Study on Algorithms of Keyword Confusion Network Generation

Lei Zhang Meimei Jia Lili Guo

College of Information and Communication Engineering, Harbin Engineering University,
Harbin, Heilongjiang, 150001, China
zhanglei@hrbeu.edu.cn mmj-family@163.com Guolili@hrbeu.edu.cn

Abstract

Keyword spotting based on large vocabulary continuous speech recognition (LVCSR) is the main researching direction of keyword spotting field. Lattice as the middle result of LVCSR, is often used in this system. But because of its big size, the performance is not efficient as we expect to be. In this paper, lattice was optimized by confusion network (CN) to achieve higher recall rate and lower error rate. And the study on algorithms of keyword confusion network generation was expanded. In the system, the proposed algorithm increased the recall rate of confusion network to 87.11%, while it was 65.46% with N-best.

Keywords: keyword spotting; confusion network; speech recognition

1. Introduction

Presently, great progress has been made in the research of LVCSR. But it still does not satisfy the demand of further application. With the development of computer and multimedia technology, audio files and text files gradually became the main method of information obtaining and storing. So how to effectively manage, classify and search these audio files of big capacity are another challenge in the field of speech recognition[1][2].

The task of keyword spotting is to detect a set of keywords in the continuous speech input. With the different detecting methods, keyword spotting can be sorted into two methods, one is based on LVCSR and another one is based on filler[3]. In the early time, keyword spotting based on filler was widely used in many researches, which satisfied the demand of real time. Unfortunately, the system should be reconstructed when keyword changing. That is because the recall rate seriously relies on the match between non-keyword and filler. At present, keyword spotting based on LVCSR becomes a prevalent method. Keyword spotting based on LVCSR referred to spotting keyword after acoustic decoding. Lattice and N-best are the two common methods of middle structure. Lattice offers enough capability in order to include sufficient candidates, but an efficient decoding algorithm is needed in lattice. So, Mangu[4] proposed transforming lattice into confusion network in 2000. However, the time complexity of clustering algorithm was high. For a lattice with T links, the time complexity for generating confusion network was O(T^3). Later, a more efficient method for confusion network generation was proposed by Jian Xun[5] *et al*. Pengyan Zhang[6] *et al* applied the clustering algorithm to keyword spotting, and the system performed well. In this paper, based on the fast confusion network algorithm of Jian Xue, an improved method of generating keyword confusion network is proposed, which is more suitable for keyword spotting.

Proceedings of the 11th Joint Conference on Information Sciences (2008)
Published by Atlantis Press

The remainder of the paper is structured as follows: in section 2, the frame of keyword spotting system based on confusion network is described. The concept of confusion network and the generation of confusion network are discussed in section 3. Section 4 gives the experiment result, and finally section 5 draws some conclusions from the proposed method.

2. Keyword spotting based on confusion network

The frame of keyword spotting system based on confusion network is described as Fig. 1.

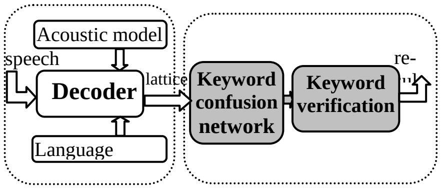

Fig. 1: Diagram of keyword spotting system based on confusion network

As shown in Fig.1, keyword spotting is divided into front-end processing and back-end spotting as a whole. In the front-end processing block, HTK is employed to train acoustic and language models [7]. Acoustic model of this system is context-dependent tri-phone model, whose topology is left-to-right with jump. Every model with five states is jointed as syllable model according to dictionary. Language model is syllable based bi-gram model, and Katz [8] approach is adopted as the smoothing algorithm.

Back-end spotting is the research emphasis of this paper. Firstly, lattice which was generated in decoding block is the input of keyword confusion network block. It generates keyword confusion network with marked scores by matching the keyword. Then, keyword will be verified in the next block. In this block, every arc of lattice has acoustic and language scores. After generating confusion network, the both scores are normalized. After adding the normalized two scores with different weight, the final combination score will be obtained, which can represented the matching between the result and speech signals. Finally, according to the final combination score, the most possible candidate can be achieved.

3. Confusion network

3.1. Concept of confusion network

Confusion network is a structure which dynamically aligns the arcs and nodes in lattice. In this structure, a set is formed by all the words which competing the same pronunciation positions. After aligning these sets according to their start time, the optimum string of words is formed by picking the most probability candidate from each set. Fig. 2 shows the diagram of comparison between lattice and confusion network. Taking ‘jing1ji4jian4she4’ as an example, its lattice and confusion network structures are respectively shown in Fig. 2-a and Fig. 2-b, where the arrangement of nodes strictly according to their start time.

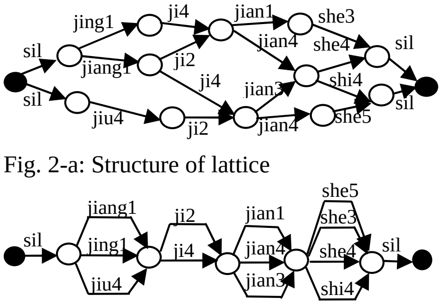

Fig. 2-a: Structure of lattice

Fig. 2-b: Structure of confusion network

Fig. 2: Diagram of lattice and confusion network

In Fig. 2, confusion network commendably solved the problem of time overlapping in the lattice. The relationship among different recognition results of the same pronunciation fragment was appeared obviously in confusion network. The recognition result of sentence is gotten by connecting the optimal recognition results, which are found under certain conditions among every confusion network. Confusion network has highlighted the competition of candidates, and provided favorable conditions for keyword spotting. In the lattice decoding process, the Maximum Posterior Probability (MAP) decoding based on sentences is usually adopted. Such an approach is known to minimize sentence error rate, but unable to minimize the word error rate. However, the confusion network algorithm in [4] successfully reduced the word error rate by picking the word of biggest posterior probability from each set.

3.2. Confusion network generation

Keyword confusion network is defined that only transforming the keyword competitions into a network which has the same start and end nodes. So the lattice which structure is word on arc is used. The algorithm of confusion network generation in [5] generated word network for every sentence. Firstly the start node in lattice is taken as the start point of confusion network. Then the posterior probability for each link is computed in lattice. At last, confusion network is generated by judging the connection of nodes. This traditional algorithm is appropriately used in LVCSR, but it is unsuitable for keyword spotting, in which, there is no need to transform all the words into confusion network. Based on the fast generation algorithm in [5], we present the following improved method to generate keyword confusion network, which is more suitable to keyword spotting.

Let $N=\{n_0,n_1,\cdots\}$ be the set of nodes and $E=\{e_0,e_1,\cdots\}$ be the set of links in the original lattice, where every node $n_i \in N$ has a time mark $t(n_i)$. Let $e_{u\to v}$ denote a link in lattice with the start and end nodes *u* and *v*. let $NS=\{N_0,N_1,\cdots\}$ be the set of node sets in the confusion network and let $E_{N_i \to N_j}$ be the set of links with start and end node sets N_i and N_j.

1. $\forall n_i \in N_i$, $n_j \in N_i$, if $t(n_i) < t(n_j)$, then $i \le j$.

2. $\forall n_i \in N_i$, $n_j \in N_i$, if $t(n_i) = t(n_j)$, then $i = j$.

3. $\forall e_{u\to v} \in E$, if $u \in N_i$ and $v \in N_j$, then $e_{u\to v}$ corresponds to a link in $E_{N_m \to N_n}$ set, where $i \le m \le n \le j$ and $n=m+1$. Furthermore, two consecutive nodes of $e_{u\to v}$ are aligned to N_m and N_n.

Based on assumptions above, algorithm of keyword confusion network can be generated as follow:

Step-1 Transforming keyword into syllable string: $K_1 \cdots K_M$ (*M* is the number of syllable, here *M*=2.)

Step-2 Traversing all the nodes in lattice to find n_k which is matched with K_1. Then assign n_k to N_{k_1}.

Step-3 Suppose N_{k_1} be the ending node.

1) If there is no link between N_{k_1} and n_{k-1}, assign n_{k-1} to N_{k_1}. Then continue to search the former node of n_{k-1}.

2) Otherwise, stop searching.

Step-4 Suppose N_{k_1} be the beginning node.

1) If there is no link between N_{k_1} and n_{k+1}, assign n_{k+1} to N_{k_1}, then continue to search the next node of n_{k+1}.

2) Otherwise, stop searching.

Step-5 For $\forall n_k \in N_k$, then $e_{k\to k+1} \in E_{k_1 \to k_2}$, $n_{k+1} \in N_{k_2}$.

Step-6 $E_{k_1 \to k_2}$ set which includes all the $e_{k \to k+1}$ forms the keyword confusion network.

Step-7 For every link $e_{u \to n_i} \in E$, we can suppose u belong to N_s and n_i belong to N_t . If $t = s + 1$, then the link is directly assigned to $E_{N_s \to N_t}$. Otherwise, the link is assigned to $E_{N_{n-1} \to N_n}$ set where *n* can be determined by the link word probability and degree of time overlap as equation (1).

$$n = \underset{s+1 \le k \le t}{\arg\max} \left\{ SIM\left(E_{N_{k-1} \to N_k}, e \right) \right\} \qquad (1)$$

Where $SIM\left(E_{N_{k-1} \to N_k}, e \right)$ is presented as below:

$$SIM\left(E_{N_{k-1} \to N_k}, e\right) = \frac{1}{\left|E_{N_{k-1} \to N_k}\right|} \times \sum_{l \in E_{N_{k-1} \to N_k}} sim\left(w(l), w(e)\right) overlap\left(E_{N_{k-1} \to N_k}, e\right) \qquad (2)$$

Where $w(l)$ and $w(e)$ represent words corresponding to *l* and *e*, $sim(.,.)$ is the phonetic similarity between two words, which is computed from the most likely phonetic base forms. $overlap\left(E_{N_{k-1} \to N_k}, e\right)$ is defined as the time overlap between $E_{N_{k-1} \to N_k}$ and e which is normalized by the sum of their lengths. Finally, the time mark $t(n_i)$ of every node $n_i \in N$ will be used as a constraint in determining search stopping.

Taking keyword ‘jing1ji4’ as an example, keyword confusion network is shown in Fig. 3.

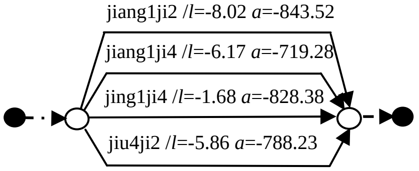
jiang1ji2 /*l*=-8.02 *a*=-843.52
jiang1ji4 /*l*=-6.17 *a*=-719.28
jing1ji4 /*l*=-1.68 *a*=-828.38
jiu4ji2 /*l*=-5.86 *a*=-788.23

Fig. 3: Diagram of keyword confusion network

In figure 3, every arc of the keyword confusion network is marked with the final result and score. *a* is the possibility likelihood score of acoustic, *l* is the possibility likelihood score of language. Based on the generated keyword confusion network, keyword will be verified next.

3.3. Keyword verification

For the generated confusion network, the acoustic and language scores for each candidate are in different orders of magnitude. So they are normalized firstly by formula (3).

$$y = \frac{\left[x - \min(Value)\right]}{\left[\max(Value) - \min(Value)\right]} \qquad (3)$$

In the formula, *Value* represents the set of acoustic or language scores in keyword confusion network. *x* is the score under normalization, and *y* is the normalized score. After score normalization, the acoustic and language scores are assigned to different weights. Then the sum which is marked for the arc value in confusion network is calculated by adding the two scores together. Taking ‘jing1ji4’ as an example, the result is shown in Fig. 4.

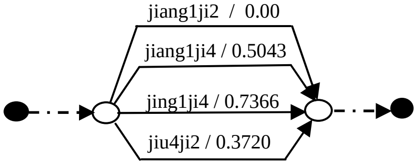

Fig. 4: Diagram of keyword verification

In Fig. 4, the candidate with the highest score can be chosen from the confusion network, which is used for keyword judging. If the word with highest score is a real keyword, it is output with time mark. Otherwise, searching will be restart in the next keyword confusion network.

4. Experiments

In this paper, the keyword spotting experiments based on N-best and keyword confusion network were conducted respectively.

In these experiments, HTK toolkit was employed to build the recognition platform. Training data is from the 863 national corpuses. Twenty keywords were chosen from a test set of 500 sentences, in which keywords totally appear 194 times. To evaluating the keyword spotting system, the recall rate and the equal error rate are used. The recall rate is defined as the number of correct keywords divided by the number of total keywords. The equal error rate is defined as the number of incorrect keywords divided by the number of detected keywords.

4.1. Experimental results

Here, the baseline system directly used N-best as the middle structure. In N-best, different *n* value leads to different detecting performances. The first experiment result about the recall rate and the error rate of different *n* values are shown as follows:

Table 1 Comparison of different *n* in N-best

N-best	Recall rate (%)	Error rate (%)
n=1	56.19	2.68
n=10	61.34	4.03
n=20	65.46	3.79

It can be seen in Table 1 that, compared with the results when *n* equals 1 and *n* equals 10, the recall rate and the error rate of keyword is obviously better when *n* equals 20. What cause this result is that the undetected keywords were probably detected with increasing candidates. The error rate of 3.79% is also represented that the system would reach the relative ideal performance when *n* equals 20 in N-best.

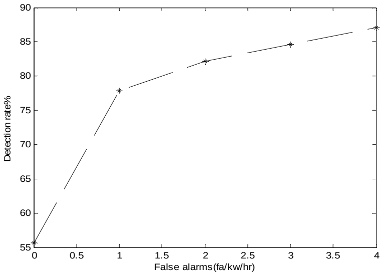

Fig. 5: ROC of keyword spotting based on CN

Furthermore, the second experiment used keyword confusion network as the middle structure, where the weight of language score and acoustic score was 0.7 and 0.3. Fig. 5 describes the ROC[9] curve of keyword spotting system based on confusion network. We can see from it that with the rising of false alarm number per hour, the detection rate of system is increased.

The recall rate and the error rate are further compared across N-best and keyword confusion network, which are shown in Table 2.

Table 2 Comparison of different methods

Method	Recall rate (%)	Error rate (%)
20-best	65.46	3.79
Keyword CN	87.11	9.14

As Table 2 shows that the recall rate of keyword confusion network increases 21.65% compared with 20-best's rate of 65.46%. That is because the method of keyword confusion network not only simultaneously considered acoustic score and language score and highlighted the competition of keyword, but also minimized word error. So keyword spotting based on confusion network has better performance than N-best. However, it can also be seen that the error rate rises 5.35% in keyword confusion network. The reason is that when the recall rate is increasing, the number of the error alarm rises too.

5. Conclusions

Keyword confusion network is a new method of keyword spotting. At present, confusion network has widely applied to the research of LVCSR. In this paper, we have proposed keyword confusion network which can be used as a novel approach of optimizing lattice. Furthermore, it can also be used to build a keyword spotting system which has better performance than N-best.

6. Acknowledgement

It is a project supported by National Natural Science Foundation of China (No.60575030 and No.60702053).

7. References

[1] Ye Liang, Wang Zhibin, Shao Qianming, "Speech Retrieval Engine Based on Relevance Feedback," *Computer Engineering*, Vol.33, No. 17, pp. 228-230, 2007.

[2] Wang Rangding, Yuan Xuhai, Xu Ji, "Novel mixing speech retrieval algorithm". *Application Research of Computers*, Vol. 25, No. 5, pp. 1349-1351, 2008.

[3] Zheng Tieran, Han Jiqing, "Study on Syllable Based Indexing Methods in Mandarin Speech Retrieval", *Proceedings of National Conference on Man-Machine Speech Communication*, 2005.

[4] L. Mangu, E. Brill, A. Stolcks. "Finding consensus in speech recognition: word error minimization and other applications of confusion network", *Computer Speech and Language*, Vol. 14, No. 4, pp.373-400, 2000.

[5] Xue Jian, Zhao Yunxin. "Improved confusion network algorithm and shortest path search from word lattice", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal*, pp.853-856, 2005.

[6] Zhang Pengyuan, Shao Jian, Zhao Qingwei, et al. "Keyword Spotting based on syllable confusion network", *The Third International Conference on Natural Computing*, 2007.

[7] Http: //htk.eng.cam.ac.uk.

[8] Joshua T. Goodman, "A bit of progress in language modeling", *Computer Speech and Language*, Vol. 15, No. 4,pp.403-434, 2001.

[9] National Institute of standards and technology. The Road Rally Word-Spotting Corpora, September 1991. Speech Disc 6-1.1.