An Efficient Algorithm to Search For A Minimum Sentence Set For Collecting Speech Database

Jin-Song Zhang and Satoshi Nakamura
ATR Spoken Language Translation Research Laboratories.
2-2-2 Hikaridai, "Keihanna Science City", Kyoto 619-0288 Japan
{jinsong.zhang, satoshi.nakamura}@atr.co.jp

Abstract

This paper introduces an efficient search algorithm to find a minimum sentence set for collecting speech database for speech study. The minimum set should have small text size to reduce the collection cost, and cover all the focused phonetic units. The method tries to select a sentence with the lowest cost from a subset corpus, which consists of all the sentences containing at least one token of the least frequent unit of the whole corpus. Compared with other conventional greedy search algorithms, the method successfully achieved a smaller objective set at significantly less computation time.

1 INTRODUCTION

A sentence-set is necessary for collecting a speech database. The design method is to some extent decided by the ratio of the number of focused phonetic units and the size of the set. If the ratio is relatively small, then the set is required not only to cover all the units but also to have an equal appearance of the units. If the ratio is relatively large, then it is difficult to include all the units in a limited set at an equal appearance, as showed in the study of [1]. With the increasing number of phonetic factors in recent speech study, for example from isolate segments to wide context dependent ones, the number of units may increase dramatically. In such cases, it becomes more necessary to find a minimum sentence set to include all units with the focused phonetic factors [2].

Taking the example of Chinese, we may focus not only the segmental events, but also the influences from tonality. There are about 60 demi-syllabic units (Initials and Finals), among them 40 Finals each may carry 5 tones. So the number of isolate segments is about 220. If the neighboring coarticulation effects need to be considered, the number of triplet units may amount to millions. Besides the method of merging factors based on acoustic characteristics as done in [3] to reduce the number of focused factors, an efficient search method is also necessary to control the size of the minimum set.

The problem to find a smallest subset containing all units of a large set is known as set-covering problem [4], which does not have a solution in polynomial time. The most popular approximate answer is known as *greedy algorithm* [4]. The standard greedy search algorithm tries to find the objective subset step by step, selecting the sentence with the lowest cost at each step. When the original set consists of a large number of sentences, the locally optimal selection by the standard greedy search may not easily find a compact enough objective subset. Furthermore, the computation cost is in exponential relation to the size of the original set. For an objective set containing N sentences, the search process approximately needs M^N evaluations of sentence costs when the original set has M sentences. In a realistic task of building Chinese tri-phone minimum sentence set, the standard greedy search usually costs tens of hours to find a result on a normal computation platform.

In [4], the authors proposed to associate meaningful weights with the units, and generalize the greedy algorithm to select sentences based on the sum of the weights of the unique units contained in a sentence. When assigning weights equal to the inverse of unit frequency in the text corpus, the authors found that the algorithm would tend to select sentences with "hard-to-get" units first, and the more frequent units are encountered as a by-product. The resulting set usually has less number of sentences than that generated by the standard greedy algorithm.

Although the weighting method in [4] has the effect of reducing the size of generated objective sentence set, it has no effect of reducing the computation cost. Based on it, we proposed in [5] a modified greedy search, named as least-to-most greedy search, which selects a sentence only from those sentences containing the unit of least frequency. The method was found to be able to generate a smaller objective set at significantly lower computation cost when compared to the standard greedy search.

This paper is designated to give a comparison study of the proposed least-to-most search and the standard and the weighted greedy algorithms. Experiments are carried out to cover 4,854 Chinese class triphones from an original set of 500,000 sentences.

2 GREEDY SEARCH ALGORITHMS

The greedy search algorithms we study here includes the following four methods:

- **Method 1**: the standard greedy algorithm.

ISBN 1-876346-48-5

  - Step 1: A = {all original sentences.}, B={null}, U={unit list to be covered.}
  - Step 2: compute covering score s_i for each sentence i according to the following formula.

$$S_i = \frac{\text{Types of uncovered units in } i}{\text{Total tokens of units in } i}$$

  - Step 3: Select the sentence s_h with the highest score and insert it into B, then delete all newly covered units in s_h from U.
  - Step 4: do step 1, 2, 3 iteratively until U becomes null or all s_i equal zero.
  - Step 5: B is the objective minimum sentence set.

- **Method 2**: the weighted greedy algorithm by the inverse of frequencies.
  - Step 1: compute unit weight w_j for each unit j in U.

$$w_j = \frac{1}{\text{Frequency of the unit } j \text{ in } A}$$

  - Step 2: compute weighted sentence score.

$$S_i = \frac{\sum_{\text{Types of uncovered units in } i} w_j}{\text{Total tokens of units in } i}$$

  - Step 3-5: the same as *Method 1.*

- **Method 3**: This method and the next one are based on the proposal to build a text subset with each sentence containing at least one token of the least frequent uncovered unit, then to do sentence selection from the subset [5].
  - Step 1: For any unit u_k in U, A_{u_k}={All sentences containing at least one token of u_k }.
  - Step 2: Put all the to-be-covered units in U to a queue in order, $Q = \{u_1, u_2, \ldots, u_w\}$, where u_1 is the least frequent unit and u_w the most frequent one in A.
  - Step 3: From the sentence subset A_{u_1}, use the *Method 1* to find a best sentence s_h and insert it into B.
  - Step 4: Delete all the newly covered units in s_h from the queue Q.
  - Step 5: Redo Step 3-4 until the queue becomes empty.

- **Method 4**: Applying the weighting method of Method 2 to Method 3. All the steps are the same as *Method 3* except the step 3,
  - Step 3: the *Method 2* is used to find the best sentence s_h.

3 EXPERIMENTAL SETUP

The experiments we carried out to test the search algorithms are to find a minimum set covering triphones of standard Chinese Putonghua. A Chinese word is composed of one to several Chinese characters, and each character is pronounced as a monosyllable with a pitch tone. The number of base syllables without tone information is about 410. Traditional Chinese phonology divides the base syllable into demi-syllable units[8]: an Initial and a Final. An Initial is a consonant and a Final may be a vowel, a diphthong, a triphthong, or vowel compound with a nasal ending. Besides, there are also 35 non-Initial syllables with only Finals. As Initials and Finals have been widely used as the basic units to build Chinese speech recognition system, and our purpose to build the minimum sentence set is to collect a training database for building Chinese acoustic models, we adopt the 21 Initials and 37 Finals as the basic units.

3.1 Merged target triphones

The possible number of triphones based on the 21 Initials, 37 Finals and one silence symbols amounts to as many as 111,625 [3]. Hence it is impossible to cover such huge number of units in a reasonable sized sentence set. To solve this problem, we do phonetically merging of the context effects of the basic units. From the phonetic studies, we know that sounds in similar ariticulatory configurations may have similar coarticulation influences on their neighboring sounds. The merge does not likely lead to loss of acoustic information, but can reduce the total number of triphones significantly. In our approach, we classify the 37 Finals into 12 anticipation groups, 10 carryover groups, and merge the 21 Initials into 8 context groups [3]. Such a merge reduced the 111,625 possible triphones into about 5,300 ones, which seems likely to be covered in a few hundreds of sentences.

3.2 Text Corpus Preparation

The text corpuses included two sources: People Daily newspaper and basic travel conversation sentences. Some preprocessing is necessary for preparing the text corpus for minimum sentence set build.

- Sentence segmentation: the original html text is divided into short sentences or phrases whose lengths are no longer than 30 characters. The divisions were basically based on a hierachical definition of punctuation marks. The final text corpus for study consists of 500,000 randomly selected distinct sentences from one year's newspaper.
- Word segmentation and pronunciation symbol conversion: the continuous Chinese sentences are inserted with word boundary information, then converted into Initials and Finals according to a pronunciation lexicon. Intra-sentence punctuation marks were converted into silence symbols.

ISBN 1-876346-48-5

- Class triphone generation: the monophone Initials and Finals are converted into class triphones based on the above-mentioned context merge.

Table 1 gives the detailed information about the text corpus used in the following experiments. As only 4,854 triphone types of 5,300 appeared in the text corpus, the following experiments are to search a minimum sentence-set which covers all the 4,854 triphone units.

Number of sentences	500,000
Number of characters	5,487,398
Characters per sentence	10.97
Number of triphones	12,157,448
Phones per sentence	24.31
Types of triphones appeared	4,854

Table 1: Statistics of the text corpus.

4 EXPERIMENTAL RESULTS

Method	# of sent.	# of char.	Avg. length
1	1,727	13,357	7.73
2	1,487	11,243	7.56
3	1,127	12,470	11.06
4	1,437	11,270	7.84

Table 2: Statistics of the resulting minimum sentence sets of the four methods.

4.1 Size of The Generated Sets

Table 2 gives the size statistics of the generated objective sentence sets by the four methods with respect to the number of sentences, the number of characters and average sentence length in characters. From it, we can see,

- The generated set by Method 1, the standard greedy algorithm, has the most sentences. Whereas the one by Method 3, the least-to-most search algorithm, has the least sentences among the four results. The other two have similar number of about 1,400 sentences, less than that of Method 1 but more than that of Method 3.
- When the total number of characters is evaluated, Method 1 got the most characters. Method 2 and Method 4 got the fewest characters. Method 3 got the intermediate.
- Sentences found by Method 1, 2, 4 have averagely similar lengths. But those by Method 3 had averagely 3 more characters per sentence.

Discussion: If least number of sentences is preferred for the minimum set, then Method 3 achieved the best performance. If least characters is desired, Method 2 or 4 got the best performances.

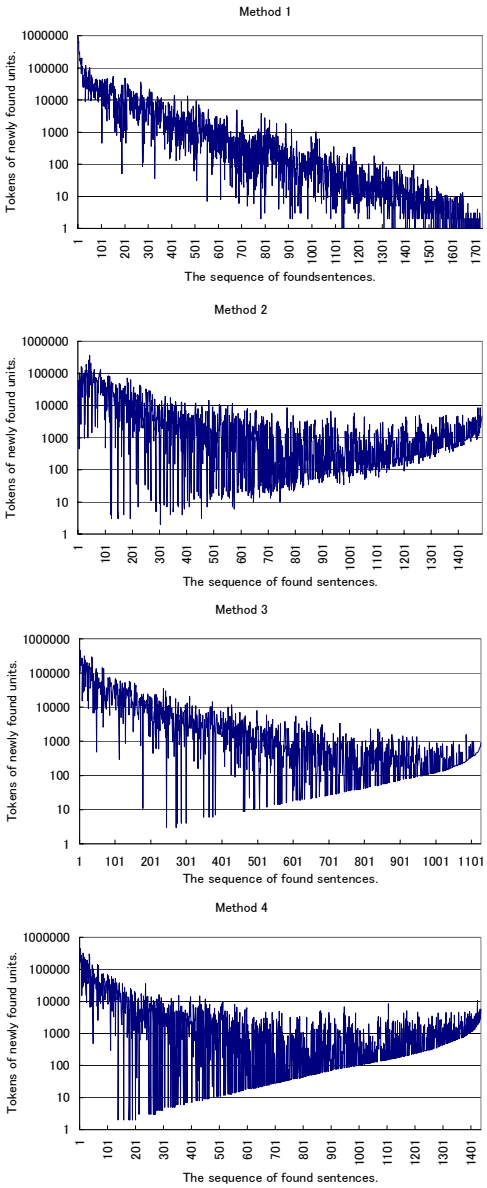

Figure 1: The relation between the found order and the unit frequencies.

4.2 Search Analysis

To understand the differences of the four methods, we plotted the histogram of newly covered units of sentences in the order of found sequence in Fig. 1. What we can see includes,

- For the standard greedy search, the figure clearly shows that the found order is potentially related to the unit frequencies in the original text corpus: more frequent earlier, less frequent later. Therefore, sentences found later are to cover hard-to-get units, possibly leading to more redundancy.
- For the other three methods, the figure clearly shows that they all tend to cover the hard-to-get units in an earlier stage. After they are covered, other frequent and easy-to-get units can be found more efficiently by less number of sentences.

ISBN 1-876346-48-5

- The histogram of Method 3 is slightly different from those of Method 2 and 4. If a sentence of high value in the histogram is considered to be more representative of the text corpus than one of low value. The methods of weighting by the inverse of token frequency, i.e., Method 2 and Method 4, tend to select the less representative sentences much earlier than Method 3.

- The histograms of Method 2 and Method 4 are similar.

Figure 2 illustrates the searching courses of the experiments, with the lower one emphasizing the first 400 sentences. We can see the differences: although the standard greedy algorithm found more units than the others in the early stage, it decreased the covering rate in the later stage. The other three methods keep on a rapid rate until it finds all units.

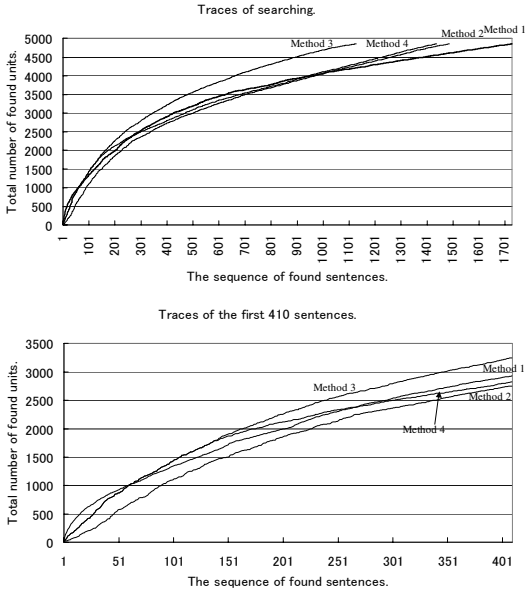

Figure 2: Illustration of the searched courses of the four methods.

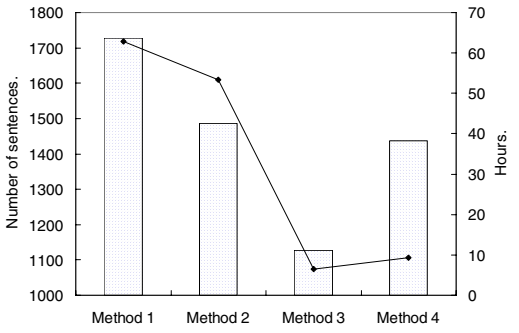

Figure 3: Number of sentences in the objective sets (Blocks) and the computation time on an 1 GHz Pentium CPU (Dots).

4.3 Computation Costs

For Method 1 and Method 2, almost all the sentences in the text corpus are evaluated to renew the covering scores at each step during the search. Whereas for Method 3 and Method 4, only a small fraction of the whole corpus are evaluated. So the computation costs for the proposed Method 3 and 4 are much less than those of Method 1 and 2. Experimental results also confirmed this as illustrated in Fig. 3. On a platform of 1 GHz Pentium CPU, the running time of Method 1 and Method 2 is more than 60 and 50 hours respectively, and those of Method 3 and 4 are less than 10 hours.

5 Conclusion

Among the four greedy methods, Method 3 achieved the least number of objective sentences, Method 2 and Method 4 got the least number of characters. Method 3 and Method 4 have much less computation than Method 1 and Method 2. Therefore, the proposed algorithm of searching from least to most was not only efficient at finding smaller sentence sets but also efficient at reducing significantly the computation costs.

Acknowledgement: This research was supported in part by the telecommunications Advancement Organization of Japan.

References

[1] J.-S. Zhang, T. Matsui and S. Nakamura, "A design of Chinese phonetically-balanced sentence set", Proc. of ASJ, March. 2001, pp.189-190.

[2] J. van Santen, "Diagnostic perceptual experiments for text-to-speech system evaluation", Proc. of ICSLP92, pp.555-558

[3] J.-S. Zhang, S.-W. Zhang, Y. Sagisaka and S. Nakamura, "A hybrid approach to enhance task portability of acoustic models in Chinese speech recognition", Proc. of Eurospeech 2001, pp.1661-1664.

[4] J. van Santen and A. Buchsbaum, "Methods for optimal text selection", Proc. of Eurospeech 97, pp.553-556.

[5] J.-S. Zhang and S. Nakamura, "Least-to-Most ordered search for minimum sentence set for collecting speech database", Proc. of ASJ, Oct. 2001, pp.145-146.

[6] H. Francois and O. Boeffard, "Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem", Proc. of Eurospeech 2001.

[7] H. Francois and O. Boefard, "The greedy algorithm and its application to the construction of a continuous speech database", Proc. of LREC 2002, pp.1420-1426.

[8] Z.-J. Wu et al,"Xian dai han yu yu yin gai yao", Sinolingua Press, Beijing, 1992.