Automatic Error Correction for Repeated Words in Mandarin Speech Recognition

Xiangdong Wang, Hong Liu, and Yueliang Qian
Institute of Computing Technology, Chinese Academy of Sciences, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Beijing, China
Email: {xdwang, hliu, ylqian}@ict.ac.cn

Xinhui Li
Tencent Inc., China
Email: hiccupli@tencent.com

***Abstract*—In this paper, an approach of automatically correcting recognition errors of repeated words is proposed by exploiting recognition results of preceding utterances. During the error correction, the words that might appear again in the following utterances are collected from the recognition results of preceding utterances. For each utterance, there are four steps involved in the correction: 1) initial recognition. In order to correct recognition error of the repeated word, character confusion network (CCN) is adopted as the result of initial Mandarin recognition, 2) detecting repeated words by computing the phonetic similarity between collected words and the CCN, 3) correcting recognition errors of repeated words automatically, and 4) extracting new words from the recognition result of the current utterance. Experiments show that more than 5% absolute character error rate (CER) reduction can be achieved using the proposed method.**

***Index Terms*—speech recognition, error correction, repeated word, confusion network**

I. Introduction

The past several decades have seen significant progress in Large Vocabulary Continuous Speech Recognition (LVCSR) technologies, and applications such as writing e-mail by speech and speech translation are becoming practical on mobile terminals. However, because the input speech is not necessarily consistent with the acoustic and language models used for LVCSR, it is impossible to avoid recognition errors. Thus, for increasing the usability of speech recognition, an effective scheme for correcting recognition errors is important in addition to improving the performance of speech recognition.

Currently, the task of LVCSR mainly focuses on three fields: speech dictation, speech transcription and dialogue speech recognition. No matter what kind the recognition task belongs to, the speech to be recognized basically has a theme and some content words related to the theme appear repeatedly in the speech. We call these words repeated words. For a repeated word, since the context and pronunciation varies much in different appearances, the recognition results of the same repeated word might be different in different utterances. Even though it is recognized correctly at its first appearance, it still might be recognized falsely in the following utterances. In the existing error correction techniques [1]-[4], each time the repeated word is recognized falsely in the task, the user will be required to correct the recognition error. If there are a lot of words which are difficult to be recognized correctly and appear repeatedly in the task, such as professional terms and name entities, correcting these errors using the existing error correction techniques will cost the user a lot of time.

In order to improve the efficiency of error correction, we propose an approach which automatically corrects recognition errors of repeated words by using their prior recognition results and correction results. In this method, when the word first appeared in the speech, its recognition result and correction result will be saved in a word template database. For each utterance, there are four steps involved in the correction: 1) recognizing the speech with a well-trained recognizer, 2) detecting repeated words by computing the phonetic similarity between recognition result and word templates, 3) correcting recognition errors of repeated words automatically, 4) extracting new words using the text after manual error correction or extract new words without any additional correction, and save these words into word template database.

Although the recognition result of the same word might be different in different utterances, the pronunciations of these different results are quite similar. For an utterance, if a word has a phonetic similarity with part of its recognition result, this word might appear in this utterance. In the proposed method, we adopt character confusion network (CCN) as the result of initial recognition. Compared to 1-best recognition result, the CCN contains a lot of intermediate results besides the best result. Thus, using the CCN to compute the phonetic similarity is more reliable than just using 1-best result.

Manuscript received December 21, 2014; revised May 8, 2015.

doi: 10.12720/joace.4.2.153-158

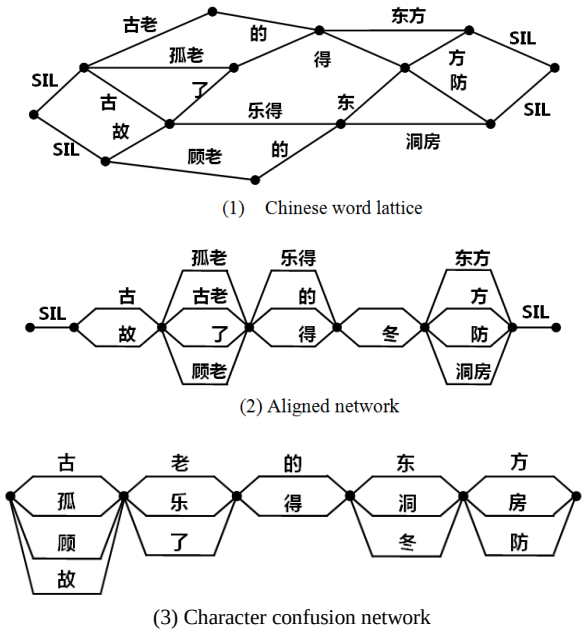
The rest of the paper is organized as follow. In Section II, the CCN used for repeated word detection and error correction is introduced. Section III presents the proposed method of automatic error correction for repeated words. Experimental results are given in Section IV. And finally, conclusions are drawn in section V.

II. Character Confusion Network

Confusion network (CN) is a sample network representing the intermediate recognition result and is obtained by condensing a huge word lattice. The CN was introduced in the word error minimization algorithm which minimizes the word error of the recognition result rather than the sentence error rate [5]-[7]. Unlike the English words, a Chinese word may be either a single Chinese character or a Chinese character string. For example, “中国” (China) is comprised of two characters and the character “中” (middle) itself could be a Chinese word. This difference will cause the Chinese word lattice to be different from the English lattice. Therefore, current CN algorithms [5]-[7] cannot be used directly to obtain a Chinese CN in mandarin LVCSR.

According to the characteristic of Chinese language, we proposed a CN algorithm [8] in our earlier work which transforms the Chinese word lattice into character confusion network (CCN). This algorithm is simply summarized as follows:

1. Compute the posterior probability of each word (link) in the Chinese word lattice by using forward-backward algorithm [5].
2. Build aligned network. The links with overlapping time and similar pronunciations of their last characters are aligned into one class.
3. Split the words of aligned network into characters, and put each character into different classes according to the starting time and phonetic similarity with each class.
4. Merge the same characters in each class.



(1) Chinese word lattice

(2) Aligned network

(3) Character confusion network

Figure 1. Example of character confusion network generation.

Fig. 1 gives an example of CCN generation. In the CCN, the characters that competed with each other are clustered into one class. Each character of the CCN has a posterior probability that expresses a possibility of being the recognition result, and the characters in each class are sorted by the value of the posterior probabilities. Besides the posterior probability, each character has a few combined probabilities which express the possibility this character combines with other character to form a word. For example, in Fig. 1 the character “古” has a combined probability with the character “老”, and this probability is equal to the posterior probability of word “古老” in aligned network. Since each character may split from a few different words, each character has a few combined probabilities. The first character of each class forms the best recognition result.

III. The Proposed Error Correction Method

The strategy of our error correction method is summarized as the flow chart in Fig. 2. For each utterance, the LVCSR system first recognizes the speech and generates a CCN. Then the phonetic similarities between word templates and the CCN are computed, and repeated words in the utterance are detected using the phonetic similarities. According to the detection result, errors of repeated words are corrected automatically. If the system offers an error correction interface, correct the recognition errors of non-repeated words with user involved, and then extract new word templates from the final result. Otherwise, new word templates are extracted directly without manual correction.

A. Word Template

For the first time when a word is recognized correctly or corrected by the user, it is saved in a database. When it appears again and is recognized falsely, the saved result can be used to correct the error automatically. In order to judge whether it appears again in the following utterances exactly, its competitive intermediate result is also saved. The combination of a word and its corresponding intermediate result is referred to as a word template. Fig. 3 shows an example of word template, where each character of the word corresponds to a list of competitive characters.

In the Chinese language, each Chinese character has a lot of homophonic characters and similar-sounding characters. Since a Chinese word is comprised of characters and its pronunciation is formed by joining the pronunciation of its characters, it can be inferred that the more characters it consists of, the less homophonic words and similar-sounding words it will have. In our method, the phonetic similarity is computed between word template and CCN. If the similar probability is greater than zero, we will say the word of the template might appear in the utterance. However, it might be a homophonic word or a similar-sounding word in the utterance, since the recognition results of homophonic words or similar-sounding words are phonetically similar. Therefore, in order to detect the repeated word exactly, the word of the template must be comprised of more than one character.

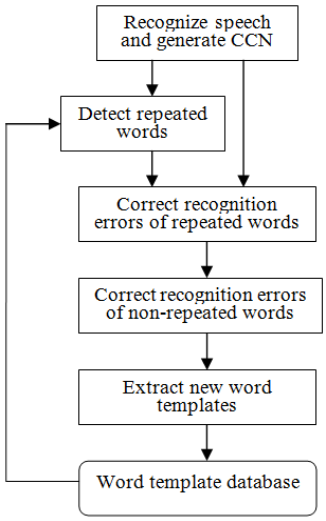


Figure 2. Flow chart of error correction for repeated word.

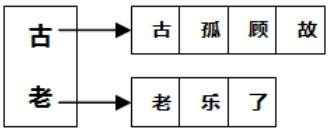


Figure 3. Example of a word template.

B. Detection of Repeated Words

After initial recognition of an utterance, the phonetic similarities between each word template and the CCN are computed. If the similar probability is greater than zero, the word of this template might be a repeated word in the utterance. The phonetic similarity function is defined as

$$p(T,CCN)=\max_{j}\prod_{i=0}^{N-1}SIM(S_i,C_{i+j}) \quad j=0,1,\cdots,M-N \tag{1}$$

where T is a word template, N is the character number of the word in T, S_i is the i-th character set in T. M is the class number of CCN, and C_{i+j} is the $(i+j)$-th class in the CCN. $SIM(,)$ is the phonetic similarity between two character sets and is defined as

$$SIM(C,C')=\frac{1}{2}\left[\frac{1}{N_1}\sum_{i=0}^{N_1-1}\delta(c_i,C')+\frac{1}{N_2}\sum_{i=0}^{N_2-1}\delta(c_i',C)\right] \tag{2}$$

where C and C' are two character sets, N_1 is the character number of C and N_2 is the character number of C', c_i is the i-th character in C and c_i' is the i-th character in C'. For a character c, if there exists a character with the same pronunciation in character set C, the value of $\delta(c, C)$ is 1, otherwise is 0.

The similar probability ranges from 0 to 1, and expresses the possibility that the word of the template appears in the utterance. In (1), the word template corresponds to the same number of continuous classes in CCN, and the repeated word might appear at the position that maximizes the similar probability. Some classes of the CCN may correspond to more than one word template, but it only allows one repeated word appearing at that position. In this situation, the word template with largest similar probability is kept.

C. Correcting Recognition Errors of Repeated Words

After repeated word detection, word templates whose words might appear in the utterance are obtained. Some words of the templates are the real repeated words that appear in the utterance. However, some words of the templates are just phonetically similar with the words that really appear in the utterance. Therefore, the words of the templates cannot be used to replace the best results in the CCN directly.

In fact, the similarity expresses the possibility that the word of the template appears in the utterance, and the meaning of this probability is the same as the posterior probability of character in CCN. Therefore, the word of the template can be viewed as intermediate recognition result and can be inserted into CCN. In order to keep the order of the CCN, the probability of the character should be recalculated and the characters should be re-ranked.

The word of template corresponds to the same number of continuous classes in the CCN, and every character of the word corresponds to one class. Insert every character of the word into its corresponding class and recalculate the probability. The new probability of original character in the class is defined as

$$p(c)=\lambda(1-p_t)p'(c)+(1-\lambda)p'(c) \tag{3}$$

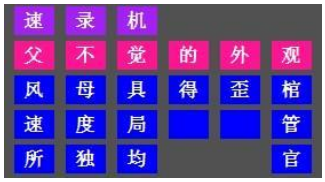
And the probability of the inserted character is defined as

$$p_t(c)=\lambda p_t \tag{4}$$

where p_t is the similar probability of word template, $p'(c)$ is the original posterior probability, and λ is the weight factor. The value of λ ranges from 0 to 1, and varies with the word template. As we know, the longer the word of template is, the fewer homophonic words it will have, and the more reliable the repeated word detection will be. Therefore, the value of λ is raised with the length of word to be inserted.

When the character is to be inserted into its corresponding class, it should be checked whether it is already in the class. If yes, combine the new probabilities of the character. Otherwise, insert the character. Finally, the characters of the class whose probabilities are changed in the CCN are re-ranked according to new probabilities. Through the re-ranking, the recognition errors of the real repeated words are corrected, and the best recognition results of words that are not the real repeated words are not changed. Fig. 4 shows a process of error correction of repeated words.

(1) The user uttered "速录机的外观", and the system generated CCN without any correction.



(2) The repeated word "速录机" was detected after repeated word detection.

(3) The recognition error was corrected automatically by re-rank the CCN.

Figure 4. Example of error correction of repeated word.

D. Extraction of New Word Templates

The key idea of our method is using previous recognition results and correction results to correct recognition errors in the following utterances. Therefore, after recognition and correction of each utterance, new word templates should be extracted from the result. Besides the recognition errors of repeated words, there are some recognition errors of non-repeated words in the best result. Some LVCSR systems offer a correction interface, and allow the user to correct errors through the interface. However, some systems don't offer any correction interface.

If the system offers a correction interface, the user can correct the recognition errors of non-repeated words through the interface. After the user involved correction, there is no error in the best result. The challenge of extracting new word templates is how to extract Chinese words from the best result. Because the best recognition result is a character string and each character can form several different words by combining with neighbor characters. However, the probability of the character combining with neighbor characters approximates to the probability of its corresponding class combining with neighbor classes. Therefore, we can compute the probability of neighbor classes, and maximize the probability to extract words.

$$w_{t,k} = \arg\max_{(s_t, s_{t+1}, \cdots, s_{t+k})} \sum_{i=1}^{num(S_t)} \sum_{j=1}^{num(S_{t+1})} \cdots \sum_{l=1}^{num(S_{t+k})} p(c_{t,i}, c_{t+1,j}, \cdots, c_{t+k,l}) \quad (5)$$

where S_t is the *t-th* class in CCN, $num(S_t)$ denotes the character number of class S_t. $c_{t,i}$ is the *i-th* character in the *t-th* class, and the probability $p(c_{t,i}, c_{t+1,j}, , c_{t+k,l})$ is the combined probability in CCN.

However, some systems do not offer any correction interface at all. In this situation, the recognition errors of non-repeated words will not be corrected and will be left in the best recognition result. Therefore, the words which were extracted from the best result by using (5) might contain errors. In order to exclude the extracted words that contain errors, the proposed method examines the character probability of the extracted word. If every character probability of the extracted word is greater than a threshold, the extracted word will be kept. Otherwise, it will be excluded.

If the left extracted word does not exist in the word template database, the word and its corresponding intermediate results will form a new word template and be added into the database. Otherwise, the intermediate results of the extracted word will be used to renew it existing word template. As the recognition continues, the size of the word template database will increase. Although it covers more words that appear in the speech, it slows the speed of error correction of repeated words. In fact, there are few words appeared repeatedly, and the interval of next appearance isn't long. Thus, the size of word template database can be set flexibly, and the least recently used (LRU) word template will be overwritten when the word template database is full.

IV. Experiments

To evaluate our approach for automatic error correction of repeated words, we conducted two experiments: An experiment without size limit of word template database, and an experiment with size limit of word template database.

In the experiment without size limit of word template database, every word template will be saved in the database and no one will be replaced. We selected 4 articles from the day's Chinese newspaper and asked 4 volunteers to read them in the order they were composed. Each volunteer read one article, and each article contains about 100 sentences. The recognition of each article was an independent task, and the utterances were recognized in three different ways. *Baseline*: all the utterances of the article are recognized without error correction of repeated word. *ECR-unsup*: the utterances are recognized with automatic error correction of repeated word by using our proposed method, but there is no user-involved correction. *ECR-sup*: the utterances are recognized with automatic error correction of repeated word by using our proposed method, and the user corrects the remaining errors after automatic error correction of repeated word.

Table I summarizes the result of the experiment without size limit of word template database. In mandarin LVCSR, the character error rate (CER) is used commonly to evaluate the performance of recognition. From Table I, it can be seen that about 1.7% absolute CER reduction is obtained by using our proposed method without user involved correction. And about 5.6% absolute CER reduction is achieved by using our proposed method with user involved correction. This is because the word templates are more reliable and comprehensive after the user corrected the remaining errors. Note that the CER of the third article reduced largely, since there are a lot of professional terms in the article which are difficult to be recognized and appear repeatedly. Therefore, using our proposed method can reduce the CER effectively especially when the recognition task is not consistent with the acoustic and language models.

TABLE I. EXPERIMENTAL RESULT WITHOUT SIZE LIMIT OF WORD TEMPLATE DATABASE

CER	Article 1	Article 2	Article 3	Article 4	Average
Baseline	21.20%	17.70%	32.10%	13.90%	21.20%
ECR-unsup	19.60%	15.50%	29.70%	13.30%	19.50%
ECR-sup	16.70%	12.00%	21.60%	12.10%	15.60%

For the experiment with size limit of word template database, we set the word template database of different sizes. When the database is full, the least recently used (LRU) word template will be replaced. In order to evaluate the influences on the performance of the proposed method with different size of word template database, the corpus should be long enough so that it can be extracted enough word templates. Therefore, this experiment was carried out on two different Chinese speech corpora, a reading speech corpus and a lecture speech corpus. Each corpus contains more than 300 sentences (about 2400 Chinese words, more than 1500 words of them are different), and each of them is relevant to a topic. In this experiment, the utterances were recognized with error correction of repeated word by using our proposed method, and we corrected the remaining errors after automatic error correction of repeated word. Fig. 5 shows the recognition accuracy with different size of word template database.

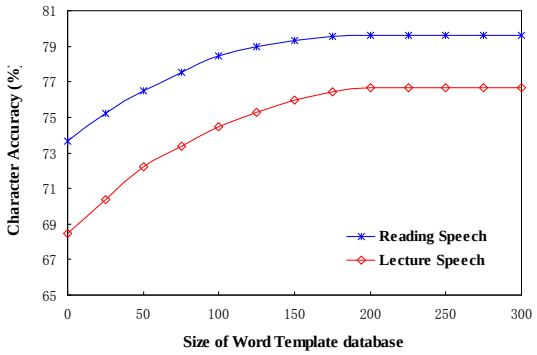


Figure 5. Experiment with size limit of word template database.

The X-axis in Fig. 5 indicates the size of word template database (the number of word templates it can save). It can be seen from this figure that the character accuracy was improved drastically at the beginning when increased the size of word template database. However, when the size reached a certain point, the character accuracy did not increase obviously. For the reading speech recognition, the system obtained a stable accuracy (79.56%) when the size is around 175 word templates. For the lecture speech recognition, the system obtained a stable accuracy (76.64%) when the size is around 200 word templates. This experiment indicates the words that appear repeatedly are very few compared to the whole different words in the task. Therefore, it is not necessary to keep a huge word template database in our method, and its size can be adjusted flexibly.

V. CONCLUSION

According to the characteristic of Chinese language, we proposed an approach to automatically correct recognition error of the repeated words by using its prior recognition result and correction result. In order to test the effectiveness of the proposed method, we conducted two experiments. The first experiment confirms that using our proposed method can reduce the character error rate effectively. In the second experiment, we conclude that the character accuracy is not always improved as the size of word template database increased.

For practical application, if the mandarin speech recognition system does not offer an error correction interface, like the automatic speech recognition (ASR) system, our proposed method without user involved correction can be used to reduce the CER. Otherwise, like some speech input applications, out proposed method with user involved correction can be used to reduce the number of user operation and speed up the speech input.

In future work, we will continue to take advantage of user feedback to improve the performance of speech recognition system. At the same time, we think that the characteristic of Chinese language can make us think out new method to improve the performance of mandarin speech recognition.

ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (61202209), Beijing Natural Science Foundation (4122079), and National Science Supported Planning (No. 2012BAH39B02).

REFERENCES

[1] B. Suhm, B. Myers, and A. Waibe, "Designing interactive error recovery methods for speech interfaces," in *Proc. ACM CHI 1996, Workshop on Designing the User Interface for Speech Recognition Applications*, 1996.
[2] B. Suhm, "Empirical evaluation of interactive multimodal error correction," in *Proc. IEEE Workshop on Speech Recognition and Understanding*, 1997, pp. 583-590.
[3] C. Karat, C, Halverson, D. Horn, and Karat, "Patterns of entry and correction in large vocabulary continuous speech recognition systems," in *Proc. CHI 1999*, 1999, pp. 568-575.
[4] J. Ogata and M. Goto, "Speech repair: Quick error correction just by using selection operation for speech input interfaces," in *Proc. Interspeech 2006*, 2006, pp. 133-136.
[5] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minization and other application of confusion network," *Computer Speech and Language*, vol. 14, no. 4, pp. 373-400, 2000.
[6] L. Mangu, "Finding consensus in speech recognition," PhD Thesis, Johns Hopkins University, 2000.
[7] J. Xue and Y.-X. Zhao, "Improved confusion network algorithm and shortest path search from word lattice," in *Proc. ICASSP 2005*, 2005, pp. 853-856.
[8] X. Li., X. Wang, Y. Qian, and S. Lin, "Candidate generation for interactive Chinese speech recognition," in *Proc. JCPC 2009*, 2009, pp. 583-588

Xiangdong Wang was born in Inner Mongolia, China in 1979. He received Doctor's degree in Computer Science at Institute of Computing Technology, Chinese Academy of Sciences in 2007. His research field includes human-computer interaction, speech recognition and audio processing. He is now an associate professor in Institute of Computing Technology, Chinese Academy of Sciences.

Xinhui Li was born in Hunan, China in 1984. He received his Master's degree in Computer Science at Institute of Computing Technology, Chinese Academy of Sciences in 2010. His research field includes speech recognition and audio processing. He is now a senior engineer in Tencent Inc, China.

Hong Liu was born in Shandong province, China in 1975. She received her Doctor's degree in Computer Science at Institute of Computing Technology, Chinese Academy of Sciences in 2007. Her research field includes human-computer interaction, multimedia technology, and video processing. She is now an associate professor in Institute of Computing Technology, Chinese Academy of Sciences.

Yueliang Qian was born in Shanghai, China in 1960. He received his Bachelor's degree in Computer Science at Fudan University, Shanghai, China in 1983. His research field includes human-computer interaction and pervasive computing. He is now a professor in Institute of Computing Technology, Chinese Academy of Sciences.