AN ALGORITHM FOR AUTOMATIC GENERATION OF MANDARIN PHONETIC BALANCED CORPUS

Jyh-Shing Shyuu and Jhing-Fa Wang

Department of Computer Science and Information Engineering
National Cheng Kung University,
Tainan, Taiwan, ROC

ABSTRACT

This paper proposed an algorithm for automatic generation of Mandarin phonetic balanced corpus. The design of phonetic balanced corpus is particularly important for the collection of continuous speech database to reduce the co-articulate effects in continuous speech recognition (CSR).[1,2,3] Traditionally, balanced corpus is generated manually or semi-automatically.[4] Our proposed algorithm tries to find a minimum number of sentences from a large text corpus set and ensures that 408 Mandarin base syllables(without tonal information) and 38*22 co-articulations between vowels and consonants are distributed in the extracted sentences. The automatic generation of balanced corpus problem can be also treated as a covering problem. In other words, the objective of the problem here is to find the set with minimum number of sentences that can cover all the syllables and co-articulations from a text corpus. If the average number of syllables in a sentence is N, it gives 2*N-1 coverings(N syllables and N-1 co-articulations). The theoretical minimum number of balanced sentences is (408+38*22) / (2*N-1). For example, N=6, the minimum number of balanced sentences is 114.

1. INTRODUCTION

It is well known that collecting of large speech database to train the acoustic model or to evaluate the system performance is the first step to develop a large vocabulary and continuous speech recognition system. The design of phonetic balanced corpus determines what speech sentences should be collected so that each acoustic recognition unit can be well trained. Hence, a phonetic balanced corpus should contain at least the following information. First, each acoustic recognition unit must appear in the balanced corpus uniformly. Secondly, the co-articulations between acoustic recognition units must be included so that the co-articulation effect can be also trained into each acoustic recognition model. For example, in Mandarin speech, 60 recognition units are commonly used, including 22 consonants and 38 vowels. Theoretically, 3600 co-articulations should be included in the training corpus. However, only 408 syllables are valid phonic combinations (consonant + vowel). Hence, 408+38*22 combinations should be included in the training corpus.

In the past, the phonetic balanced corpus was designed manually or semi-automatically[4]. In collecting of speech database, it is preferable to provide many training corpus sets so that different co-articulations can be collected. If the balanced corpus is designed by manually, one may take a lot of efforts in designing the training corpus. Hence, automatic generation of balanced corpus is necessary. The automatic generation of balanced corpus problem can be also considered as a covering problem. In other words, the objective of the problem is to find the corpus set with minimum number of sentences from a large text corpus so that the balanced set can cover all the co-articulations between acoustic recognition units.

This paper is organized as follows. In Section 2, we go through details on our proposed algorithm for automatic generation of balanced corpus. In Section 3, we show our experimental results. Conclusion is given in Section 4.

2. ALGORITHM FOR AUTOMATIC GENERATION OF BALANCED CORPUS

Because most of the available corpora are composed of text sentences, we convert each text sentence into a phonetic string by a word segmentation algorithm as shown in Fig. 1. Basically, the word segmentation algorithm uses a Viterbi searching to determine the most likely word sequence based on a bigram language model. The phonetic taggings are then found from the word dictionary.

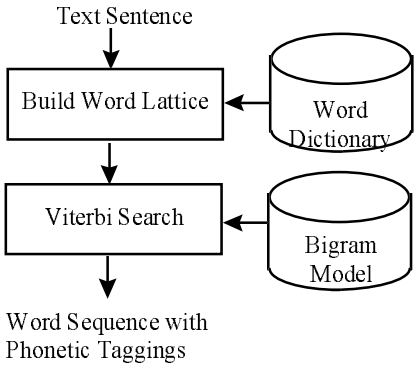

Figure1: Word Segmentation Algorithm

As the phonetic-tagged sentences can be obtained from the text corpus by the word segmentation algorithm, we proposed an algorithm and try to find a minimum number of phonetic-balanced sentences. The algorithm is described as follows,

1. Set up a covering table that represents the covering status for currently selected balanced sentences.

2. Find essential sentences in the corpus
   2.1. Scan all sentences in the corpus and find the essential sentences. For a given syllable or co-articulation, if there is only one sentence to cover the given syllable or co-articulation, the sentence is called an essential sentence. For a given syllable or co-articulation, if there is no sentence to cover it, the text corpus is insufficient to cover all balanced information. In such case, we append the text corpus with sentences that covers the given syllable or co-articulation and go back to step 2.1.

3. Randomly select sentences into balanced corpus to form a cover
   3.1. Select the essential sentences into the balanced corpus and update the covering table.
   3.2. Randomly select non-essential sentence into the balanced corpus if its redundancy is less than a threshold. The redundancy of a sentence is defined in Step 5.1.
   3.3. Repeat step 3.2 until the balanced corpus covers all the balanced information.
4. Select a balanced corpus with minimum number of sentences
   4.1. Repeat Step3 as many times as possible to get numbers of balanced corpora. In the experiments, we construct 1000 balanced corpora.
   4.2. Select the balanced corpus with minimum number of sentences from numbers of balanced corpora for further processing.
5. Replace redundant sentences in the balanced corpus
   5.1. For the remaining sentences that are not in the balanced corpus, estimate the redundancy (or overhead) and try to replace a sentence in the balanced corpus if its redundancy is less than one of the sentences in the balanced corpus. The detailed procedure is given by the following program code.

```
for k=1 to balanced-sentences
        temporary remove balanced-sentence(k) from the
        balanced corpus and update covering table
        compute overhead(balanced-sentence(k)) for balance-
        sentence(k)
        re-insert balanced-sentenced(k) into the balanced corpus
        and update the covering table
loop k
sort balanced sentences by the overhead value
return

overhead(sentence(s))
{
overhead=0
for i=1 to number of syllables in sentence s
        if number of distribution for syllable(i) in covering table
        <> 0 then
                overhead = overhead + 1
        endif
loop i
for i=1 to (number of syllables in sentence s) - 1
        if number of distribution of (vowel(i),consonant(i+1)) in
        covering table <> 0 then
                overhead = overhead +1
        endif
loop i
return(overhead)
}
```

6. Remove redundant sentences from the balanced corpus
   6.1 Scan each sentence in the balanced corpus and try to remove the sentence if the remaining sentences still form a cover. Eventually, the set of the remaining sentences is the balanced corpus.

It is noted that in Step 4 we use a heuristic method to construct many balanced corpora. This is because that the number of balanced sentences is highly influenced by the parsing order of the text sentence. If we construct only one balanced corpus in Step 4, the algorithm must be executed recursively to reduce the effect of the sentence parsing order, i.e., by first parsing the balanced sentences obtained from each iteration then the large text corpus. In fact we can combine the heuristic method and the recursive method together to further reduce the number of balanced sentences.

In Step 5, balanced sentences are sorted by the overhead value. In calculating the overhead of a sentence in the balanced corpus, we must first remove out the sentence, and compute the overhead value with the remaining sentences in the balanced or unbalanced corpus.

3. EXPERIMENTS AND DISCUSSIONS

In the experiments, we try to change the number of balanced corpora so that we can determine a reasonable number of balanced corpora used in Step 4. The experimental results is shown in Table 1.

	Number of Balanced Corpora				
	400	600	800	1000	1200
Number of Balanced Sentences	280	215	185	140	140

Table 1: Number of balanced corpora used in Step 4, and number of balanced sentences obtained by the algorithm

Based on the algorithm we have found 140 balanced sentences from a text corpus with 378,964 sentences. As we have mentioned above, the theoretical minimum number of balanced sentence is 114. The average utility rate for balanced sentences

is 81% in our balanced corpus. The utility rate can be further increased by enlarging the text corpus size. For example, a text corpus size with more than 1,000,000 sentences can give higher utility rate because there may include additional co-articulations that do not appear in the small text corpus. The balanced sentences are shown in the Appendix.

As a different balanced corpus can be constructed by changing the text corpus, we can generate many different balanced-corpus sets for collecting the speech database for continuous speech recognition.

4. CONCLUSION

In this paper, a new algorithm for automatic generation of balanced corpus is proposed. The experimental result shows that our algorithm can generate an acceptable balanced corpus with 81% average utility rate. Besides, it is very easy to modify the algorithm to generate a balanced corpus with each balanced information more than one distribution. It is particularly useful for collection of speaker independent speech database. The algorithm can also freely generate numbers of different balanced corpus for different purposes, such as one for training database and the other for testing database. Moreover, our algorithm can be easily applied to other language (English, Japanese, German, etc) by replacing the text corpus, the phonetic table.

5. REFERENCE

1. Ching-Hsiang, Lin-Shan Lee, " An Initial Study on Large Vocabulary Continuous Mandarin Speech Recognition", *Proceedings of ICS 1990*, pp.981-986, DEC 1990.

2. L.R. Rabiner, J.G. Wilpon, and B-H. Juang, "A segmental K-means training procedure for connected word recognition," *AT&T Technical Journal*. 65(3), 21-31, 1986.

3. Lin-Shan Lee, Chiu-Yu Tseng, Hun-yan Gu, Fu-Hua Liu, Chen-Hao Chang, Yueh-Hing Lin, Yumin Lee, Shih-Lung Tu, Shew-Heng Hsieh, and Chian-Hung Chen, " Golden Mandarin (I)-A Real-Time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary ", *IEEE Trans. Speech and Audio Processing*, Vol.1, No.2, pp158-179, 1993.

4. S.M. Wu, J.S. Liau, "On the Creation of Mandarin Phonetic Balanced Sentences, " *Telecommunication Journal*, Vol 19, No 1. Pp.79-87, MAR. 1990.

APPENDIX

Balanced Sentences Listing

這種人不配當嘉義市長
擅長發上網球的諾瓦娜
鑽進賓士車的引擎蓋內
這些文化遺跡將被發掘
害怕再懷孕等心理因素
香蕉不要放進冰箱冷藏
使台北高雄兩市的發展
假資料的情況就非常少
立委質疑法務部掃黑結果是黑槍越掃越多
溝通的第一步當然要先讓所有的人認識你
終場仍未打破廿八點六元關卡
漢翔民營化說明會四日舉行
嫩江大堤雖經多年修護加固
約需費用四到八萬美元不等
現爲台大外文所博士候選人
給玩者不同選擇的多重歡樂
國內業者也曾經嘗試要成立
燙金印刷冒牌純金製品泛濫
軍方過去曾發生飛彈快艇在澎湖海域擱淺
其中規模最大的是香港團
警察抓強盜經常處於下風
謝兩人無緣參與此屆世界錦標賽
顯現逐漸有民眾將廢紙和廢塑膠
是爲配合推動小組三百萬人上網之目標
突波導電陳品豪無大礙
艾瑪跟可愛的外星寶寶
在挪威首都奧斯陸頒發
並要各位主管不要顧忌
傳統布袋戲偶全高約三十公分
免費資源區留言板與討論群
並採用最新實際案例教學法
中華日報服務網搜尋中文台
在籌備工作上花費不少心思
雌虎約每二年至二年半交配
庫長春叢書目錄長春叢書小說類目錄小
火苗迅速從褲腳往上竄燒
北京調整對台策略的原因
章思統乃寧海縣桑州坑口村人
從沒有人畫出一隻逼真的怒虎
這也算是拉斯維加斯深度之旅
開發基金目前並無缺錢的窘境
光華投資及華夏投資擬投資瑞軒科技普通股案
壓迫性骨折併發脊椎損傷性截癱的
另用二大匙油炒芹菜段
爲何海軍會發生爆炸呢
新聞論下新聞論下集錢
部份民眾也丟雞蛋洩憤
不是缺製作軟體就是缺美感
雙方僵持不下談判暫告破裂
各種情況的急救處理狀況等
再加上平均每一位受害兒童
在香蕉山海拔八百公尺的二層坪
但是網路八卦婆告訴大家一個好地方
當局還將透過全面打擊走私等活動來全面

對他們而言更是深具魅力
味道極佳且鴨肉肉質甜美
網友只要填寫好個人資料
轉存央行的轉存款如果要釋出
學雜費基數則依各院所而不同
民進黨長期對許家班的尊敬已經完全破滅
中華隊此次雖在預賽即慘遭淘汰
要求廠方回饋每村五百萬元
有朋自遠方來跨越語言障礙
曾文溪口賞鳥稀世珍鳥
理論與分析的思考能力
白雪公主的條件去篩選合適的對象
噴灑消毒工作由安平區育平里
最難渾沌初開時最難渾沌初開時彭欣予著
國外訂單如雪片滾滾而來
那宗訓那宗訓先生北平人
是您刊登廣告最佳的選擇
等膾炙人口耶誕頌歌跨年音樂會
藤枝森林遊樂區位於高雄縣桃源鄉森濤村旁
歐美學界對神經衰弱的見解
我的兒子和兒媳都非常孝順
在一波三折的第三次投票後
陳亞南執筆元老人病的治療
沖泡牛奶忌用沸水沖泡
病患由於病程變化快速
並附有心臟疾病解剖圖
樊雪春樊雪春得獎作品
想了解各縣市童軍團的概況嗎
家庭美食梁瓊白家庭美食梁瓊白梁瓊白著
中華民國國民創作的短篇小說
配合祠外延平公園的池榭垂柳
而且每天都有最新行情
劉唐綿劉唐綿女士是散文名家劉俠
怎樣養個乖巧快樂的寶寶
十月兩次調降存款準備率後
第二日安排日月潭湖泊之旅
對保持皮膚光滑潤澤很有幫助
還是從通訊錄翻都令人頭昏腦脹
預計明年元月開播後全天候播出
要尋找某個特定的人也是個大困擾
彌陀寺奉祀釋迦牟尼佛
性別男男性別女女主題
另化工股三晃持續下挫
這種恩情與親情似海深
將你的愛心與關懷上網
印尼及越南新娘則在三
希望能藉由網路內容了解資訊界消費新聞
爲全球華人搭起友誼橋樑
莫非也是胡炫亂酷的下場
並宣稱如果開城門後城內平穩
沒幾年在大湖山莊買了樓中樓
爲中華文化增添多采多姿的篇章
而蝴蝶博物館內陳列中外蝴蝶標本
稱讚中華民國行政院長蕭萬長所提
尤其對工作繁忙的父母
德州熱浪造成列車出軌
五戰機才在花蓮擦撞墜毀
她用鋪絨布的提籃代替懷抱
共同促進區域的安全與穩定
散戶往往是扮演被抗殺角色
只有穿著紅色衣服的紅衫軍才能參加
而今雖然各地的孔廟每年也都舉行祭孔大典
學生家長可以選擇最合適的方法與老師對話
如何滿足資訊時代新生活的需求社會化
有些人下定決心好好建檔保存
中華日報用花草美化居家環境
大門用鐵鎖由外面鎖住
甚至與村民們發生衝突
註冊的會員數就高達人
而這也是陳癸淼第三度擔任此項職務
就可讓你如置身蒙地卡羅的一般
橫跨波濤洶湧的吼門水道
詳洽台中干城站南投客運埔里往翠峰
包括不同品牌的產品介紹評薦
但最近連續二場均被對手完封
由於日圓近期持續下降等因素
至今有三十二所學校參加試辦
金融局將儘快公布審查結果
如僅是偶而發生且情況不嚴重
有的作姦犯科淪爲竊賊搶匪
農業局曾輔導北門地區走向休閒觀光漁場
與後來趕到的婆婆緊緊相擁哭泣
聲援遭凌虐的印尼華婦
台灣自一九九六年開始
其中艾旺及文雄各八次
因此額外繳納一些費用