Fast decoding for open vocabulary spoken term detection

[1]B. Ramabhadran,[1]A. Sethy, [2]J. Mamou*[1] B. Kingsbury, [1] U. Chaudhari

[1]IBM T. J. Watson Research Center
Yorktown Heights,NY

[2]IBM Haifa Research Labs
Mount Carmel,Haifa

Abstract

Information retrieval and spoken-term detection from audio such as broadcast news, telephone conversations, conference calls, and meetings are of great interest to the academic, government, and business communities. Motivated by the requirement for high-quality indexes, this study explores the effect of using both word and sub-word information to find in-vocabulary and OOV query terms. It also explores the trade-off between search accuracy and the speed of audio transcription. We present a novel, vocabulary independent, hybrid LVCSR approach to audio indexing and search and show that using phonetic confusions derived from posterior probabilities estimated by a neural network in the retrieval of OOV queries can help in reducing misses. These methods are evaluated on data sets from the 2006 NIST STD task.

1 Introduction

Indexing and retrieval of speech content in various forms such as broadcast news, customer care data and on-line media has gained a lot of interest for a wide range of applications from market intelligence gathering, to customer analytics and online media search. Spoken term detection (STD) is a key information retrieval technology which aims open vocabulary search over large collections of spoken documents. An approach for solving the out-of-vocabulary (OOV) issues (Saraclar and Sproat, 2004) consists of converting speech into phonetic, syllabic or word-fragment transcripts and representing the query as a sequence of phones, syllables or word-fragments respectively. Popular approaches include subword decoding (Clements et al., 2002; Mamou et al., 2007; Seide et al., 2004; Siohan and Bacchiani, 2005) and representations enhanced with phone confusion probabilities and approximate similarity measures (Chaudhari and Picheny, 2007).

2 Fast Decoding Architecture

The first step in converting speech to a searchable index involves the use of an ASR system that produces word, word-fragment or phonetic transcripts. In this paper, the LVCSR system is a discriminatively trained speaker-independent recognizer using PLP-derived features and a quinphone acoustic model with approximately 1200 context dependent states and 30000 Gaussians. The acoustic model is trained on 430 hours of audio from the 1996 and 1997 English Broadcast News Speech corpus (LDC97S44, LDC98S71) and the TDT4 Multilingual Broadcast News Speech corpus (LDC2005S11).

The language model used for decoding is a trigram model with 84087 words trained on a collection of 335M words from the following data sources: Hub4 Language Model data, EARS BN03 closed captions and GALE Broadcast news and conversations data. A word-fragment language model is built on this same data after tokenizing the text to fragments using a fragment inventory of size 21000. A greedy search algorithm assigns the longest possible matching fragment first and iteratively uses the next longest possible fragment until the entire pronunciation of the OOV term has been represented

*The work done by J. Mamou was partially funded by the EU projects SAPIR and HERMES

Proceedings of NAACL HLT 2009: Short Papers, pages 277–280,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

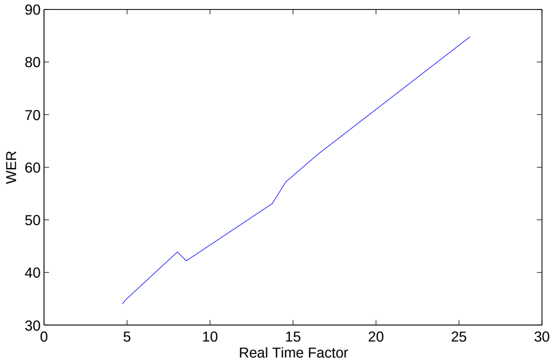

Figure 1: Speed vs WER

by sub-word units.

The speed and accuracy of the decoding are controlled using two forms of pruning. The first is the standard likelihood-based beam pruning that is used in many Viterbi decoders. The second is a form of Gaussian shortlisting in which the Gaussians in the acoustic model are clustered into 1024 clusters, each of which is represented by a single Gaussian. When the decoder gets a new observation vector, it computes the likelihood of the observation under all 1024 cluster models and then ranks the clusters by likelihood. Observation likelihoods are then computed only for those mixture components belonging to the top maxL1 clusters; for components outside this set a default, low likelihood is used. To illustrate the trade-offs in speed vs. accuracy that can be achieved by varying the two pruning parameters, we sweep through different values for the parameters and measure decoding accuracy, reported as word error rate (WER), and decoding speed, reported as times faster than real time (xfRT). For example, a system that operates at 20xfRT will require one minute of time (measured as elapsed time) to process 20 minutes of speech. Figure 1 illustrates this effect on the NIST 2006 Spoken Term Detection Dev06 test set.

3 Lucene Based Indexing and Search

The main difficulty with retrieving information from spoken data is the low accuracy of the transcription, particularly on terms of interest such as named entities and content words. Generally, the accuracy of a transcript is measured by its word error rate (WER), which is characterized by the number of substitutions, deletions, and insertions with respect to the correct audio transcript. Mamou (Mamou et al., 2007) presented the enhancement in recall and precision by searching on word confusion networks instead of considering only the 1-best path word transcript. We used this model for searching in-vocabulary queries.

To handle OOV queries, a combination of word and phonetic search was presented by Mamou (Mamou et al., 2007). In this paper, we explore fuzzy phonetic search extending Lucene[1], an Apache open source search library written in Java, for indexing and search. When searching for these OOVs in word-fragment indexes, they are represented phonetically (and subsequently using word-fragments) using letter-to-phoneme (L2P) rules.

3.1 Indexing

Each transcript is composed of basic units (e.g., word, word-fragment, phones) associated with a begin time, duration and posterior probability. An inverted index is used in a Lucene-based indexing scheme. Each occurrence of a unit of indexing u in a transcript D is indexed on its timestamp. If the posterior probability is provided, we store the confidence level of the occurrence of u at the time t that is evaluated by its posterior probability $Pr(u|t, D)$. Otherwise, we consider its posterior probability to be one. This representation allows the indexing of different types of transcripts into a single index.

3.2 Retrieval

Since the vocabulary of the ASR system used to generate the word transcripts is known, we can easily identify IV and OOV parts of the query. We present two different algorithms, namely, exact and fuzzy search on word-fragment transcripts. For search on word-fragment or phonetic transcripts, the query terms are converted to their word-fragment or phonetic representation.

Candidate lists of each query unit are extracted from the inverted index. For fuzzy search, we retrieve several fuzzy matches from the inverted index for each unit of the query using the edit distance weighted by the substitution costs provided by the confusion matrix. Only the matches whose weighted

[1] http://lucene.apache.org/

edit distance is below a given threshold are returned. We use a dynamic programming algorithm to incorporate the confusion costs specified in the matrix in the distance computation. Our implementation is fail-fast since the procedure is aborted if it is discovered that the minimal cost between the sequences is greater than a certain threshold.

The score of each occurrence aggregates the posterior probability of each indexed unit. The occurrence of each unit is also weighted (user defined weight) according to its type, for example, a higher weight can be assigned to word matches instead of word-fragment or phonetic matches. Given the nature of the index, a match for any query term cannot span across two consecutively indexed units.

3.3 Hybrid WordFragment Indexing

For the hybrid system we limited the word portion of the ASR system's lexicon to the 21K most frequent (frequency greater than 5) words in the acoustic training data. This resulted in roughly 11M (3.1%) OOV tokens in the hybrid LM training set and 1127(2.5%) OOV tokens in the evaluation set. A relative entropy criterion described in (Siohan and Bacchiani, 2005) based on a 5-gram phone language model was used to identify fragments. We selected 21K fragments to complement the 21K words resulting in a composite 42K vocabulary. The language model text (11M (3.1%) fragment tokens and 320M word tokens) was tokenized to contain words and word-fragments (for the OOVs) and the resulting hybrid LM was used in conjunction with the acoustic models described in Section 2.

4 Neural Network Based Posteriors for Fuzzy Search

In assessing the match of decoded transcripts with search queries, recognition errors must be accounted for. One method relies on converting both the decoded transcripts and queries into phonetic representations and modeling the confusion between phones, typically represented as a confusion matrix. In this work, we derive this matrix from broadcast news development data. In particular, two systems: HMM based automatic speech recognition (ASR) (Chaudhari and Picheny, 2007) and a neural network based acoustic model (Kingsbury, 2009), are used to analyze the data and the results are compared to produce confusion estimates.

Let $X = \{x_t\}$ represent the input feature frames and $\mathcal{S}$ the set of context dependent HMM states. Associated with $\mathcal{S}$ is a many to one map $\mathbf{M}$ from each member $s_j \in \mathcal{S}$ to a phone in the phone set $p_k \in \mathcal{P}$. This map collapses the beginning, middle, and end context dependent states to the central phone identity. The ASR system is used to generate a state based alignment of the development data to the training transcripts. This results in a sequence of state labels (classes) $\{\mathbf{s}_t\}$, $\mathbf{s}_t \in \mathcal{S}$, one for each frame of the input data. Note that the aligned states are collapsed to the phone identity with $\mathbf{M}$, so the frame class labels are given by $\{c_t\}$, $c_t \in \mathcal{P}$.

Corresponding to each frame, we also use the state posteriors derived from the output of a Neural Network acoustic model and the prior probabilities computed on the training set. Define $X_t = \{\ldots, x_t, \ldots\}$ to be the sub-sequence of the input speech frames centered around time index t. The neural network takes X_t as input and produces

$$l_t(s_j) = y(s_j|X_t) - l(s_j), s_j \in \mathcal{S}$$

where y is the neural network output and l is the prior probability, both in the log domain. Again, the state labels are mapped using $\mathcal{M}$, so the above posterior is interpreted as that for the collapsed phone:

$$l_t(s_j) \equiv l_t(\mathcal{M}(s_j)) = l_t(p_j), p_j = \mathbf{M}(s_j).$$

The result of both analyses gives the following set of associations:

$$\begin{array}{c} c_0 \leftrightarrow l_0(p_0), l_0(p_1), l_0(p_2), \ldots \\ c_1 \leftrightarrow l_1(p_0), l_1(p_1), l_1(p_2), \ldots \\ . \\ . \\ c_t \leftrightarrow l_t(p_0), l_t(p_1), l_t(p_2), \ldots \end{array}$$

Each log posterior $l_i(p_j)$ is converted into a count

$$n_{i,j} = ceil[N \times e^{l_i(p_j)}],$$

where N is a large constant, i ranges over the time index, and j ranges over the context dependent states. From the counts, the confusion matrix entries are computed. The total count for each state is

$$n_j(k) = \sum_{i:c_i=p_j} n_{i,k},$$

where k is an index over the states.

$$\begin{bmatrix} n_1(1) & n_1(2) & \dots \\ n_2(1) & n_2(2) & \dots \\ & \cdot & \\ & \cdot & \end{bmatrix}$$

The rows of the above matrix correspond to the reference and the columns to the observations. By normalizing the rows, the entries can be interpreted as ”probability” of an observed phone (indicated by the column) given the true phone.

5 Experiments and Results

The performance of a spoken term detection system is measured using DET curves that plot the trade-off between false alarms (FAs) and misses. This NIST STD 2006 evaluation metric used Actual/Maximum Term Weighted Value (ATWV/MTWV) that allows one to weight FAs and Misses per the needs of the task at hand (NIST, 2006).

Figure 2 illustrates the effect of speed on ATWV on the NIST STD 2006 Dev06 data set using 1107 query terms. As the speed of indexing is increased to many times faster than real time, the WER increases, which in turn decreases the ATWV measure. It can be seen that the use of word-fragments improves the performance on OOV queries thus making the combined search better than simple word search. The primary advantage of using a hybrid decoding scheme over a separate word and fragment based decoding scheme is the speed of transforming the audio into indexable units. The blue line in the figure illustrates that when using a hybrid setup, the same performance can be achieved at speeds twice as fast. For example, with the combined search on two different decodes, an ATWV of 0.1 can be achieved when indexing at a speed 15 times faster than real time, but with a hybrid system, the same performance can be reached at an indexing speed 30 times faster than real time. The ATWV on the hybrid system also degrades gracefully with faster speeds when compared to separate word and word-fragment systems. Preliminary results indicate that fuzzy search on one best output gives the same ATWV performance as exact search (Figure 2) on consensus output. Also, a closer look at the retrieval results of OOV terms revealed that many more OOVs are retrieved with the fuzzy search.

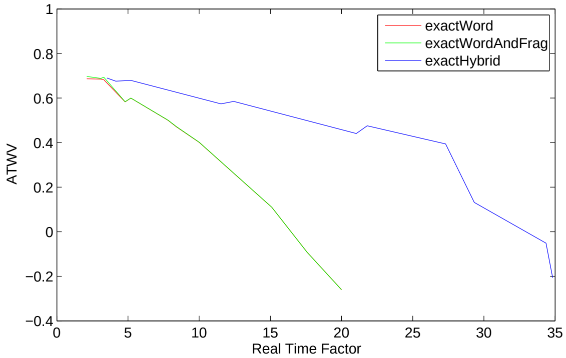

Figure 2: Effect of WER on ATWV. Note that the cuves for exactWord and exactWordAndFrag lie on top of each other.

6 CONCLUSION

In this paper, we have presented the effect of rapid decoding on a spoken term detection task. We have demonstrated that hybrid systems perform well and fuzzy search with phone confusion probabilities help in OOV retrieval.

References

U. V. Chaudhari and M. Picheny. 2007. Improvements in phone based audio search via constrained match with high order confusion estimates. In *Proc. of ASRU*.

M. Clements, S. Robertson, and M. S. Miller. 2002. Phonetic searching applied to on-line distance learning modules. In *Proc. of IEEE Digital Signal Processing Workshop*.

B. Kingsbury. 2009. Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In *Proc. of ICASSP*.

J. Mamou, B. Ramabhadran, and O. Siohan. 2007. Vocabulary independent spoken term detection. In *Proc. of ACM SIGIR*.

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

M. Saraclar and R. Sproat. 2004. Lattice-based search for spoken utterance retrieval. In *Proc. HLT-NAACL*.

F. Seide, P. Yu, C. Ma, and E. Chang. 2004. Vocabulary-independent search in spontaneous speech. In *Proc. of ICASSP*.

O. Siohan and M. Bacchiani. 2005. Fast vocabulary independent audio search using path based graph indexing. In *Proc. of Interspeech*.