人耳的聽覺效應。在本研究中，我們使用 SPTK (SPTK Working Group, 2009) 工具抽取 24 階廣義梅爾倒頻譜係數，設定 γ=0 以及 α=0.5，音檔取樣頻率為 20kHz，所使用的分析音框為 25ms（500 個資料點）的漢明窗（Hamming window），音框位移為 5ms（100 個資料點）。另外，抽取基頻參數則使用 Wavesurfer 工具中的 ESPS 方法求取 (Sjlander and Beskow, 2000)，分析音框大小（window size）為 7.5ms，而音框位移（window size）為 5ms。

2.1.2 文本標示 (label)

文本標示提供訓練 CDHMM 及 state duration 的文脈相關語言參數，或在合成時挑選適當的 CDHMM 及 state duration 模型。訓練 CDHMM 時依照文本標示提供的文脈相關資訊對聲學參數作訓練，文本標示的文脈相關參數會影響 HMM 單元本身的頻譜及韻律變化，也會影響 HMM 單元之間連接的狀況，如連音現象、詞首詞尾和句首句尾明顯的音高差異及音節伸長縮短。本系統使用的文脈資訊如表 2：

表2. 文脈相關語言參數

p_{n-1}, p_n, p_{n+1}	Previous(PRE)/current(CUR)/following(FOL) Initial/Final/SP
ST_{n-1}, ST_n, ST_{n+1}	Lexical tones of PRE/CUR/FOL syllable
PW_1 / PW_2	Syllable position in a lexical word (LW) (forward/backward)
PS_1 / PS_2	Syllable position in a sentence (forward/backward)
PM	Punctuation mark after the current syllable
$WL_{n-2}, WL_{n-1}, WL_n, WL_{n+1}, WL_{n+2}$	Lengths of PRE-PRE/PRE/CUR/FOL/FOL-FOL LWs in syllable
$WP_{n-2}, WP_{n-1}, WP_n, WP_{n+1}, WP_{n+2}$	POSs of PRE-PRE/PRE/CUR/FOL/FOL-FOL LWs
SL_{n-1}, SL_n, SL_{n+1}	Lengths of PRE/CUR/FOL sentences in syllable

由於我們將長靜音以及短靜音視爲 HMM 的訓練單元，長靜音就是在音檔開始和結束的靜音部份，而短靜音則定義爲語句中音節間大於 25ms 靜音停頓，所以在文本標示中，對於短靜音也給予文脈相關資訊，在訓練時也會學習到不同文脈相關資訊下的停頓長度。

2.1.3 隱藏馬可夫模型之訓練

文本標示的文脈相關資訊組合相當多，每一種組合都是個別的 CDHMM，在訓練語料不夠充足的情況下，多數組合的 CDHMM 訓練資料量過少，使得訓練出來的模型會不夠準確造成過度訓練（overfitting），因此本研究使用標準的 Tree-based CDHMM 訓練方法 (Zen *et al*., 2007; Yoshimura, 2002)，以決策樹搭配適當的問題集來分群作訓練，以語言學的知識爲基礎設計出合理的問題集，對於某些資料量較少的模型可以合併在一起訓練以增加訓練的資料，如此可訓練出較強健的模型。在合成時，輸入文本標示依據決策樹上每個節點的問題，可找出適當的 CDHMM 串接，進而產生聲學以及韻律參數。以下爲問題集的概述：

✧ 依據前一個、現在、後一個聲母或韻母的發音方法、發音位置、送氣不送氣以及清音濁音設定問題集。

✧ 依據前一個、現在、後一個音節聲調的調值特性作分類，設定問題集，如一聲和二聲以高調值(H)爲結尾、一聲和四聲以高調值爲開始、二聲和三聲以中(M)或低調(L)值開始。

✧ 考慮現在音節所在的詞長和詞的位置，將主要會影響韻律特性的位置和詞長合併，設定爲問題集，如現在音節是否在詞首或詞尾、詞長是否大於四字詞等。

✧ 考慮前後及現在詞的詞類，將中研院 46 類詞類依實詞虛詞、八大詞類及其他特殊詞類集合合併，產生問題集。

✧ 考慮現在音節所在的句長和句子的位置,將主要會影響韻律特性的位置和句長合併，設定爲問題集，如現在音節是否在句首或句尾、句長是否大於十個字等。

由上列問題集概述的考量，本研究所設定的問題集共約 2100 個左右。

2.2 基於決策樹之停頓預估

由於在訓練時把靜音停頓也視爲一個 CDHMM 來作訓練，其存在與否可由語音切割資訊來決定(短靜音定義爲語句中音節間大於 25ms 靜音停頓),靜音停頓的長度(state duration)可由標準的 Tree-based CDHMM 訓練後得到的決策樹依輸入的文本標示決定。但在合成時，靜音停頓存在與否，只能由文本標示的文脈相關語言參數資訊去預估。本研究分別對於不同語速的語料獨自訓練其靜音停頓決策樹模型，目標爲預估音節間是否有靜音停頓。由不同語速語料靜音停頓的觀察，發現快速語料的靜音停頓較中速少，而中速語料又比慢速少，利用這種語速語靜音停頓多寡的的關係，在合成時將決策樹對於每個音節間是否有靜音停頓求出機率值，即爲有靜音停頓的機率和沒有靜音停頓的機率，分別利

用快中慢的決策樹預估出三組機率，再利用權重值乘以相對應的機率值相加，以達到不同語速下預估靜音停頓的目的。

本研究只考慮詞和詞之間的靜音停頓，假設詞內音節間無靜音停頓，所以預估處理的單元爲詞邊界，所使用的文脈相關資訊如訓練 CDHMM 的文本標示一樣，但問題集只考慮表二之中詞以上的語言參數，而決策樹的分裂條件爲 maximum information gain。

3. 多語速文字轉語音系統

3.1 Text Analysis

文字分析(Text analysis)是文字轉語音系統的第一級，傳統的國語斷詞器使用的是長詞優先及構詞規則，最著名的是中央研究院的中文斷詞系統。但自 2000 年起，由於 conditional random field (CRF)方法 (Lafferty *et al.*, 2001) 被提出，並有效的使用在自然語言處理中的各個問題，都被證實較傳統規則法或其他統計式方法爲佳 (Jiang *et al.*, 2006)。因此，本系統的 Text analysis 的斷詞、base-phrase chunker 及詞類標記部分，便是採用 CRF 的方法做爲核心，其系統架構如下圖 2 所示，其中包含了(1) symbol normalization、(2) word segmentation、(3) POS(part-of-speech) tagger、(4) Word construction、(5) base-phrase chunker 及(6) grapheme to phone 六部分。

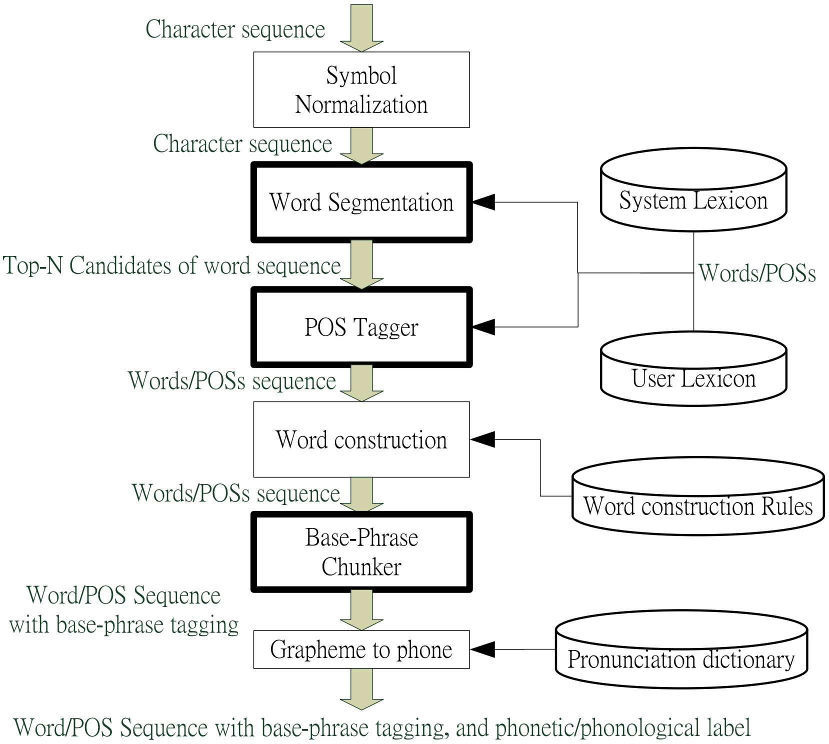


圖 2. Text analysis 之系統方塊圖。

1. Symbol normalization：在此級中將輸入的字串如有 ASCII 的部分，要轉換爲 BIG5，另外，有很多標點符號是屬於同一種標點符號類別，我們將這些同義異形的標點符號正規劃爲其中一種作爲代表。
2. Word segmentation：由於中文文章沒有標示詞的邊界，我們必須將詞的邊界識別出來以得到語音合成需要的語言參數，本系統是以 CRF 以每一個中文字做爲 input feature，要預估的目標爲每個中文字後的標示：{ B1, B2, B3, M, E, S}，其中 B1、B2、B3 分別表示該字位於一個詞的前三字位置，M 代表該字位於詞中第四個字元之後但非詞尾的位置，E 代表字位於詞尾，S 代表單字詞，另外我們也可以使用 user define 的外掛字典輔助斷詞。
3. POS tagger：利用 CRF 以詞、詞對應可能的 POS 爲 input feature，預估每個詞對應到的 POS。
4. Word construction：在這一級我們以規則法，將符合構詞規則的詞由前級斷詞和標示 POS 的結果來構成更具語法和語義的詞，這些詞包括定量複合詞、重複詞等等。
5. Base-Phrase chunker：在這一級我們利用斷出的詞和詞類，以 CRF 將一些基本語法片語標記出來，這些基本語法片語包含 VP：述詞詞組、NP：名詞詞組、GP：方位詞詞、PP：介詞詞組、AP/ADVP：形容詞詞組及副詞詞組。
6. Grapheme to phone：此級爲文字分析器的最後一級，將前級所斷出的詞以一個十二萬詞的發音字典標記上發音和聲調，另外我們也以規則法處理了一些常見的破音字，使其發音和聲調正確。

表 3 爲 word segmentation、POS tagging 以及 base-phrase chunker 效能的評估，其實驗語料的設定皆爲十分之九的訓練及十分之一的測試。由表所示的數據顯示，本文字分析的效能十分優良。

表 3. 文字分析器效能評估

實驗	實驗語料	accuracy	precision	recall	FB1
Word segmentation	Bakeoff-2004	98.30	95.95	96.79	96.37
POS tagging	中央研究院 漢語平衡語料庫	94.73	94.73	94.73	94.73
Base-phrase chunker	中研院 sinica treebank3.0	93.16	92.18	92.27	92.22

3.2 Weight Interpolation

爲了達成多語速合成，本系統具有兩組權重值，一組權重值爲調整預測靜音停頓決策樹的比重，調整此權重影響最大的是 SR，當權重值調成接近慢速，預估的靜音停頓會越來越多，利用決策樹對於每個音節間是否有靜音停頓求出機率值，即爲有靜音停頓的機率和沒有靜音停頓的機率，分別利用快中慢的決策樹預估出三組機率，在利用權重值乘以

相對應的機率值相加，以達到調整權重決定靜音停頓的目的，如下式：

$$sp_n^* = \arg\max_{sp_n} \sum_{i=1}^{3} w_i \times P_i(sp_n \mid L_n) \tag{1}$$

其中 i 爲決策樹的 index (i=1：慢，i=2：中，i=3：快)； w_i 爲第 i 個決策樹模型的權重值； $sp_n \in \{$靜音停頓, 非靜音停頓$\}$ 爲第 n 個詞後面的靜音停頓與否； L_n 爲文脈語言資訊； $P_i(sp_n \mid L_n)$ 爲經由文脈語言資訊 (L_n) 組成決策樹問題集後，由第 i 個決策樹結構裡，找尋到對應之葉節點下 (leaf node) 靜音停頓和非靜音停頓的機率值。

而第二組權重值影響著頻譜、音長及基頻，在語料分析中發現語速越快不僅音長變短，基頻也會隨著拉高，直接影響到隱藏式馬可夫模型的參數，很直觀地，當調整權重值越靠近快速語速語音之隱藏式馬可夫模型時，相對於僅使用慢速語音之隱藏式馬可夫模型，每個音節的音長會變短且音頻會提高。以不同權重值內差三種語速之模型參數方法如下式 (Yoshimura *et al*., 2000; Iwano *et al*., 2002)：

$$\boldsymbol{\mu} = \sum_{i=1}^{3} a_i \times \boldsymbol{\mu}_i \tag{2}$$

$$\mathbf{U} = \sum_{i=1}^{3} a_i^2 \times \mathbf{U}_i \tag{3}$$

其中 i 爲 CDHMM 模型的 index(i=1：慢，i=2：中，i=3：快)； a_i 爲第 i 個 CDHMM 模型的權重值， $\boldsymbol{\mu}_i$ 及 $\mathbf{U}_i$ 分別爲 CDHMM state 之 mean vector 及 covariance matrix。

第一組權重值影響靜音停頓的變化，而第二組權重值影響了音長、頻譜及基頻，在自然的語音訊號中，慢速語料靜音停頓較多，音節音長也會拉長，快速語料則相反。在給定權重值也需要按照語速的規則，當想要合成語速較快的語音訊號時，增加快速語速之靜音停頓決策樹的比重，使靜音停頓預估出的數量較少，只在適合的位置給定靜音停頓，同時，我們也調整隱藏式馬可夫模型權重，增加快速語料的比重，而可以產生出較短的音長及較高的基頻，這兩組權重值需要有正相關才會匹配，兩組不匹配的權重值會合成出不自然的語音訊號，因此在不同語速下兩組權重值的匹配是相當重要的，在之後的實驗會對這兩組權重值匹配作主觀測試的實驗。

3.3 Label Construction

欲合成的文字經由文本分析後，可得到對應文脈相關的語言參數資訊，使用之前以內插靜音停頓決策樹所預估之靜音停頓，放入文本標示(label)，最終產生的 label 具有欲合成文本中每個聲母、韻母及靜音停頓的文脈相關語言參數。

3.4 Parameter Generation from HMM

在 label construction 步驟後產生文本標示 (label)，依據文本標示使用三種語速之 CDHMM 模型、state duration 模型及 CDHMM 模型參數權重，由文本相關決策樹找到適當的模型，首先預估出每個聲母、韻母或靜音停頓的長度，再利用 maximum likelihood

法 (Tokuda *et al*., 2000) 產生每個音框的 logF0 及 MGC 頻譜參數。

3.5 Excitation Generation and Synthesis Filter

將上一步得到的每個音框之 logF0 和 MGC 頻譜參數輸入至 MSLA filter (Mel-Log Spectrum Approximation filter) (Imai, 1983)產生合成語音。

4. 實驗結果及討論

實驗語料已於 1.4 中介紹，對於每種語速取其約 328 個語句爲訓練語料，另 20 句爲測試語料，爲了評估本可變速漢語語音合成系統的效能，我們分別對合成語音進行客觀及主觀的測試。在客觀測試方面，我們量測了靜音停頓決策樹的預估正確性，另外，也量測了整個系統合成語音和目標語音的量化誤差。而在主觀測試方面，對系統兩組權重值匹配的狀況作主觀測試的實驗，合成音檔的展示請連結 http://140.113.144.71。

4.1 客觀測試

在第一個實驗中，我們對靜音停頓決策樹的效能進行評估，計算合成音檔和目標語句的靜音停頓預估的正確率及混淆程度，因爲只有單純三種語速的目標語句，沒有實際介於這三種語速的目標語句，所以只有對於三種不同語速目標語句的預估結果作觀察，以合成快速語音爲例，當測試快速語料時，我們調整快速的靜音停頓預估決策樹權重值爲 1，其他語速之權重爲 0，中慢速測試亦同。表 4 爲預估靜音停頓對於快中慢語速的結果。

*表4. 不同語速下預估靜音停頓的結果，**XX***代表預測爲靜音停頓或非靜音停頓（以百分比表示），**Total** 爲 **Non-SP** 或 **SP** 的總個數。*

慢	Inside			Outside			
	Non-SP*	SP*	Total		Non-SP*	SP*	Total
Non-SP	90.05	9.95	28108	Non-SP	89.66	10.34	1885
SP	30.19	69.81	20486	SP	33.57	66.43	1415
中	Inside			Outside			
	Non-SP*	SP*	Total		Non-SP*	SP*	Total
Non-SP	92.77	7.23	29119	Non-SP	91.55	8.45	1977
SP	37.81	62.19	19314	SP	39.61	60.39	1323
快	Inside			Outside			
	Non-SP*	SP*	Total		Non-SP*	SP*	Total
Non-SP	96.34	3.66	35380	Non-SP	94.83	5.17	2496
SP	49.5	50.5	11613	SP	52.74	47.26	804

由實驗結果發現，對於快速合成語音預估靜音停頓的結果是最差的，錯誤大多是在預測目標語句有靜音停頓的部份，主要原因可能是因爲快速語料裡音節間的靜音停頓較少，所以造成了決策樹學習到音節間無靜音停頓的機率較大，在預測結果也是偏向沒有靜音停頓，另外可能的原因，是考慮到快速語料不論是 AR 和 SR 變化都是最大的，語句和語句間語速有較大的差異，因爲語速和靜音停頓的多寡有關係，語速的差異潛在會造成快速語料靜音停頓預估上的困難。在慢速語料上雖然在非靜音停頓預測上略輸快速語料，但在有靜音停頓預測上比快速語料準得多，可能是因爲語者於朗讀慢速語料時，會將詞或韻律詞的結構清楚念出，所以在語料上產生較一致性的靜音停頓，較容易從語言參數學習到規則，因此準確度比快速要高的多。

第二個客觀測試，我們分別測量合成和目標語句其基頻、停頓靜音的長度以及音節的長度的誤差，使用均方根誤差（Root Mean Square Error, RMSE）用來評估誤差值，因語料庫只有三種語速，所以在測量快速語料的 RMSE 時，預測靜音停頓決策樹的權重和隱藏馬可夫式模型的權重，均設定快速權重值爲 1，其他權重設爲 0，中慢速語料也使用同樣的方法測量，表 5 爲實驗結果。

由整體來看 Inside test 的 RMSE 都低於 Outside test 這是因爲過度訓練（overfitting）的關係,經觀察發現靜音停頓音長的預測不論 Outside test 和 Inside test 在語速快的 RMSE 均爲最低，由於快速語速在靜音停頓的音長並不長，就算靜音停頓沒有正確預估出來，誤差也不會太大，而慢速的靜音停頓就不一樣，靜音停頓音長較長，沒有正確預估到靜音停頓誤差就會較大，我們觀察音節音長的 RMSE 也看到同樣的結果，在語速快的音節音長 RMSE 均爲最低，因爲快速語料音節音長都較短，計算誤差也不會太大。

表 5. 快中慢語料作測試之 RMSE 值

測試項目	語速		
	Fast	Median	Slow
Inside F0 (Hz)	36.28	34.38	35.21
Outside F0 (Hz)	42.66	42.78	45.23
Inside sp duration (ms)	44.97	64.19	84.17
Outside sp duration (ms)	56.55	60.02	85.55
Inside syllable duration (ms)	37.53	41.44	44.19
Outside syllable duration (ms)	39.23	42.66	47.08

4.2 主觀測試

主觀測試目的爲測試系統兩組權重值不同的組合，以主觀測試判別合成語音的自然度，對於快中慢兩組權重值設爲：1-0-0、0-1-0、0-0-1、0-0.5-0.5（x-x-x 中的 x 順序代表慢速、中速、以及快速權重值），因爲考慮的組合數量過多，而且慢速跟中速語料依據 SR、 AR 以及基頻的統計差異並不大，因此不考慮 0.5-0.5-0 這個權重值組合，所以本實驗只有 16 種靜音停頓-CDHMM 權重的組合。

每一個合成文本均爲 outside test 的語句，一個文本依據兩組權重值變化會產生 16 種不同語速變化的合成音檔，各分爲四組作測試，以同樣隱藏馬可夫模型的權重值爲同一組，目的爲固定一組權重值，觀察不同權重預測靜音停頓的匹配程度。主觀測試中語音自然度的評分爲五分制，分數爲一至五，一爲最不自然，五爲最自然，總共對 6 人作主觀測試，每個測試者由九句文本中選聽兩句文本的語句，其中一句文本與另一個測試者重複，因爲每文本有 16 種語速權重組合，所以每個人聽 32 句測試語句，整個測試語句共有 192 句，測試結果如表 6 所示。

表6. 主觀測試的平均值±一個標準差，x-x-x中的x順序代表慢、中、快權重值

預測靜音停頓權重值	隱藏馬可夫權重值			
	1-0-0	0-1-0	0-0.5-0.5	0-0-1
1-0-0	2.33±0.61	3.08±1.36	2.79±0.98	2.21±0.70
0-1-0	2.54±0.88	3.38±0.96	3.25±0.391	2.21±0.52
0-0.5-0.5	2.67±0.60	3.08±0.99	3.67±0.79	2.54±1.43
0-0-1	2.83±0.88	2.88±1.00	3.71±0.93	3.25±1.66

由主觀實驗發現，兩組權重值必須有正相關的關係，合成出的語音才會自然，當兩組權重值不相匹配的時候，合成出的語音大多不自然，因此通常在表六對角線附近會有最大的自然度，但當靜音停頓決策樹權重值爲 0-0-1 和隱藏馬可夫權重值爲 1-0-0 是比較令人訝異的結果，猜測在隱藏馬可夫權重值爲 1-0-0 時語速很慢，造成測試者聽得厭煩，由表六固定隱藏馬可夫權重值 1-0-0 觀察，發現無論靜音停頓決策樹權重如何調整，受測者所給予的自然度都偏低，在這種權重值組合下，受測者覺得厭煩分數都給的較低。

5. 結論與未來方向

本系統爲可變速中文文字轉語音，經由權重值調整所合成出的語音，合成出來的語音基本上尙佳，在主觀實驗中兩組權重值皆調爲 0-0.5-0.5 所合成出的語音自然度也是令人滿意的，其語速介於中速及快速之間，系統可預測出適當的靜音停頓、頻譜及其他韻律參數，達到合成出不同語速的自然語音。由客觀實驗發現快速的靜音停頓預估結果較差，未來的研究會以慢速爲基準預估其他語速的靜音停頓，因爲在慢速時讀稿人會完整分析詞和韻律詞結構後念出語句，考慮相對靜音停頓的變化，如某些音節間或詞間不管在慢速還是快速都需要靜音停頓，而有些靜音停頓在快速時反而消失，考慮這些相對的變化再進一步進行靜音停頓預估是需要的。

靜音停頓決策樹是由語言參數預估詞之間靜音停頓的出現與否，雖然本系統所預估的靜音停頓結果尙佳，但以這個系統所預估出來靜音停頓特性並沒有考慮實際靜音停頓

的長度，是這個預估靜音停頓系統的重大盲點，要改進此靜音停頓預估系統，必須分析靜音停頓長度隨語速變化的特性，依照這些特性設計出更適合的預估系統。

Reference

Chou, F.-C., Tseng, C.-Y., & Lee, L.-S. (2002). A Set of Corpus-Based Text to Speech Synthesis Technologies for Mandarin Chinese. *IEEE Trans. on Speech and Audoio Processing*, 10(7) , 481-494.

Chiang, C.-Y, Tang, C.-C., Yu, H.-M., Wang, Y.-R., & Chen, S.-H. (2009). An Investigation on the Mandarin Prosody of a Parallel Multi-Speaking Rate Speech Corpus. In *Proc. of Oriental COCOSDA 2009*, 148-153.

Huang, C.-R., Chen, K.-J., Chen, F.-Y., Gao, Z.-M., & Chen, K.-Y. (2000). Sinica Treebank: Design criteria, annotation guidelines, and pn-line interface. In *Proc. of the Second Chinese Language Processing Workshop 2000*, 29-37.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., & Tokuda, K. (2007) The HMM-based speech synthesis system version 2.0. In *Proc. 6th ISCA Workshop Speech Synth.*, 294-299.

Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale. In *Proc. of ICASSP*, 93-96.

Iwano, K., Yamada, M., Togawa, T., & Furui, S. (2002). Speech-rate-variable HMM-based Japanese TTS system. In *Proc. of IEEE TTS Workshop 2002*, 219-222.

Jiang, W., Guan, Y., & Wang, X.-L. (2006) Conditional Random Fields Based Label Sequence and Information Feedback. *Lecture Notes in Computer Science of Natural Language Processing and Expert Systems*, (4114), 677-689.

Li, A.-J., & Zu, Y.-Q. (2008). Speaking Rate Effects on Discourse Prosody in Standard Chinese. In *Proc. of the Speech Prosody2008*, 449-452.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, 282-289.

Sjlander, K. & Beskow, J. (2000). Wavesurfer - an open source speech tool. In *Proc. of the ICSLP 2000*, 4, 464-467.

SPTK Working Group. (2009). Reference Manual for Speech Signal Processing Toolkit Ver 3.3. available at http://sp-tk.sourceforge.net/

Tokuda, K., Kobayashi, T., Masuko, T. & Imai, S. (1994). Mel- generalized cepstral analysis - A unified approach to speech spectral estimation. In *Proc. of ICSLP'94*, 1043-1046

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-Based speech synthesis. In *Proc. of ICASSP*, 1315-1318.

Tamura, M., Masuko, T., Tokuda, K., & Kobayashi, T. (2001). Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In *Proc of ICASSP*, 805-808.

Tseng, C.-Y. (2008). Corpus Phonetic Investigations of Discourse Prosody and Higher Level Information. Language and Linguistics, Institute of Linguistics, 9(3).

Yoshimura, T., Masuko, T., Tokuda, K., Kobayashi, T., & Kitamura, T. (2000). Speaker interpolation for HMM-based speech synthesis system. *J. Acoust. Soc. Jpn.* (E), 21(4), 199-206.

Yoshimura, T. (2002) Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems. *Ph.D thesis, Nagoya Institute of Technology*.

Yu, J., Huang, L.-X., Tao, J.-H., & Wang, X. (2007). Modeling Incompletion Phenomenon in Mandarin Dialog Prosody. *In Proc. of the Interspeech2007*, 462-465.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. C. (2006). The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK.

中央研究院中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw/, last visit 2009/09/09

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 1, March 2012, pp. 43-64

The Polysemy Problem, an Important Issue in a Chinese to Taiwanese TTS System

Ming-Shing Yu[*] and Yih-Jeng Lin[+]

Abstract

This paper brings up an important issue, polysemy problems, in a Chinese to Taiwanese TTS (text-to-speech) system. Polysemy means there are words with more than one meaning or pronunciation, such as “我們” (we), “不” (no), “你” (you), “我” (I), and “要” (want). We first will show the importance of the polysemy problem in a Chinese to Taiwanese (C2T) TTS system. Then, we will propose some approaches to a difficult case of such problems by determining the pronunciation of “我們” (we) in a C2T TTS system. There are two pronunciations of the word “我們” (we) in Taiwanese, */ghun/* and */lan/*. The corresponding Chinese words are “阮” (we$_1$) and “咱” (we$_2$). We propose two approaches and a combination of the two to solve the problem. The results show that we have a 93.1% precision in finding the correct pronunciation of the word “我們” (we). Compared to the results of the layered approach, which has been shown to work well in solving other polysemy problems, the results of the combined approach are an improvement.

Keywords: Polysemy, Taiwanese, Chinese to Taiwanese TTS System, Layered Approach

1. Introduction

Besides Mandarin, Taiwanese is the most widely spoken dialect in Taiwan. According to Liang *et al.* (2004), about 75% of the population in Taiwan speaks Taiwanese. Currently, it is government policy to encourage people to learn one’s mother tongue in schools because local languages are a part of local culture.

Researchers (Bao *et al.*, 2002; Chen *et al.*, 1996; Lin *et al.*, 1998; Lu, 2002; Shih *et al.*, 1996; Wu *et al.*, 2007; Yu *et al.*, 2005) have had outstanding results in developing Mandarin

[*] Department of Computer Science and Engineering, National Chung-Hsing University, Taichung 40227, Taiwan.

[+] Department of Information Management, Chien-Kuo Technology University, Chang-hua 500, Taiwan.
E-mail: yclin@ctu.edu.tw

text-to-speech (TTS) systems over the past ten years. Other researchers (Ho, 2000; Huang, 2001; Hwang, 1996; Lin *et al.*, 1999; Pan & Yu, 2008; Pan, Yu, & Tsai, 2008; Yang, 1999; Zhong, 1999) have just begun to develop Taiwanese TTS systems. There are no formal characters for Taiwanese, so Chinese characters are officially used in Taiwan. Consequently, many researchers have focused on Chinese to Taiwanese (C2T) TTS systems. This means that the input of a so-called Taiwanese TTS system is Chinese text. Yang (1999) developed a method based on machine translation to help solve this problem. Since there are differences between Mandarin and Taiwanese, a C2T TTS system should have a text analysis module that can solve problems specific to Taiwanese. For instance, there is only one pronunciation for "我們" (we) in Chinese, but there are two pronunciations for "我們" (we) in Taiwanese.

Figure 1 shows a common structure of a C2T TTS system. In general, a C2T TTS system should contain four basic modules. They are (1) a text analysis module, (2) a tone sandhi module, (3) a prosody generation module, and (4) a speech synthesis module. A C2T TTS system also needs a text analysis module like that of a Mandarin TTS system. This module requires a well-defined bilingual lexicon. We also find that text analysis in a C2T TTS system should have functions not found in a Mandarin TTS system, such as phonetic transcription, digit sequence processing (Liang *et al.*, 2004), and a method for solving the polysemy problem. Solving the polysemy problem is the most complex and difficult of these. There has been little research on solving the polysemy problem. Polysemy means that a word has two or more meanings, which may lead to different pronunciations. For example, the word "他" (he) has two pronunciations in Taiwanese, */yi/* and */yin/*. The first pronunciation */yi/* of "他" (he) means "he," while the second pronunciation /yin/ of "他" (he) means "second-person possessive". The correct pronunciation of a word affects the comprehensibility and fluency of Taiwanese speech.

Many researchers have studied C2T TTS systems (Ho, 2000; Huang, 2001; Hwang, 1996; Lin *et al.*, 1999; Pan & Yu, 2008; Pan, Yu, & Tsai, 2008; Yang, 1999; Zhong, 1999). Nevertheless, none of the researchers considered the polysemy problem in a C2T TTS system. We think that solving the polysemy problem in a C2T TTS system is a fundamental task. The correct meaning of the synthesized words cannot be determined if this problem is not solved properly.

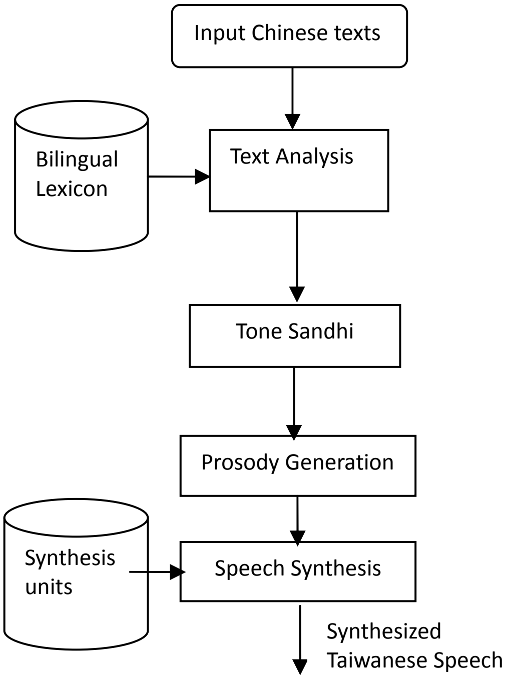


Figure 1. A Common module structure of a C2T TTS System.

The remainder of this paper is organized as follows. In Section 2, we will describe the polysemy problem in Taiwanese. We will give examples to show the importance of solving the polysemy problem in a C2T TTS system. Determining the correct pronunciation of the word “我們” (we) is the focus of the challenge in these cases. Section 3 is the description of the layered approach, which has been shown to work well in solving the polysemy problem (Lin *et al.*, 2008). Lin (2006) has also shown that the layered approach works very well in solving the polyphone problem in Chinese. We will apply the layered approach in determining the pronunciation of “我們” (we) in this section. In Section 4 and Section 5, we use two models to determine the pronunciation of the word “我們” (we) in sentences. The first approach in Section 4 is called the word-based unigram model (WU). The second approach, which will be applied in Section 5, is the word-based long-distance bigram model (WLDB). We also make some new inferences in these two sections. Section 6 shows a combination of the two models discussed in Section 4 and Second 5 for a third approach to solving the polysemy problem. Finally, in Section 7, we summarize our major findings and outline some future works.

2. Polysemy Problems in Taiwanese

Unlike in Chinese, the polysemy problem in Taiwanese appears frequently and is complex. We will give some examples to show the importance of solving the polysemy problem in a C2T TTS system.

The first examples feature the pronouns “你” (you), “我” (I), and “他” (he) in Taiwanese. These three pronouns have two pronunciations, each of which corresponds to a different meaning. Example 2.1 shows the pronunciations of the word “我” (I) and “你” (you) in Taiwanese. The two pronunciations of “我” (I) are */ghua/* with the meaning of “I” or “me” and */ghun/* with the meaning of “my”. The two pronunciations of “你” (you) are /*li*/ with the meaning of “you” and /*lin*/ with the meaning of “your”. If one chooses the wrong pronunciation, the utterance will carry the wrong meaning.

Example 2.1 我/ghua/過一會兒會拿幾本有關台語文化的書到你/*lin*/家給你/*li*/，你/*li*/可以不必到我*/ghun/*家來找我/*ghua*/拿。 (I will bring some books about Taiwanese culture to your house for you later; you need not come to my home to get them from me.)

Example 2.2 shows the two different pronunciations of “他” (he). They are /*yi*/, with the meaning of “he” or “him,” and /*yin*/, with the meaning of “his”.

Example 2.2 我看到他/yi/拿一盆蘭花回他/*yin*/家給他/*yin*/爸爸。 (I saw him bring an orchid back to his home for his father.)

The following examples focus on “不” (no), which has six different pronunciations. They are /*bho*/, /*m*/, /*bhei*/, /*bhuaih*/, /*mai*/, and /*but*/. Examples 2.3 through 2.6 show four of the six pronunciations.

Example 2.3 一般人並不/*bho*/容易看出它的重要性。 (It is not easy for a person to see its importance.)

Example 2.4 不/m/知浪費了多少國家資源。 (We do not know how many national resources were wasted.)

Example 2.5 讓人聯想不/*bhei*/到他與機械的關係。 (One would not come to the proper conclusion regarding the relationship between that person and machines.)

Example 2.6 華航使用之航空站交通已不/*but*/如從前方便。 (The traffic at the airport is not as convenient as it was in the past for China Airlines.)

Examples 2.7 through 2.9 are examples of pronunciations of the word “上” (up). The word “上” (up) has three pronunciations. They are /*ding*/, /*siong*/, and /*jiunn*/. The meaning of the word “上” (up) in Example 2.7 has the sense of “previous”. Example 2.8 shows a case where “上” (up) means “on”. Example 2.9 is an example of the use of “上” (up) to mean, “get on”.

Example 2.7 我上/*ding*/個月花了好多錢去買有關台語的教科書。(Last month, I spent so much money on buying Taiwanese textbooks.)

Example 2.8 我是在這地圖上/*siong*/的哪裡？ (Where am I on this map?)

Example 2.9 我上/*jiunn*/了公車後才發現我搭錯車了。 (After I got on the bus, I realized that I boarded the wrong one.)

Another word we want to discuss is “下” (down). The word “下” (down) has four pronunciations. They are /*ha*/, /*ao*/, /*loh*/, and /*ei*/. Examples 2.10–2.13 are some examples of pronunciations of the word “下” (down). The meaning of “下” (down) in Example 2.10 is “close” or “end”. Example 2.11 shows how the same word can mean “next”. Example 2.12 illustrates the meaning “falling”. Example 2.13 shows another example of it used to mean “next”.

Example 2.10 我今天將在十點下/*ha*/課。 (I will finish my class at ten o’clock today.)

Example 2.11 台中下/*ao*/星期有甚麼音樂會？ (What concerts are scheduled for next week in Taichung?)

Example 2.12 彰化已經開始下/*loh*/大雨了。 (It has begun to rain heavily in Changhua.)

Example 2.13 請問下/*ei*/一列火車何時開出？ (Excuse me. Could you please tell me when the next train will depart?)

We have proposed a layered approach in predicting the pronunciations “上” (up), “下” (down), and “不” (no) (Lin *et al.*, 2008). The layered approach works very well in solving the polysemy problems in a C2T TTS system. A more difficult case of the polysemy problem will be encountered in this paper.

In addition to the above words, another difficult case is “我們” (we). Taiwanese speakers arrive at the correct pronunciation of the word “我們” (we) by deciding whether to include the listener in the pronoun.

Unlike Chinese, “我們” (we) has two pronunciations with different meanings when used in Taiwanese. This word can include (1) both the speaker and listener(s) or (2) just the speaker. These variations lead to two different pronunciations in Taiwanese, */lan/* and */ghun/*. The Chinese characters for */lan/* and */ghun/* are “咱” (we) and “阮” (we), respectively. The following example helps to illustrate the different meanings. More examples to illustrate these differences will be used later in this section.

Assume first that Jeffrey and his younger brother, Jimmy, ask their father to take them to see a movie then go shopping. Jeffrey can say the following to his father:

Example 2.14 爸爸你要記得帶我們一起去看電影，我們看完電影後，再一起去逛街。 (Daddy, remember to take us to see a movie and go shopping with us after we see the movie.)

The pronunciation of the first word “我們” (we) in Example 2.14 is */ghun/* in Taiwanese since the word “我們” (we) does not include the listener, Jeffrey’s father. The second instance of “我們” (we), however, is pronounced /*lan*/ since this instance includes both the speaker and the listener.

The pronunciation of “我們” (we) in Example 2.15 is */ghun/* in Taiwanese since the word “我們” (we) includes Jeffrey and Jimmy but does not include the listener, Jeffrey’s father.

Example 2.15 爸爸，我要和弟弟去看電影，我們看完電影後，會一起去逛街。 (Daddy, I will go to see a movie with my younger brother, and the two of us will go shopping after seeing the movie.)

If a C2T TTS system cannot identify the correct pronunciation of the word “我們” (we), we cannot understand what the synthesized Taiwanese speech means. In a C2T TTS system, it is necessary to decide the correct pronunciation of the Chinese word “我們” (we) in order to have a clear understanding of synthesized Taiwanese speech.

Distinguishing different kinds of meanings of “我們” (we) is a semantic problem. It is a difficult but important issue to be overcome in the text analysis module of a C2T TTS system. As there is only one pronunciation of “我們” (we) in Mandarin, a Mandarin TTS system does not need to identify the meaning of the word “我們” (we).

To compare this work with the research in Hwang *et al.* (2000) and Yu *et al.* (2003), determining the meaning of the word “我們” (we) may be more difficult than solving the non-text symbol problem. A person can determine the relationship between the listeners and the speaker then determine the meaning of the word “我們” (we). It is more difficult, however, for a computer to recognize the relationship between the listeners and speakers in a sentence.

Since determining whether listeners are included is a context-sensitive problem, we need to look at the surrounding words, sentences, or paragraphs to find the answer.

Let us examine the following Chinese sentence (Example 2.16) to help clarify the problem.

Example 2.16 我們必須加緊腳步改善台北市的交通狀況。 (We should press forward to improve the traffic of Taipei City.)

It is difficult to determine the Taiwanese pronunciation of the word “我們” (we) in Example 2.16 from the information in this sentence. To get the correct pronunciation of the word “我們” (we), we need to expand the sentence by adding words to the subject, *i.e.*, look forward, and predicate, *i.e.*, look backward. Assume that, when we add words to the subject and the predicate, we have a sentence that looks like Example 2.17:

Example 2.17 台北市長馬英九在接見美國記者時指出:「我們必須加緊腳步改善台北市的交通狀況。」 (Taipei city mayor Ma Ying-Jeou said that we should press forward to improve the traffic of Taipei city when he received some reporters from the USA.)

As the reporters from the USA have no obligation to improve the traffic of Taipei, we can conclude that “我們” (we) does not include them. Therefore, it is safe to say that the correct pronunciation of the word “我們” (we) in Example 2.17 should be */ghun/*.

On the other hand, if the sentence reads as in Example 2.18 and context is included, the pronunciation of the word “我們” (we) should be */lan/*. We can find some important keywords such as “台北市長” (the Taipei city mayor) and “市府會議” (a meeting of the city government).

Example 2.18 台北市長馬英九在市府會議中指出:「我們必須加緊腳步改善台北市的交通狀況。」 (In a meeting of the city government, the Taipei city mayor, Ma Ying-Jeou, said that we should press forward to improve the traffic of Taipei City.)

When disambiguating the meaning of some non-text symbols, such as “/”, “:”, and “-” the keywords to decide the pronunciation of the special symbols may be within a fixed distance from the given symbol. Nevertheless, the keywords can be at any distance from the word “我們” (we), as per Example 2.19. Some words that could be used to determine the pronunciation of “我們” (we), such as “市府會議” (a meeting of the city government), “台北市長” (the Taipei city mayor), and “馬英九” (Ma Ying-Jeou), are at various distances from “我們” (we).

Example 2.19 在今天的市府會議中，台北市長馬英九提到關於台北市的交通問題時，馬市長說:「我們必須加緊腳步改善台北市的交通狀況。」 (In a meeting of the city government, the Taipei city mayor, Ma Ying-Jeou, talked about the problem of the traffic in Taipei city. Mayor Ma said that we should press forward to improve the traffic of Taipei city.)

These examples illustrate the importance of determining the proper pronunciation for each word in a C2T TTS system. Compared to other cases of polysemy, determining the proper pronunciation of the word “我們” (we) in Taiwanese is a difficult task. We will focus on solving the polysemy problem of the word “我們” (we) in this paper.

3. Using the Layered Approach to Determine the Pronunciation of “我們” (we)

Lin (2006) showed that the layered approach worked very well in solving the polyphone problem in Chinese. Lin (2006) also showed that using the layered approach to solve the polyphone problem is more accurate than using the CART decision tree. We also show that using the layered approach in solving the polysemy problems of other words has worked well in our research (Lin *et al.*, 2008). We will apply the layered approach in solving the polysemy problem of “我們” (we) in Taiwanese.

3.1 Description of Experimental Data

First, we will describe the experimental data used in this paper. The experimental data is comprised of over forty thousand news items from eight news categories, in which 1,546 articles contain the word “我們” (we). The data was downloaded from the Internet from August 23, 2003 to October 21, 2004. The distribution of these articles is shown in Table 1. We determined the pronunciation of each “我們” (we) manually.

Table 1. Distribution of experimental data

News Category	Number of News Items	Number of News Items Containing the word "我們"	Percentage
International News	2242	326	14.5%
Travel News	9273	181	1.9%
Local News	6066	95	1.5%
Entertainment News	3231	408	12.6%
Scientific News	3520	100	2.8%
Social News	4936	160	3.2%
Sports News	2811	193	6.9%
Stock News	8066	83	1.0%
Total Number of News Items	40145	1546	3.9%

As shown in Table 2, in the 1,546 news articles, “我們” occurred 3,195 times. In our experiment, 2,556 samples were randomly chosen for the training data while the other 639 samples were added to the test data. In the training data, there were 1,916 instances with the pronunciation of */ghun/* for the Chinese character “阮” and 640 instances with the pronunciation of */lan/* for the Chinese character “咱”.

Table 2. Distribution of training and testing data.

Frequency of “我們”	Pronunciation */lan/*	Pronunciation */ghun/*	Total Frequency
Training data	640	1,916	2,556
Test data	160	479	639
Token frequency of “我們”	800	2,395	3,195

3.2 Description of Layered Approach

Figure 2 shows the layered approach to the polysemy problem with an input test sentence. We use Example 3.1 to illustrate how the layered approach works.

Example 3.1 爸爸 告訴 我們 過 馬路 要 小心。 (Dad told us to be careful when crossing the street.)

Example 3.1 is an utterance in Chinese with segmentation information. Spaces were used to separate the words in Example 3.1. We want to predict the correct pronunciation for the word “我們” (we) in Example 3.1.

As depicted in Figure 2, there are four layers in our approach. We set $(w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2})$ as (爸爸,告訴,我們,過,馬路). This pattern (爸爸,告訴,我們,過,馬路) will be the input for Layer 4. Nevertheless, as this pattern is not found in the training data, we cannot decide the pronunciation of “我們” (we) with this pattern. We then use two patterns $(w_{-2}, w_{-1}, w_0, w_{+1})$ and $(w_{-1}, w_0, w_{+1}, w_{+2})$ to derive (爸爸,告訴,我們,過) and (告訴,我們,過,馬路), respectively, as the inputs for Layer 3. Since we cannot find any patterns in the training data that match either of these patterns, the pronunciation cannot be decided in this layer.

Three patterns are used in Layer 2. They are (爸爸,告訴,我們), (告訴,我們,過), and (我們,過,馬路). We find that the pattern (爸爸,告訴,我們) has appeared in training data. The frequencies are 2 for pronunciation */ghun/* and 1 for */lan/*. Thus, the probabilities for the possible pronunciations of “我們” (we) in Example 3.1 are 2/3 for */ghun/* and 1/3 for */lan/*. We can conclude that the predicted pronunciation is */ghun/*. The layered approach terminates in Layer 2 in this example. If the process did not terminate prematurely, as in this example, it would have terminated in Layer 1, as shown by the dashed lines in Figure 2.

3.3 Results of Using the Layered Approach

We used the experimental data mentioned in 3.1. There are 3,159 samples in the corpus. We used 2,556 samples to train the four layers. The other 639 samples form the test data. Table 3 shows the accuracy of using the layered approach based on word patterns. Thus, the features in the layered approach are words. The results show that the layered approach does not work well. The overall accuracy is 77.00%.

Table 3. Results of using the layered approach with word pattern.

	Number of test samples	Number of correct samples	Accuracy rate
/ghun/	479	445	92.90%
/lan/	160	47	29.38%
Total	639	492	77.00%

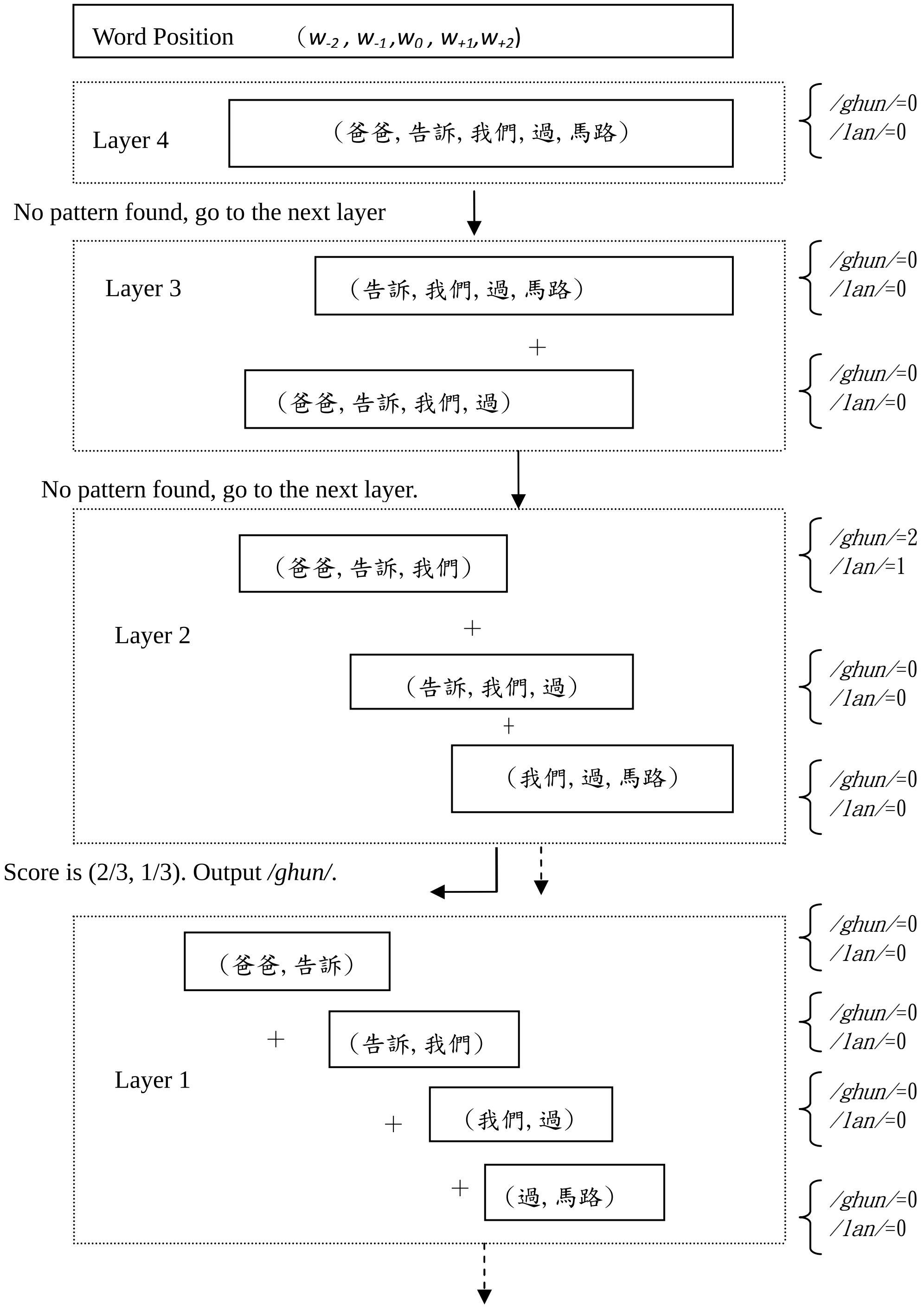


Figure 2. An example applying the layered approach.

4. Word-based Unigram Language Model

In this section, we propose a word-based unigram language model (WU). Two statistical results are needed in this model. Statistical results were compiled for (1) the frequency of appearance for words that appear to the left of “我們” (we) in the training data and (2) the frequencies for words that appear to the right. Each punctuation mark was treated as a word. Each testing sample looks like the following:

$$w_{-M}\ w_{-(M-1)} \ldots w_{-2}\ w_{-1}\ \boxed{\text{我們}}\ w_{+1}\ w_{+2} \ldots w_{+(N-1)}\ w_{+N}$$

where w_{-i} is the i^{th} word to the left of “我們” (we) and w_i is the i^{th} word to the right. The following formulae were used to find four different scores for each testing sample: $S_{uL}(/lan/)$, $S_{uR}(/lan/)$, $S_{uL}(/ghun/)$, and $S_{uR}(/ghun/)$.

$$S_{uL}(/lan/) = \sum_{j=1}^{M} \frac{\dfrac{C(/lan/\&w_{-j})}{T_{uL}(/lan/)}}{\dfrac{C(/lan/\ \&w_{-j})}{T_{uL}(/lan/)} + \dfrac{C(/ghun/\&w_{-j})}{T_{uL}(/ghun/)}} \quad (1)$$

$$S_{uR}(/lan/) = \sum_{j=1}^{N} \frac{\dfrac{C(/lan/\&w_{+j})}{T_{uR}(/lan/)}}{\dfrac{C(/lan/\&w_{+j})}{T_{uR}(/lan/)} + \dfrac{C(/ghun/\&w_{+j})}{T_{uR}(/ghun/)}} \quad (2)$$

$$S_{uL}(/ghun/) = \sum_{j=1}^{M} \frac{\dfrac{C(/ghun/\&w_{-j})}{T_{uL}(/ghun/)}}{\dfrac{C(/lan/\ \&w_{-j})}{T_{uL}(/lan/)} + \dfrac{C(/ghun/\ \&w_{-j})}{T_{uL}(/ghun/)}} \quad (3)$$

$$S_{uR}(/ghun/) = \sum_{j=1}^{N} \frac{\dfrac{C(/ghun/\ \&w_{+j})}{T_{uR}(/ghun/)}}{\dfrac{C(/lan/\&w_{+j})}{T_{uR}(/lan/)} + \dfrac{C(/ghun/\&w_{+j})}{T_{uR}(/ghun/)}} \quad (4)$$

where

$$T_{uL}(/lan/) = \sum_{l=1}^{uL} C(/lan/\&w_{-l}) \quad (5)$$

$$T_{uL}(/ghun/) = \sum_{p=1}^{uL} C(/ghun/\&w_{-p}) \quad (6)$$

$$T_{uR}(/lan/) = \sum_{l=1}^{uR} C(/lan/\&w_{+l}) \quad (7)$$

$$T_{uR}(/ghun/) = \sum_{p=1}^{uR} C(/ghun/\&w_{+p}) \quad (8)$$

uL different kinds of words appear on the left side of “我們” (we) in the training corpus. $T_{uL}(/lan/)$ is the total frequency of these *uL* words in the training data where the pronunciation of “我們” (we) is */lan/*. Similarly, $T_{uL}(/ghun/)$ represents the total frequency of *uL* words where “我們” (we) is pronounced */ghun/*. *uR* is the number of different words that appear to the right side of “我們” (we) in the training corpus. $T_{uR}(/lan/)$ and $T_{uR}(/ghun/)$ are the total frequencies of these *uR* words in the training data where pronunciation of “我們” (we) is */lan/ and /ghun/*, respectively. $C(/ghun/\&w_p)$ is the frequency that the word w_p appears in the training corpus where the pronunciation of “我們” (we) is */ghun /*. $\frac{C(/lan/\&w_{-j})}{T_{uL}(/lan/)}$ in (1) means the significance of pronunciation */lan/* of word w_{-j} in training data.

Formulae (1) through (4) were applied to each test sample to produce four scores. The scores were $S_{uL}(/lan/)$ for the words to the left of “我們” (we) when the pronunciation was */lan/*, $S_{uR}(/lan/)$ for the words to the right when the pronunciation was */lan/*, $S_{uL}(/ghun/)$ for the words to the left of “我們” (we) when the pronunciation was */ghun/*, and $S_{uR}(/ghun/)$ for the words to the right when the pronunciation was */ghun/*. The pronunciation of “我們” (we) is */lan/* if $S_{uL}(/lan/)+ S_{uR}(/lan/) > S_{uL}(/ghun/) + S_{uR}(/ghun/)$. The result is */ghun/* otherwise.

The experiments were inside and outside tests. First, we applied WU with the training data mentioned in Section 3.1 to find the best ranges in determining the pronunciation of “我們” (we). We defined a window as (*M, N*), where *M* was number of words to the left of “我們” (we) and *N* was the number of words to the right. Three hundred and ninety nine (20*20-1=399) different windows were applied when using the WU model. As shown in Table 4, the best result from an inside test was 87.00%, with a window of (17, 10).

The best result when the correct pronunciation of “我們” (we) was */ghun/* was 94.01%, achieved when the window was (12, 6). Nevertheless, the results when the pronunciation was */lan/* and the window was the same were not good. The highest accuracy achieved was 45.48%. Also, as shown in 4th row of Table 4, the best result when applying WU when the pronunciation was */lan/* was just 77.88%, when the window was (19, 14). This shows that WU did not work well when the pronunciation of “我們” (we) was */lan/*.

Table 4. The results of the inside test of applying WU.

Window Size (*M*, *N*)	Accuracy when the pronunciation is */ghun/*	Accuracy when the pronunciation is */lan/*	Overall accuracy
(17, 10)	91.04%	74.92%	87.00%
(12,6)	94.01%	45.48%	81.85%
(19, 14)	88.75%	77.88%	86.03%

We applied WU with a window of (17, 10) for testing data. The overall accuracy of the outside tests was 75.59%. The accuracies were 90.40% and 31.25% when the pronunciations were */ghun/* and */lan/*, respectively.

5. Word-based Long Distance Bigram Language Model

We will bring up the word-based long-distance bigram language model (WLDB) in this section. According to Section 2 of this paper, there are two different meanings for “我們” (we). The two meanings are different in that one includes the listener(s) and the other does not. We propose a modification of the WU model by having two words appear together in the text to clarify the relationship between the speaker and listener(s). Examples of this modification are “台北市長” (the Taipei city mayor) and “美國記者” (the reporter(s) from the USA) in Example 2.17 and “台北市長” and “市府會議” (a city government meeting) in Examples 2.18 and 2.19.

For each testing sample,

$$w_{-M}\ w_{-(M-1)} \ldots w_{-2}\ w_{-1}\ \boxed{\text{我們}}\ w_{+1}\ w_{+2} \ldots w_{+(N-1)}\ w_{+N}\ .$$

The following formulae were used to find four scores for each testing sample, $S_{bL}(/lan/)$, $S_{bR}(/lan/)$, $S_{bL}(/ghun/)$, and $S_{bR}(/ghun/)$.

$$S_{bL}(/lan/) = \sum_{i=1}^{M}\sum_{j=i}^{M} \frac{\dfrac{C(/lan/\,\&w_{-i}\,\&\,w_{-j})}{T_{bL}(/lan/)}}{\dfrac{C(/lan/\,\&w_{-i}\,\&\,w_{-j})}{T_{bL}C(/lan/)} + \dfrac{C(/ghun/\,\&w_{-i}\,\&\,w_{-j})}{T_{bL}C(/ghun/)}} \tag{9}$$

$$S_{bR}(/lan/) = \sum_{i=1}^{N}\sum_{j=i}^{N} \frac{\dfrac{C(/lan/\,\&w_{i}\,\&w_{j})}{T_{bR}(/lan/)}}{\dfrac{C(/lan/\,\&w_{+i}\,\&w_{+j})}{T_{bR}(/lan/)} + \dfrac{C(/ghun/\,\&w_{+i}\,\&w_{+j})}{T_{bR}(/ghun/)}} \tag{10}$$

$$S_{bL}(/ghun/) = \sum_{i=1}^{M}\sum_{j=i}^{M} \frac{\dfrac{C(/ghun/\,\&w_{-i}\,\&w_{-j})}{T_{bL}(/ghun/)}}{\dfrac{C(/ghun/\,\&w_{-i}\,\&w_{-j})}{T_{bL}(/ghun/)} + \dfrac{C(/lan/\,\&w_{-i}\,\&w_{-j})}{T_{bL}(/lan/)}} \tag{11}$$

$$S_{bR}(/ghun/) = \sum_{i=1}^{N}\sum_{j=i}^{N} \frac{\dfrac{C(/ghun/\,\&w_{i}\,\&w_{j})}{T_{bR}(/ghun/)}}{\dfrac{C(/ghun/\,\&w_{+i}\,\&w_{+j})}{T_{bR}(/ghun/)} + \dfrac{C(/lan/\,\&w_{+i}\,\&w_{+j})}{T_{bR}(/lan/)}} \tag{12}$$

where

$$T_{bL}(/lan/) = \sum_{l=1}^{bL}\sum_{k=l}^{bL} C(/lan/\,\&w_{-l}\,\&w_{-k}) \tag{13}$$

$$T_{bR}(/lan/) = \sum_{l=1}^{bR}\sum_{k=l}^{bR} C(/lan/\,\&w_{+l}\,\&w_{+k}) \tag{14}$$

$$T_{bL}(/ghun/) = \sum_{l=1}^{bL}\sum_{k=l}^{bL} C(/ghun/\,\&w_{-l}\,\&w_{-k}) \tag{15}$$

$$T_{bR}(/ghun/) = \sum_{l=1}^{bR} \sum_{k=l}^{bR} C(/ghun/\&w_l \& w_k) \tag{16}$$

We assume that *bL* different words appear to the left of “我們” (we) in the training corpus and *bR* different words appear to the right. Formulae 9, 10, 11, and 12 were applied to each test sample, and they produced four scores. $C(/lan/\&w_i\&w_j)$ in (9) is the frequency at which words w_i and w_j appear in the training corpus when the pronunciation of “我們” (we) is */lan/*. $S_{bL}(/lan/)$ is the score for the words to the left of “我們” (we) when the pronunciation is */lan/*, and $S_{bR}(/lan/)$ is the score for the words to the right. Similarly, $S_{bL}(/ghun/)$ and $S_{bR}(/ghun/)$ represent the scores for the words to the left and right, respectively, when “我們” (we) is pronounced */ghun/*. In summary, the pronunciation of the word “我們” (we) is */lan/* if $S_{bL}(/lan/) + S_{bR}(/lan/) > S_{bL}(/ghun/) + S_{bR}(/ghun/)$. The pronunciation is */ghun/* otherwise.

We applied WLDB with the training data mentioned in Section 3.1 to find the best ranges in determining the pronunciation of “我們” (we). We defined a window of (*M*, *N*), where *M* was the number of words to the left and *N* was number of words to the right. Three hundred and sixty (19*19-1=360) different windows were applied in the analysis of using the WLDB model. As shown in the 2nd row of Table 5, the best result of the inside test was 94.25% with the best range being 11 words to the left of “我們” (we) and 7 words to the right.

The best result when the correct pronunciation of “我們” (we) was */lan/* was 99.87%, when the window was (11, 5). Nevertheless, the result for */ghun/* with the same window was not good. The highest accuracy achieved was 89.69%. As shown in the 3rd row of Table 5, the best result when applying WLDB when the pronunciation was */ghun/* was 93.48%, when the window was (4, 13). This shows that WLDB does not work well when the pronunciation of “我們” (we) is */ghun/*.

Table 5. The results of the inside test of applying WLDB.

Window Size (k_L, k_R)	Accuracy when the pronunciation is */ghun/*	Accuracy when the pronunciation is */lan/*	Overall accuracy
(11,7)	93.33%	97.04%	94.25%
(4, 13)	93.48%	93.61%	93.52%
(11,5)	89.69%	99.87%	92.15%

We applied the WLDB model to the test data using a window of (11, 7). The overall accuracy of outside tests was 85.72%. The accuracies were 83.26% and 93.10% when the pronunciations were */ghun/* and */lan/*, respectively.

6. The combined Approach

Based on the results from the two models, WU and WLDB, we can draw the following conclusions: the word-based long distance bigram language model is good when the pronunciation is */lan/*, while the word-based unigram language model works well when the pronunciation is */ghun/*. In this section, we propose combining the models to achieve better results.

According to the inside experimental results shown in Table 4 and Table 5, we will combine the WU model with a window of (12, 6) and the WLDB model with a window of (11, 5) as our combined approach. This combination of WU and WLDB is similar to the approach used by Yu and Huang. We will try to find the possibility of making a correct choice when using WU or WLDB, which will be termed “confidence”. We will adopt the output of the method with higher confidence.

6.1 Confidence Measure

The first step in this process is to find a confidence curve for each model. The goal is to estimate the confidence for each approach and assess the difference. The higher score is more likely to be the correct answer. To do so, we measure the accuracy of each division and use a regression to estimate the confidence measure.

Algorithm 1, below, will be used to find the confidence curve for the word-based unigram language model. As the total number of words in each input sample is not constant, we must first normalize the scores $Su_i(/lan/)$ and $Su_i(/ghun/)$. We will find the precision rates (PR_k) in the interval [0, 1] for $|NSu_i(/ghun/)\text{-}\ NSu_i(/lan/)|$ in Step 2 of Algorithm 1 for each *i*. We then find a regression curve for the PR_k. The regression curve is used to estimate the probability of making a correct decision when using WU. Therefore, it follows that, the higher the probability is, the greater the confidence we can have in the results from WU.

Algorithm 1: Finding the confidence curve of WU.

Input: The score for each training sample, $Su_i(/lan/)$ and $Su_i(/ghun/)$, where i=1,2,3, …, n and n is the number of training samples.

Output: A function for the confidence curve for the given $Su_i(/lan/)$ and $Su_i(/ghun/)$, i=1,2,3, …, n.

Algorithm:

Step 1: Normalize $Su_i(/lan/)$ and $Su_i(/ghun/)$ for each training sample i using the following formula:

$$NSu_i(/lan/)=Su_i(/lan/)/(Total\ number\ of\ words\ in\ training\ sample\ i)$$

$$NSu_i(/ghun/)=Su_i(/ghun/)/(Total\ number\ of\ words\ in\ training\ sample\ i)$$

Step 2: Let $d_i=|\ NSu_i(/ghun/)\text{-}\ NSu_i(/lan/)|$ and let $D=\{d_1, d_2,\ldots,d_n\}$. Find the accuracy rate for each interval using the following formula:

$$PR_k= C_k/N_k,\ k=1,\ 2,\ \ldots,\ 18$$

Here, C_k is the number of correct conjectures of training sample i with $(k\text{-}1)/18 \leqq d_i < (k+1)/18$, and N_k is the number of training sample i with $(k\text{-}1)/18 \leqq d_i < (k+1)/18$.

Step 3: Find a regression curve for PR_1, PR_2, …, PR_{18}. Output the function of the regression curve.

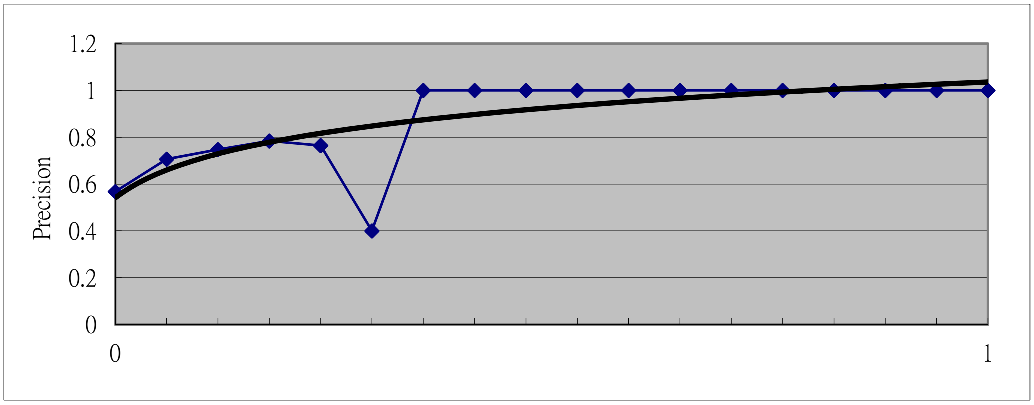


Figure 3. Estimate the confidence curve using WU. The function we attained is f(x)=0.1711*ln(x)+1.0357.

The confidence curve for WU is the black line in Figure 3. The function derived was *f(x)=0.1711*ln(x)+1.0357*, where *x* is the absolute value of the difference between the normalized $Su_i(/lan/)$ and $Su_i(/ghun/)$.

Algorithm 2 is used to find the confidence curve for the word-based long-distance bigram language model (WLDB). We began by normalizing the scores of pronunciation $Sb_i(/lan/)$ and $Sb_i(/ghun/)$. In Step 2, we find the precision rates (PR_k) in the interval [0, 1] then calculate a regression curve for the PR_k. The regression curve will be used to estimate the probability of making a correct decision. Again, it follows that, the higher the probability, the more confidence in the results from using WLDB.

The confidence curve of WLDB is the black line in Figure 4, in which the function is *f(x) = 0.2346*ln(x) + 1.0523*, where *x* is the difference between the normalized $Sp_i(/lan/)$ and $Sp_i(/ghun/)$.

```
Algorithm 2: Find the confidence curve of WLDB
    Input:  The score of each training sample, named $Sb_i(/lan/)$ and $Sb_i(/ghun/)$, where $i$=1, 2, 3, …, $n$, and $n$ is the number of training samples.
    Output: A function for the confidence curve for the given $Sb_i(/lan/)$ and $Sb_i(/ghun/)$, $i$=1, 2, 3, …, $n$.
Algorithm:
    Step 1: Normalize $Sb_i(/lan/)$ and $Sb_i(/ghun/)$ for each training sample $i$ using the following formula:
        $NSb_i(/lan/)=Sb_i(/lan/)/(Total\ number\ of\ words\ in\ training\ sample\ i)^2$
        $NSb_i(/ghun/)=Sb_i(/ghun/)/(Total\ number\ of\ words\ in\ training\ sample\ i)^2$
    Step 2: Let $d_i$=| $NSb_i(/ghun/)$- $NSb_i(/lan/)$| and let $D=\{d_1, d_2,\ldots,d_n\}$. Find the accuracy rate for each interval using the following formula:
        $PR_k= C_k/N_k$, k=1, 2, …, 13
        where $C_k$ is the number of correct conjectures of training samples $i$ with $(k-1)/13 \leqq d_i<(k+1)/13$ and $N_k$ is the number of training samples $i$ with $(k-1)/13\leqq d_i<(k+1)/13$.
```

Step 3: Find a regression curve for PR_1, PR_2, …, PR_{13}. Output the function of the regression curve.

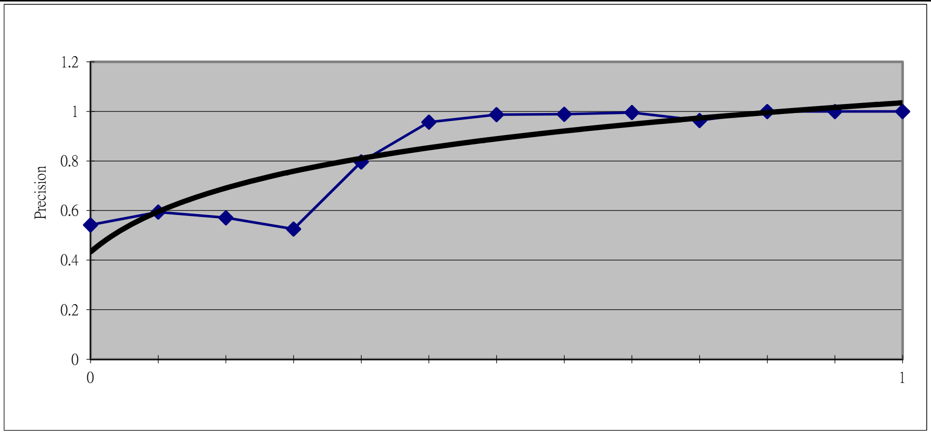


Figure 4. Estimate the confidence curve of WLDB. The function we attained is $f(x)=0.2346*\ln(x)+1.0523$.

6.2 Determining the Pronunciation for “我們” (we)

After the functions for the confidence curves for the two models have been derived, the combined approach can be applied. The two models are used to determine the pronunciation of “我們” (we) for a given input text. The two functions for the confidence curves, derived in Section 6.1, are applied to evaluate the degree of confidence in the two models. Let the confidence curves of the two models be C_{WU} for WU and C_{WLDB} for WLDB. We will use the results obtained using WU under the condition $C_{WU} > C_{WLDB}$. Otherwise, we will use the results obtained from using the WLDB model.

Consider Figure 4, which is derived from the training data. The *x*-axis is the normalized difference between the two scores. The *y*-axis is the percentage of correct decisions. Take the example sentence “如果花旗希望繼續做我們的大股東，我們還是很歡迎”. We want to predict the pronunciation of the first “我們” (we) in the above sentence. Its confidences were 0.875 for the WU model (choosing */ghun/*) and 0.761 for the WLDB model (choosing */lan/*). Since the confidence of the WU model was higher than that of the WLDB model, we adopted */ghun/* as the pronunciation.

6.3 Experimental Results Using Combined Models

We used the 639 testing samples described in Section 3.1. Among the 639 testing samples, there were 479 samples with the pronunciation */ghun/* and 160 samples with the pronunciation */lan/*.

We used the test data mentioned in 3.1 as the experimental data. The overall accuracy rate from applying the combined approach was 93.6%. The accuracy rate was 95.00% when the answer was */lan/*, and the accuracy rate was 93.1% when the answer was */ghun/*. Based on these results, it can be concluded that the combination of the two models works very well in determining the pronunciation of the word “我們” (we) for a given Chinese text.

The three approaches, WU, WLDB, and combined, are compared in Table 6. As shown in Table 6, the word-based long-distance bigram language model (WLDB) worked well in the case of */lan/* and achieved an accuracy rate of 93.10%. The word-based unigram language (WU) model worked well in the case of */ghun/* and achieved an accuracy rate of 90.40%. The combined approach, however, achieved higher accuracy rates in both cases, achieving accuracy as high as 93.6%.

Table 6. Comparison - WU, WLDB, and Combined approach.

	Accuracy using WU	Accuracy using WLDB	Accuracy combing the two models
/ghun/	90.40%	83.26%	93.10%
/lan/	31.25%	93.10%	95.00%
Total	75.59%	85.72%	93.60%

There is an important issue in the combined approach. When we use a language model like WLDB, we may encounter the problem of data scarcity. If data is scarce, the combined approach will use the result of the word-based unigram language model.

6.4 Discussion

Table 7 compares the accuracy of the approaches used in this paper. The findings show that the combined approach (CP) performed the best. We can conclude that layered approach does not work well in determining the pronunciation of “我們” (we) in Taiwanese. It also shows that the polysemy problem caused by “我們” (we) is more difficult and quite different from that caused by the words “上” (up), “下” (down), and “不” (no). This also shows that the viewpoints we gave in Section 2 are reasonable.

Table 7. A comparison of the proposed methods and the layered approach. CP refers to the combined approach, while LP refers to the layered approach. The combined approach achieved the highest accuracy.

	WU	WLDB	LP	CP
/ghun/	90.40%	83.26%	92.90%	93.10%
/lan/	31.25%	93.10%	29.38%	95.00%
Total	75.59%	85.72%	77.00%	93.60%

For our approaches, we might encounter the problem of data sparseness, especially with WLDB. It seems that this cannot be avoided in processing languages like Taiwanese, for which corpora are rare. We have tried to use part-of-speech information as the features in our approaches. The experimental results are not good. We also find that most cases can be solved by using WU or WLDB, and only about 5% are solved by using default values. This shows that our approach is suitable for the current data size. We have shown that our combined approach is promising.

7. Conclusion and Future Works

This paper proposes an elegant approach to determine the pronunciation of “我們” (we) in a C2T TTS system. Our methods work very well in determining the pronunciations of the Chinese word “我們” (we) in a C2T TTS system. Experimental results also show that the model used is better than the layered approach, the WU model, and the WLDB model. Polysemy problems in translating C2T are very common and it is imperative that they are solved in a C2T TTS system. We will continue to focus on other important polysemy problems in a C2T TTS system in the future.

The polysemy problem of “我們” (we) is more difficult than that of other words in Taiwanese. We have proposed a combined approach for this problem. If more training data can be prepared, the proposed approach can be expected to achieve better results. Nevertheless, as the training data needs to be processed manually, we will attempt to propose unsupervised approaches in the future.

To build a quality C2T TTS system is a long-term project because of the many issues in the text analysis phase. In contrast to a Mandarin TTS system, a C2T TTS system needs more textual analysis functions. In addition, two imperative tasks are the development of solutions for the polysemy problem and the tone sandhi problem.

Reference

Bao, H., Wang, A., & Lu, S. (2002). A Study of Evaluation Method for Synthetic Mandarin Speech, in *Proceedings of ISCSLP 2002, The Third International Symposium on Chinese Spoken Language Processing*, 383-386.

Chen, S. H., Hwang, S. H., & Wang, Y. R. (1996). A Mandarin Text-to-Speech System, *Computational Linguistics and Chinese Language Processing*, 1(1), 87-100.

Ho, C. C. (2000). *A Hybrid Statistical/RNN Approach to Prosody Synthesis for Taiwanese TTS*, Master thesis, Department of Communication Engineering, National Chiao Tung University.

Hunag, J. Y. (2001). *Implementation of Tone Sandhi Rules and Tagger for Taiwanese TTS*, Master thesis, Department of Communication Engineering, National Chiao Tung University.

Hwang, C. H. (1996). *Text to Pronunciation Conversion in Taiwanese*, Master thesis, Institute of Statistics, National Tsing Hua University.

Hwang, F. L., Yu, M. S., & Wu, M. J. (2000). The Improving Techniques for Disambiguating Non-Alphabet Sense Categories, in P*roceedings of ROCLING XIII*, 67-86.

Liang, M. S., Yang, R. C., Chiang, Y. C., Lyu, D. C., & Lyu, R. Y. (2004). A Taiwanese Text-to-Speech System with Application to Language Learning, in *Proceedings of the IEEE International Conference on Advanced Learning Technologies, 2004.*

Lin, C. J. & Chen, H. H. (1999). A Mandarin to Taiwanese Min Nan Machine Translation System with Speech Synthesis of Taiwanese Min Nan, *International Journal of Computational Linguistics and Chinese Language Processing*, 14(1), 59-84.

Lin, Y. C. (2006). *The Prediction of Pronunciation of Polyphonic Characters in a Mandarin Text-to-Speech System*, Master thesis, Department of Computer Science and Engineering, National Chung Hsing University.

Lin, Y. J. & Yu, M. S. (1998). An Efficient Mandarin Text-to-Speech System on Time Domain, *IEICE Transactions on Information and Systems*, E81-D(6), June 1998, 545-555.

Lin, Y. J., Yu, M. S., Lin, C. Y., & Lin, Y. T. (2008). A Multi-Layered Approach to the Polysemy Problems in a Chinese to Taiwanese TTS System, in *Proceeding of 2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, June, 2008, 428-435.

Lu, H. M. (2002). *An Implementation and Analysis of Mandarin Speech Synthesis Technologies*, M. S. Thesis, Institute of Communication Engineering, National Chiao-Tung University, June 2002.

Pan, N. H. & Yu, M. S. (2008). Improving Intonation Modules in Chinese TTS Systems, in *The 13th Conference on Artificial Intelligence and Applications (TAAI 2008)*, 329-336, Nov. 21-22, 2008, Yilan, Taiwan.

Pan, N. H., Yu, M. S., & Tsai, C. M. (2008). A Mandarin Text to Taiwanese Speech System, in *The 13th Conference on Artificial Intelligence and Applications (TAAI 2008)*, 1-5, Nov. 21-22, 2008, Yilan, Taiwan.

Shih, C. & Sproat, R. (1996). Issues in Text-to-Speech Conversion for Mandarin, *Computational Linguistics and Chinese Language Processing*, 1(1), 37-86.

Wu, C. H., Hsia, C. C., Chen, J. F., & Wang, J. F. (2007). Variable-Length Unit Selection in TTS Using Structural Syntactic Cost, *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4), 1227-1235.

Yang, Y. C. (1999). *An Implementation of Taiwanese Text-to-Speech System*, Master thesis, Department of Communication Engineering, National Chiao Tung University, 1999.

Yu, M. S., Chang, T. Y., Hsu, C. H., & Tsai, Y. H. (2005). A Mandarin Text-to-Speech System Using Prosodic Hierarchy and a Large Number of Words, in *Proc. 17th Conference on Computational Linguistics and Speech Processing, (ROCLING XVII)*, 183-202, Sep. 15-16, 2005, Tainan, Taiwan.

Yu, M. S. & Huang, F. L. (2003). Disambiguating the Senses of Non-Text Symbols for Mandarin TTS Systems with a Three-Layer Classifier, *Speech Communication*, 39(3-4), 191-229.

Zhong, X. R. (1999). *An Improvement on the Implementation of Taiwanese TTS System*, Master thesis, Department of Communication Engineering, National Chiao Tung University.

Computational Linguistics and Chinese Language Processing
Vol. 17, No. 1, March 2012, pp. 65-86

A Comparative Study of Methods for Topic Modeling in Spoken Document Retrieval

Shih-Hsiang Lin[+,*] **and Berlin Chen**[#,*]

Abstract

Topic modeling for information retrieval (IR) has attracted significant attention and demonstrated good performance in a wide variety of tasks over the years. In this paper, we first present a comprehensive comparison of various topic modeling approaches, including the so-called document topic models (DTM) and word topic models (WTM), for Chinese spoken document retrieval (SDR). Moreover, different granularities of index features, including words, subword units, and their combinations, are also exploited to work in conjunction with various extensions of topic modeling presented in this paper, so as to alleviate SDR performance degradation caused by speech recognition errors. All of the experiments were performed on the TDT Chinese collection.

Keywords: Information Retrieval, Document Topic Models, Word Topic Models, Spoken Document Retrieval.

1. Introduction

Due to the advances in computer technology and the proliferation of Internet activity, huge volumes of multimedia data, such as text files, broadcast radio and television programs, lectures, and digital archives, are continuously growing and filling networks. Development of intelligent and efficient information retrieval techniques to provide people with easy access to all kinds of information is now becoming more and more emphasized. Meanwhile, with the rapid evolution of speech recognition technology, substantial efforts and very encouraging results on spoken document retrieval (SDR) also have been demonstrated in the recent past. Although most retrieval systems participating in the TREC-SDR evaluations claimed that speech recognition errors do not seem to cause much adverse effect on SDR performance

[+] Voice Division Research Center, Delta Electronics
E-mail: shlin@csie.ntnu.edu.tw

[#] Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: berlin@csie.ntnu.edu.tw

[*] Corresponding authors.

when merely using imperfect recognition transcripts derived from one-best recognition results from a speech recognizer (Garofolo *et al.*, 2000; Chelba *et al.*, 2008), this is probably attributed to the fact that the TREC-style test queries tend to be quite long and contain different words describing similar concepts that can help the queries match their relevant spoken documents. Furthermore, a query word (or phrase) may occur repeatedly (more than once) within a relevant spoken document, and it is not always the case that all of the occurrences of the word would be misrecognized totally as other words. We, however, believe that SDR would still present a challenge in situations where the queries are relatively short and there exists severe deviation in word usage between the queries and spoken documents.

Among several promising information retrieval approaches, statistical language modeling (LM) (Ponte & Croft, 1998), aiming to capture the regularity in human natural language and quantify the acceptability of a given word sequence, has continuously been a focus of active research in the last decade (Miller *et al.*, 1999; Hofmann, 2001). The basic idea is that each individual document in the collection is treated as a probabilistic language model for generating a given query. A document is deemed to be relevant to a query if its corresponding document language model generates the query with higher likelihood. In practice, the relevance measure for the LM approach is usually computed by two different matching strategies, namely, *literal term matching* and *concept matching* (Lee & Chen, 2005). The unigram language model (ULM) is perhaps the most representative example for literal term matching strategy (Miller *et al.*, 1999). In the ULM approach, each document is interpreted as a generative model composed of a mixture of unigram (multinomial) distributions for observing a query, while the query is regarded as observations, expressed as a sequence of indexing words (or terms).

Nevertheless, these approaches would suffer from the problems of word usage diversity, which might make the retrieval performance of the system degrade severely as a given query and its relevant documents are using quite a different set of words. In contrast, the concept matching strategy tries to explore the topic information conveyed in the query and documents. Based on this, the retrieval process is performed. The probabilistic latent semantic analysis (PLSA) (Hofmann, 2001) and the latent Dirichlet allocation (LDA) (Blei *et al.*, 2003) are often considered to be two basic representatives of this category. They both introduce a set of latent topic variables to describe the "*word-document*" co-occurrence characteristics. More specifically, the relevance between a query and a document is not computed directly based on the frequency of the query words occurring in the document, but instead based on the frequency of these words appearing in the latent topics as well as the likelihood that the document generates those respective topics, which exhibits some sort of concept matching. Further, although there have been many follow-up studies and extensions of PLSA and LDA, it has been shown that more sophisticated (or complicated) topic models, such as the pachinko allocation model (PAM) and correlated topic model (CTM), do not necessarily offer further retrieval benefits (Zhai, 2008; Blei & Lafferty, 2009). On the other hand, rather than treating each document as a whole as a document topic model (DTM), such as PLSA and LDA, the word topic model (WTM) (Chen, 2009) attempts to discover the long-span co-occurrence dependence "*between words*" through a set of latent topics, while each document in the collection consequently can be represented as a composite WTM model in an efficient way for predicting an observed query. Interested readers can refer to Griffiths *et al.* (2007), Zhai (2008), and Blei and Lafferty (2009) for a thorough and updated overview of the major topic-based language models that have been successfully developed and applied to various IR tasks.

Although most of the above approaches can be equally applied to both text and spoken documents, the latter presents unique difficulties, such as speech recognition errors, problems posed by spontaneous speech, and redundant information. A straightforward remedy, apart from the conventional approaches target at improving recognition accuracy, is to develop more robust representations of spoken documents for spoken document retrieval (SDR). For example, multiple recognition hypotheses, beyond the top scoring ones, are expected to provide alternative representations for the confusing portions of the spoken documents (Chelba *et al.*, 2008; Chia *et al.*, 2008). Another school of thought attempts to leverage subword units, as well as the combination of words and subword units, for representing the spoken documents, which also has been shown beneficial for SDR. The reason for the fusion of word- and subword-level information is that incorrectly recognized spoken words often include several subword units that are correctly recognized. Hence, the retrieval process based on subword-level representations may take advantage of partial matching (Lin & Chen, 2009).

With the above inspiration in mind, we first compare the structural characteristics of various topic models for Chinese SDR, including PLSA and LDA, as well as WTM. The utility of these models is thoroughly examined using both long and short test queries. Moreover, different granularities of index features, including words, subword units, and their combinations, are also exploited to work in conjunction with various extensions of topic modeling presented in this paper, so as to alleviate SDR performance degradation caused by imperfect recognition transcripts. To our knowledge, there is little literature on leveraging various topic decompositions together with various granularities of index features for topic modeling in SDR.

The rest of this paper is structured as follows. Section 2 elucidates the structural characteristics of the different types of topic models for the retrieval purpose. Section 3 discusses two different extensions of topic modeling. Section 4 describes the spoken document collection used in this paper, as well as the experimental setup. A series of experiments and associated discussions are presented in Section 5. Finally, Section 6 concludes this paper and suggests possible avenues for future work.

2. Topic Models

In this section, we first describe the probabilistic generative framework for information retrieval. We then briefly review the document topic models (DTM), including the probabilistic latent semantic analysis (PLSA) (Hofmann, 2001) and the latent Dirichlet model (LDA) (Blei *et al.*, 2003; Wei & Croft, 2006), followed by an introduction to the word topic model (WTM) (Chen, 2009), as well as the word Dirichlet topic model (WDTM).

2.1 Probabilistic Generative Framework

When the language modeling approach is applied to IR, it basically makes use of a probabilistic generative framework for ranking each document D in the collection given a query Q, which can be expressed by $P(D|Q)$. By applying Bayes' theorem, this ranking criterion can be approximated by the likelihood of Q generated by D, *i.e.*, $P(Q|D)$, when we assume that the prior probability of each document $P(D)$ is uniformly distributed. For this idea to work, each document D is treated as a probabilistic language model M_D for generating the query. Furthermore, if the query Q is treated as a sequence of words (or terms), $Q = w_1 w_2 \ldots w_N$, where the query words are assumed to be conditionally independent given the document model M_D and their order is also assumed to be of no importance (*i.e.*, the so-called "*bag-of-words*" assumption), the relevance measure $P(Q|D)$ can be further decomposed as a product of the probabilities of the query words generated by the document:

$$P(Q|D) = \prod_{w_i \in Q} P(w_i|\mathrm{M}_D)^{c(w_i,Q)}, \tag{1}$$

where $c(w_i,Q)$ is the number of times that each distinct word w_i occurs in Q. The document ranking problem has now been reduced to the problem of constructing the document model $P(w_i|\mathrm{M}_D)$.

The simplest way to construct $P(w_i|\mathrm{M}_D)$ is based on literal term matching, or using the unigram language model (ULM), where each document of the collection can respectively offer a unigram distribution for observing a query word, i.e., $P_{\mathrm{ULM}}(w_i|\mathrm{M}_D)$, which is estimated on the basis of the words occurring in the document:

$$P_{\mathrm{ULM}}(w_i \mid \mathrm{M}_D) = \frac{c(w_i,D)}{|D|}, \tag{2}$$

where $c(w_i,D)$ is the number of times that word w_i occurs in the document D and $|D|$ is the number of words in the document. In order to avoid the problem of zero probability, the ULM is usually smoothed by a unigram distribution estimated from a general collection, *i.e.*, $P_{\mathrm{ULM}}(w_i|\mathrm{M}_C)$:

$$\hat{P}_{\text{ULM}}\left(w_i\middle|D\right)=\lambda\cdot P_{\text{ULM}}\left(w_i\middle|\text{M}_D\right)+(1-\lambda)\cdot P_{\text{ULM}}\left(w_i\middle|\text{M}_C\right), \tag{3}$$

where λ is a weighting parameter. It turns out that a document with more query words occurring in it would tend to receive a higher probability; further, the use of $P_{\text{ULM}}\left(w_i\middle|\text{M}_C\right)$ to some extent can help deemphasize common (non-informative) words but instead put more emphasis on discriminative (or informative) words for the purpose of document ranking (Zhai, 2008). In the following, $P_{\text{ULM}}\left(w_i\middle|\text{M}_D\right)$ and $P_{\text{ULM}}\left(w_i\middle|\text{M}_C\right)$ will be termed the document model and the background model, respectively.

2.2 Document Topic Model (DTM)

As mentioned earlier, there probably would be word usage mismatch between a query and a spoken document, even if they are topically related to each other. Therefore, instead of constructing the document model based on the literal term information, we can exploit probabilistic topic models to represent each spoken document through a latent topic space (Blei *et al.*, 2010). In this spectrum of research, each document D is regarded as a document topic model (DTM), consisting of a set of K shared latent topics $\{T_1,\ldots,T_k,\ldots,T_K\}$ with document-specific weights $P\left(T_k\middle|\text{M}_D\right)$, where each topic T_k in turn offers a unigram distribution $P\left(w_i\middle|T_k\right)$ for observing an arbitrary word of the language. For example, in the PLSA model, the probability of a word w_i generated by a document D is expressed by:

$$P_{\text{PLSA}}\left(w_i\middle|\text{M}_D\right) = \sum_{k=1}^{K} P\left(w_i\middle|T_k\right)P\left(T_k\middle|\text{M}_D\right). \tag{4}$$

The key idea we wish to illustrate here is that, for PLSA, the relevance measure of a query word w_i and a document D is not computed directly based on the frequency of w_i occurring in D, but instead based on the frequency of w_i in the latent topic T_k as well as the likelihood that D generates the respective topic T_k, which in fact exhibits some sort of concept matching. A document is believed to be more relevant to the query if it has higher weights on some topics and the query words also happen to appear frequently in these topics.

In the practical implementation of PLSA, the corresponding DTM models are usually trained in an unsupervised way by maximizing the total log-likelihood of the document collection $\mathbf{D}$ in terms of the unigram $P_{\text{PLSA}}\left(w_i\middle|\text{M}_D\right)$ of all words w_i observed in the document collection, or, more specifically, the total likelihood of all documents generated by their own DTM models:

$$\begin{aligned} L_{\text{PLSA}} &= \prod_{D\in\mathbf{D}} P_{\text{PLSA}}\left(D\middle|\text{M}_D\right) \\ &= \prod_{D\in\mathbf{D}}\prod_{w_i\in D} P_{\text{PLSA}}\left(w_i\middle|\text{M}_D\right)^{c(w_i,D)}. \end{aligned} \tag{5}$$

We can first use the *K*-means algorithm to partition the entire document collection into K topical classes. Hence, the initial topical unigram distribution $P\left(w_i \middle| T_k\right)$ for a topical cluster can be estimated according to the underlying statistical characteristics of the document being assigned to it and the probabilities for each document generating the topics, *i.e.*, $P\left(T_k \middle| \mathrm{M}_D\right)$, are measured according to its proximity to the centroid of each respective cluster. Then, (5) can be iteratively optimized by the following three expectation-maximization (EM) (Dempster *et al.*, 1977) updating equations:

- **E (Expectation) Step**

$$P\left(T_k \mid w_i, \mathrm{M}_D\right) = \frac{P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_D\right)}{\sum_{T_k'} P\left(w_i \mid T_k'\right) P\left(T_k' \mid \mathrm{M}_D\right)}, \tag{6}$$

- **M (Maximization) Step**

$$\hat{P}\left(w_i \mid T_k\right) = \frac{\sum_D c\left(w_i, D\right) P\left(T_k \mid w_i, \mathrm{M}_D\right)}{\sum_w \sum_D c\left(w, D\right) P\left(T_k \mid w, \mathrm{M}_D\right)}, \tag{7}$$

$$\hat{P}\left(T_k \mid \mathrm{M}_D\right) = \frac{\sum_w c\left(w, D\right) P\left(T_k \mid w, \mathrm{M}_D\right)}{\sum_{w'} c\left(w', D\right)}, \tag{8}$$

where $P\left(T_k \mid w_i, \mathrm{M}_D\right)$ is the probability that the latent topic T_k occurs given the word w_i and the document model M_D, which is computed using the probability quantities $P\left(w_i \mid T_k\right)$ and $P\left(T_k \mid \mathrm{M}_D\right)$ obtained in the previous training iteration.

On the other hand, LDA, having a formula analogous to PLSA for document ranking, is regarded as a generalization of PLSA and has enjoyed considerable success in a wide variety of natural language processing (NLP) tasks. LDA differs from PLSA mainly in the inference of model parameters: PLSA assumes the model parameters are fixed and unknown; while LDA places additional *a priori* constraints on the model parameters, *i.e.*, thinking of them as random variables that follow Dirichlet distributions. In other words, the total log-likelihood of all documents generated by LDA models is defined as:

$$L_{\mathrm{LDA}} = \iint \prod_{z=1}^{K} P\left(\varphi_z \mid \beta\right) \prod_{D \in \mathbf{D}} p\left(\theta_D \mid \alpha\right) \left(\prod_{i=1}^{|D|} \sum_{k=1}^{K} P\left(w_i \middle| T_k, \varphi_z\right) P\left(T_k \middle| \theta_D\right) \right) d\theta d\varphi \tag{9}$$

where θ_d and φ_z are multinomial distributions with Dirichlet parameter α and β, respectively, and $|D|$ is the number of words in the document D. LDA possesses fully consistent generative semantics by treating the topic mixture distribution as a K-parameter hidden random variable rather than a large set of individual parameters that are explicitly linked to the training set (Blei *et al.*, 2003). Compared to PLSA, LDA overcomes the problem of overfitting and the problem of generating new documents incurred by PLSA.

Since LDA has a more complex form for model optimization, which is difficult to be solved by exact inference, several approximate inference algorithms, such as the variational Bayes approximation (Blei *et al.*, 2003), the expectation propagation method (Ypma *et al.*, 2002), and the Gibbs sampling algorithm (Griffiths, 2004), have been proposed in the literature for estimating the model parameters of LDA. In this paper, we adopt the Gibbs sampling algorithm, where θ and φ are marginalized out and only the latent variables T_k are sampled, to infer the model parameters. Then, the probability of a word w_i generated by a document D in the LDA model is expressed by:

$$P_{\mathrm{LDA}}\left(w_i \middle| \hat{\phi}, \hat{\theta}, \mathrm{M}_D\right) = \sum_{k=1}^{K} P\left(w_i \middle| T_k, \hat{\phi}\right) P\left(T_k \middle| \hat{\theta}, \mathrm{M}_D\right), \tag{10}$$

where $\hat{\varphi}$ and $\hat{\theta}$ are the posterior estimates of θ and φ, respectively. We refer the readers to Griffiths and Steyvers (2004) for a better understanding of the detailed inference procedure.

2.3 Word Topic Model (WTM)

Rather than treating each document in the collection as a document topic model, we can regard each word w_j of the language as a word topic model (WTM). To get to this point, all words are assumed to share the same set of latent topic distributions but have different weights over these topics. The WTM model of each word w_j for predicting the occurrence of a particular word w_i can be expressed by:

$$P_{\mathrm{WTM}}\left(w_i \mid \mathrm{M}_{w_j}\right) = \sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_{w_j}\right), \tag{11}$$

where $P\left(w_i \middle| T_k\right)$ and $P\left(T_k \middle| \mathrm{M}_{w_j}\right)$ are the probability of a word w_i occurring in a specific latent topic T_k and the probability of the topic T_k conditioned on M_{w_j}, respectively. Then, each document naturally can be viewed as a composite WTM, while the relevance measure between a word w_i and a document D can be expressed by:

$$P_{\mathrm{WTM}}\left(w_i \middle| \mathrm{M}_D\right) = \sum_{w_j \in D} P_{\mathrm{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right) P_{\mathrm{ULM}}\left(w_j \middle| \mathrm{M}_D\right), \tag{12}$$

The resulting composite WTM model for D, in a sense, can be thought of as a kind of language model for translating words in D to w_i.

The model parameters of WTM can be inferred by unsupervised training as well. More precisely, each WTM model M_{w_j} can be trained by concatenating those words occurring in the vicinity of (or a context window of size S around) each occurrence of w_j, which are postulated to be relevant to w_j, to form a relevant observation sequence O_{w_j} for training M_{w_j}. The words in O_{w_j} are also assumed to be conditionally independent, given M_{w_j}.

Therefore, the WTM models of the words in the vocabulary set $\mathbf{w}$ can be estimated by maximizing the total likelihood of their corresponding relevant observation sequences generated by themselves:

$$L_{\mathrm{WTM}}$$

$$= \prod_{w_j \in \mathbf{w}} P_{\mathrm{WTM}}\left(O_{w_j} \middle| \mathrm{M}_{w_j}\right) = \prod_{w_j \in \mathbf{w}} \prod_{w_i \in O_{w_j}} P_{\mathrm{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right)^{c\left(w_i, O_{w_j}\right)}, \tag{13}$$

Then, the parameters of each WTM model can be estimated using the following EM updating formulae:

- **E (Expectation) Step**

$$P\left(T_k \mid w_i, \mathrm{M}_{w_j}\right) = \frac{P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_{w_j}\right)}{\sum_{T_k'} P\left(w_i \mid T_k'\right) P\left(T_k' \mid \mathrm{M}_{w_j}\right)}, \tag{14}$$

- **M (Maximization) Step**

$$\hat{P}\left(w_i \mid T_k\right) = \frac{\sum_{w_j \in \mathbf{w}} c\left(w_i, O_{w_j}\right) P\left(T_k \mid w_i, \mathrm{M}_{w_j}\right)}{\sum_{w_l \in \mathbf{w}} \sum_{w_n \in O_{w_l}} c\left(w_n, O_{w_l}\right) P\left(T_k \mid w_n, \mathrm{M}_{w_l}\right)}, \tag{15}$$

$$\hat{P}\left(T_k \mid \mathrm{M}_{w_j}\right) = \frac{\sum_{w \in O_{wj}} c\left(w, O_{wj}\right) P\left(T_k \mid w, \mathrm{M}_{w_j}\right)}{\sum_{w'} c\left(w', O_{wj}\right)}. \tag{16}$$

Along a similar vein to the LDA model, word Dirichlet topic model (WDTM) can be derived as well. WDTM essentially has the same ranking formula as WTM, except that it further assumes the model parameters are governed by some Dirichlet distributions.

2.4 Analytic Comparisons between DTM and WTM

DTM (PLSA or LDA) and WTM (WTM or WDTM) can be analyzed from several perspectives. First, DTM models the co-occurrence relationship between words and documents, while WTM models the co-occurrence relationship between words in the collection. More explicitly, we may compare DTM and WTM through nonnegative (or probabilistic) matrix factorizations, as depicted in Figure 1. For DTM models, each column of Matrix $\mathbf{A}$ denotes the probability vector of a document in the collection, which offers a probability for every word occurring in the document. For WTM models, each column of Matrix $\mathbf{B}$ is the probability vector of a word's vicinity, which offers a probability for observing every other word occurring in its vicinity. Both Matrices $\mathbf{A}$ and $\mathbf{B}$ can be decomposed into two matrices standing for the topic mixture components and the topic mixture weights, respectively.

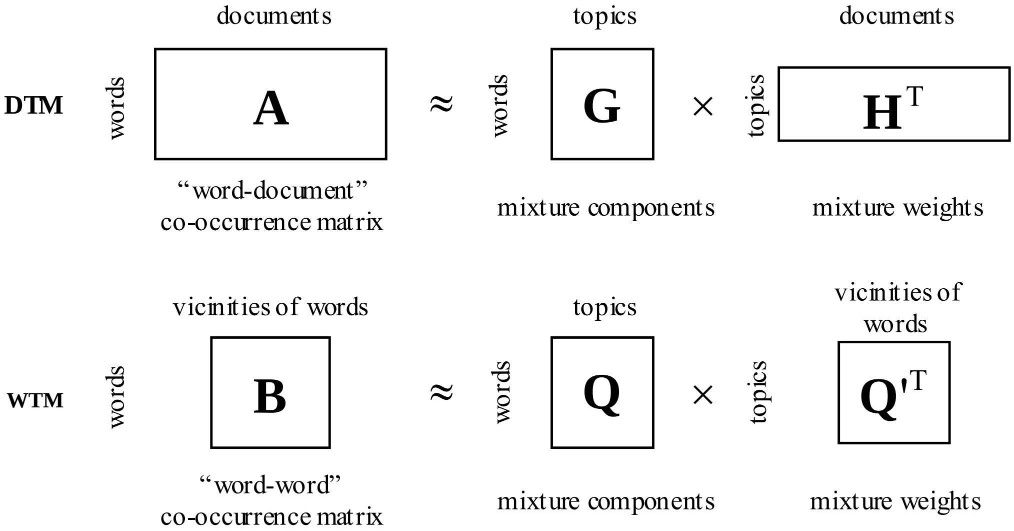


Figure 1. A schematic illustration for the matrix factorizations of DTM and WTM.

Furthermore, the topic mixture weights of DTM for a new document have to be estimated online using EM or other more sophisticated algorithms, which would be time-consuming; on the contrary, the topic mixture weights of WTM for a new document D can be obtained on the basis of the topic mixture weights of all words involved in the document without using a complex inference procedure.

Finally, if the context window for modeling the vicinity information of WTM is reduced to one word ($S = 1$), WTM can be either degenerated to a unigram model as the latent topic number K is set to 1, or viewed as analogous to a bigram model (as $K = V$) or an aggregate Markov model (as $1 < K < V$). Thus, with some appropriate values of S and K being chosen, we can show that WTM seems to be a good method of approximating the bigram or skip-bigram models for sparse data (Chen, 2009).

3. Extensions of Topic Modeling

3.1 Hybrid of DTM and WTM

As mentioned in the previous section, DTM and WTM are different from each other in their fundamental premises to determine a hidden topical decomposition of the document collection through the exploration of the topical information underlying the "word-document" or "word-word" co-occurrence relationships, respectively. Thus, we may fuse the results of the two different topical decompositions from DTM and WTM together for better ranking of spoken documents.

One possible method is to train each of these two models individually and linearly combine their respective document-ranking scores in the log-likelihood domain subsequently (called "Individual Topics" hereafter). Nevertheless, this approach could not arrive at the same set of topic components (*i.e.*, $P(w_i|T_k)$, $k = 1,\ldots,K$) that are potentially associated with the spoken document collection. Alternatively, we may seek to conduct a single (or unique) topical decomposition of the spoken document collection by simultaneously exploiting these two types of co-occurrence relationships (called "Shared Topics" hereafter). This approach tries to estimate the DTM and WTM model parameters by jointly maximizing the total likelihood of words occurring in the spoken documents and the total likelihood of the words occurring in the vicinities of arbitrary words in the vocabulary. A pictorial representation for the probabilistic matrix decomposition of the spoken document collection with this approach is illustrated in Figure 2, where each column of the left hand side matrix denotes either the probability vector of a document in the collection, which offers a probability for every word occurring in the document (*i.e.*, DTM), or the probability vector of the vicinity of a word in the vocabulary, which offers a probability for observing every other word occurring in the vicinity (*i.e.*, WTM). Then, this matrix can be decomposed into two matrices standing for the topic mixture components (*i.e.*, $\mathbf{F}$) and the topic mixture weights (*i.e.*, $\mathbf{H}$ and $\mathbf{Q}'$), respectively.

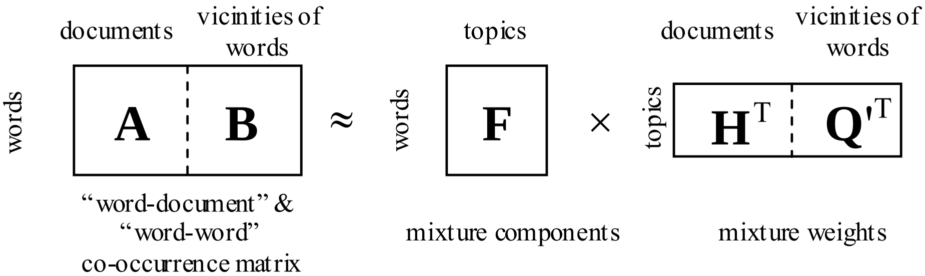


Figure 2. A schematic illustration for the matrix factorization of hybrids of DTM and WTM.

3.2 Topic Modeling with Subword-level Units

In this paper, we also investigate leveraging subword-level information cues for topic modeling in Chinese SDR. To do this, syllable pairs are taken as basic units for indexing instead of words. In the following paragraphs, we will elucidate the reasons for using syllable-level features for the retrieval purpose before describing how they can be integrated into the DTM and WTM models.

Mandarin Chinese is phonologically compact; an inventory of about 400 base syllables provides full phonological coverage of Mandarin audio if the differences in tones are disregarded. On the other hand, an inventory of about 13,000 characters provides full textual coverage of written Chinese. Each word is composed of one or more characters, and each character is pronounced as a monosyllable and is a morpheme with its own meaning. As a result, new words are generated easily by combining a few characters. Such new words also include many proper nouns, like personal names, organization names, and domain-specific terms. The construction of words from characters is often quite flexible. One phenomenon is that different words describing the same or similar concepts can be constructed of slightly different characters. Another phenomenon is that a longer word can be arbitrarily abbreviated into a shorter word. Moreover, there is a many-to-many mapping between characters and syllables; a foreign word can be translated into different Chinese words based on its pronunciation, while different translations usually have some syllables in common, or may have exactly the same syllables. Statistical evidence also shows that, in the Chinese language, about 91% of the top 5,000 most frequently used polysyllabic words are bi-syllabic, i.e., they are pronounced as a segment of two syllables. Therefore, such syllable segments (or syllable pairs) definitely carry a plurality of linguistic information and make great sense to be used as important index terms.

The characteristics of the Chinese language mentioned above lead to some special considerations for SDR. Word-level index features possess more semantic information than syllable-level ones; thus, word-based retrieval enhances the precision. On the other hand, syllable-level index features are more robust against the Chinese word tokenization ambiguity, Chinese homophone ambiguity, open vocabulary problem, and speech recognition errors; therefore, the syllable-level information would enhance the recall. Accordingly, there is good reason to fuse the information obtained from index features of different levels. It has been shown that using syllable pairs as the index terms is very effective for Chinese SDR, and the retrieval performance can be further improved by incorporating the information from word-level index features.

In this paper, both the manual transcript and the recognition transcript of each spoken document, in the form of a word stream, were automatically converted into a stream of overlapping syllable pairs. Then, all of the distinct syllable pairs occurring in the spoken document collection were identified to form an indexing vocabulary of syllable pairs. Topic modeling with the syllable-level information can be fulfilled in two ways. One is to simply use syllable pairs, as a replacement for words, to represent the spoken documents and to construct the associated probabilistic latent topic distributions for DTM and WTM accordingly. The other is to jointly utilize both words and syllable pairs, as two types of index terms, to represent the spoken documents, as well as to construct the associated probabilistic latent topic distributions. To this end, each spoken document is represented virtually with a spliced text stream, consisting of both words and syllable pairs. Figure 3 takes DTM as an example to graphically illustrate such an attempt, which is expected to discover correlated topic patterns of the spoken document collection when using both word- and syllable-level index features simultaneously.

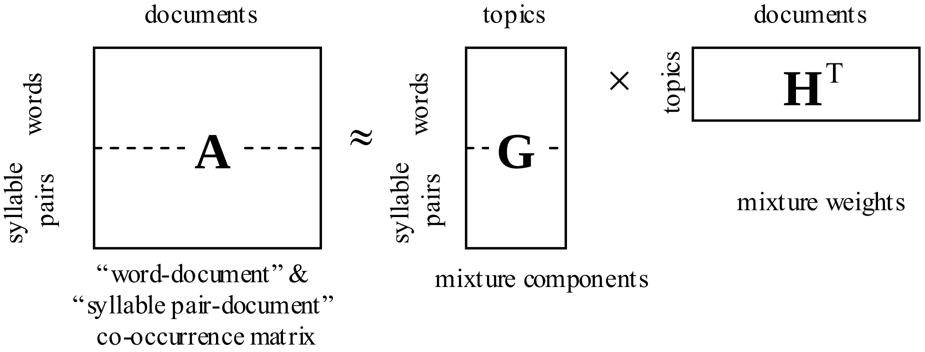


Figure 3. A schematic illustration for the matrix factorization of DTM, jointly using words and syllable pairs as the index terms.

4. Experimental Setup

4.1 Corpus and Evaluation Metric

We used the Topic Detection and Tracking (TDT-2) collection for the SDR task (LDC, 2000). TDT is a DARPA sponsored program where participating sites tackle tasks, such as identifying the first time a news story is reported on a given topic or grouping news stories with similar topics from audio and textual streams of newswire data. Both the English and Mandarin Chinese corpora have been studied in the recent past. The TDT corpora have also been used for cross-language spoken document retrieval (CLSDR) in the Mandarin English Information (MEI) Project (Meng *et al.*, 2004). In this paper, we used the Mandarin Chinese collections of the TDT corpora for the retrospective retrieval task, such that the statistics for the entire document collection was obtainable. Chinese text news stories from Xinhua News Agency were compiled to form the test queries (or query exemplars). More specifically, in the following experiments, we will either use a whole text news story as "*long*" query or merely extract the title field from a text news story to form a relatively "*short*" query.

The Mandarin news stories (audio) from Voice of America news broadcasts were used as the spoken documents. All news stories were exhaustively tagged with event-based topic labels, which merely serve as the relevance judgments for performance evaluation and will not be utilized in the training of topic models (*cf*. Section 2). Table 1 shows some basic statistics about the corpus used in this paper. The Dragon large-vocabulary continuous speech recognizer provided Chinese word transcripts for our Mandarin audio collections. To assess the performance level of the recognizer, we spot-checked a fraction of the spoken document collection set (about 40 hours), and obtained error rates of 35.38% (in word), 17.69% (in character), and 13.00% (in syllable). Since Dragon's lexicon is not available, we augmented the LDC Mandarin Chinese Lexicon with 24,000 words extracted from Dragon's word recognition output, and used the augmented LDC lexicon (about 51,000 words) to tokenize the manual transcripts for computing error rates. We also used this augmented LDC lexicon to tokenize the text queries in the retrieval experiments.

Table 1. Statistics for TDT-2 Collections Used for Spoken Document Retrieval

# Spoken documents	2,265 stories 46.03 hours of audio			
# Distinct test queries	16 Xinhua text stories (Topics 20001~20096)			
	Min.	Max.	Med.	Mean
Document length (in characters)	23	4841	153	287
Length of long query (in characters)	183	2623	329	533
Length of short query (in characters)	8	27	13	14
# Relevant documents per test query	2	95	13	29

The retrieval results are expressed in terms of non-interpolated mean average precision (mAP) following the TREC evaluation (Harman, 1995), which is computed by the following equation:

$$\text{mAP} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{j}{r_{i,j}}, \tag{17}$$

where L is the number of test queries, N_i is the total number of documents that are relevant to query Q_i, and $r_{i,j}$ is the position (rank) of the *j*-th document that is relevant to query Q_i, counting down from the top of the ranked list.

4.2 Model Implementation

Topic models, such as DTM and WTM, introduce a set of latent topics to cluster concept-related words and match a query with a document at the level of these word clusters. Although document ranking based merely on DTM or WTM tends to increase recall, using just one of them is liable to hurt the precision for SDR. Specifically, they offer coarse-grained concept clues about the document collection at the expense of losing discriminative power among concept-related words in finer granularity. Therefore, in this paper, when either DTM or WTM was employed in evaluating the relevance between a query Q and a document D, we additionally incorporated the unigram probabilities of a query word (or term) occurring in the document $P_{\text{ULM}}(w_i \mid \text{M}_D)$ and a general text corpus $P_{\text{ULM}}(w_i \mid \text{M}_C)$ with the topic model $P_{\text{Topic}}(w_i \mid \text{M}_D)$ (either DTM or WTM), for probability smoothing and better performance. For example, the probability of a query word generated by one specific topic model of a document (*cf.* (4), (10), and (12)) was modified as follows:

$$\begin{aligned} P(w_i \mid D) &= \alpha \cdot \left[\beta \cdot P_{\text{Topic}}(w_i \mid \text{M}_D) + (1-\beta) \cdot P_{\text{ULM}}(w_i \mid \text{M}_D) \right] \\ &\quad + (1-\alpha) \cdot P_{\text{ULM}}(w_i \mid \text{M}_C) \end{aligned} \tag{18}$$

where $P_{\text{Topic}}(w_i \mid \text{M}_D)$ can be the probability of a word w_i generated by PLSA or LDA (*cf.* (4) or (10)) or WTM (*cf.* (12)); the values of the interpolation weights α and β can be empirically set or further optimized by other optimization techniques (Zhai, 2008). A detailed account of this issue will be given in Section 5.2. On the other hand, the Gibbs sampling algorithm (Griffiths, 2004) is used to infer the parameters of LDA and WDTM.

5. Experimental Results

5.1 Baseline Experiments

The baseline retrieval results obtained by the ULM model are shown in Table 2. The retrieval results, assuming manual transcripts for the spoken documents to be retrieved (denoted TD, text documents) are known, are listed for reference and are compared to the results when only erroneous recognition transcripts generated by speech recognition are available (denoted SD, spoken documents). As can be seen, the performance gap between the TD and SD cases was about 7% absolute in terms of mAP when using either long or short queries, although the word error rate (WER) for the spoken document collection was higher than 35%. On the other hand, retrieval using short queries degraded the performance approximately 45% relative to retrieval using long queries. This is due to the fact that a long query usually contains a variety of words describing similar concepts. Even though some of these words might not be correctly transcribed in the relevant spoken documents, they, in the ensemble, still provide plenty of clues for literal term matching. From now on, unless otherwise stated, we will only report the retrieval results for the SD case.

Table 2. Baseline retrieval results (in mAP) achieved by ULM.

Query Type	TD	SD
Long	0.639	0.562
Short	0.370	0.293

5.2 Experiments on DTM and WTM

In the next set of experiments, we assessed the utility of various topic models for SDR, including PLSA, LDA, and WTM, as well as WDTM. The corresponding retrieval results are shown in Table 3. It is worth mentioning that all of these topic models were trained without supervision and had the same number of latent topics, which was set to 32 in this study. A detailed analysis for the impact of the model complexity of PLSA and WTM on SDR performance can be found in Chen (2009). On the other hand, both WTM and WDTM had the same context window size S set to 21. Since this project set out to investigate the effectiveness of various topic models for SDR, the interpolation weights α and β defined in (18) were optimized for each respective topic model with a two-dimensional grid search over the range from 0 to 1 and in increments of 0.1. Consulting Table 3, we find that all of these topic models give moderate but consistent improvement over the baseline ULM model when long queries are evaluated. One possible explanation is that the information need already might have been stated fully in a long query, whereas additional incorporation of the topical information into the document language model does not seem to offer many extra clues for document ranking. On the contrary, the retrieval performance receives great boosts from the additional use of the topical information when the queries are short. This implies that incorporating the topical information with the literal term information for document modeling is especially useful when the query is inadequate to address the information need.

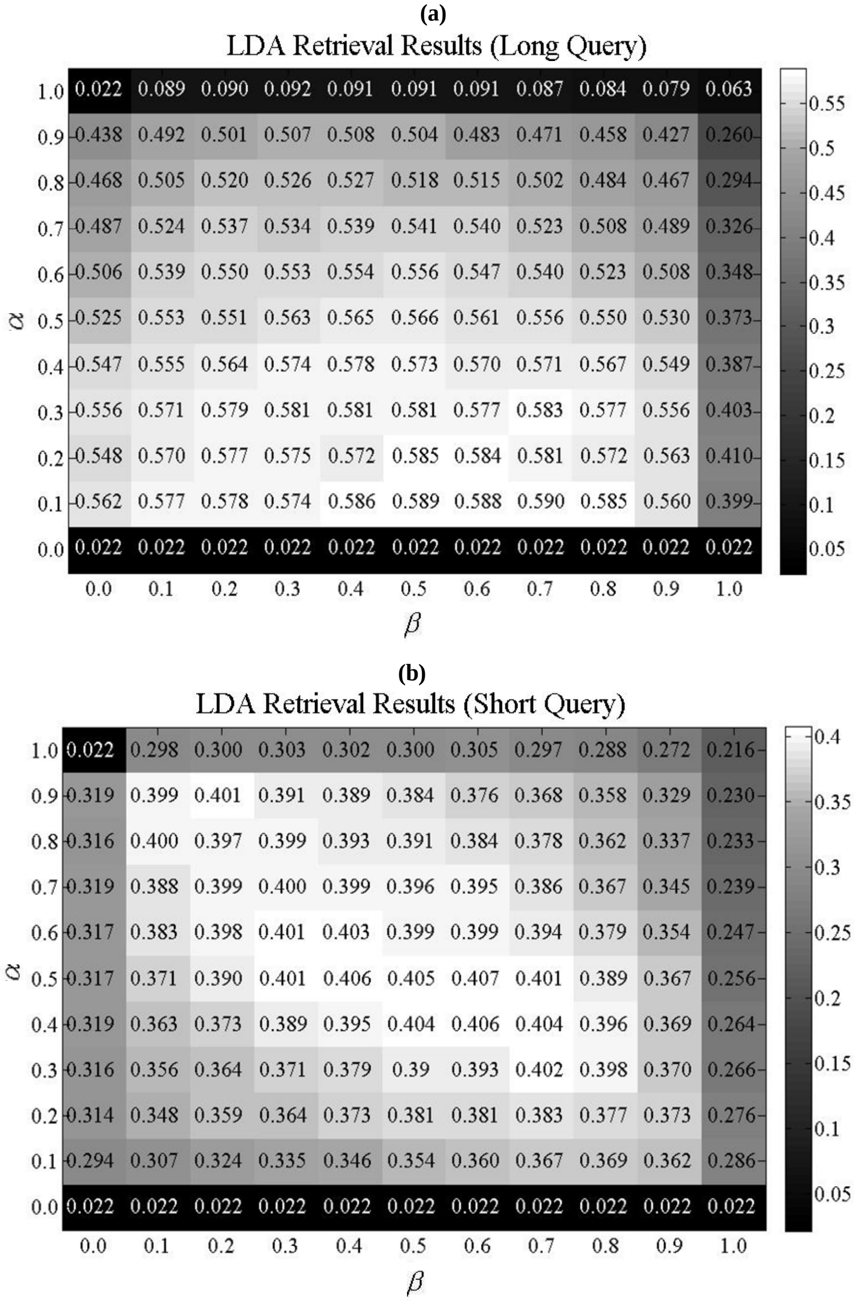
Table 3. Spoken document retrieval results achieved by various topic models.

Method	Long Query	Short Query
ULM	0.562	0.293
PLSA	0.569	0.374
LDA	0.590	0.407
WTM	0.573	0.351
WDTM	0.574	0.377
LDA+WDTM (Individual Topics)	0.592	0.418
LDA+WDTM (Shared Topics)	0.595	0.415

We then turned our attention to compare the following topic models. 1) LDA outperforms PLSA, and WDTM outperforms WTM. This finding supports the argument that constraining the latent topic distributions with Dirichlet priors will lead to better model estimation. 2) LDA is the best among these topic models. As compared to the baseline ULM model, it yielded about 5% and 39% relative improvements for long and short queries, respectively. Moreover, we investigated the effectiveness of the fusion of DTM and WTM to the retrieval performance (*cf.*, the last two rows of Table 3). Here, we took LDA and WDTM as the training example since they achieved better retrieval performance in the previous experiment. It is also worth mentioning that the row “LDA+WDTM (Individual Topics)” shown in Table 3 indicates that each topic model was trained individually and their respective document-ranking scores were combined in the log-likelihood domain. On the contrary, the row “LDA+WDTM (Shared Topics)” in Table 3 denotes the hybrid of DTM and WTM in both model training and testing (*cf.* Section 3.1). As is evident, the fusion of LDA and WDTM (*i.e.*, with either individual sets of topics or a shared set of topics) is beneficial to the retrieval performance. This provides an additional 1% absolute improvement for the case of using short queries, as compared to that using LDA alone. Nevertheless, the joint exploration of “word-document” and “word-word” latent topic information (*i.e.*, with a shared set of topics) in the training phrase does not provide any added benefit compared to the results obtained by training LDA and WDTM individually (*i.e.*, with individual sets of topics). This is an interesting phenomenon and awaits further exploration. Readers may refer to Chen, *et al.* (2010) for an attempt that applies a similar idea to the speech recognition task.

To go a step further, we attempted to investigate the more subtle interaction effects among the topic model $P_{\text{Topic}}(w_i|\mathrm{M}_D)$, the document model $P_{\text{ULM}}(w_i|\mathrm{M}_D)$, and the background model $P_{\text{ULM}}(w_i|\mathrm{M}_C)$ in (18) by varying the values of the interpolation weights α and β. Here, LDA was taken as an example topic model since it exhibits the best performance among the topic models compared in this paper. The retrieval results are graphically illustrated in Figure 4, where the horizontal and vertical axes denote the values of α and β, respectively. As seen in the results revealed in Figure 4, additional incorporation of $P_{\text{ULM}}(w_i|\mathrm{M}_D)$ and $P_{\text{ULM}}(w_i|\mathrm{M}_C)$ into LDA is beneficial for retrieval. In an extreme case, when both the values of α and β are set to one, as shown in the top right corner of Figure 4, the retrieval model is based merely on the topical information, which has poor retrieval performance, especially for the case using long queries. One possible reason is that a long query may contain several common non-informative words and using the topical information alone will let the query become biased away from representing the true theme of the information need, probably due to these non-informative words. This argument again can be verified by examining the rightmost columns of Figure 4, where using the background model $P_{\text{ULM}}(w_i|\mathrm{M}_C)$ can absorb the contributions of the common (or non-informative) words made to document ranking, thus giving better retrieval performance.

Looking at each row of Figure 4, we see that smoothing LDA with the document model $P_{\text{ULM}}(w_i|\mathrm{M}_D)$ is also useful. This is attributed to the fact that discriminative (or informative) words will occur repeatedly in a specific document; $P_{\text{ULM}}(w_i|\mathrm{M}_D)$ gives more emphasis on these words. On the other hand, Figure 4 also reflects that smoothing LDA with the background model $P_{\text{ULM}}(w_i|\mathrm{M}_C)$ is necessary when the query is long, but it does not seem to be helpful for the case of using a relatively short query. This is mainly because the information need stated by the short query is already concise, and the importance of the role that $P_{\text{ULM}}(w_i \mid \text{M}_C)$ plays in filtering out or deemphasizing common (or non-informative) words is less pronounced.



(a)

LDA Retrieval Results (Long Query)

α \ β	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	0.022	0.089	0.090	0.092	0.091	0.091	0.091	0.087	0.084	0.079	0.063
0.9	0.438	0.492	0.501	0.507	0.508	0.504	0.483	0.471	0.458	0.427	0.260
0.8	0.468	0.505	0.520	0.526	0.527	0.518	0.515	0.502	0.484	0.467	0.294
0.7	0.487	0.524	0.537	0.534	0.539	0.541	0.540	0.523	0.508	0.489	0.326
0.6	0.506	0.539	0.550	0.553	0.554	0.556	0.547	0.540	0.523	0.508	0.348
0.5	0.525	0.553	0.551	0.563	0.565	0.566	0.561	0.556	0.550	0.530	0.373
0.4	0.547	0.555	0.564	0.574	0.578	0.573	0.570	0.571	0.567	0.549	0.387
0.3	0.556	0.571	0.579	0.581	0.581	0.581	0.577	0.583	0.577	0.556	0.403
0.2	0.548	0.570	0.577	0.575	0.572	0.585	0.584	0.581	0.572	0.563	0.410
0.1	0.562	0.577	0.578	0.574	0.586	0.589	0.588	0.590	0.585	0.560	0.399
0.0	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022

(b)

LDA Retrieval Results (Short Query)

α \ β	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	0.022	0.298	0.300	0.303	0.302	0.300	0.305	0.297	0.288	0.272	0.216
0.9	0.319	0.399	0.401	0.391	0.389	0.384	0.376	0.368	0.358	0.329	0.230
0.8	0.316	0.400	0.397	0.399	0.393	0.391	0.384	0.378	0.362	0.337	0.233
0.7	0.319	0.388	0.399	0.400	0.399	0.396	0.395	0.386	0.367	0.345	0.239
0.6	0.317	0.383	0.398	0.401	0.403	0.399	0.399	0.394	0.379	0.354	0.247
0.5	0.317	0.371	0.390	0.401	0.406	0.405	0.407	0.401	0.389	0.367	0.256
0.4	0.319	0.363	0.373	0.389	0.395	0.404	0.406	0.404	0.396	0.369	0.264
0.3	0.316	0.356	0.364	0.371	0.379	0.39	0.393	0.402	0.398	0.370	0.266
0.2	0.314	0.348	0.359	0.364	0.373	0.381	0.381	0.383	0.377	0.373	0.276
0.1	0.294	0.307	0.324	0.335	0.346	0.354	0.360	0.367	0.369	0.362	0.286
0.0	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022	0.022

Figure 4. Detailed spoken document retrieval results achieved by LDA with respect to different types of queries.

5.3 Experiments on using Subword-level Index features

In the fourth set of experiments, we evaluated the performance of the topic models when syllable pairs were utilized instead as the index terms. Here, we took LDA and WDTM as the example topic models, and the corresponding models are denoted by Syl_LDA and Syl_WDTM, respectively. The fusion of words and syllable pairs for topic modeling was investigated as well. Notice that Word_LDA denotes LDA using words as the index terms, which was termed LDA in the previous sections.

The retrieval results of Syl_LDA and Syl_WDTM are shown in Table 4, where the results achieved by ULM and using syllable pairs as the index terms (denoted by Syl_ULM) are also depicted for comparison. Several observations can be made from Table 4. First, the topic models (Syl_LDA and Syl_WDTM) again are superior to the unigram language model when the syllable-level information is used in place of the word-level information (denoted by Syl_ULM). Syl_LDA results in absolute improvements of about 8% and 3% over Syl_ULM when evaluated using the long and short queries, respectively. Second, the topic models with the syllable-level information perform worse than those with the word-level information. This may be due simply to the fact that syllable pairs are not as good as words in representing the semantic content of the queries and the documents. Third, the fusion of the word- and syllable-information for topic modeling (each topic model was trained individually beforehand) demonstrates much better retrieval results (*cf.* the last two rows of Table 4) as compared to that of the topic models with merely the word-level information (*cf.* Table 3).

Table 4. Spoken document retrieval results achieved by LDA and WDTM, respectively, using syllable pairs along with the combination of words and syllable pairs.

Method	Long Query	Short Query
Syl_ULM	0.492	0.274
Syl_LDA	0.571	0.302
Syl_WDTM	0.536	0.299
Word_LDA+Syl_LDA	0.613	0.412
Word_WDTM+Syl_WDTM	0.575	0.383

Finally, we examined the contributions made by modeling the correlated topic patterns of the spoken document collection when jointly using words and syllable pairs in the construction of the latent topic distributions. We took the LDA model as an example to study the effectiveness of such an attempt, and the associated results are shown in Table 5. The results reveal that, when only syllable pairs are used as the index terms for the final document ranking, modeling the correlated topic patterns, namely, jointly using words and syllable pairs in the construction of the latent topic distributions for LDA (denoted by Syl_LDA (Corr.)) is better than that only using syllable pairs to construct the latent topic distributions (denoted by Syl_LDA). On the other hand, such an attempt slightly hurts the performance of LDA using words for the final document ranking (denoted by Word_LDA (Corr.)). This phenomenon seems to be reasonable because the semantic meanings carried by words would probably see interference from syllable pairs when we attempt to splice these two distinct index term streams together for constructing the latent topic distributions of LDA. It can be observed that Syl_LDA (Corr.) significantly outperforms all other topic models in the case of using long queries (*cf*. Tables 3, 4, and 5). This demonstrates the potential benefit of using the syllable-level information in topic modeling for SDR if we can carefully delineate the syllable-level information. Nevertheless, in the case of using short queries, Syl_LDA (Corr.) does not perform as well as LDA using words as the index terms to construct the latent topic distributions (denoted by Word_LDA). We conjecture that one possible reason is that the topical information inherent in a short query cannot be unambiguously depicted with limited syllable pairs. In order to mitigate this deficiency, we combined Word_LDA with Syl_LDA (Corr.) to form a new retrieval model (denoted by Word_LDA + Syl_LDA (Corr.)), which yields the best results of 0.636 and 0.431 for long and short queries, respectively. One should keep in mind that these results were obtained using the erroneous speech transcripts of the spoken documents (*i.e*., the SD case). This also reveals that Word_LDA + Syl_LDA (Corr.) can make retrieval using the speech transcripts achieve almost the same performance as ULM using the manual transcripts (*i.e*., the TD case) when the queries are long, and can perform even better than the latter for short queries.

Table 5. Spoken document Retrieval results achieved by correlated LDA, using words (Word_LDA(Corr.)), syllable pairs (Syl_LDA(Corr.)), and their combination (Word_LDA + Syl_LDA(Corr.)).

Method	Long Query	Short Query
Word_LDA (Corr.)	0.577	0.349
Syl_LDA (Corr.)	0.618	0.356
Word_LDA+Syl_LDA (Corr.)	0.636	0.431

6. Conclusions

In this paper, we have investigated the utility of two categories of topic models, namely, the document topic models (DTM) and the word topic models (WTM), for SDR. Moreover, we have leveraged different levels of index features for topic modeling, including words, syllable pairs, and their combinations, so as to prevent the performance degradation facing most SDR tasks. The proposed models indeed demonstrated significant performance improvements over the baseline model on the Mandarin SDR task. Our future research directions include: 1) training the topic models in a lightly supervised manner through the exploration of users' click-through data, 2) investigating discriminative training of topic models, 3) integrating the topic models with the other more elaborate representations of the speech recognition output (Yi and Allan, 2009; Chelba *et al.*, 2008) for larger-scale SDR tasks, and 4) utilizing speech summarization techniques to help estimate better document models and topic models.

Acknowledgement

This work was sponsored in part by "Aim for the Top University Plan" of National Taiwan Normal University and Ministry of Education, Taiwan, and the National Science Council, Taiwan, under Grants NSC 101-2221-E-003 -024 -MY3, NSC 99-2221-E-003-017-MY3, NSC 98-2221-E-003-011-MY3, NSC 100-2515-S-003-003, and NSC 99-2631-S-003-002.

Reference

Blei, D.M., Ng, A.Y., & Jordan, M. I., (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blei, D. & Lafferty, J., (2009). Topic models. In A. Srivastava and M. Sahami, (eds.), *Text Mining: Theory and Applications*. Taylor and Francis, 2009.

Blei, D., Carin, L., & Dunson, D., (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6), 55-65.

Chelba, C., Hazen, T. J., & Sarclar, M., (2008). Retrieval and browsing of spoken content. *IEEE Signal Processing Magazine*, 25(3), 39-49.

Chen, B., (2009). Word topic models for spoken document retrieval and transcription. *ACM Transactions on Asian Language Information Processing*, 8(1), Article 2.

Chia, T. K., Sim, K. C, Li, H. Z. & Ng, H. T., (2008). A lattice-based approach to query-by-example spoken document retrieval. In *Proceeding the ACM SIGIR Conference on R&D in Information Retrieval*, 363-370.

Dempster, A. P., Laird, N. M., & Rubin, D. B., (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* 39 (1): 1-38.

Garofolo, J., Auzanne, G., & Voorhees, E., (2000). The TREC spoken document retrieval track: A success story. In *Proceeding the 8th Text REtrieval Conference*. NIST, 107-129.

Griffiths, T. L. & Steyvers, M., (2004). Finding scientific topics. In *Proceeding of the National Academy of Sciences*, 5228-5235.

Griffiths, T. L., Steyvers, M. & Tenenbaum, J. B., (2007). Topics in semantic representation. *Psychological Review*, 114, 211-244.

Harman D., (1995). Overview of the Fourth Text Retrieval Conference (TREC-4). In *Proceeding the Fourth Text Retrieval Conference*, 1-23.

Hofmann, T., (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, 177-196.

LDC, (2000). Project topic detection and tracking. Linguistic Data Consortium. http://www.ldc.upenn.edu/Projects/TDT/.

Lee, L. S. & Chen B., (2005). Spoken document understanding and organization. *IEEE Signal Processing Magazine*, 22(5), 42-60.

Meng, H., Chen, B., Khudanpur, S., Levow, G. A., Lo, W. K., Oard, D., Schone, P., Tang, K., Wang, H. M., & Wang, J., (2004). Mandarin-English information (MEI): investigating translingual speech retrieval. *Computer Speech and Language*, 18(2), 163-179.

Miller, D. R. H., Leek, T., & Schwartz, R., (1999). A hidden Markov model information retrieval system. In *Proceeding ACM SIGIR Conference on R&D in Information Retrieval*, 214-221.

Ponte, J. M. & Croft, W. B., (1998). A language modeling approach to information retrieval. In *Proceeding the ACM SIGIR Conference on R&D in Information Retrieval*, 275-281.

Wei, X., & Croft, W. B., (2006). LDA-based document models for ad-hoc retrieval. In *Proceeding the ACM SIGIR Conference on R&D in Information Retrieval*, 178-185.

Lin, S. H. & Chen B., (2009). Topic modeling for spoken document retrieval using word- and syllable-level information. In *Proceedings of the third workshop on Searching spontaneous conversational speech*, 3-10.

Chen, K. Y., Chiu, H. S. & Chen B., (2010). Latent topic modeling of word vicinity information for speech recognition. In *Proceeding of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5394-5397.

Ypma, J., Basten, T. & Lafferty, J., (2002). Expectation-propagation for the generative aspect model. In *Proceeding Conference on Uncertainty in Artificial Intelligence*, 352-359.

Zhai, C. X., (2008). Statistical language models for information retrieval (Synthesis Lectures Series on Human Language Technologies). Morgan & Claypool Publishers.

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member：	NT$ 1,000 (US$50.-)
group membership：	NT$20,000 (US$1,000.-)

life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#： ________________

Name： __________________________________ Date of Birth： _______________

Country of Residence： ___________________ Province/State： ___________________

Passport No.： ___________________________ Sex: _______________________

Education(highest degree obtained)： _______________________________________

Work Experience： ____

____

Present Occupation： ____

Address： ____

____

Email Add： ____

Tel. No： ____________________________ Fax No： __________________________

Membership Category：□ Regular Member □ Life Member

Date： _____/____/____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member	：	US$ 50.- （NT$ 1,000）
Life Member	：	US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）

（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目

（三）收集國內外有關計算語言學知識之圖書及最新發展之資料

（四）發行有關之學術刊物，論文集及通訊

（五）研定有關計算語言學專用名稱術語及符號

（六）與國際計算語言學學術機構聯繫交流

（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

會員類別	□終身 □個人 □學生	會員編號	（由本會填寫）		
姓　　名		性別		出生日期	年　月　日
				身分證號碼	
現　　職		學　　歷			
通訊地址	□□□				
戶籍地址	□□□				
電　　話		E-Mail			
申請人:　　　　（簽章） 中　華　民　國　　年　　月　　日					

審查結果:

1. 年費：

   終身會員：　10,000.-

   個人會員：　1,000.-

   學生會員：　500.-（限國內學生）

   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)

   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638

   E-mail：aclclp@hp.iis.sinica.edu.tw　　網址: http://www.aclclp.org.tw

   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

PAYMENT FORM

Name :_______________________________________(Please print) Date: __________________

Please debit my credit card as follows: US$______________________

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD Issue Bank:_________________________

Card No.: __________- __________ - __________ - __________ Exp. Date:_______________

3-digit code: ______________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE : ___________________________________

Tel.: __________________________________ E-mail: __________________________________

Add: ____

PAYMENT FOR

US$ _____ ❑ Computational Linguistics & Chinese Languages Processing (CLCLP)

Quantity Wanted: ______

US$ _____ ❑ Publications:___________________________________

US$ _____ ❑ Text Corpora: _________________________________

US$ _____ ❑ Speech Corpora:________________________________

US$ _____ ❑ Others: _______________________________________

US$ __________ ❑Life Member Fee ❑ New Member ❑Renew

US$ __________ = Total

Fax : 886-2-2788-1638 or Mail this form to :

ACLCLP
% Institute of Information Science, Academia Sinica
R502, 128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中華民國計算語言學學會

信用卡付款單

姓名: _________________________________(請以正楷書寫)　日期:：　_______________

卡別：□ VISA CARD □ MASTER CARD □ JCB CARD　發卡銀行：_________________

卡號:__________-__________-__________-　________　有效日期：__________________

卡片後三碼：______________________（卡片背面簽名欄上數字後三碼）

持卡人簽名：　___________________________________(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：________________________　E-mail：__________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$___________□ 中文計算語言學期刊(IJCLCLP)

NT$___________□ 中研院詞庫小組技術報告

NT$___________□ 中文（新聞）語料庫

NT$___________□ 平衡語料庫

NT$___________□ 中文詞庫八萬目

NT$___________□ 中文句結構樹資料庫

NT$___________□ 平衡語料庫詞集及詞頻統計

NT$___________□ 中英雙語詞網

NT$___________□ 中英雙語知識庫

NT$___________□ 語音資料庫_______________________________

NT$___________□ 會員年費　□續會　□新會員　□終身會員

NT$___________□ 其他: ___________________________________

NT$________________=　合計

填妥後請傳真至 02-27881638 或郵寄至:
115台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

		AIR	AIR			
		Surface	(US&EURP)	(ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息為本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT： □ Credit Card (Preferred)
  □ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01　「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義 （中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期） 年份：______ （過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※　此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人： 黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：________________________ 收據抬頭：________________________

地　　址：____

電　　話：________________________ E-mail: ________________________

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright：It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.
1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.
2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.
3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.
4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).
5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript
6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.
7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).
Please visit the following websites for details.
(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)
(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

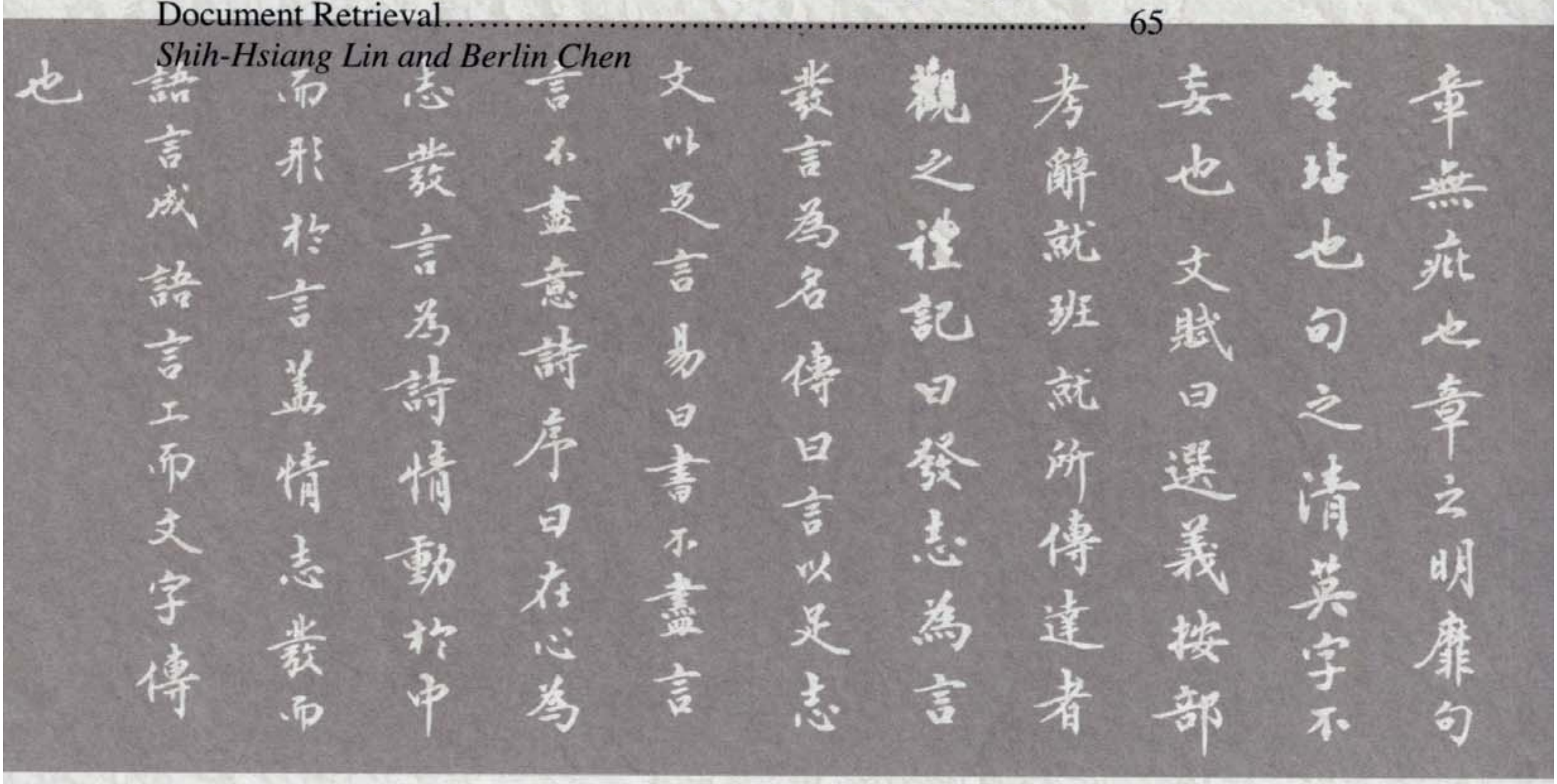
Online Submission: http://www.aclclp.org.tw/journal/submit.php

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.17 no.1 March 2012

Contents

Papers



ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing