SUBWORD-BASED SPOKEN TERM DETECTION IN AUDIO COURSE LECTURES

[1]*Richard Rose,* [1]*Atta Norouzian,* [1]*Aarthi Reddy,* [1]*Andre Coy,* [2]*Vishwa Gupta,* [3]*Martin Karafiat*

[1]Department of ECE
McGill University
Montreal, Canada

[2]Centre de Recherche
Informatique de Montreal
Montreal, Canada

[3]Brno University of Technology
Brno, Czech Republic

ABSTRACT

This paper investigates spoken term detection (STD) from audio recordings of course lectures obtained from an existing media repository. STD is performed from word lattices generated offline using an automatic speech recognition (ASR) system configured from a meetings domain. An efficient STD approach is presented where lattice paths which are likely to contain search terms are identified and an efficient phone based distance is used to detect the occurrence of search terms in phonetic expansions of promising lattice paths. STD and ASR results are reported for both in-vocabulary (IV) and out-of-vocabulary (OOV) search terms in this lecture speech domain.

Index Terms— Speech recognition, spoken term detection

1. INTRODUCTION

The ability to search online media is of value for many applications including access to recorded lectures, broadcast news, voice mail messages, and conversational telephone speech. Spoken document retrieval (SDR) and spoken term detection (STD) have been active areas of research over the last decade [1, 2, 3, 4, 5, 6]. Many applications, for example, search of recorded audio lectures, involve locating audio segments within potentially hundreds of hours of audio in response to queries entered by a user. The requirements of achieving subsecond response times to these queries for an unlimited vocabulary of search terms limits the range of STD approaches that might be considered practical in this scenario.

Many of the systems developed for these applications begin with a large vocabulary automatic speech recognition (ASR) system generating word transcriptions or word lattices from spoken audio documents [1, 4, 5, 3, 6, 7]. It is generally acknowledged that lattice based techniques for STD can yield better recall performance, especially when the ASR word accuracty (WAC) is low [5, 3]. In word based STD and SDR systems, deriving scores for search terms from decoded occurrences in ASR word lattices has been shown to improve performance over systems that rely on a single recognized string [5]. Furthermore, there have been several proposals for extending these approaches to open vocabulary STD by using efficient means for re-scoring phone level lattices [5].

The interest in this work is in applications which require subsecond response latencies for locating search terms in potentially very large audio repositories. It is expected that any scheme involving exhaustive re-scoring of lattice hypotheses as done, for example, in [7] will have unreasonable computational complexity. To address this issue, an efficient STD approach is presented in Section 3. Lattice paths that are likely to contain search terms are identified and a fast phone based distance measure is used to detect the occurence of search terms in phonetic expansions of promising lattice paths. This approach is evaluated in a lecture speech domain where recorded course lectures stored in a variety of formats are made available to users via an online multimedia server [8]. A description of the recorded audio lectures and the lecture speech indexing task is provided in Section 2. The experimental study is presented in Section 5.

The performance of ASR, SDR, and STD in the lecture speech domain has proven to be problematic relative to other application domains [2, 3, 9]. Both ASR word error rates (WER) and languange model (LM) perplexity (PPL) for the 2006 Rich Transcription lecture speech track were higher than more general spontaneous speech domains [9]. Poorer ASR and STD performance for lecture speech as compared to other domains were also found in [3]. To address this issue, acoustic and language modeling techniques for ASR are described in Section 4 and are shown to have considerable impact on both ASR and STD performance for the lecture speech task described in Section 2.

2. TASK DOMAIN

The task domain used for this study consists of audio recordings of course lectures obtained from the McGill COurses OnLine (COOL) repository [8]. There are a large number of course lectures and public speeches in the repository and, as with many collections of this type, they are collected in a variety of lecture halls often times with microphone and recording equipment provided by the lecturer. This lack of control over the acoustics and recording equipment results in huge variability in quality. One can find lectures ranging in quality from being nearly inaudible in some cases to having reasonably high signal-to-noise ratio (SNR) in others. A collection

This work was partially supported by NSERC, the Heritage Canada Foundation, CRIM, European AMIDA project FP6-033812, and Czech Ministry of Education project No. MSM0021630528

of lectures, each slightly over an hour in length, recorded using lapel microphones were chosen for study in this work.

The STD techniques being investigated here are being applied to a search engine that accepts typed search terms as input from a user and returns a list of audio segments of lectures from the COOL website for review by the user. A search engine has been developed based on a two pass procedure where searchable indices are first created off-line from audio files and then search is performed by locating phonetic expansions of query terms in these indices. Searchable indices were created for a collection of approximately 20 chemistry lectures available on the McGill COOL website. These indices form the searchable representation of the lecture audio for an online search engine hosted by CRIM [10].

3. SPOKEN TERM DETECTION

This section presents an efficient lattice based approach for estimating word level scores in open vocabulary STD. It is motivated by lattice re-scoring STD approaches proposed in [5] and relies on a simple phone based distance measure similar to that proposed in [6]. ASR is performed offline as described in Section 4 to produce word lattices for each audio segment. STD search is performed using an efficient two step procedure. First, for a given query term, Q, individual paths in ASR word lattices are selected for further evaluation based on their proximity to Q. Second, a phonetic expansion is obtained for the closest scoring path in the word lattice and this is searched for instances of phone sequences that are a close match to the phonetic expansion of the query term.

After ASR lattices have been generated off line, inverted indices are created. For each word, W_i, in the lattice, there is a list of paths that contain W_i along with the likelihoods for those paths

$$W_i : (p_{i,1}, L_{i,1}), (p_{i,2}, L_{i,2}), (p_{i,3}, L_{i,3}), \ldots \quad (1)$$

where $p_{i,j}$ is the jth path index for word W_i and $L_{i,j}$ is the path likelihood for the jth path that contains W_i.

For in-vocabulary (IV) search terms, finding the lattice paths that are "close" to the query in the first step of the search procedure is straight-forward. For an in-vocabulary query term, Q, index entries $\hat{i}$ are found such that $W_{\hat{i}} = Q$. For all $j = 1, 2, 3, \ldots$, path likelihoods, $L_{\hat{i},j}$, are incremented for paths containing $W_{\hat{i}}$ by an empirically chosen "boosting factor", B, to obtain boosted path likelihoods $L'_{\hat{i},j} = L_{\hat{i},j} + mB$, where m is the number of occurrences of W_i in $p_{\hat{i},j}$. The new highest ranking lattice path is identified based on L' and a phonetic expansion is obtained for this path.

In the second step of STD search, search for occurrences of Q in the phonetic expansion of the top scoring path is performed using a constrained phonetic string alignment. A score is computed for the phonetic expansion of the query, $\mathbf{Q} = \{q_0, q_1, \ldots, q_{n-1}\}$, with respect to each phone index, k, in the phonetic expansion of the re-ranked path. The score is computed for phone sequences of length n beginning in the top ranking phone string at phone index, k: $\mathbf{H}_k = \{h_k, \ldots, h_{k+n-1}\}$,

$$\mathcal{M}(\mathbf{Q}, \mathbf{H}_k) = \frac{1}{n} \sum_{l=0}^{n-1} p(q_l | h_{k+l}). \quad (2)$$

A normalized distance is computed from this score as $\mathcal{D}_k(\mathbf{Q}) = \mathcal{M}(\mathbf{Q}, \mathbf{H}_k)/\mathcal{M}(\mathbf{Q}, \mathbf{Q})$. The probabilities $p(q|h)$ are approximated by normalized counts taken from phone confusion matrices. These are computed using time aligned decoded and reference phoneme transcriptions obtained from training speech taken from the lecture domain.

For OOV search terms, the first step of finding the lattice paths that are close to the search term differs from the IV case since the process of associating the search term with the lattice index entries is not as straight-forward. When a search term, Q, is entered by the user, the index entry, $W_{\hat{i}}$, is found such that $\hat{i} = \arg\max_i \mathcal{M}(\mathbf{Q}, \mathbf{V}_i)$, where $\mathbf{V}_i$ is the phonetic expansion of W_i. The same process as described above for IV search terms is performed for the OOV terms. Boosted likelihoods are obtained for the lattice paths associated with $W_{\hat{i}}$, the paths are re-ordered based on the boosted likelihoods, and phonetic search is performed on the phonetic expansion of the top scoring path. In Section 5, STD results are reported for this approach on a subset of the lecture data described in Section 2.

4. ACOUSTIC / LANGUAGE MODELING FOR ASR

This section describes the ASR system and its application to the lecture speech task described in Section 2. The acoustic and language modeling techniques and the speech and text corpora used to train them were developed under the AMI project [9].

4.1. Acoustic Modeling

Acoustic modeling is performed in a hybrid feature space [9]. Perceptual linear prediction (PLP) based acoustic analysis is performed with first, second, and third difference coefficients concatenated with static coefficients to obtain a 52 dimensional feature vector. An HLDA transformation is performed to reduce the feature vector dimensionality to 39 components. Posterior features are obtained using neural network based phoneme state posterior estimators. The posterior features are transformed to a feature dimension of 25 and concatenated with the PLP features resulting in a combined 64 component feature vector.

Speaker normalization is performed using vocal tract length normalization. Speaker and environment adaptation is performed using unsupervised constrained maximum likelihood linear regression (CMLLR) which is applied as a feature space transformation. Recognition is performed in multiple passes. Lattices are generated using a bigram language model and then the lattices are rescored using the trigram LM described in the following section. The acoustic hidden Markov model (HMM) was trained using discriminative minimum phone error (MPE) training from approximately 100 hours of meetings conducted at several sites participating in the AMI

project [9]. Most of the speakers in the training corpus are non-native speakers of English.

4.2. Language Modeling

The baseline language model was trained by interpolating language models from many different sources including meetings transcriptions and transcriptions of telephone conversations and news broadcasts [9]. A test set perplexity of 148 was measured on the lecture data. It was found in [9] that the perplexity of this LM when measured on a variety of meetings scenarios was approximately 100. However, it was also found that perplexities as high as 140 were obtained when perplexity was evaluated for a similar language model on data taken from the NIST RT07 development set lecture scenario. This agrees with observations made in Section 1 suggesting that the lectures very often correspond to highly specialized domains that are not well modeled by data collected from more general domains.

The following scenario was considered for adapting the above LM to the lecture domain data described in Section 2. First, additional text data from the chemistry domain was obtained by locating an online glossary of chemistry related terms. Second, ten lectures from chemistry courses at McGill were transcribed by human transcribers. This additional text amounted to a total of 135,400 words and was used for training a separate "domain specific" trigram LM. A domain adapted (DA) LM was obtained by interpolating this new LM with the baseline LM described above. The total vocabulary size of the original LM is 50,000 words and the vocabulary size of the DA LM is 52,800 words. The OOV rate for the original LM measured on the test transcriptions is 12.2% which is very high. The OOV rate for the DA LM is 11.2%, representing only a small reduction. The results presented in Section 5 show that the reduction in test set perplexity is also fairly small.

5. EXPERIMENTAL STUDY

This section investigates spoken term detection performance for the COOL lecture speech domain.

5.1. Test Set and Evaluation Metrics

The test set consists of two of the recorded lectures taken from the task domain described in Section 2. The two lectures contain a total of 131 minutes of speech data recorded over a lapel microphone from a single male speaker who speaks English as his third language. Time aligned reference transcriptions were produced by an automatic segmentation procedure. These speech segments were then processed separately by the ASR system. Segments ranged from several seconds to over two minutes in length. It is expected that any reasonable automatic segmentation procedure would result in similar performance.

The search terms were chosen based on their frequency of occurence in the lecture utterances. A set of the 184 most frequently occurring non-function words were chosen from the test text transcriptions and used as keywords in the STD evaluation. Only 150 of these keywords correspond to words contained in the ASR vocabulary and the remaining 34 words are out of vocabualry (OOV). There are a total of 2004 keyword occurrences out of a total of 17,914 words in the test set text transcriptions.

ASR performance is presented below as word accuracy (WAC). The STD performance is reported as the recall rate or the probability of keyword detection, $P_d = N_d/N_t$, where N_d correponds to the number of correctly decoded keywords and $N_t = 2004$ corresponds to the total number of keywords in the test utterances. This is reported with respect to the total number of false detections per keyword that are generated for the entire test set normalized by the test set duration, $T = 2.2$ hours.

Figure 1 presents the performance as a plot of P_d versus false alarms per keyword per hour (fa/kw/hr) which is generated by applying a threshold to the normalized distance, $\mathcal{D}_k(\mathbf{Q})$, defined in Section 3. A single value for STD performance is reported in Table 1 corresponding to P_d evaluated at 10 fa/kw/hr. This is one of many measures used to evaluate STD performance including the NIST actual term weighted value measure (ATWV) which is a weighted average of detection and false alarm probabilities [4]. The ATWV was not used here since there is an interest in looking at the term based recall and precision performance separately.

5.2. ASR Performance

A study was performed to evaluate the impact of the acoustic and language modeling techniques described in Section 4 on both ASR and STD performance. Table 1 summarizes the experimental results according to three different measures computed on the test set described in Section 5.1 for multiple configurations of the ASR system. ASR performance is reported as LM perplexity and WAC, and STD performance is reported as P_d evaluated at 10 fa/kw/hr.

ASR	PPL	WAC	IV P_d at 10 fa/kw/hr
Baseline	148	41.7	-
Enhanced AM	148	53.1	49.8
DA-LM	143	54.5	50.5
DA-LM-LR	143	46.5	61.7

Table 1. Performance for enhanced acoustic model (AM), domain adapted (DA) LM, and lattice-rescored (LR)

The impact of the acoustic modeling techniques on ASR WAC can be seen by comparing rows one and two of Table 1. The baseline system uses PLP acoustic features, maximum likelihood trained acoustic HMM models, and the same trigram LM used for the "enhanced" acoustic model (AM). Clearly, the combination of the enhanced feature representation, speaker normalization, speaker and environment adaptation, and discriminative training has significant effect on both measures. However, the relatively low ASR WAC of 53.1%

reflects the inherent difficulty of the task. This is consistent with results reported elsewhere on lecture domain tasks [3, 9].

Comparison of rows two and three of the table shows that the domain adapted (DA) LM described in Section 4.2 provides only a small reduction in perplexity and a small increase in WAC relative to the baseline. It is suspected that this is partly due to the highly specialized domain and partly due to the relatively small size of the available domain specific text. However, it also suggests that improving ASR WAC for this domain by incorporating more relevant LM training text is much harder that we expected. The fourth row of Table 1, labeled as DA-LM-LR, shows the STD performance obtained using the lattice re-scoring (LR) procedure described in Section 3. This has the most significant effect on STD performance, resulting in an increase of 22% in P_d at 10 fa/kw/hr. Not surprisingly, the WAC for this condition actually decreases due to the high number of added insertion and substitution errors.

5.3. STD Performance

Recall (P_d) versus false detection (fa/kw/hr) curves are plotted in Figure 1 for both IV and OOV search terms. The curves labeled "DA-IV" and "DA-OOV" give the detection characteristic when the phone based distance measure, $\mathcal{D}_k(\mathbf{Q})$, is applied to the most likely word string obtained using the DA-LM ASR system from Table 1. The left-most point of the DA-IV curve shows that the ASR system obtains a 34.1% recall rate at 1 fa/kw/hr for the IV words. The DA-IV curve shows that applying a decision threshold to $\mathcal{D}_k(\mathbf{Q})$ yields an increase in recall (P_d) of approximately 13% with respect to the 34.1% recall rate obtained by the ASR system at only 3 fa/kw/hr. The curves labeled "DA-LR-IV" and "DA-LR-OOV" give the detection characteristic when $\mathcal{D}_k(\mathbf{Q})$ is applied after the lattice re-scoring procedure described in Section 3. For the IV case, the recall obtained for DA-LR-IV is over 10% higher than that obtained for DA-IV, but only for false alarm rates greater than approximately 7 fa/kw/hr.

The DA-OOV and DA-LR-OOV curves in Figure 1 follow similar trends as the IV curves; however, the best recall rate for OOV queries is about half that obtained for the IV queries. It is interesting to note that similar gains are obtained for OOV terms as are obtained for IV terms by applying $\mathcal{D}_k(\mathbf{Q})$ after lattice re-scoring.

6. SUMMARY AND CONCLUSION

A study of ASR and STD performance has been presented for a task involving lecture speech from audio recordings obtained from an online media repository. An efficient approach to STD was presented where lattice paths that are likely to contain search terms are identified and a fast phone based search algorithm is used for term detection. The approach is suitable for vocabulary independent tasks that require extremely fast response times in searching potentially very large indices. It was found that searching ASR word lattices, rather than single best decoded word strings, provides significant increase in query term recall when higher rates of false term detections are acceptable.

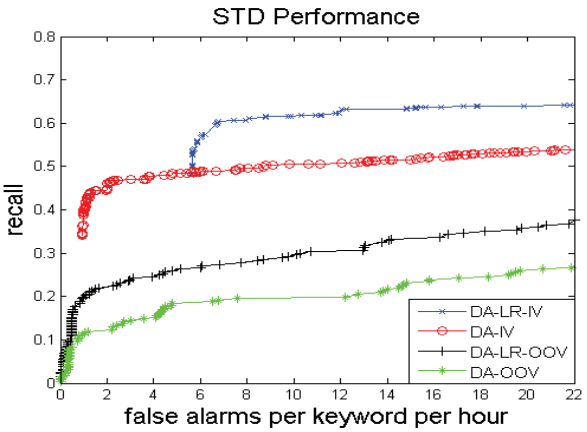

Fig. 1. Spoken term detection (STD) performance for in-vocabulary (IV) and out-of-vocabulary (OOV) search terms plotted as recall (P_d) versus false detection (fa/kw/hour).

7. REFERENCES

[1] "The spoken term detection (STD) 2006 evaluation plan (NIST)," http://www.nist.gov/speech/tests/std/docs/std06-evalplan-v10.pdf, 2006.

[2] C. Chelba and A. Acero, "Speech Ogle: indexing uncertainty for spoken document search," in *Annual Meeting - Assoc. for Computational Linguistics*, 2005, vol. 43, p. 41.

[3] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proc. 30th annual international ACM SIGIR conf. on Research and Development in Information Retrieval*. New York, NY, 2007, pp. 615–622.

[4] B. Matthews, U. Chaudhari, and B. Ramabhadran, "Fast audio search using vector space modelling," in *IEEE Workshop on Automatic Speech Recog. & Understanding*, 2007, pp. 641–646.

[5] P. Yu and F. Seide, "Fast two-stage vocabulary-independent search in spontaneous speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05)*, 2005, vol. 1.

[6] UV Chaudhari and M. Picheny, "Improvements in phone based audio search via constrained match with high order confusion estimates," in *IEEE Workshop on Automatic Speech Recognition & Understanding, 2007. ASRU*, 2007, pp. 665–670.

[7] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," *Proceedings of HLT-NAACL*, pp. 129–136, 2004.

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

[9] 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*. Springer, 2007.

[10] "CRIM-McGill-COOL," http://cool.crim.ca/coolMcGill/.