OPEN-VOCABULARY SPOKEN TERM DETECTION USING GRAPHONE-BASED HYBRID RECOGNITION SYSTEMS

Murat Akbacak, Dimitra Vergyri, Andreas Stolcke

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA 94025, USA
{murat,dverg,stolcke}@speech.sri.com

ABSTRACT

We address the problem of retrieving out-of-vocabulary (OOV) words/queries from audio archives for spoken term detection (STD) task. Many STD systems use the output of an automatic speech recognition (ASR) system which has a limited and fixed vocabulary, and are not capable of detecting rare words of high information content, such as named entities. Since such words are often of great interest for a retrieval task it is important to index spoken archives in a way that allows a user to search an OOV query/term.[1] In this work, we employ hybrid recognition systems which contain both words and subword units (graphones) to generate hybrid lattice indexes. We use a word-based STD system as our baseline, and present improvements by employing our proposed hybrid STD system that uses words plus graphones on the English broadcast news genre of the 2006 NIST STD task.

Index Terms— spoken term detection, audio indexing, voice search, open vocabulary

1. INTRODUCTION

Information search in audio recordings (e.g., audio broadcasts, archives from digital libraries, audio content on the Internet) is becoming more popular every day, and is expanding at an increasing rate as more audio data becomes available in different languages.

Different audio search applications are presented in the literature, such as keyword spotting and spoken document retrieval (SDR). NIST has organized evaluations in the past to evaluate state-of-the-art SDR systems [1] where the goal was to retrieve relevant audio documents in response to a query representing a topic. Recently, NIST defined a new task, spoken term detection (STD) [2], in which the goal is to locate a specified *term* rapidly and accurately in large heterogeneous audio archives, to be used ultimately as input to more sophisticated audio search systems. Unlike SDR, the STD task is formulated as a detection task. The evaluation metric has two important characteristics: (1) missing a term is penalized more heavily than having a false alarm for that term, (2) detection results are averaged over all query terms rather than over their occurrences, i.e., the performance metric considers the contribution of each term equally. Therefore, although the OOV rates are typically low in recognition systems and the OOV words are mostly infrequent words, it is still important to be able to retrieve OOV terms since they have the same impact on the performance metric as in-vocabulary (IV) terms. Furthermore, the impact of OOV on real-life applications could be substantial since OOVs tend to be associated with names, as well as with recent and/or rare events.

[1] OOV query/term is a sequence of words where at least one word is OOV.

Results of the NIST 2006 STD evaluation have shown that systems based on word recognition have an accuracy advantage over systems based on subword recognition (although they typically pay a price in run time). Yet, word recognition system are usually based on a fixed vocabulary, resulting in a word-based index that does not allow text-based searching for OOV words. To retrieve OOVs, as well as misrecognized IV words, audio search based on subword units (such as syllables and phone N-grams) has been employed in many systems [3, 4, 5, 6, 7]. During recognition, shorter units are more robust to errors and word variants than longer units, but longer units capture more discriminative information and are less susceptible to false matches during retrieval. In order to move toward solutions that address the problem of misrecognition (both IV and OOV) during audio search, previous studies have employed fusion methods [4, 6, 8, 9] to recover from ASR errors during retrieval.

Here, we propose a hybrid STD system that uses words and subword units together in the recognition vocabulary. The ASR vocabulary is augmented by graphone units as in [14]. We extract from ASR lattices a hybrid index, which is then converted into a regular word index by a post-processing step that joins graphones into words. It is important to represent ASR lattices with only words (with an expanded vocabulary) rather than with words and subword units since the lattices might serve as input to other information processing algorithms, such as for named entity tagging or information extraction, which assume a word-based representation.

In Sections 2 and 3, we provide an overview of the STD task and SRI's STD system, respectively. Recognition results and evaluation of the proposed retrieval algorithm are presented in Section 4. Discussion and future work are presented in Section 5, with conclusions in Section 6.

2. THE STD TASK

2.1. Data

The test data consists of audio waveforms, a list of regions to be searched, and a list of query terms. NIST provided development (dev06) and dry-run (dry06) test sets; however, the audio was common to both sets. The development set consisted of about 3 h of broadcast news, 3 h of conversational telephone speech, and 2 h of meetings. The dev06 set contained 1107 query terms respectively. The speech-to-text (STT) components of the system were trained using corpora available from the Linguistic Data Consortium (LDC). However, data generated after December 2003 was excluded from training STT and STD components to comply with evaluation requirements. A complete set of results for different genres can be found in [10]. For expedience we focus in this study on English and the genre with the highest OOV rate, broadcast news (BN).

2.2. Evaluation Metric

Since this is a detection task, performance can be characterized by detection error tradeoff (DET) curves of miss (P_{miss}) versus false alarm (P_{fa}) probabilities, or by a weighted function of the two probabilities. For the NIST STD06 evaluation the primary evaluation metric was the actual term-weighted value (ATWV), which is defined as follows [2]:

$$ATWV = 1 - \frac{1}{T}\sum_{t=1}^{T}(P_{miss}(t) + \beta P_{fa}(t)) \quad (1)$$

$$P_{miss}(t) = 1 - \frac{N_{corr}(t)}{N_{true}(t)}\ ,\ P_{fa}(t) = \frac{N_{spurious}(t)}{Total - N_{true}(t)} \quad (2)$$

where T is the total number of terms, β is set to approximately 1000, N_{corr} and $N_{spurious}$ are the total number of correct and spurious term detections, N_{true} is the total number of true term occurrences in the corpus, and $Total$ is the duration (in seconds) of the indexed audio corpus.

3. STD SYSTEM DESCRIPTION

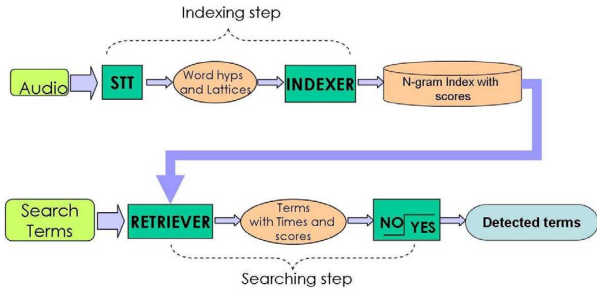

Fig. 1. *SRI spoken term detection system*

Indexing consists of two major steps in our system, as seen in Figure 1. First, audio input is run through the STT system that produces word or word+graphone recognition hypotheses and lattices. These are converted into a candidate term index with times and detection scores (posteriors). When hybrid recognition (word+graphone) is employed, graphones in the resulting index are combined into words. To be able to do this, we keep word start/end information with a tag in the graphone representation (e.g., "[[.]", "[.]]","[.]" indicate a graphone at the beginning, or end, or in the middle of a word, respectively). During the retrieval step, first the search terms are extracted from the candidate term list, and then a decision function is applied to accept or reject the candidate based on its detection score. In this section we describe in detail each of the components of the STD system.

3.1. Speech-to-Text System

3.1.1. Recognition Engine

The STT system used for this task was a sped-up version of the STT systems used in the NIST evaluation for 2004 Rich Transcription (RT-04) [11]. We are using multipass [12] systems (bigram decoding with within-word models generating word lattices, lattice expansion with a higher-order N-gram, followed by word posterior generation from expanded lattices, rescoring of expanded lattices with adapted cross-word models, and updating of word posteriors), and thus generate STT outputs with different accuracies at different stages of computation. In [10], ATWV results are presented by using these multiple-pass system outputs to show the trade-off between accuracy and speed.

STT is using SRI's Decipher(TM) speaker-independent continuous speech recognition system, which is based on continuous density, state-clustered hidden Markov models (HMMs), with a vocabulary optimized for the BN genre. More details about the BN STT engine can be found in [10, 13].

In our BN system, for acoustic training we used data distributed by LDC: 1996 and 1997 Hub-4 (200 h), TDT4 (274 h), TDT2 (272 h) and EARS BNr1234 (2300 h). We trained gender-independent within-word and cross-word models with about 500K Gaussians each. For the language model (LM) the training data was partitioned based on source. Separate component LMs were generated from each partition, and then interpolated for the final LM, using a vocabulary of 60K words.

3.1.2. Graphone-based Hybrid Recognition

To compensate for OOV words during retrieval, we used an approach and tool presented in [14] where subword units called *graphones* are used to model OOV words. The underlying assumption used in this model is that, for each word, its orthographic form and its pronunciation are generated by a common sequence of graphonemic units. Each graphone is a pair of a letter sequence and a phoneme sequence of possibly different lengths. In this data driven approach, M-gram graphone models are trained using a pronunciation dictionary. In our experiments, we used 50K words (excluding the 10K most frequent ones in our vocabulary) to train the graphone module, with maximum window length, M, set to 4. The LM training data was represented in terms of the reduced vocabulary by replacing OOV words with their graphone representations. A hybrid word+graphone LM was estimated and used for recognition. Following is an example of an OOV word modeled by graphones:

abromowitz: [[abro] [mo] [witz]]

where graphones are represented by their grapheme strings enclosed in brackets, and "[[" and "]]" tags are used to mark word boundary information that is later used to join graphones back into words for indexing. The graphone vocabulary and pronunciations are automatically inferred from the dictionary of in-vocabulary words, using the method described in [14].

3.2. N-gram Indexing

Since the lattice structure provides additional information about the correct hypothesis could appear, to avoid misses (which have a higher cost in the evaluation score than false alarms) several studies have used the whole hypothesized word lattice [6, 8] to obtain the searchable index. We used the *lattice-tool* in SRILM [15] (version 1.5.1) to extract the list of all word/graphone N-grams (up to $N = 5$ for a word-only (W) STD system, $N = 8$ for a hybrid (W+G) STD system). The term posterior for each N-gram is computed as the forward-backward combined score (acoustic, language, and prosodic scores were used) through all the lattice paths that share the N-gram nodes. We used a 0.5 second time tolerance to merge same N-grams with different times. All N-gram terms with posterior score greater than 0.001 were sorted alphabetically and incorporated into the term index.

Table 1. *OOV statistics for different vocabulary sizes.*

Vocab. Size	60K	20K	10K
OOV_{doc}	0.49%	1.57%	3.30%
OOV_{query}	0.03%	0.06%	0.18%
$Num.Query_{OOV}$	50	100	186

3.3. Term Retrieval

The term retrieval was implemented using the Unix `join` command, which concatenates the lines of the sorted term list and the index file for the terms common to both. No effort was spent on optimizing the runtime of the retrieval component. The computational cost of this simple retrieval mechanism depends only on the size of the index.

Each putative retrieved term is marked with a hard decision (YES/NO). Our decision-making module relies on the posterior probabilities generated by the STT system. One of two techniques were employed during the decision-making process. The first one determines a global threshold for posterior probability (GL-TH) by maximizing the ATWV score, which for this task was found to be 0.4 and 0.0001 for word-based and hybrid systems respectively. An alternative strategy can be formulated that computes a term-specific threshold (TERM-TH), which has a simple analytical solution [16]. Based on decision theory the optimal threshold θ for each candidate should satisfy

$$\theta \cdot V_{hit} - (1-\theta) \cdot C_{fa} = 0 \quad \Longleftrightarrow \quad \theta = \frac{C_{fa}}{V_{hit} + C_{fa}} \qquad (3)$$

where V_{hit} is the value of a correct detection and C_{fa} is the cost for a false alarm. For the ATWV metric we have

$$V_{hit} = \frac{1}{N_{true}(t)} \quad , \quad C_{fa} = \frac{\beta}{Total - N_{true}(t)}. \qquad (4)$$

Since the number of true occurrences of the term is unknown we approximate it for the calculation of the optimal θ by the sum of the posterior probabilities of the term in the corpus.

4. EXPERIMENTAL RESULTS

We evaluated the baseline and proposed algorithms on the devset of the English BN task, which consists of 3 h of speech data and 1107 query terms. Since our STT system has an unrealistically low OOV rate on the given test data, and to study the effect of varying OOV rates, we prepared systems with three different word vocabulary sizes: 60K words, 20K words, and 10K words. In the hybrid (W+G) STD system there are approximately 15K graphones added to the recognition vocabulary. Table 1 shows the OOV statistics for different vocabulary sizes, which are calculated for the document list and the term list separately.

Table 2 shows the word error rate (WER) values for these different recognition systems. The first row shows the WER when a word-only (W) recognizer is used. In terms of the impact of OOV words on WER, an observation from previous studies can also be made here: on average every OOV word results in two errors (itself, and one for a neighboring word because of incorrect context in the language model scoring). In the second row, WER results for the graphone-based hybrid (W+G) recognition system are shown. Almost half the errors resulting from OOV words are recovered by using a hybrid system (e.g., going from 60K vocabulary to 20K vocabulary leads to a 3.2% increase in WER, and the hybrid system brings this number down to 1.3%). Note that the improvements in oracle-WER (which is more correlated with STD performance since lattice indexes are used for search in our system) are expected to be even greater.

Table 2. *WER values for different vocabulary sizes using word-only recognizer (W) vs. hybrid recognizer using words and graphones (W+G).*

Vocab. Size	60K	20K	10K
WER_W	15.1%	18.3%	21.0%
WER_{W+G}	14.8%	16.4%	18.2%

Table 3. *ATWV scores calculated (via GL-TH and TERM-TH) for IV queries and for all queries, using a word-only recognizer (W) with different vocabulary sizes.*

Vocab. Size	60K	20K	10K
$ATWV_{IV}$ - (GL-TH)	0.868	0.878	0.830
$ATWV_{all}$ - (GL-TH)	0.835	0.808	0.730
$ATWV_{IV}$ - (TERM-TH)	0.901	0.908	0.887
$ATWV_{all}$ - (TERM-TH)	0.867	0.836	0.753

Tables 3 and 4 show ATWV scores for word-only (W) and hybrid (W+G) systems respectively. Table 4 shows how the ATWV score (computed with both GL-TH and TERM-TH) changes for the reduced vocabulary systems. In both word-only and hybrid systems, term-specific thresholding (TERM-TH) consistently yields higher ATWV scores in compare to global-thresholding scheme (GL-TH). When the corresponding rows from Tables 3 and 4 are compared, you can observe how the hybrid system compensates for some of the performance loss of the word systems when the OOV rate is high.

An interesting observation is that even for IV terms the hybrid (W+G) STD yields better performance than the word-only (W) STD system. This is because hybrid recognition improves both IV-word and OOV-word recognition, resulting in better retrieval performance for IV and OOV words at the same time.

Table 4. *ATWV scores calculated (via GL-TH and TERM-TH) for IV queries, OOV queries, and all queries, using a hybrid recognizer (W+G) with different vocabulary sizes.*

Vocab. Size	60K	20K	10K
$ATWV_{IV}$ - (GL-TH)	0.873	0.882	0.841
$ATWV_{OOV}$ - (GL-TH)	0.201	0.256	0.328
$ATWV_{all}$ - (GL-TH)	0.842	0.828	0.764
$ATWV_{IV}$ - (TERM-TH)	0.911	0.914	0.892
$ATWV_{OOV}$ - (TERM-TH)	0.245	0.288	0.359
$ATWV_{all}$ - (TERM-TH)	0.889	0.872	0.785

5. DISCUSSION AND FUTURE WORK

Further improvements can be obtained from different parameter settings during graphone extraction. In the current system, this step is optimized to yield the lowest 1-best WER, but it can be optimized for the ATWV metric directly. Also, during recognition we used a fixed setup, but one can also use settings that might be more suitable for a hybrid system (e.g., using higher-order LM in the first pass as well as in the rescoring step). Although we evaluated the graphone-based system only in English and obtained improvements, we believe that morphologically rich languages (e.g., Arabic) will benefit from the hybrid STD approach even more since vocabulary size expands more quickly in these languages. Future work will evaluate the hybrid STD system in such languages.

6. CONCLUSION

In this paper, we have presented improvements to SRI's STD system, focusing on the problem of handling OOV queries and terms. We showed that a hybrid word/subword recognition framework using "graphones" works well in this context and gave substantial reduction on the NIST 2006 spoken term detection task. By limiting the size of the word vocabulary we investigated this approach at different OOV word rates. At document OOV rates of 0.5%, 1.6%, and 3.3%, we observed relative improvements in the detection cost metric of between 2.5% and 4.3%.

7. ACKNOWLEDGMENTS

We thank Christian Gollan and Ralf Schlüter of RWTH Aachen for helping with in the implementation of the graphone method in our recognition system. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

8. REFERENCES

[1] J. Garofolo et al., "TREC-6 1997 Spoken Document Retrieval Track Overview and Results", *NIST Special Publication*, vol. 500, no. 240, pp. 83–92, 1998.

[2] NIST, "The Spoken Term Detection (STD) 2006 Evaluation Plan, `http://www.nist.gov/speech/tests/std/docs/std06-evalplan-v10.pdf`, 2006.

[3] M. Witbrock and A. Hauptmann, "Using Words and Phonetic Strings for Efficient Information Retrieval from Imperfectly Transcribed Spoken Documents", *Proc. 2nd ACM Int. Conf. on Digital Libraries*, pp. 30-35, 1997.

[4] M.G. Brown et al., "Open-vocabulary Speech Indexing for Voice and Video Mail Retrieval", *Proc. ACM Multimedia*, pp. 307-316, Boston, 1996.

[5] G.J.F. Jones et al., "Retrieving spoken documents by combining multiple index sources", *Proc. SIGIR 96*, pp. 3038, Zurich, 1996.

[6] D.A. James and S.J. Young, "A Fast Lattice-Based Approach to Vocabulary Independent Word spotting", *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1029-1032, Istanbul, Turkey, 2000.

[7] K. Ng and V.W. Zue, "Subword-based Approaches for Spoken Document Retrieval", *Speech Communication*, vol. 32, no. 3, pp. 157-186, October 2000.

[8] C. Allauzen et al., "General Indexation of Weighted Automata - Application to Spoken Utterance Retrieval", *Proc. HLT-NAACL Conf.*, 2004.

[9] M. Saraclar and R. Sproat, "Lattice-based search for Spoken Utterance Retrieval", *Proc. HLT-NAACL Conference*, Boston, 2004.

[10] D. Vergyri et al., "The SRI/OGI 2006 Spoken Term Detection System", *Proc. of Interspeech Conf.*, Belgium, 2007.

[11] A. Stolcke et al., "STT Research and Developement at SRI-ICSI-UW", `http://www.sainc.com/richtrans2004/uploads/monday/SRI-ISCI-UW.ppt`, 2004.

[12] A. Stolcke et al., "Recent Innovations in Speech-to-Text Transcription at SRI-ICSI-UW", *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1729-1744, 2006.

[13] J. Zheng and A. Stolcke, "Improved Discriminative Training Using Phone Lattices", *Proc. of Interspeech Conf.*, pp. 2125-2128, Portugal, 2005.

[14] M. Bisani and H. Ney, "Open Vocabulary Speech Recognition with Flat Hybrid Models", *Proc. of Interspeech*, pp. 725–728, 2005.

[15] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit", *Proc. Intl. Conf. Spoken Language Processing*, Denver, Colorado, September 2002 .

[16] D. Miller et al., "Rapid and Accurate Spoken Term Detection", *Proc. of Interspeech Conf.*, Belgium, 2007.