OPEN-VOCABULARY SPOKEN UTTERANCE RETRIEVAL USING CONFUSION NETWORKS

Takaaki Hori[*], *I. Lee Hetherington*[**], *Timothy J. Hazen*[**], *and James R. Glass*[**]

[*]NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
hori@cslab.kecl.ntt.co.jp
[**]MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139, USA
{ilh,hazen,glass}@csail.mit.edu

ABSTRACT

This paper presents a novel approach to open-vocabulary spoken utterance retrieval using confusion networks. If out-of-vocabulary (OOV) words are present in queries and the corpus, word-based indexing will not be sufficient. For this problem, we apply phone confusion networks and combine them with word confusion networks. With this approach, we can generate a more compact index table that enables robust keyword matching compared with typical lattice-based methods. In the retrieval experiments with speech recordings in MIT lecture corpus, our method using phone confusion networks outperformed lattice-based methods especially for OOV queries.

Index Terms— Spoken Utterance Retrieval, Confusion Network, Audio Indexing

1. INTRODUCTION

In the last decade, it has become much easier for end users to generate, store, and browse large amounts of audio-visual materials. Text and multimedia documents are increasingly available on the Web. With this rapid growth of multimedia data, development of a technology to quickly find something interesting for individuals in a large amount of data is a key issue.

Recently, spoken document retrieval (SDR) has intensively been investigated in the SDR track of Text REtrieval Conference (TREC) [1], in which the task is to retrieve broadcast news stories relevant to some specific keywords. Usually each document is previously transcribed by automatic speech recognition (ASR), and indexed with the transcribed text. Thus, the relevant documents can be retrieved as well as text document retrieval.

Recent work on SDR has reported that it is adequate to use the single-best ASR output for indexing. This may be true when using such news stories that are relatively long, contain redundant information, and can be recognized with low error rate. However, when retrieving short utterances with relatively high error rate, it is not enough to use only the single-best output because recognition errors more seriously affect the retrieval performance.

Spoken utterance retrieval (SUR) is an important technology to find short snippets that contain given keywords or phrases since it is useful for browsing unstructured long recordings including different topics. As mentioned above, the performance of SUR tends to be affected by the speech recognition accuracy. Therefore it is necessary to make the retrieval system more robust for recognition errors by using multiple hypotheses of ASR. In addition, how to handle queries including out-of-vocabulary words (OOV queries) is also important because it is more difficult to retain the retrieval performance for short utterances only using in-vocabulary (IV) words.

Several retrieval techniques dealing with multiple hypotheses from an ASR system have been proposed, in which word and/or subword lattices are used to index each utterance [2] [3]. Subword-based representations such as phone lattices are crucial especially for OOV words. It is also effective to combine word and subword lattices to achieve high retrieval performance for both IV and OOV queries since subword-based indices generally yield a lower precision for IV queries compared with word-based ones.

However, incorporation of subword lattices dramatically increases the size of the index table. In our experiments, it was more than ten times larger than that of the single-best word hypotheses. Thus indexing with subword lattices is expensive in time for building the table and in space for storing them. In addition, since lattices have only paths allowed by the ASR grammar, matching word/phone sequences with lattices is also restricted by the grammar. This is less flexible for robust keyword matching.

In this paper, we incorporate confusion networks in SUR. The confusion network is the most compact representation of multiple hypotheses, which can be transformed from a lattice [4]. Since it has more paths than the original lattice, it should achieve more robust keyword matching for OOV words and errorful recognition hypotheses. Therefore the confusion network is a reasonable choice for SUR. We also propose a method to combine word and phone confusion networks to improve the performance for both IV and OOV queries. For combining the two networks, we employ composition operation for weighted finite-state transducers (WFSTs) to align the two networks. In addition, we also prune phone arcs that overlap high confidence words to reduce the size of the index table.

In SUR experiments with lecture speech recordings from MIT lecture corpus [5], we compared several indexing methods including word/phone confusion networks and the combined networks.

(a) Lattice

(b) Confusion network

Fig.1. Lattice vs. confusion network. A, B, …, J indicate hypothesized symbols. @ is a label that stands for allowing null transitions.

2. SPOKEN UTTERANCE RETRIEVAL

The purpose of spoken utterance retrieval (SUR) is to find all short segments (utterances) containing a given keyword or phrase in a large spoken archive. The keyword or phrase called *query* is given in text.

In general SUR, the process is separated into two parts. The first one is *indexing* and the second one is *search*. Indexing is an offline process in which linguistic information is extracted from all speech data in the archive, and an index table is built. A speech recognizer is used to extract such linguistic information. The index table maps each extracted linguistic symbol (word, subword or phrase) to a set of utterances that the symbol matches. Search is an online process in which users' queries are accepted and utterances that each query matches are found. The system finds the target utterances efficiently using the previously built table. The index table helps to search with a desirable speed that is almost independent of the size of the archive.

In [2], a lattice-based SUR method has been proposed. The system makes an index table from multiple hypotheses represented in lattice format where each lattice is assumed to be a weighted automaton. To search with OOV queries, phone lattices have also been utilized.

In that work, an inverted index data structure for the index table has been adopted, in which each symbol (word, phone, etc.) is linked to a set of lattice arcs labeled with the same symbol. By letting each arc have an ID number indicating the utterance that the arc belongs to, the system can quickly find a set of utterances that the query word matches. In the table, an index file is prepared for each symbol, in which each arc a has information: (i, $p[a]$, $n[a]$, $w[a]$, $f(p[a])$), where i is the utterance ID number, $p[a]$ is the source state, $n[a]$ is the destination state, $w[a]$ is the weight of the arc, and $f(p[a])$ is the potential weight from the initial state to $p[a]$, which is used to normalize the weight of the arc or path starting from the arc so that the (accumulated) weight becomes a posterior probability. By pruning the matched utterances based on their posterior probabilities, the degree of extraction can be controlled, which is reflected in the precision and recall rates.

For phrase queries, utterances including all symbols are retrieved and then connectivity for each arc pair in the same utterance is checked using its source and destination states. If there is a path matching the query, the utterance becomes a candidate of SUR. For open-vocabulary SUR, i.e. when using subword lattices, this phrase-based search is performed through conversion of each query word to its subword representation.

3. INDEXING WITH CONFUSION NETWORKS

In this study, we introduce confusion networks into spoken utterance retrieval. The confusion network can be converted efficiently from a lattice using Mangu's algorithm [4].

Since the confusion network has the most compact structure representing multiple hypotheses while keeping the order of symbols (phones/words) along the time axis, the space for the index table can be reduced. In addition, the confusion network essentially has more paths than the original lattice that has only paths allowed by the recognizer. The confusion network has many additional paths on which any connection of hypothesized symbols is allowed unless breaking their original order. Fig.1 shows an example of a lattice (a) and the confusion network (b) that is converted from the lattice. Thus it is compact and potentially achieves more robust keyword matching for OOV words and errorful recognition hypotheses.

A confusion network is a finite state network, and therefore the same approach as the lattice-based method can be applied. However, in a confusion network, arcs are aligned to columns as in Fig. 1 (b) and the weight of each arc has already been normalized so that the sum of weights in each column becomes 1. As the result, $f(p[a])$ is always 1, so this element can be eliminated. In addition, $n[a]$ can also be eliminated since $n[a]$ is necessarily located right next to $p[a]$.

4. WORD-PHONE-COMBINED INDEXING

Subword-based indexing is effective especially for OOV queries but generally yield a lower precision for IV queries than that of word-based indexing. Therefore some combination techniques between word and subword hypotheses have been proposed. In [3], word and subword lattices are connected in parallel or combined in the ASR grammar. This work has reported that the combination is effective to achieve high retrieval performance for both IV and OOV queries.

We propose a method to combine word and phone confusion networks to improve performance for both IV and OOV queries. Suppose there are word and phone confusion networks for the same utterance as in Fig.2 (a) and (b) where each network is represented as an automaton. Our method generates a combined network as in Fig.2 (c). In addition, we reduce the size of the network using a limited graph traversal technique by which phone arcs overlapping high confidence words can be detected and pruned. The method is performed by the following steps.

1. Generate a phone-to-word transducer from the word confusion network using the pronunciation lexicon where

each word is replaced with arcs representing its pronunciation as in Fig.3. The transducer has the starting and ending arcs of each word marked with starting symbol 's' or ending symbol 'e' to each word symbol like 'As' and 'Ae'.

2. Convert a symbol on each arc in the phone confusion network to a pair of input and output symbols so that the input symbol includes the original phone, its source and destination state IDs, and the output symbol has just the phone. For example, symbol 'a3' is replaced with 'a3(2-3) : a3'.
3. Apply composition operation to the two transducers generated in steps 1 and 2.
4. Generate a combined network by adding word arcs to the original phone confusion network according to each arc of the transducer composed in step 3. Each arc of the transducer has a pair of a phone label with its source and destination states, and possibly a word start or word end label such as 'a1(0-1) : As', 'a3(2-3) : Ae', etc. Accordingly, for example, a word arc with symbol A from state 0 to 3 can be added to the phone confusion network as in Fig.2(c).
5. Finally prune unnecessary phone arcs in the combined network. A limited graph traversal method is performed, which takes phone arcs at each state only when any high-confidence word arcs (i.e. their posterior probabilities are above a predefined threshold) are not found from the state. After that, the unvisited arcs are pruned.

When using the combined index table, only OOV query words are replaced with their phone sequences. Query phrases can include both word and phone sequences in series.

5. UTTERANCE SEARCH

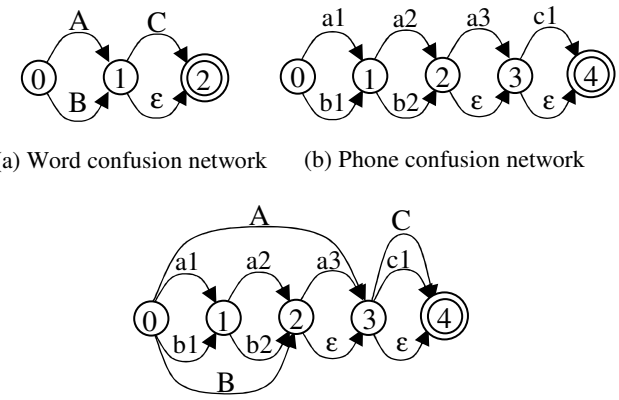
Given a query, our retrieval system finds a set of utterances which possibly contain the query. The query is translated into an automaton representing a single word, a word sequence, a phone sequence, or a word-phone-mixed sequence. By using an automaton as a query, multiple pronunciations and mixed sequences can be easily represented.

The search process consists of two passes. The first pass searches for utterances and arcs which have labels in the query while ignoring proximity of the arcs. A logical AND search is performed on the inverted index table to find utterances including all labels in the query sequence. A logical OR search is also used if the query automaton has branches.

The second pass checks the proximity of arcs for each utterance found in the first pass. To check the proximity, an automaton for each utterance is dynamically constructed, which has only the arcs with the labels in the query. When using confusion networks or word-phone-combined networks, epsilon arcs also need to be included in the utterance automaton.

Proximity between arcs can be checked by automata intersection. If intersection of the query and utterance automata results in null, we can consider that the utterance does not match the query. To compare the query with any part of the utterance, all states of the utterance automaton need to be marked as both initial and final before performing the intersection. The accumulated weight through the best path in the intersection automaton can be used for ranking and pruning the utterances that match the query.

Although the main computation is devoted in the second pass, the computation amount for the automata intersection and the best-path search is not a big problem because both the query and utterance automata are usually small. Since the utterance automata have only the arcs with the labels in the query, they are much smaller than the original networks generated by the speech recognizer.



(a) Word confusion network (b) Phone confusion network

(c) Word-phone combined network

Fig.2. Combining word and phone confusion networks: A, B, and C are words and a1, a2, a3, b1, b2, and c1 are corresponding phones to A, B, and C.

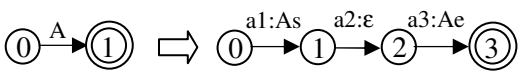


Fig.3. Conversion into phone-to-word transducer: As and Ae stand for the starting and ending arcs of A.

6. EXPERIMENTS

We conducted spoken utterance retrieval experiments with lecture speech recordings in MIT lecture corpus [5]. The corpus is a collection of audio-visual recordings of lectures and seminars presented at MIT, which is approximately 300 hours containing lectures from eight different courses and from 80 seminars given on a variety of topics.

For speech recognition, we used the SUMMIT segment-based recognizer [6] that efficiently drives weighted finite-state transducers (WFSTs) representing phonological rules, a lexicon, and a language model for speech recognition. The recognizer we used for lecture speech had a 16K-word vocabulary including a single OOV word model [7] and worked with a 2-pass search strategy based on bigram and trigram language models trained with the transcripts of the corpus and a Computer Science textbook used in the lectures.

We prepared two test sets for evaluating several indexing methods. One is the small data collection that consists of approximately 6 hours containing 4 lectures from a course of computer science. The word error rate is 37.2%. The other is the middle-size data collection that consists of approximately 22 hours containing 20 lectures [8]. The word error rate is 43.9%.

We constructed index tables using 1-best speech recognition results (1-best), word lattices (WLAT), phone lattices (PLAT), word confusion networks (WCN), and phone confusion networks (PCN), respectively. All the lattices were optimized with WFST operations before making the tables. The phone lattices were

generated using the word-based recognizer for high recognition accuracy, where phone sequences were restricted by the lexicon but the OOV model helped to accept OOV phone sequences.

For the small collection, we prepared 115 test queries (2.3 words per query on the average). OOV words were included in 15 queries (13%) for which any word-based index does not match. The size of each index table and the retrieval performance in F-score are shown in Table 1.

PCN yielded the best F-score (80.2%) with a relatively small table size (5.0MB). We also calculated F-scores separately for the in-vocabulary (IV) queries and the out-of-vocabulary (OOV) queries. The results show that PCN improved F-score especially for OOV queries while achieving a comparable score with the others for IV queries.

For the middle-size collection, we prepared 185 test queries (1.8 words per query on the average), most of which are extracted from the real index of the textbook used in [8]. Since the textbook was used to construct the language models, all the queries were IV words. To evaluate performance for OOV queries we added 30 OOV queries to the original query set. The results are summarized in Table 2. In this data collection, word-based methods outperformed the phone-based methods. To obtain the maximum F-scores, we had to adjust the threshold to decide the degree of extraction. This fact indicates that the number of incorrect matches increases when retrieving in a large data set. For OOV queries, PCN also yielded the best F-score.

Finally we evaluated our word-phone-combined indexing method. In table 3, the size of the index tables was reduced by changing the threshold with little degradation of F-score. The threshold was used to choose high confidence words, i.e. only word arcs with a posterior probability over this threshold can be passed through in the graph traversal to detect redundant phone arcs. This combined network achieved high F-score for both IV and OOV queries. We looked at the result for each query, and confirmed that the in-vocabulary words in the OOV queries contributed to the improvement of F-score for the OOV queries.

Table 1. Index table size and retrieval performance for the small data collection.

	1-best	WLAT	WCN	PLAT	PCN
Table size [MB]	0.8	6.7	2.4	14.0	5.0
F-score [%]	70.3	77.1	77.6	78.5	**80.2**
IV queries	73.1	80.0	80.4	79.8	**81.0**
OOV queries	0	0	0	54.1	**66.7**

Table 2. Index table size and retrieval performance for the middle-size data collection.

	1-best	WLAT	WCN	PLAT	PCN
Table size [MB]	3.0	26.7	9.5	59.4	20.8
Max F-score [%]	82.5	83.8	83.8	74.9	**74.5**
IV queries	85.0	86.3	86.2	76.4	**75.9**
OOV queries	0	0	0	27.9	**38.4**

Table 3. Index table size and retrieval performance with word-phone combined indexing for the middle-size data collection.

Threshold	-	0.95	0.8
Table size [MB]	51.6	36.1	29.2
Max F-score [%]	85.1	85.0	84.8
IV queries	86.2	86.2	86.2
OOV queries	52.0	51.7	48.4

7. CONCLUSION

We have applied phone confusion networks to spoken utterance retrieval and proposed a method to combine phone and word confusion networks. In retrieval experiments with MIT lecture corpus, the confusion-network-based method generated more compact index tables and yielded better retrieval performance compared with the typical lattice-based method especially for OOV queries. In future work, we plan to evaluate our approach using a larger data collection over 100 hours, and compare it with other approaches [9] [10].

8. REFERENCES

[1] S. Renals et al., "Indexing and retrieval of broadcast news," *Speech Communication*, vol. 32, no. 1-2, pp. 5—20, 2000.

[2] M. Saraclar and R. Sproat, "Lattice-Based Search for Spoken Utterance Retrieval," *Proc. HLT-NAACL*, 2004.

[3] P. Yu and F. Seide, "A hyprid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech," *Proc. Interspeech 2004*, pp. 293—296, 2004.

[4] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, 14, pp. 373—400, 2000.

[5] J. Glass, T. Hazen, L. Hetherington and C. Wang, "Analysis and Processing of Lecture Audio Data: Preliminary Investigations," *Proc. HLT-NAACL Workshop: Interdisciplinary Approaches to Speech Indexing and Retrieval*, Boston, 2004.

[6] J. Glass, "A Probabilistic Framework for Segment-Based Speech Recognition," *Computer Speech and Language*, 17, 2003.

[7] I. Bazzi and J. Glass, "A Multi-Class Approach for Modelling Out-of-Vocabulary Words," Proc. *ICSLP 2002*, pp. 1613—1616, 2002.

[8] A. Park, T. J. Hazen, and J. Glass, "Automatic processing of audio lectures for information retrieval: vocabulary selection and language modeling," *Proc. ICASSP 2005*, Vol. I, pp. 497—500, 2005.

[9] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," Proc. *ACL 2005*, pp. 443-450, 2005.

[10] C. Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata – application to spoken utterance retrieval," *Proc. HLT-NAACL*, 2004.