AN INCREMENTAL APPROACH TO SELECTION OF WELL BALANCED CORPUS

Li Ming[1] , Jochen Junkawitsch[2] , Tiecheng Yu[1]
[1]Institute of Acoustics, Chinese Academy of Sciences, P.R.China
[2]Siemens AG, Corporate Technology, 81730 Munich, Germany

ABSTRACT: While achieving a large vocabulary speech recognition system, we often need appropriate corpora for training, test and initialization. Designing those corpora manually is very time consuming. So they are usually generated automatically. But by the common generation method, there are always some phonemes which are not balanced well. In this paper, we present a novel approach by which we can get very good corpora which are well balanced for all phonemes. In this approach we adopt an incremental strategy to obtain the corpus, namely obtain the whole corpus part by part. We also put forward a new method to evaluate sentences in the data source, by which we can select data more effectively. In our experiments, this approach achieves a significant improvement for the quality of the selected corpus.

1. INTRODUCTION

For large vocabulary speech recognition research, it is often necessary to obtain all kinds of corpora, such as training corpus, test corpus, initial corpus and so on. Usually different kinds of corpus have different request for the distribution of phonemes. For example, test corpus for recognition should have natural distribution which is the same as that of the original huge data source, while initial corpus which is used to initialize HMMs should have even distribution for all phonemes. In the following, we will discuss how to get evenly balanced corpus in others' paper first, then we will give the method how to get other distribution corpus.

In (Wang, 1993), (Gao, 1995) and (Shyuu, 1998), they have done some research on this respect. In (Gao, 1995), first they define a weight for every phoneme as Equation (1).

$$w(u_i) = 1 - p(u_i) \qquad (1)$$

where, u_i is the i^{th} phoneme in the phonemes set $U = \{ u_1, u_2, \ldots , u_L \}$, there are L kinds of phonemes, $p(u_i)$ is the percentage of phoneme u_i in the data source, then they evaluate all sentences in the data source by an evaluation equation which is similar to (2). The score shows the average rareness degree of all phonemes in a sentence. The higher the score is, the rarer those phonemes are.

$$E(s) = \frac{1}{K} \sum_{j=1}^{K} w(m_j) \qquad (2)$$

Where, s is the sentence to be evaluated, K is the number of all phonemes in the sentence, m_j is the j^{th} phoneme in the sentence. Finally, they get adequate sentences according to their scores by one selection. But we found that the corpus obtained by this method is always not well balanced for some phonemes. We think it seems impossible to get a well balanced corpus by once selection. Therefore, we put forward an incremental approach which obtains the whole corpus part by part. In this approach, each time it finishes choosing part of data, it will automatically adjust the weight of every phoneme according to its occurrence frequency in the current corpus and change the selective emphasis on the phonemes which are relatively inadequate in the current corpus. In this way, we can get the entire corpus step by step and the final corpus will be well balanced for all phonemes.

This paper is organized as following. In section 2, we give details about our incremental approach. In section 3, we describe the refined evaluation method. Experiment results are presented in section 4. The summary is given in section 5.

2. INCREMENTAL APPROACH FOR SELECTION OF WELL BALANCED CORPUS

Our approach can be described briefly as follows. First we initialize the weights of all phonemes according to their data amount in the data source. Then we obtain the corpus part by part. We get the evaluation scores of all sentences by an evaluation equation. After that we sort sentences by their scores in descending order. Then we select sentences from top to bottom until we get enough data for this time. After this selection, if all wanted data is obtained, it stops. Otherwise it adjusts the weight of every phoneme and goes to next selection. For example, if we want to get a 100-sentence corpus from a 1000-sentence data source, we can select 50, 30, 20 sentences from the data source for the first, second and third selection respectively. So we obtain the 100-sentence corpus by three selection. Fig.1 shows the main steps of our method.

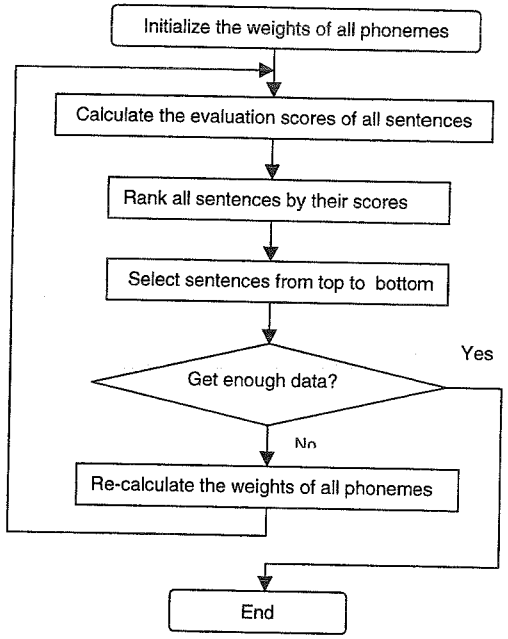

Fig.1 The flowchart of the incremental approach

2.1 Initialization of the weights of all phonemes

Usually we need some corpus have different distribution. For example, we expect the initial corpus has the even distribution and test corpus for recognition has the same distribution as that of the data source. Equation (3) is to initialize the weights of all phonemes which can help us to get the corpus with the expected distribution.

$$w(u_i) = \left[\frac{r_{\max}}{r(u_i)} \right]^{\varepsilon} \qquad (3)$$

Where, $r(u_i) = p(u_i) - g(u_i) + \alpha$, $g(u_i)$ is the distribution you want, if you need even distribution, all phonemes should have the same values of $g(u_i)$, α is a constant which prevents $p(u_i)$ below zero, $r_{\max} = \max_{1 \le i \le L} r(u_i)$, ε is a constant between 0.5 and 0.8 which is chosen empirically.

2.2 Calculating the evaluation scores of all sentences
We still use (2) as our evaluation equation. The sum of all phonemes in the sentence is divided by the number of phonemes. So the evaluating scores are irrelevant with the length of sentences.

2.3 Ranking all sentences by their scores and choose top part of data
We sort all sentences by their scores in descending order. Then the top sentences of the data source are picked out and added to the current corpus.

In practice, the data source is very big so that it takes much time to sort all sentences but we only need the rank order of the top ones. Thus, it's not necessary to rank them all. It will do if we obtain the rank order of the top N sentences. N should be big enough to ensure we can get enough data. After current selection, if all needed data is obtained, it stops, otherwise it goes to next selection. When each selection is finished, the selected sentences will be removed from the data source.

2.4 Adjusting the weights of phonemes
Now we can adjust the weights according to their data amount in the current corpus so that we can change our selection emphasis. Those phonemes whose weights are quite high already become relatively adequate, so their weights should decline. While those phonemes which are not sufficient relatively in the current corpus should increase their weights. The adjusting weight equation is similar with the initial weight equation (3). But the statistical range of phonemes number is within the current corpus instead of the data source.

$$w(u_i) = \left[\frac{r_{\max}^{(k)}}{r^{(k)}(u_i)} \right]^{\varepsilon} \qquad (4)$$

Where, $r^{(k)}(u_i) = p^{(k)}(u_i) - g(u_i) + \alpha$, $p^{(k)}(u_i)$ is the percentage of phoneme u_i in the corpus after k^{th} selection, $r_{\max}^{(k)} = \max_{1 \le i \le L} r^{(k)}(u_i)$.

3. REFINMENT OF EVALUATION METHOD
In our experiments, we find that some sentences which contain several rare phonemes don't have high evaluation scores. These sentences don't have high possibility to be selected, which makes the result not very good. Observing these sentences, it can be found that most of these sentences are quite long. Though a phoneme which has a high weight, it will still account for a small proportion in the evaluation score because the evaluation score is the average weight of phonemes in a sentence. Therefore we should try to enhance their impact on the evaluation scores. But if we simply increase the weights of those rare phonemes, our experiment shows it doesn't work well. So we try to expand the proportion of rare phonemes' weights in the evaluation scores. Our method is like as flollowing. First we sort all phonemes in a sentence by their weights in descending order, then distribute the proportion of each phoneme in the evaluation score according to its rank order in the sentence. Finally, we sum up all weights and divide the sum by the factor Q, then we get the score. Our method can be described as Equation (5).

$$E(s) = \frac{1}{Q} \sum_{j=1}^{K} w(m_j) q^{Rank(m_j)} \qquad (5)$$

Where, q is a constant q = 0.55 ~ 0.85 which is also chosen empirically, $Q = \sum_{n=1}^{K} q^n$, Rank(m_j) is the rank order of phoneme m_j according to its weight in descending order in the sentence. In this way, those phonemes which have high weights will always have great effect on the evaluation scores.

4. EXPERIMENT RESULT

In our experiments, we would like to get a even distribution corpus comprising of about 300 sentences derived from a data source which contains about 6000 sentences. In the experiments, we use standard deviation of phonemes occurrence frequency in the selected corpus as the indicator of the quality of the corpus we obtain.

Let $$\overline{p} = \frac{1}{L}\sum_{l=1}^{L} p(u_l) \qquad (6)$$

Where, $p(u_i)$ is the percentage of phoneme u_i in the final corpus.

Let $$\sigma = \left\{\frac{1}{L}\sum_{l=1}^{L}[p(u_i) - \overline{p}]^2\right\}^{1/2} \qquad (8)$$

σ is used to measure the quality of the corpus. The smaller σ indicates the better corpus is obtained.

As for the number of selection, empirically the selected data would be balanced better if we get the final corpus by more times' selection. But it takes too long time to get the corpus after many times' selection. We think five times is a good compromise. Table1 shows selection percentage of the corpus we obtain every time.

Selection No.	1	2	3	4	5
Percentage	40%	15%	15%	15%	15%

Table1 corpus selection percentage at every time

Fig.2 shows our experiments result. The value of σ at 0^{th} selection is the standard deviation of the original data source. The figure indicates the value of σ decreases after every selection, in other words, the quality of the corpus become better and better.

In order to compare with the previous method, we also made experiments using the old method(Gao, 1995). We use Equation (2) as the evaluation equation and get the corpus by once selection. Fig.3 shows we get a significant improvement by about 20% reduction of standard deviation compared with the previous method.

5 SUMMARY

For the purpose of obtaining well balanced corpus, this paper proposes a new method to get the corpus by incremental strategy. This method changes the selective emphasis after every selection and get the corpus part by part. We get very good result by this approach. This approach can not only be used for the selection of training corpus and test corpus for speech recognition but also be used for the selection of corpus for speech synthesis research.

ACKNOWLEDGEMENTS
We would like to thank Dr. Hoege Harald for his good advice for this paper. We would also like to thank Dr. Tropf Herbert, Josef G. Bauer and Ute Ziegenhain for their kind help in this research.

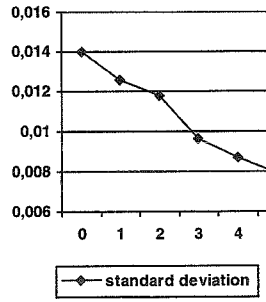

Fig. 2 Standard deviation of the phonemes occurrence frequency of the corpus after every selection

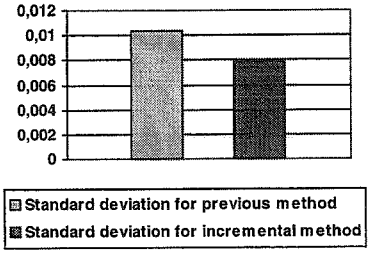

Fig.3 Comparison between the previous method and the incremental method

REFERENCES

Gao Yuqing, Hsiao-Wuen Hon, Zhiwei Lin, Gareth Loudon, S. Yogananthan and Baosheng Yuan (1995) "Tangerine: A Large Vocabulary Mandarin Dictation System", ICASSP'95, Vol.1, pp. 77-80, Detroit, Michigan, USA, May 1995

Jyh-Shing Shyuu and Jhing-Fa Wang (1998) "An Algorithm for Automatic Generation of Mandarin Phonetic Balanced Corpus", ICSLP'98, Vol.7, pp.3175-3178, Sydney Australia, November 1998

Wang Hsin-Min, Chang Yuen-Chen and Lee Lin-Shan (1993) "Automatic Selection of Chinese Syllable-Balanced Sentences from Chinese Text Corpus", pp.195, ROCLING-IV, 1993