FACILITATING OPEN VOCABULARY SPOKEN TERM DETECTION USING A MULTIPLE PASS HYBRID SEARCH ALGORITHM

Atta Norouzian, Richard Rose

Department of ECE, McGill University, Montreal, Canada

ABSTRACT

This paper presents an efficient approach to spoken term detection (STD) from unstructured audio recordings using word lattices generated off-line from an automatic speech recognition (ASR) system. The approach facilitates open vocabulary STD and focuses specifically on reducing the difference between detection performance obtained for within-vocabulary (IV) and out-of-vocabulary (OOV) search terms. Improved OOV detection performance is obtained by using a two-pass search procedure. Candidate audio segments are retrieved from an index of word lattice paths in the first pass. Locations of OOV search terms are detected in the second pass from a constrained alignment of phonemic expansions of the query terms with phoneme sequences obtained from acoustic segments using an unconstrained neural network based phone decoder. It is found that the combination of first pass segment retrieval and second pass term verification significantly increases STD performance for OOV query terms with no increase in search time for utterances taken from a lecture speech domain.

Index Terms— Spoken term detection, spoken utterance retrieval, automatic speech recognition

1. INTRODUCTION

There are many multimedia search applications involving large media repositories containing unstructured recordings of lectures, voice mail, and conversational telephone speech [1, 2, 3, 4, 5]. Spoken term detection involves fast search of large repositories of audio documents from query terms entered by a user. The particular interest in this work is in search applications that provide subsecond response latencies for users entering completely unrestricted queries to search audio collections containing many hundreds of hours of speech. State–of–the–art ASR word accuracy (WAC) in many of the above task domains can be as low as 50 to 60 percent, making the problem of detecting unrestricted queries more difficult. This poor performance results from poor acoustic conditions, ill-formed utterances, and a large percentage of the words in the utterances being out of the ASR vocabulary.

In order to achieve this level of efficiency in search while maintaining good STD performance, a number of multi-stage approaches have been proposed. Most of these approaches begin with the generation of word or phone lattices for short audio segments extracted from the audio recordings [1, 2, 3, 6, 7]. The search process associated with STD is made far more efficient by generating these lattices prior to performing search. However, even when these ASR lattices have been generated off-line, it has been found that approaches requiring exhaustive search of the lattices associated with each audio segment do not scale well to very large collections.

This work was partially supported by NSERC and by FQRNT

To deal with this issue, it is possible to index the lattices associated with each audio segment [1, 2, 4, 6, 7, 8]. After off-line lattice generation and index creation, search is performed in two passes. First, index terms are identified from the user's query and lattices or lattice paths associated with the indexing term are accessed from the index. In the second pass, a detailed search for the query term is performed on the lattice paths retrieved from the index. Using word-based indexing strategies are problematic for OOV query terms. Subword-based indexing strategies have been used to facilitate open vocabulary queries [2, 4, 6, 7]. However, using subword based indexing and subword search strategies generally results in a significant reduction in STD performance for within vocabulary search terms relative to word based approaches.

This paper presents an STD approach designed for both IV and OOV search terms. The approach follows from previous work involving a hybrid indexing strategy for efficient open vocabulary STD [9]. The strategy, summarized in Section 2, relies on word-based indexing terms, but uses a phoneme-based distance measure to associate the user's query with index entries. This allows lattice paths to be retrieved for both IV and OOV queries during the first pass search so that more sensitive measures for detecting search term occurrences from the retrieved lattice paths can be applied in the second pass search.

While detection performance reported in [9] for IV queries was quite good, it was found that there was still a significant gap between the detection performance obtained for IV and OOV terms. The techniques presented in Section 2 significantly reduce this performance gap by using a more powerful strategy for verifying OOV term occurrences from lattice paths retrieved from the hybrid index. Phone sequences derived from unconstrained neural network based phone posteriors are used to verify OOV term occurrences hypothesized from the offline large vocabulary speech recognition constrained search. The STD performance is presented in Section 4 for lecture speech utterances taken from recorded course lectures stored on an online media server [10]. The results demonstrate the importance of using a combination of the efficient index based first pass search and an unconstrained neural network phoneme decoder based second pass search to obtain good overall STD performance.

2. HYBRID MULTI-PASS SEARCH ALGORITHM

This section describes the hybrid spoken term detection procedure particularly as it applies to detecting OOV words. While the procedure was developed to provide a consistent approach to STD for both IV and OOV query terms, the focus in this paper is mainly on issues related to OOV queries. Prior to search, continuous audio files are segmented into short (20 - 30 second) segments, $\mathcal{S} = s_1, \ldots, s_P$, and ASR word lattices, $\mathcal{L} = l_1, \ldots, l_P$, are generated for each segment. Section 2.1 describes how a word based index of lattice paths is constructed from these segment based lattices and how this in-

dex can also accommodate retrieving lattice paths from OOV search terms. Once a query term has been entered by the user, a first pass search is performed to identify candidate audio segments that are likely to contain the search term. Section 2.2 describes how candidate lattice paths associated with these segments can be generated from an OOV query using the word based lattice index. The occurrence of a query term is verified in these candidate segments using a second pass phonemic search. Section 2.3 gives a description of how this search is performed from a linguistically constrained phoneme sequence for IV search terms and an unconstrained phoneme network for OOV search terms.

2.1. Offline - Building an Index of Lattice Paths

Given an ASR lexicon, $\mathbf{V} = \{V_1, \ldots, V_N\}$, the goal of index construction strategy in this work is to construct an inverted index so that for each word $V_i \in \mathcal{V}$, referred to here as indexing terms, there is a list of lattice paths that are likely to contain V_i and their associated lattices:

$$V_i : (p_{i,1}, l_{i,1}), (p_{i,2}, l_{i,2}), (p_{i,3}, l_{i,3}), \ldots \quad (1)$$

In Equation 1, $p_{i,j}$ is the jth path for word V_i and $l_{i,j}$ is the lattice associated with that path.

Index construction refers to the process of identifying the paths associated with each V_i. The criterion used for choosing path, $p_{i,j}$ from lattice $l_{i,j}$, for indexing term V_i is that the path have a high probability of containing V_i relative to other paths in the lattice. The following strategy for selecting a list of paths from a lattice to be added to the index for a given index term is designed to satisfy this criterion. For index term V_i find all the paths p in lattice j which contain V_i. Increment the likelihoods, $L_{i,j}$, of all these paths by an empirically chosen "scaling factor", $B = 200$, to obtain updated path likelihoods

$$L'_{i,j} = L_{i,j} + mB, \quad (2)$$

where m is the number of times V_i appears in $p_{i,j}$. The update in Equation 2 reflects an expected increase in prior probability for V_i relative to the probability predicted by the language model. Once the likelihoods of the paths containing the index term are scaled, all the paths of l_j are re-ranked and then if the highest ranking path contains V_i that path is added to the index for V_i. The motivation and a detailed description of this indexing method can be found in [9].

2.2. First Pass Search - Retrieving Segments

Given an input query term, Q, the goal of first pass search is to retrieve candidate audio segments which are most likely to contain occurrences of the query term. This corresponds to identifying lattice paths from the index described in Section 2.1 that are likely to contain Q. This process is carried out in two steps. First, the index terms that most closely match the query are identified. Second, the lattice paths associated with these index terms are retrieved for use in the second pass search for verifying the occurrence of query terms in the candidate audio segments.

The acoustic similarity between a query term Q and an index term V_i is measured based on the phonemic distance between the phonemic expansion of the query term, $\mathbf{Q} = \{q_1, \ldots, q_n\}$, and that of the index term, $\mathbf{V}_i = \{v_{i,1}, \ldots, v_{i,m_i}\}$. The phonemic expansions of IV query terms are obtained from the ASR pronunciation lexicon and phoneme expansions of OOV query terms are generated automatically. A constrained alignment is performed to obtain the phonemic distance between a query term and an index term. For the case where $\mathbf{V}$ is longer than $\mathbf{Q}$, the distance between $\mathbf{Q}$ and $\mathbf{V}_i[k]$, the subsequence of $\mathbf{V}_i$ beginning at phoneme index k, is computed as

$$\mathcal{M}(\mathbf{Q}, \mathbf{V}_i[k]) = \frac{1}{n} \sum_{l=0}^{n-1} p(q_l | v_{k+l}), \quad (3)$$

where the probabilities, $p(q|v)$, are approximated by normalized counts obtained from a phone confusion matrix. This matrix is created from time aligned decoded and reference phoneme transcriptions obtained from training utterances taken from the lecture domain. For the case where $\mathbf{V_i}$ is shorter than $\mathbf{Q}$, the order of $\mathbf{V_i}$ and $\mathbf{Q}$ in Equation 3 is reversed. The score given to each index term with respect to a query term is obtained from $\arg\max_k \mathcal{M}(\mathbf{Q}, \mathbf{V}_i[k])$.

Once the scores, $\mathcal{M}(\mathbf{Q}, \mathbf{V}_i[k])$, are computed for all the index terms with respect to a query term, Q, a set of top scoring index terms $\mathcal{I}_Q$ is identified. The lattice paths associated with these index terms are obtained for use in the second pass search. It is important to note that this strategy facilitates the use of a word based index even when the query terms are not contained in the ASR lexicon. It is also important that the first pass search only requires on the order of $|\mathcal{V}|$ string matches of the type shown on Equation 3 for each query where $|\mathcal{V}|$ is the ASR vocabulary size.

2.3. Second Pass Search - Term Detection

Identifying the candidate lattice paths in the first pass search reduces the number of acoustic segments that must be subjected to a detailed search for the query term. This has the effect of reducing the computational complexity of the search relative to performing an exhaustive detailed search of all audio segments. It will be shown in Section 4 that it also serves to significantly improve the OOV term detection performance when combined with the second pass query term verification. The occurrence of a query term in a candidate segment is verified from the query term phonemic expansion, $\mathbf{Q}$, by using the same phonemic distance as given in Equation 3.

Two different query term verification procedures are performed for IV and OOV query terms. For IV query terms, a phonemic expansion is obtained using the ASR lexicon for path $p_{i,j}$ associated with index entry $i \in \mathcal{I}_Q$ and segment s_j where $\mathcal{I}_Q$ is the set of top scoring index terms for query Q. This phonemic expansion is given by $\mathbf{H}_{i,j}$. A score is then computed for query phoneme expansion $\mathbf{Q}$ with respect to the sub-string, $\mathbf{H}_{i,j}[k]$, beginning at phoneme index, k, in the phonemic expansion of the path. A normalized score is computed for all phone offsets, all paths, and all index terms $i \in \mathcal{I}_Q$ as

$$\mathcal{D}(i,j)_k(\mathbf{Q}) = \mathcal{M}(\mathbf{Q}, \mathbf{H}_{i,j}[k]) / \mathcal{M}(\mathbf{Q}, \mathbf{Q}). \quad (4)$$

For OOV terms, a phoneme string for segment s_j is obtained from unconstrained phonemic decoding of the audio segment without applying any of the lexical or language constraints associated with ASR. The unconstrained phone decoder used for this purpose is based on a neural network (NN) phone classifier whose outputs are estimates of 135 probabilities associated with the states of a reference phone class hidden Markov model. The NN is based on the temporal pattern (TRAP) architecture whose features correspond to long temporal contexts of mel-filterbank energies [11]. The TRAP NN parameters are trained on approximately 100 hours of speech from the AMI Meetings corpus. The posterior probabilities obtained from this NN are used in a hybrid HMM/NN phone decoder with no lexical, phonotactic, or language constraints applied in decoding [12]. A phone accuracy of 44.2% was measured for the HMM/NN decoder on the lecture speech data associated with the test speech corpus described in Section 3.

Query term verification for OOV query terms is performed using the phonemic expansion of segment s_j given by $\mathbf{K}_j$ obtained using the above unconstrained HMM/NN decoder. The phonemic distance given by Equation 3 is used for finding the similarity between the phonemic expansion of the query term $\mathbf{Q}$ and $\mathbf{K}_j$. Following the procedure for IV query verification, the normalized score for the single phoneme sequence, $\mathbf{K}_j[k]$, with respect to $\mathbf{Q}$ is computed as

$$\mathcal{D}(j)_k(\mathbf{Q}) = \mathcal{M}(\mathbf{Q}, \mathbf{K}_j[k])/\mathcal{M}(\mathbf{Q}, \mathbf{Q}). \quad (5)$$

For both IV and OOV query term verification, the scores, $\mathcal{D}_k(\mathbf{Q})$, obtained for each value of k are compared to a threshold to accept or reject the hypothesized term occurrence. In order to report the exact location of the query term in the audio segment, the phoneme sequences corresponding to the candidate segments need to be time aligned. To locate hypothesized detections of OOV query term occurrences, the time alignment of the unconstrained phoneme sequence, $\mathbf{K}_j$, is generated directly by the HMM/NN phone decoder. Detected IV query term occurrences are located by Viterbi alignment of the $\mathbf{H}_{i,j}$ sequences using the LVCSR acoustic model. The STD detection performance for the above IV and OOV query term verification procedures is presented in Section 4.

3. TASK DOMAIN

The task domain used for this study consists of audio recordings of course lectures obtained from the McGill COurses OnLine (COOL) repository [10]. This repository includes a large number of audio lectures recorded in lecture room environments using a variety of microphones. For development and evaluation purposes, several of these recordings were randomly chosen and manually transcribed. The evaluation speech data consists of two lectures containing a total of 131 minutes of speech recorded through a lapel microphone from a single male speaker who speaks English as his third language. The two test lectures were segmented into 387 audio segments with an average segment length of 20 seconds. The manual transcriptions for these lectures contain a total of 17914 words.

An LVSCR system, originally developed and evaluated under the AMI project [13], was configured for this task as described in [1]. A word accuracy of 56.5 and a language model test set perplexity of 143 were measured on the test set. With a vocabulary of 52,800 words, the rate of occurrence of OOV words in the test lectures is a relatively high 11.2 percent.

To evaluate STD performance, a set of query terms were chosen from the words in the test set transcriptions based on their frequency of occurrence. After removing function words, a set of 176 of the most frequent words in the test set were chosen as query terms for the study in Section 4. The query terms consist of 142 IV words and 34 OOV words with respect to the LVCSR vocabulary. The length of the phonemic expansion of the query terms ranges from as few as 2 phonemes for "*ear*" to 17 phonemes for the term "*phenylpropanolamine*." There are a total of 1441 occurrences of IV and OOV query terms in the test set transcriptions out of which 1199 are IV and 242 are OOV occurrences.

4. EXPERIMENTAL STUDY

This section presents experimental results evaluating the performance of the two-pass search procedure presented in Section 2 for the set of IV and OOV search terms described in Section 3. The performance of the first pass search procedure is evaluated in terms of the richness of the segments retrieved using the lattice index in response to a query term. The performance of the second pass query term verification process is evaluated in terms of its ability to detect the locations of query terms in the retrieved audio segments. Additional discussion providing analysis of the word based lattice index construction is provided in [9].

4.1. Segment Retrieval Performance

In Table 1 the performance of the first search pass is evaluated at segment level for IV and OOV queries in terms of average precision-recall values. Precision is measured as the number of retrieved segments that contain occurrences of the query term relative to the total number of retrieved segments. Recall is defined as the number of retrieved segments that contain a given query term relative to the total number of segments wherein that term occurs. Table 1 shows that by performing the first pass search 79% of segments containing IV terms are retrieved, while the number of retrieved segments containing OOV terms is as low as 58.37%. A close examination of the segments containing OOV terms which were not identified in the first pass search indicated that there are two main reasons for the observed reduction in recall. The first, results from errors in the automatically derived pronunciations obtained for OOV terms. Second, there are errors resulting from the inexact phonemic distance used for associating query terms with index terms. One can re-interpret

Query Term	Segment Level		Term Level
	Precision	Recall	Recall
IV	61.89	79.21	91.16
OOV	12.83	58.37	71.49

Table 1. Performance of the candidate segments identified in the first pass search for in-vocabulary (IV) and out-of-vocabulary (OOV) query terms.

the performance of the first pass search in terms of the percentage of actual search term occurrences that are contained in the retrieved segments. This is illustrated in column three of Table 1. It can be seen that the term level recall is considerably higher than the segment level recall. This is due to the fact that many of the retrieved segments contain multiple term occurrences.

4.2. Overall STD Performance

The overall performance of the STD system is measured in terms of probability of detection, P_d, and the false alarm per query term per hour (fa/qt/hr).The probability of detection, $P_d = N_d/N_t$, is defined as the number of correctly detected query terms, N_d, normalized by the total number of actual occurrences of the query term in the test set, N_t. A query term detection is considered to be "correct" if it occurs within a 2 second window surrounding the starting time of the actual occurrence of the term in the test data. Otherwise it is labeled as a false alarm.

The curves plotted in Figure 1 are obtained by varying the threshold applied to the phone distance score described in Section 2.3. The curve labeled "IV" in Figure 1 is the average detection characteristic obtained for IV query terms using the constrained phonemic expansion of retrieved lattice paths. It is clear from Figure 1 that an overall detection rate of approximately 80% is obtained for the IV query terms across a wide range of false detections. This is quite high given that only 91% of total occurrences of the IV query terms are retrieved in the first pass search.

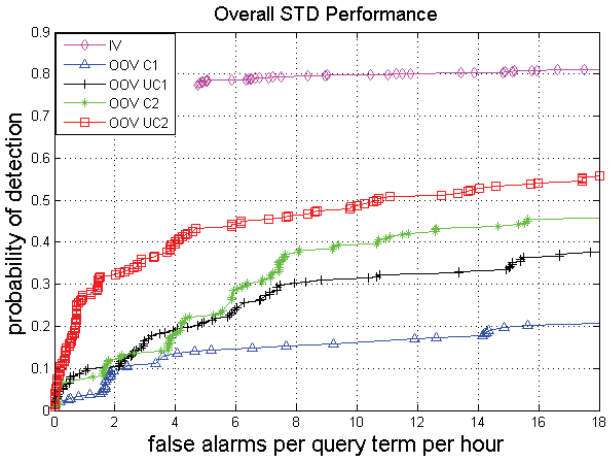

Fig. 1. STD performance for in-vocabulary (IV) and out-of-vocabulary (OOV) query terms obtained from constrained (C) and unconstrained (UC) phonemic representations.

Two scenarios were investigated for evaluating detection performance for OOV query terms. In the first case detection performance was measured without performing any pre-selection of segments. In the second scenario performance is reported for the full two-pass search strategy described in Section 2.

The curves labeled "OOV C1" and "OOV UC1" in Figure 1 correspond to scenarios where no segment pre-selection is performed. For "OOV C1" a search is performed for phonemic expansion of query terms in a phone string corresponding to the 1-best ASR hypothesis of the entire test set. For "OOV UC1" a search is performed in a phonemic representation of the entire test set generated by the unconstrained HMM/NN decoder. It is clear from the figure that search through the phone string generated by the constrained decoder resulted in significantly higher detection performance than searching through the phone string obtained from the constrained decoder.

The curves labeled "OOV C2" and "OOV UC2" in Figure 1 correspond to the two-pass scenario where term detection is preceded by segment pre-selection. The "OOV C2" curve corresponds to search through the constrained ASR generated phone strings in the retrieved segments. The "OOV UC2" corresponds to searching through the phone strings generated by the unconstrained decoder in the retrieved segments. Comparing "OOV C2" with "OOV UC2" again shows that the detection performance obtained from the phone strings generated by the unconstrained decoder is significantly higher than for the constrained case. Furthermore, and most important, it is clear that the combination of the first pass segment pre-selection and the second pass unconstrained query term verification gives the best overall OOV term detection performance.

5. SUMMARY & CONCLUSION

An efficient two pass approach to open vocabulary STD has been presented with emphasis on reducing the gap between detection performance obtained for in-vocabulary and out-of-vocabulary search terms. For OOV search terms, it was found that a combination of an efficient first pass search for retrieving relevant acoustic segments and a second pass search for verifying term occurrences from unconstrained decoded phone sequences resulted in significantly improved STD performance. The decoded phone sequences used in the second pass search were obtained from a hybrid NN/HMM decoder. A reduction in the difference between OOV and IV detection performance of approximately 25 percent over a range of false detection rates was observed relative to previous approaches to verifying OOV query terms. It is expected that this performance gap can be further reduced by incorporating additional acoustic and linguistic knowledge sources for improving OOV query term detection.

6. REFERENCES

[1] R. Rose, A. Norouzian, A. Reddy, A. Coy, V. Gupta, and M. Karafiat, "Subword-based spoken term detection in audio course lectures," in *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 5282 – 5285.

[2] Cyril Allauzen, Mehryar Mohri, and Murat Saraclar, "General indexation of weighted automata - application to spoken utterance retrieval," in *In Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval (HLT/NAACL*, 2004, pp. 33–40.

[3] Igor Szoke, Martin Karafiat, Petr Schwarz, Ilya Oparin, and Pavel Matejka, "Search in speech for public security and defense," in *IEEE Workshop on Signal Processing Applications for Public Security and Forensics, SAFE.*, 2007, pp. 1–7.

[4] U.V. Chaudhari and M. Picheny, "Improvements in phone based audio search via constrained match with high order confusion estimates," in *ASRU*. IEEE, 2007, pp. 665 –670.

[5] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, et al., "An audio indexing system for election video material," in *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE Computer Society, 2009, vol. 0, pp. 4873–4876.

[6] O. Siohan and M. Bacchiani, "Fast vocabulary-independent audio search using path-based graph indexing," in *Ninth European Conference on Speech Communication and Technology*. ISCA, 2005.

[7] P. Yu and F. Seide, "Fast two-stage vocabulary-independent search in spontaneous speech," in *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2005, vol. 5, pp. 481–484.

[8] D. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S.A. Lowe, R.M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *In Proceedings of Interspeech*, 2007, pp. 314–317.

[9] A. Norouzian and R. Rose, "An efficient approach for two-stage open vocabulary spoken term detection," in *IEEE Workshop on Spoken Language Technology Proc., 194-199, 2010.* IEEE, 2010, pp. 194–199.

[10] "COOL Courses Online," http://www.cool.mcgill.ca.

[11] P. Schwarz, P. Matějka, and J. Černockỳ, "Towards lower error rates in phoneme recognition," in *Text, Speech and Dialogue*. Springer, 2004, pp. 465–472.

[12] H. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*, Springer, 1994.

[13] T. Hain, L. Burget, J. Dines, M. Karafiat, D. van Leeuwen, M. Lincoln, G. Garau, and V. Wan, "The 2007 AMI (DA) system for meeting transcription," in *Proc. NIST RT07 Workshop*.