Spoken Document Retrieval Based on Confusion Network with Syllable Fragments

Zhang Lei¹*, Yoshihiko Gotoh² and Muhammad Usman Ghani Khan²

1 College of Information and Communication Engineering, Harbin Engineering University, Harbin, PRC
2 Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

* Corresponding author E-mail: zhanglei@hrbeu.edu.cn

Received 5 Jun 2012; Accepted 17 Aug 2012

DOI: 10.5772/52454

© 2012 Lei et al.; licensee InTech. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract This paper addresses the problem of spoken document retrieval under noisy conditions by incorporating sound selection of a basic unit and an output form of a speech recognition system. Syllable fragment is combined with a confusion network in a spoken document retrieval task. After selecting an appropriate syllable fragment, a lattice is converted into a confusion network that is able to minimize the word error rate instead of maximizing the whole sentence recognition rate. A vector space model is adopted in the retrieval task where tf-idf weights are derived from the posterior probability. The confusion network with syllable fragments is able to improve the mean of average precision (MAP) score by 0.342 and 0.066 over one-best scheme and the lattice.

Keywords spoken document retrieval, lattice, confusion network, syllable fragment

1. Introduction

Speech processing techniques can be utilized with robots in many ways, such as controlling robots by commands [1], a conversation system between a robot and a human [2] and so on. Maybe one day robots will friends who help us to find valuable information from huge archives of spoken messages or recordings. That is an information retrieval task. For a high intelligence quotient robot, information retrieval is an essential function which may happen under noisy conditions. Our aim is to solve the problem for noisy condition retrieval tasks which is also important for future high IQ robots.

Many of these applications in robots involve speech recognition. Speech recognition technologies have made impressive progress in recent years, but there are still some problems in real applications, especially under noisy conditions. One study showed a performance reduction of up to 70% in a noisy environment from the ideal quiet one [3]. There are many kinds of noise, but to simulate channel distortion in robot communication this paper uses speech from the wireless radio and cable video, aiming to address the problem associated with the information retrieval task under conditions of channel noise.
This paper is concerned with two problems. One is to explore a suitable unit for speech recognition. The other is related to the output form of a recognition system. Typically words, characters, syllables or sub-syllables can be selected as a basic unit in speech recognition. In Mandarin Chinese, syllables focus on pronunciation while characters pay more attentions on writing. The character unit is helpful for solving the homophone problem, i.e., the same syllable but different character meanings. As for words, they are normally composed of several syllables, but not all composition of syllables is a word. In Mandarin Chinese, more than 80,000 commonly used words and more than 10,000 commonly used characters exist, hence it is hard to build up vocabulary based on words or characters for an ‘open’ application, although words can make full use of language information. In order to handle the OOV (out of vocabulary) problem, a syllable could be a sound choice for a basic unit [4]. Taking both the use of language information and ability to solve the OOV problem into consideration, we combine syllable and syllable fragments (syllable combination, but not limited to words) to form a recognition unit. As for speech recognizer output, three forms, one-best, N-best and lattice, can be selected. A low recognition rate in a noisy environment means that the one-best result cannot perform well in most conditions. The lattice may be a better choice for representing final speech recognition outputs [5]. The lattice is in fact a graph which can reduce the impact of speech recognition errors to a certain extent by providing multiple hypotheses. We have shown the improvement of lattice for spoken document classification task in [6]. The confusion network [7,8] is a more recent development for the presentation of speech recognition outputs. It builds on the lattice with more compact representation and is able to extract words with higher precision than the lattice.

This paper proposes a confusion network combined with syllable fragments for spoken document retrieval tasks, which can also tolerate noise disturbance.

2. System Framework

The whole system consists of three stages as shown in Figure 1. The pre-stage extracts the useful syllable fragment. This stage is off-line and the syllable fragment is not necessarily limited to words. It can also include syllable combinations such as ‘qing2shen1’, though this is not meaningful as words. Syllable fragments and syllables will compose the whole recognition unit. Speech recognizer is trained using the HTK [9] and the lattice structure is created with syllable fragments and syllables on the arc. One example of the lattice with syllable fragments is shown in Figure 2. Since the lattice structure is complicated and occupies large storage, the confusion network is derived from the lattice at the index stage. Finally in the retrieval stage, a text query can be retrieved based on the index.

There are three corpses used for this task. The first corpus contains more than 20 hours of speech and is used for training a hidden Markov model in speech recognition. It includes recordings of conversations from the wireless radio and cable video programmes, network chatting data and the 863 corpus[1]. The second corpus is for the retrieval task where conversation segments are selected from seven video programmes, lasting roughly one hour. The last corpus is used for syllable fragment extraction. Its contents are similar to those of the speech corpus for retrieval, assuring syllable fragments are useful for the retrieval task.

![Figure 1. The framework in three stages with a confusion network (CN) and syllable fragments (SF)](image)

3. Lattice with Syllable Fragment Unit

Although a syllable can deal with both the OOV and homophones problems, it neglects the relations in the language model. In some cases, identification of a word group, phrase or a combination of several syllables is more robust than recognizing each syllable and then combining them together. In this work a basic syllable unit incorporates a syllable fragment, a combination of several syllables which may not be a word group or phrase.

3.1 Extraction syllable fragments

In Chinese text a corpus is a continuous syllable stream. To obtain syllable fragments, the forward maximum matching method is utilized for segmentation. To evaluate the distance between two terms, the normalized mutual information is calculated as follows [10]:

\[
\hat{M}(v, w) = \frac{\log \frac{p(v, w)}{p(v)p(w)}}{c(v, w)}
\]

1 This is a special corpus for speech recognition in the 863 Project in China.
where, \( v \) and \( w \) are two terms in the vocabulary. \( p(v) \) and \( p(w) \) are the probabilities of terms \( v \) and \( w \) in the corpus. \( p(v, w) \) represents the probability that terms \( v \) and \( w \) are the close neighbour. \( c(v, w) \) is introduced as the sum of syllable lengths for \( v \) and \( w \) so that long terms are penalized.

Since the remaining task is related to spoken document retrieval, tf-idf is adopted to filter syllable fragments, keeping ones with the high discriminative capacity in the vocabulary. It is computed by Equation (2):

\[
tfidf(t, d) = tf(t, d)idf(t, D)
\]

where \( tf(t, d) \) is the frequency of term \( t \) occurring in document \( d \) and \( idf(t, D) \) is the inverse frequency:

\[
idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\} | + 1}
\]

Here, \(| \cdot |\) is the element number of the set.

The algorithm of syllable fragment extraction is as follows:

Step 1: initialization—setting a set of syllables as the initial vocabulary;

Step 2: segmentation—segmenting the continuous syllable stream based on the vocabulary.

Step 3: combination—computing the mutual information between terms as Equation (1); if the mutual information is larger than a threshold, combine two and add to the vocabulary.

Step 4: go back to Step 2 and repeat until no more new term is generated.

Step 5: filtering—use tf-idf weight to select those with better discriminative ability; terms with \( N \) highest weights are kept in the final vocabulary.

| length | Number of unit | Example |
|--------|----------------|---------|
| 1      | 1707           | ni3     |
| 2      | 1076           | yi2-ge4 |
| 3      | 79             | hen3-duo1-ren2 |
| 4      | 20             | fa3-zhi4-ri4-bao4 |

Table 1. The number of syllable fragments with length 1,2,3 and 4

The above algorithm is costly because the mutual information has to be computed for any two connected terms. However, the calculation is an off-line procedure and we only do this once. In the experiments, the syllable fragment extraction algorithm was able to generate 1175 fragments. Table 1 shows the number of fragments with lengths up to four. Fragments with more than four syllables are ignored. Combining 1175 syllable fragments with 1707 syllables, we are able to avoid the OOV problem because any words can be composed with these syllables. Additionally, more stable recognition results can be obtained with two to four syllable fragments. Table 2 presents the recognition rate for one-best result with and without syllable fragments. It can be seen that there is an improvement of roughly 3% by adding syllable fragments.

| One-best result without SF | One-best result with SF |
|---------------------------|------------------------|
| 58.6%                     | 61.7%                  |

Table 2. Comparison of one-best results with and without syllable fragments (SF)

3.2 Lattice with syllable fragment unit

Figure 2 illustrates a lattice with syllable fragments. Not only syllables, but also syllable fragments are assigned on the transition arc. Because syllable fragments are more stable than syllables, more significant results can be reached using the former. Using HTK, the starting times and the ending times, the probabilities of acoustic model and language model can be obtained. A lattice can provide multi-path results for recognition, however, a few problems exist if we use a lattice for indexing. For example, the same term (syllable or syllable fragment) may occur at several positions in the lattice. Time overlap between two connected terms can be quite complex. Furthermore, the storage requirement for a lattice can be very large when the pruning threshold is not carefully selected. A noisy environment may require many possible results to be kept in a lattice. Use of more flexible pruning can result in a larger number of nodes and arcs (sometimes more than 10 times). Under this condition a lattice is not the best way for indexing. Instead a confusion network is explored to handle this problem.
4. Confusion Network with Syllable Fragments

4.1 Generation of confusion network with syllable fragments

In a lattice, such as Figure 2, there are many paths between the start and the end nodes, with each path producing one recognition result. A traditional approach selects the best path with the largest posterior probability for a whole sentence. On the other hand, in many applications, such as a retrieval task, a word error rate is the main concern because a word (not a sentence) is the basic processing unit. The main idea for a confusion network is to cluster the competition nodes or arcs into the same confusion set and linearly connect those confusion sets together. As shown in Figure 3, arcs between two adjacent nodes belong to the same confusion set. During clustering, this algorithm keeps the same partial order of arcs as those in the lattice as much as possible. Different from the conventional confusion network generation algorithm, influence from syllable fragments should be considered. Multiple confusion sets may be overcrossed (e.g., ‘fá3zhi4’ and ‘xīn1wén2’ in Figure 3). In order to make full use of the conventional confusion network generation algorithm, two additional steps (Step 1 and Step 5 below) are added to handle the syllable fragments. The whole algorithm is as follows:

Step 1: Split a syllable fragment into a string of syllables (Figure 4). The duration and the probability of the syllable fragment are evenly divided into each syllable. For example, one arc in lattice, ‘bei3jing1shi4’, is split into three arcs with ‘bei3’, ‘jing1’ and ‘shi4’ before converting into a confusion network.

![Figure 4. Splitting a syllable fragment into a string of syllables](image)

Step 2: Initialization of a confusion set. Arcs with the same starting, ending time and the same label are included in a single confusion set. Then the lattice can be divided into many confusion sets according to the labels on the arcs which need to be combined further.

Step 3: Combining confusion sets with the same label, but different starting and ending time. Only the confusion sets with no partial order can be combined. At each iteration, the most similar confusion sets are combined to form a new set. As in [11], in order to handle the insertion error in a lattice, the start and end time of the new confusion set is defined by the earliest starting time and the latest ending time of all arcs in it. The similarity between confusion sets $E_i$ and $E_j$ is calculated as:

$$sim(E_i,E_j) = \max_{\alpha,\beta} overlap(e_{\alpha},e_{\beta}) p(e_{\alpha})p(e_{\beta})$$  \hspace{1cm} (4)

Where $overlap(e_{\alpha},e_{\beta})$ is defined as the time overlap between the two links normalized by the sum of their durations. The posteriors probabilities $p(e_{\alpha})$ and $p(e_{\beta})$ make the measure in Equation (4) less sensitive to unlikely candidates.

Step 4: Combining confusion sets with different labels. Because the confusion sets to be combined are with different labels, we adopt one minus the edit distance, instead of overlap in Equation (4), times posterior probability to represent the similarity between two confusion sets.

Step 5: Assigning the arc with syllable fragments to different nodes according to the starting and ending time, which will cross several confusion nets.

|          | Storage space | Nodes number | Arcs number |
|----------|---------------|--------------|-------------|
| Lattice  | 88.21 kB      | 346          | 1263        |
| CN       | 3.89 kB       | 33           | 86          |

Table 3. Complexity comparison between the lattice and the confusion network (CN)

During the generation of a confusion network, we keep the partial order in the lattice and simplify the representation by clustering. A confusion network is able to provide more compact representation than a lattice. Table 3 presents the complexity comparison between the lattice and the confusion network. The values in this table are the average of 1000 speech segments selected...
randomly. It can be seen that a confusion network is more suitable for indexing, especially with a complex condition with a wide pruning threshold.

Table 4 shows the recognition result based on the confusion network. Using the confusion network the best hypotheses is selected by a local search over a small number of candidates in a confusion set. This minimizes the word error rate instead of maximizing the posterior probability for a whole sentence. Syllable fragments introduced in a confusion network provide an additional choice for the final result. As shown in Figure 3, in addition to the best candidate (the top arc) in each confusion set, there are several arcs crossing the confusion sets (crossing arcs). There are three conditions when searching for a final outcome: firstly for the crossing arcs, if there are overlaps, we choose the arc with the best posterior probability. Secondly for the crossing arcs and the candidates in each confusion set it crosses, we compare the posterior probability of this crossing arc with the sum of the best candidates in each confusion set it crosses and keep the better one as the final result. Lastly we choose the best candidate as the result with no crossing arcs. The result is noted as CN-one-best in Table 4, which shows approximately 1.4% enhancement for the speech recognition task.

| One-best with SF | CN-one-best with SF |
|------------------|---------------------|
| 61.7%            | 63.1%               |

Table 4. Comparison of one-best and confusion network one-best result with syllable fragments (SF)

4.2 Retrieval system based on confusion network

A vector space model is adopted for the retrieval system, which presents each spoken document or a text query in the form of a vector in the high dimensional space. It is easy to convert a text query into a vector space model. For a spoken document based on a confusion network, the approach to extracting word information was discussed in [11]. Crossing arcs are also treated as a word (term). A spoken document can be represented as \( d_j = \{ t_1, \ldots, t_N \} \) where \( t_i \) represents the term weight. Normally the term weight \( t_i \) is a tf-idf measure in Equation (2). However, for a spoken document, the term is not necessarily found in the sentence. There is a little trick when computing tf-idf weight by the posterior probability instead of the term frequency. We compute \( tf(t,d) \) in Equation (2) as:

\[
  tf(t,d) = \frac{P_{td}}{\sum_i P_{td}} \tag{5}
\]

where \( P_{td} \) is the posterior probability for term \( t \) in spoken document \( d \). In order to reduce the adverse effect caused by the spoken document length, the tf-idf is normalized as follows:

\[
  w_{id} = \frac{tf(t,d) \cdot idf(t,D)}{\sqrt{\sum_i tf(i,d) \cdot idf(t,D)}} \tag{6}
\]

For the query \( q \), the cosine distance is calculated by

\[
  \text{sim}(d_j,q) = \frac{\sum_i w_{td}w_{ij}}{\sqrt{\sum_i w_{id}^2 \sum_j w_{ij}^2}} \tag{7}
\]

5. Experiments

After proving the effectiveness of syllable fragment (Table 2) and the confusion network (Table 4) in the speech recognition task, in this section, we focus on the effects of the proposed approach on the retrieval task. Since normally the retrieval task is conducted on a lattice and the one-best result of speech recognition, here these two approaches with syllable fragments are compared with the proposed approach.

Roughly 30 queries were manually selected from the retrieval corpus to create a query set. Table 5 shows a few examples from these queries.

Based on 30 queries, Table 6 presents the MAP (mean of average precision) scores for the one-best, the lattice and the confusion network with syllable fragments. A comparison of MAP scores clearly shows that the confusion network with syllable fragments could achieve the best performance, with improvements of 0.066 and 0.342 absolute over the lattice and the one-best. Additionally, the significant improvement from the one-best to the lattice-based system was not very surprising. The lattice could provide multiple candidates, which is a great advantage for a retrieval task because the correct solution would not necessarily be the top candidate. The confusion network-based approach resulted in further improvement over the lattice because it was able to improve the word recognition rate rather than the whole sentence recognition rate, which was more suitable for the retrieval task based on words (and not on sentences).

For almost each query the performance of CN+SF is better than Lattice+SF. Some effects are distinct, such as query 0 and query 30, which mean from the lattice to the confusion network the words generated are quite different and the confusion network does extract the correct ones. For some queries, this effect is not obvious enough to make it reflect on the retrieval result, such as query 2 in Table 5. Despite the better performance of the proposed approach based on the confusion network, the confusion network could provide much more compact representation, which was the additional advantage over the lattice.
6. Conclusion

In speech recognition in noisy conditions, the pruning threshold needs to be set flexibly in order to accommodate the correct result. This in turn makes the lattice size increase dramatically, leading to an unsuitably large lattice size for the retrieval task. In this paper we introduced a confusion network with syllable fragments. We showed the enhanced retrieval results due to the compact structure associated with the confusion network and the stability provided by syllable fragments.

7. Acknowledgments

This work is supported by Young Teacher Support Plan by Heilongjiang Province and Harbin Engineering University in China (no. 1155G17), and Fundamental Research Funds for the Central Universities in China.

8. References

[1] Yamamoto, S., et al. (2005) Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory. Proceeding of Int. Conf. on Robotics and Automation (ICRA), Barcelona, Spain 2005.

[2] Martinson E., Brock D. (2007) Improving human-robot interaction through adaptation to the auditory scene. HRI’07 proceedings of ACM/IEEE international conference on Human-robot interaction.

[3] Lockwood L. Boudy J. (1992) Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov models and the projection, for robust speech recognition in cars. Speech communication, 11(2-3): 215-228.

[4] B. Chen, H.M. Wang, and L.S. Lee(2004). A Discriminative HMM/N-gram-based Retrieval Approach for Mandarin Spoken Documents. ACM Trans. Asian Lang. Inform. Processing, 3(2) : 128–145.

[5] Siniscalchi, S.M. and Lee, C.H.(2009) A Study on Integrating Acoustic-phonetic Information into Lattice Rescoring for Automatic Speech Recognition. Speech communication. 51(11): 1139-1153

[6] Lei Z, Zhuo Z, Xuezhi X.(2011) Key Information Expansion Applied in Spoken Document Classification based on Lattice. Journal of Computers. 6(5): 923-930.

[7] Mangu, L. and Brill, E. and Stolcke, A.(2000) Finding consensus in speech recognition: word error minimization and other applications of confusion networks. Computer Speech & Language. 14(4): 373–400

[8] Liu, Y. and Shriberg, E. and Stolcke, A., Hillard, D., Ostendorf, M., Harper, M.(2006) Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. Audio, Speech, and Language Processing, IEEE Transactions on. 14(5): 1526-1540

[9] Young, S.J.(1993) HTK reference manual. Cambridge university engineering department.

[10] Tieran Z, Jiqing H, Haiyang L. (2011) Study on Performance Optimization for Chinese Speech Retrieval. Journal on Communications. 30(3): 84-88.

[11] Lei Z, Guoxing C, Xuezhi X, Jingxin C.(2011) Topic Mining based on Word Posterior Probability in Spoken Document. Journal of Software. 6(11): 2292-2299.