USING REFERENCE TO TUNE LANGUAGE MODEL FOR DETECTION OF READING MISCUES

Changliang Liu, FuPing Pan, Fengpei Ge, Bin Dong, Yonghong Yan

ThinkIT Speech Lab., IOA, CAS

ABSTRACT

For a reading tutor, the reference content which the reader reads is known beforehand. This apriori information is very important in automatic detection of reading miscues. This paper proposed two methods to incorporate the reference information into LVCSR framework to improve the performance of miscue detection. The two methods both tune the n-gram Language Model (LM) probabilities dynamically in the decoding process based on the analysis of current reference sentence. The first method weighs the LM probability directly if current n-gram exists in the reference, and the second method takes a liner combination of the original LM probability and the reference probability. The experiments on a Chinese Mandarin reading corpus proved the effectiveness of both methods. The detection error rate and false alarm rate are decreased by 33.1% and 35.5% respectively for the best method.

Index Terms— CALL, LVCSR, reading tutor, miscue detection

1. INTRODUCTION

In recent years, the research on reading tutors has attracted a lot of efforts. A reading tutor is a tool for helping language learners (i.e. children, foreigners) to learn a language more easily. It listens to the learners when they are reading the content aloud on the computer screen, and provides helps when the learner needs[1].

The key technology of reading tutors is speech recognition, however, not conventional speech recognition. In a conventional recognizer, the main task is to recognize what sequence of words the speaker speaks. In contrast, in a reading tutor, the main task is to detect where the reader departs from it, given the text. In other words, it is designed to detect the reading miscues of readers. The reading miscues consist of word substitutions, repetitions, self-corrections, omissions, insertions, etc..

Several research groups have developed different reading tutor systems[1, 2, 3]. All these systems are based on the automatic speech recognition framework and adopt different methods to improve the performance of miscue detection. [1] reconstructed two primary knowledge sources of Sphinx-II [4]: the dictionary of pronunciation and language model. Its

dictionary only contains words of the given text (reference) and some truncations derived from the original words. The refined pronunciation dictionary constrains the search space of the decoder and leads to higher accuracy of recognition, but it will fail when the speaker reads words outside of the dictionary. [2] used a general dictionary and a small language model which was trained only on the reading material (usually a passage). The small language model reduces the perplexity of the search space and acquires good performance of miscue detection. However, if the reading material changes, the language model has to be retrained, which is very inconvenient.

In this paper, we describe the development of a Chinese Mandarin Reading Miscue Detector, the core of a reading tutor, based on large vocabulary continuous speech recognition (LVCSR) system. In the detector, general dictionary and language model are used for reason of generalization and convenience. However, because of inaccuracy of LVCSR, the baseline performance is not satisfactory. The reference which is known beforehand is important apriori information. However, the conventional speech recognizer doesn't consider this knowledge source in its core process. In this paper, two methods to incorporate the reference into the decoder of LVCSR are attempted. As a pilot study, all the methods are implemented by tuning the original language model.

The rest of this paper is structured as follows: section 2 describes the architecture of the baseline detector system; section 3 describes two methods of using the reference to tune the language model; some experiments and results for these methods are presented in section 4; and finally in section 5 some conclusions are given.

2. THE BASELINE SYSTEM

The Chinese Mandarin reading miscue detection system is built on a conventional LVCSR system shown in figure 1. The decoder translates the speech into text, given the three knowledge sources: acoustic model (AM), language model (LM), and pronunciation lexicon (PL). In this system, the text is not a simple sequence of words. Because the decoder can reserve much rich information about the real content in the speech and this information is very helpful in detecting reading miscues, we use confusion network—a more complex structure

as the output of LVCSR to keep the rich information. It is a compact lattice format whose structure is called consensus hypothesis [7]. The complete system architecture is shown in Figure 1.

Fig. 1. The architecture of baseline reading miscue detector

In order to generate final detection results, a dynamic program is used to align the reference sequence and the confusion network. Then, the omissions, insertions and substitutions can be found in the aligned path. The rich information in the confusion network provides more aligning choices in dynamic program, thus can lead to a more accurate aligned path and better detecting performance.

3. USING REFERENCE IN DECODING

A conventional LVCSR has three knowledge sources[6]: the pronunciation lexicon (PL), the acoustic model (AM) and the language model (LM). The decoding process is to find the best path that matches the speech stream in the circular lexicon tree based on the knowledge of acoustic model and language model.

For a reading tutor, the reference content which is expected to be read is known. Though the real content the reader reads may not match the reference exactly because of reading miscues, it is reasonable to assume that most of words match. Therefore, the words in the reference have higher probability to be read. These words should be highlighted by the decoder. The decoding process should be restricted near away from the reference. In a new architecture, the reference is incorporated as another knowledge source besides the PL, AM and LM. The new architecture of reading miscue detector is shown in Figure 2.

Fig. 2. The new architecture of reading miscue detector

The probability of a word in decoding process can be de-

scribed in equation (1).

$$P(w|o) = P_{AM}(o|w) \cdot P_{LM}(w) \quad (1)$$

Where, $P_{AM}(o|w)$ is the acoustic model probability of word w given the observation vector o , and $P_{LM}(w)$ is the language model probability of word w .

In order to highlight the reference words in decoding process, a simple method is to strengthen their LM probabilities. The decoder checks whether the current word exists in the reference, if yes, strengthen it; if no, do nothing. To strengthen the words, two methods are proposed in this paper.

The first method is to weigh LM probability of word w by directly multiplying a coefficient, as is shown in equation 2. In the decoding process, if current word w exists in the reference, the LM probability $P_{LM}(w)$ will be changed to $\alpha \cdot P_{LM}(w)$. α can be decided by experiments. This method is referred to as method-1 below.

$$P_{LM}(w) = \begin{cases} \alpha \cdot P_{LM}(w) & w \in \text{reference} \\ P_{LM}(w) & w \notin \text{reference} \end{cases} \quad (2)$$

In method-1, the completeness of language model was not considered. After multiplying α to the LM probability of w , the whole probability of all words does not equal to 1. The probability (the frequency) of w in the reference was not considered either. Considering the two items described above, the second method is proposed and given in equation 3.

$$P_{LM}(w) = (1 - \alpha)P_{LM}(w) + \alpha P_{ref}(w) \quad (3)$$

Where, $P_{ref}(w)$ is the probability of w in the reference, and α is a coefficient between 0 and 1.

$$P_{ref}(w) = \frac{\text{count of } w \text{ in reference}}{\text{count of all words in reference}}$$

This method is referred to as method-2 below.

Generally, the language model is structured as n-grams. When checking words in the reference, we also use n-grams as basic unit. In practice, for reason of convenience, we extract all n-grams in the reference firstly and generate an n-gram list before decoding. Then the decoder checks if the current n-gram exists in this list during decoding. Because of reading miscues, some words in the reference may not be read completely, in other words, some characters (syllables) of the word are misread. Only extracting the word n-grams is not enough to cover these conditions. Therefore, when extracting the n-grams, not only the word grams are considered, but also the characters. The process of extracting the n-grams from the reference is as follows. Firstly, split the whole sentence of reference into word sequence according to the lexicon; secondly, expand the word into characters and merge them into the word sequence and construct a reference word graph; thirdly, traverse the graph, generate all the n-grams. Figure 3 shows the process with an example how a reference sentence is transferred into a tri-gram list. In this example, the tri-grams are extracted. In the experiments of this paper, only tri-gram is tested.

Fig. 3. Transferring reference into a tri-gram list

4. EXPERIMENTS

4.1. Corpus and method

The speech corpus with reading miscues is not easy to collect and to transcript the miscues is a time-consuming process. For a pilot study of the detection system, a special testing corpus was constructed. We select some sentences from Compendium of Chinese Mandarin Level Test (PSC)[5]. And, for each sentence, some insertions or deletions were manually added at random positions to simulate the reading miscues. Then, these sentences with miscues were read by 15 males and 13 females. All the readers are Chinese native speakers. Finally, a corpus with about 5000 sentences was collected. There are about 10% insertion miscues and 10% omission miscues in this corpus. The task of our detection system is to detect the simulated reading miscues. We can have a glance at the performance of the system from this simulated database even though we have not a speech corpus with real reading miscues.

To evaluate the performance of the proposed methods, three measures are used: Character Error Rate (CER), Miscue Detection Error Rate (MDerr) and False Alarm Rate (FARate). CER is a common measure for evaluating the performance of LVCSR. MDerr is defined as the number of miscues which have not been detected divided by the total number of miscues; and FARate is defined as the number of words erroneously detected as read incorrectly divided by the total number of miscues[2].

4.2. Experiment Setup

The front end of the system extracts 39 dimensions of MFCC feature, including 12 dimensions of static cepstrum and 1 dimension of energy, with their 1st and 2nd derivatives. The HMM acoustic model was trained on about three hundred hours of speech from Chinese native speakers. The final model has about five thousand states and each state has 32

Gauss mixtures. The language model is a tri-gram model which was trained on about 2G bits materials from news and web. The dictionary has about forty thousand words, including most words of Chinese language.

4.3. Performance of using reference

The two methods proposed to incorporate the reference into decoding in section 3 are investigated in this section. A tri-gram language model was used in these experiments.

For method-1, we tested different α to find a best value for detecting miscues. Figures 4 show its performance for different α . It can be seen in Figure 4(a) that when α changes from 0 to 10^6 , the CER decreases acutely. This proves the effectiveness of this method to restrict the decoding path around the reference. Figure 4(b) shows that MDerr and FARate both decrease when α is changing from 0 to 10^4 . After 10^4 , the MDerr begins to rise again. This is not difficult to explain. We highlight the reference in the decoding process to restrict the decoding path. However, the readers may not read the reference exactly. There are some miscues such as insertions and omissions in their utterance. Only the reference can't describe these miscues. When the restriction is appropriate, the improvement of the recognizing accuracy makes the alignment more accurate, thus can decrease the MDerr and FARate at the same time. When the restriction is too tight, the miscues may be largely smoothed. Thus, when α is too large, the MDerr begins to rise again. It should achieve a tradeoff when using the restriction of reference.

Fig. 4. Performance of method-1 at different α

For method-2, different α was also tested in the experiments. The performance is shown in figures 5. It is similar to that with method-1.

Table 1 compares the performance of method-1 and method-2. Method-2 is much better than method-1. That's because method-2 considers the completeness of language model and the probabilities of words in the reference. It is theoretically more reasonable. And it uses more information of the reference than method-1.

Fig. 5. Performance of method-2 at different α

Table 1. Performance of different methods

	CER	MDerr	FArate	Relative Improvement (MDerr/FArate)
Baseline	29.6%	11.13%	8.89%	—
Method-1	7.4%	9.1%	6.24%	18.2%/29.8%
Method-2	6.2%	7.45%	5.73%	33.1%/35.5%

5. CONCLUSION

In this paper, we describe how to use the important apriori information of reference in the detection of reading miscues. Our reading miscue detector system is based on a conventional LVCSR system. By only aligning the reference with the output of the LVCSR, this apriori information is not sufficiently used. We proposed two methods to use the reference as another knowledge source of decoder. The first is to weigh the LM probability directly when the current n-gram exists in the reference. The second method linearly combines the original LM probability and appearance probability of the reference and it considers the completeness of LM. The second achieves better performance than the first. The experiments on a Chinese Mandarin reading corpus proved the effectiveness of both methods. The detection MDerr and FArate are decreased by 33.1% and 35.5% respectively for the second method.

6. ACKNOWLEDGEMENT

This work is partially supported by The National High Technology Research and Development Program of China (863 program, 2006AA010102, 2006AA01Z195), MOST (973 program, 2004CB318106), National Natural Science Foundation of China (10574140, 60535030).

7. REFERENCES

[1] Jack Mostow, Steven F. Roth, Alexander G. Hauptmann, and Matthew Kane, "A Prototype Reading Coach that

Listens," *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, American Association for Artificial Intelligence, Seattle, WA, August 1994.

- [2] Katherine Lee, Andreas Hagen, Nicolas Romanynshyn, Sean Martin, and Bryan Pellom, "Analysis and detection of reading miscues for interactive literacy tutors," *Proc. International Conference on Computational Linguistics*, Coling, Geneva, Switzerland, Aug. 2004.
- [3] Acques Duchateau, Mari Wigham, Kris Demuyne and Hugo Van hamme, "A Flexible Recogniser Architecture in a Reading Tutor for Children," *Proc. ITRW on Speech Recognition and Intrinsic Variation*, Toulouse, France, pp. 59-64, May 2006.
- [4] X. D. Huang, F. Alleva, H. W. Hon, M. Y. Hwang, K. F. Lee, and R. Rosenfeld, "The SPHINX-II speech recognition system: An overview," *Computer Speech and Language*, 7(2), pp. 137-148.
- [5] *Compendium of Chinese Mandarin Level Test*, The Commercial Press, China, 2006.
- [6] Jian SHAO, Ta LI, Qingqing ZHANG, Qingwei ZHAO, and Yonghong YAN, "A One-Pass Real-Time Decoder Using Memory-Efficient State Network," *IEICE TRANSACTIONS on Information and Systems*, Vol.E91-D No.3 pp.529-537.
- [7] L. Mangu, E. Brill and A. Stolcke, "Finding onsensus in speech recognition: word error minimization and other applications of confusion," *Computer Speech and Language* 14(4), 373-400, October 2000.