Lattice-Based Search for Spoken Utterance Retrieval

Murat Saraclar
AT&T Labs – Research
180 Park Ave. Florham Park, NJ 07932
murat@research.att.com

Richard Sproat
University of Illinois at Urbana-Champaign
Urbana, IL 61801
rws@uiuc.edu

Abstract

Recent work on spoken document retrieval has suggested that it is adequate to take the single-best output of ASR, and perform text retrieval on this output. This is reasonable enough for the task of retrieving broadcast news stories, where word error rates are relatively low, and the stories are long enough to contain much redundancy. But it is patently not reasonable if one's task is to retrieve a short snippet of speech in a domain where WER's can be as high as 50%; such would be the situation with teleconference speech, where one's task is to find if and when a participant uttered a certain phrase.

In this paper we propose an indexing procedure for spoken utterance retrieval that works on lattices rather than just single-best text. We demonstrate that this procedure can improve F scores by over five points compared to single-best retrieval on tasks with poor WER and low redundancy. The representation is flexible so that we can represent both word lattices, as well as phone lattices, the latter being important for improving performance when searching for phrases containing OOV words.

1 Introduction

Automatic systems for indexing, archiving, searching and browsing of large amounts of spoken communications have become a reality in the last decade. Most such systems use an automatic speech recognition (ASR) component to convert speech to text which is then used as an input to a standard text based information retrieval (IR) component. This strategy works reasonably well when speech recognition output is mostly correct or the documents are long enough so that some occurrences of the query terms are recognized correctly.

Most of the research has concentrated on retrieval of Broadcast News type of spoken documents where speech is relatively clean and the documents are relatively long. In addition it is possible to find large amounts of text with similar content in order to build better language models and enhance retrieval through use of similar documents.

We are interested in extending this to telephone conversations and teleconferences. Our task is locating occurrences of a query in spoken communications to aid browsing. This is not exactly spoken document retrieval. In fact, it is more similar to word spotting. Each document is a short segment of audio.

Although reasonable retrieval performance can be obtained using the best ASR hypothesis for tasks with moderate ($\sim 20\%$) word error rates, tasks with higher ($40 - 50\%$) word error rates require use of multiple ASR hypotheses. Use of ASR lattices makes the system more robust to recognition errors.

Almost all ASR systems have a closed vocabulary. This restriction comes from run-time requirements as well as the finite amount of data used for training the language models of the ASR systems. Typically the recognition vocabulary is taken to be the words appearing in the language model training corpus. Sometimes the vocabulary is further reduced to only include the most frequent words in the corpus. The words that are not in this closed vocabulary – the out of vocabulary (OOV) words – will not be recognized by the ASR system, contributing to recognition errors. The effects of OOV words in spoken document retrieval are discussed by Woodland et al. (2000). Using phonetic search helps retrieve OOV words.

This paper is organized as follows. In Section 2 we give an overview of related work, focusing on methods dealing with speech recognition errors and OOV queries. We present the methods used in this study in Section 3.

Experimental setup and results are given in Section 4. Finally, our conclusions are presented in Section 5.

2 Related Work

There are commercial systems including Nexidia/FastTalk (www.nexidia.com), Virage/AudioLogger (www.virage.com), Convera (www.convera.com) as well as research systems like AT&T DVL (Cox et al., 1998), AT&T ScanMail (Hirschberg et al., 2001), BBN Rough'n'Ready (Makhoul et al., 2000), CMU Informedia (www.informedia.cs.cmu.edu), SpeechBot (www.speechbot.com), among others.

Also between 1997 and 2000, the Test REtrieval Conference (TREC) had a spoken document retrieval (SDR) track with many participants (Garofolo et al., 2000). NIST TREC-9 SDR Web Site (2000) states that:

> The results of the TREC-9 2000 SDR evaluation presented at TREC on November 14, 2000 showed that retrieval performance for sites on their own recognizer transcripts was virtually the same as their performance on the human reference transcripts. Therefore, retrieval of excerpts from broadcast news using automatic speech recognition for transcription was deemed to be a solved problem - even with word error rates of 30%.

PhD Theses written on this topic include James (1995), Wechsler (1998), Siegler (1999) and Ng (2000).

Jones et al. (1996) describe a system that combines a large vocabulary continuous speech recognition (LVCSR) system and a phone-lattice word spotter (WS) for retrieval of voice and video mail messages (Brown et al., 1996). Witbrock and Hauptmann (1997) present a system where a phonetic transcript is obtained from the word transcript and retrieval is performed using both word and phone indices. Wechsler et al. (1998) present new techniques including a new method to detect occurrences of query features, a new method to estimate occurrence probabilities, a collection-wide probability re-estimation technique and feature length weighting. Srinivasan and Petkovic (2000) introduce a method for phonetic retrieval based on the probabilistic formulation of term weighting using phone confusion data. Amir et al. (2001) use indexing based on confusable phone groups and a Bayesian phonetic edit distance for phonetic speech retrieval. Logan et al. (2002) compare three indexing methods based on words, syllable-like particles, and phonemes to study the problem of OOV queries in audio indexing systems. Logan and Van Thong (2002) give an alternate approach to the OOV query problem by expanding query words into in-vocabulary phrases while taking acoustic confusability and language model scores into account.

Of the previous work, the most similar approach to the one proposed here is that of Jones et al. (1996), in that they used phone lattices to aid in word spotting, in addition to single-best output from LVCSR. Our proposal might be thought of as a generalization of their approach in that we use lattices as the sole representation over which retrieval is performed. We believe that lattices are a more natural representation for retrieval in cases where there is a high degree of uncertainty about what was said, which is typically the case in LVCSR systems for conversational speech. We feel that our results, presented below, bear out this belief. Also novel in our approach is the use of *indexed lattices* allowing for efficient retrieval. As we note below, in the limit where one is using one-best output, the indexed lattices reduce to the normal inverted index used in text retrieval.

3 Methods

In this section we describe the overall structure of our system and give details of the techniques used in our investigations. The system consists of three main components. First, the ASR component is used to convert speech into a lattice representation, together with timing information. Second, this representation is indexed for efficient retrieval. These two steps are performed off-line. Finally, when the user enters a query the index is searched and matching audio segments are returned.

3.1 Automatic Speech Recognition

We use a state-of-the-art HMM based large vocabulary continuous speech recognition (LVCSR) system. The acoustic models consist of decision tree state clustered triphones and the output distributions are mixtures of Gaussians. The language models are pruned backoff trigram models. The pronunciation dictionaries contain few alternative pronunciations. Pronunciations that are not in our baseline pronunciation dictionary (including OOV query words) are generated using a text-to-speech (TTS) frontend. The TTS frontend can produce multiple pronunciations. The ASR systems used in this study are single pass systems. The recognition networks are represented as weighted finite state machines (FSMs).

The output of the ASR system is also represented as an FSM and may be in the form of a best hypothesis string or a lattice of alternate hypotheses. The labels on the arcs of the FSM may be words or phones, and the conversion between the two can easily be done using FSM composition. The costs on the arcs are negative log likelihoods. Additionally, timing information can also be present in the output.

3.2 Lattice Indexing and Retrieval

In the case of lattices, we store a set of indices, one for each arc label (word or phone) l, that records the lat-

tice number $L[a]$, input-state $k[a]$ of each arc a labeled with l in each lattice, along with the probability mass $f(k[a])$ leading to that state, the probability of the arc itself $p(a|k[a])$ and an index for the next state. To retrieve a single label from a set of lattices representing a speech corpus one simply retrieves all arcs in each lattice from the label index. The lattices are first normalized by weight pushing (Mohri et al., 2002) so that the probability of the set of all paths leading *from* the arc to the final state is 1. After weight pushing, for a given arc a, the probability of the set of all paths containing that arc is given by

$$p(a) = \sum_{\pi \in L: a \in \pi} p(\pi) = f(k[a])p(a|k[a])$$

namely the probability of all paths leading into that arc, multiplied by the probability of the arc itself. For a lattice L we construct a "count" $C(l|L)$ for a given label l using the information stored in the index $\mathcal{I}(l)$ as follows,

$$\begin{aligned} C(l|L) &= \sum_{\pi \in L} p(\pi) C(l|\pi) \\ &= \sum_{\pi \in L} \Big(p(\pi) \sum_{a \in \pi} \delta(a, l) \Big) \\ &= \sum_{a \in L} \Big(\delta(a, l) \sum_{\pi \in L: a \in \pi} p(\pi) \Big) \\ &= \sum_{a \in \mathcal{I}(l): L[a] = L} p(a) \\ &= \sum_{a \in \mathcal{I}(l): L[a] = L} f(k[a]) p(a|k[a]) \end{aligned}$$

where $C(l|\pi)$ is the number of times l is seen on path π and $\delta(a, l)$ is 1 if arc a has the label l and 0 otherwise. Retrieval can be thresholded so that matches below a certain count are not returned.

To search a multilabel expression (e.g. a multi-word phrase) $w_1 w_2 \dots w_n$ we seek on each label in the expression, and then for each (w_i, w_{i+1}) join the output states of w_i with the matching input states of w_{i+1}; in this way we retrieve just those path segments in each lattice that match the entire multi-label expression. The probability of each match is defined as $f(k[a_1])p(a_1|k[a_1])p(a_2|k[a_2]) \dots p(a_n|k[a_n])$, where $p(a_i|k[a_i])$ is the probability of the ith arc in the expression starting in arc a_1. The total "count" for the lattice is computed as defined above.

Note that in the limit case where each lattice is an unweighted single path — i.e. a string of labels — the above scheme reduces to a standard inverted index.

The count $C(l|L)$ can be interpreted as a lattice-based confidence measure. Although it may be possible to use more sophisticated confidence measures, use of (posterior) probabilities allows for a simple factorization which makes indexing efficient.

3.3 Indexing Using Sub-word Units

In order to deal with queries that contain OOV words we investigate the use of sub-word units for indexing. In this study we use phones as the sub-word units. There are two methods for obtaining phonetic representation of an input utterance.

1. Phone recognition using an ASR system where recognition units are phones. This is achieved by using a phone level language model instead of the word level language model used in the baseline ASR system.

2. Converting the word level representation of the utterance into a phone level representation. This is achieved by using the baseline ASR system and replacing each word in the output by its pronunciation(s) in terms of phones.

Both methods have their shortcomings. Phone recognition is known to be less accurate than word recognition. On the other hand, the second method can only generate phone strings that are substrings of the pronunciations of in-vocabulary word strings. An alternative is to use hybrid language models used for OOV word detection (Yazgan and Saraclar, 2004).

For retrieval, each query word is converted into phone string(s) by using its pronunciation(s). The phone index can then be searched for each phone string. Note that this approach will generate many false alarms, particularly for short query words, which are likely to be substrings of longer words. In order to control for this a bound on minimum pronunciation length can be utilized. Since most short words are in vocabulary this bound has little effect on recall.

3.4 Using Both Word and Sub-word Indices

Given a word index and a sub-word index, it is possible to improve the retrieval performance of the system by using both indices. There are many strategies for doing this.

1. *combination:*
   Search both the word index and the sub-word index, combine the results.

2. *vocabulary cascade:*
   Search the word index for in-vocabulary queries, search the sub-word index for OOV queries.

3. *search cascade:*
   Search the word index,
   if no result is returned search the sub-word index.

In the first case, if the indices are obtained from ASR best hypotheses, then the result combination is a simple union of the separate sets of results. However, if indices

are obtained from lattices, then in addition to taking a union of results, retrieval can be done using a combined score. Given a query q, let $C_w(q)$ and $C_p(q)$ be the lattice counts obtained from the word index and the phone index respectively. We also define the normalized lattice count for the phone index as

$$C_p^{\text{norm}}(q) = (C_p(q))^{\frac{1}{|\text{pron}(q)|}}$$

where $|\text{pron}(q)|$ is the length of the pronunciation of query q. We then define the combined score to be

$$C_{wp}(q) = C_w(q) + \lambda C_p^{\text{norm}}(q)$$

where λ is an empirically determined scaling factor.

In the other cases, instead of using two different thresholds we use a single threshold on $C_w(q)$ and $C_p^{\text{norm}}(q)$ during retrieval.

4 Experiments

4.1 Evaluation Metrics

For evaluating ASR performance we use the standard word error rate (WER) as our metric. Since we are interested in retrieval we use OOV rate by type to measure the OOV word characteristics. For evaluating retrieval performance we use precision and recall with respect to manual transcriptions. Let $\text{Correct}(q)$ be the number of times the query q is found correctly, $\text{Answer}(q)$ be the number of answers to the query q, and $\text{Reference}(q)$ be the number of times q is found in the reference.

$$\text{Precision}(q) = \frac{\text{Correct}(q)}{\text{Answer}(q)}$$

$$\text{Recall}(q) = \frac{\text{Correct}(q)}{\text{Reference}(q)}$$

We compute precision and recall rates for each query and report the average over all queries. The set of queries Q consists of all the words seen in the reference except for a stoplist of 100 most common words. The measurement is not weighted by frequency – i.e. each query $q \in Q$ is presented to the system only once, independent of the number of occurences of q in the transcriptions.

$$\text{Precision} = \frac{1}{|Q|} \sum_{q \in Q} \text{Precision}(q)$$

$$\text{Recall} = \frac{1}{|Q|} \sum_{q \in Q} \text{Recall}(q)$$

For lattice based retrieval methods, different operating points can be obtained by changing the threshold. The precision and recall at these operating points can be plotted as a curve.

In addition to individual precision-recall values we also compute the F-measure defined as

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

and report the maximum F-measure (maxF) to summarize the information in a precision-recall curve.

4.2 Corpora

We use three different corpora to assess the effectiveness of different retrieval techniques.

The first corpus is the DARPA Broadcast News corpus consisting of excerpts from TV or radio programs including various acoustic conditions. The test set is the 1998 Hub-4 Broadcast News (hub4e98) evaluation test set (available from LDC, Catalog no. LDC2000S86) which is 3 hours long and was manually segmented into 940 segments. It contains 32411 word tokens and 4885 word types. For ASR we use a real-time system (Saraclar et al., 2002). Since the system was designed for SDR, the recognition vocabulary of the system has over 200K words. The pronunciation dictionary has 1.25 pronunciations per word.

The second corpus is the Switchboard corpus consisting of two party telephone conversations. The test set is the RT02 evaluation test set which is 5 hours long, has 120 conversation sides and was manually segmented into 6266 segments. It contains 65255 word tokens and 3788 word types. For ASR we use the first pass of the evaluation system (Ljolje et al., 2002). The recognition vocabulary of the system has over 45K words. For these words the average number of pronunciations per word is 1.07.

The third corpus is named *Teleconferences* since it consists of multiparty teleconferences on various topics. The audio from the legs of the conference are summed and recorded as a single channel. A test set of six teleconferences (about 3.5 hours) was transcribed. It contains 31106 word tokens and 2779 word types. Calls are automatically segmented into a total of 1157 segments prior to ASR, using an algorithm that detects changes in the acoustics. We again use the first pass of the Switchboard evaluation system for ASR.

In Table 1 we present the ASR performance on these three tasks as well as the OOV Rate by type of the corpora. It is important to note that the recognition vocabulary for the Switchboard and Teleconferences tasks are the same and no data from the Teleconferences task was used while building the ASR systems. The mismatch between the Teleconference data and the models trained on the Switchboard corpus contributes to the significant increase in WER.

4.3 Using ASR Best Word Hypotheses

As a baseline, we use the best word hypotheses of the ASR system for indexing and retrieval. The performance

Task	WER	OOV Rate by Type
Broadcast News	∼20%	0.6%
Switchboard	∼40%	6%
Teleconferences	∼50%	12%

Table 1: Word Error Rate (WER) and OOV Rate (by type) of various LVCSR tasks

of this baseline system is given in Table 2. As expected, we obtain very good performance on the Broadcast News corpus. It is interesting to note that when moving from Switchboard to Teleconferences the degradation in precision-recall is the same as the degradation in WER.

Task	WER	Precision	Recall
Broadcast News	∼20%	92%	77%
Switchboard	∼40%	74%	47%
Teleconferences	∼50%	65%	37%

Table 2: Precision Recall for ASR 1-best

4.4 Using ASR Word Lattices

In the second set of experiments we investigate the use of ASR word lattices. In order to reduce storage requirements, lattices can be pruned to contain only the paths whose costs (i.e. negative log likelihood) are within a threshold with respect to the best path. The smaller this cost threshold is, the smaller the lattices and the index files are. In Figure 1 we present the precision-recall curves for different pruning thresholds on the Teleconferences task.

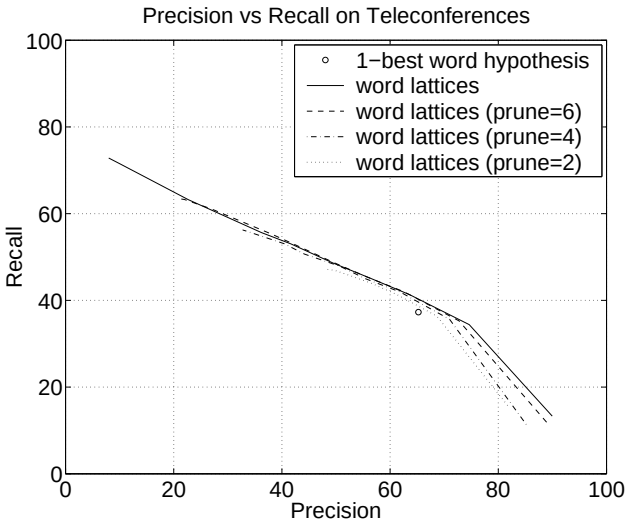

Figure 1: Precision Recall using word lattices for teleconferences

In Table 3 the resulting index sizes and maximum F-measure values are given. On the teleconferences task we observed that cost=6 yields good results, and used this value for the rest of the experiments. Note that this increases the index size with respect to the ASR 1-best case by 3 times for Broadcast News, by 5 times for Switchboard and by 9 times for Teleconferences.

Task	Pruning	Size (MB)	maxF
Broadcast News	nbest=1	29	84.0
Broadcast News	cost=6	91	84.8
Switchboard	nbest=1	18	57.1
Switchboard	cost=6	90	58.4
Teleconferences	nbest=1	16	47.4
Teleconferences	cost=2	29	49.5
Teleconferences	cost=4	62	50.0
Teleconferences	cost=6	142	50.3
Teleconferences	cost=12	3100	50.1

Table 3: Comparison of index sizes

4.5 Using ASR Phone Lattices

Next, we compare using the two methods of phonetic transcription discussed in Section 3.3 – phone recognition and word-to-phone conversion – for retrieval using only phone lattices. In Table 4 the precision and recall values that yield the maximum F-measure as well as the maximum F-measure values are presented. These results clearly indicate that phone recognition is inferior for our purposes.

Source for Indexing	Precision	Recall	maxF
Phone Recognition	25.6	37.3	30.4
Conversion from Words	43.1	48.5	45.6

Table 4: Comparison of different sources for the phone index on the Teleconferences corpus

4.6 Using ASR Word and Phone Lattices

We investigated using the strategies mentioned in Section 3.4, and found strategy 3 – search the word index, if no result is returned search the phone index – to be superior to others. We give a comparison of the maximum F-values for the three strategies in Table 5.

Strategy	maxF
1.*combination*	50.5
2.*vocabulary cascade*	51.0
3.*search cascade*	52.8

Table 5: Comparison of different strategies for using word and phone indices

In Figure 2 we present results for this strategy on the Teleconferences corpus. The phone indices used in these experiments were obtained by converting the word lattices into phone lattices. Using the phone indices obtained by phone recognition gave significantly worse results.

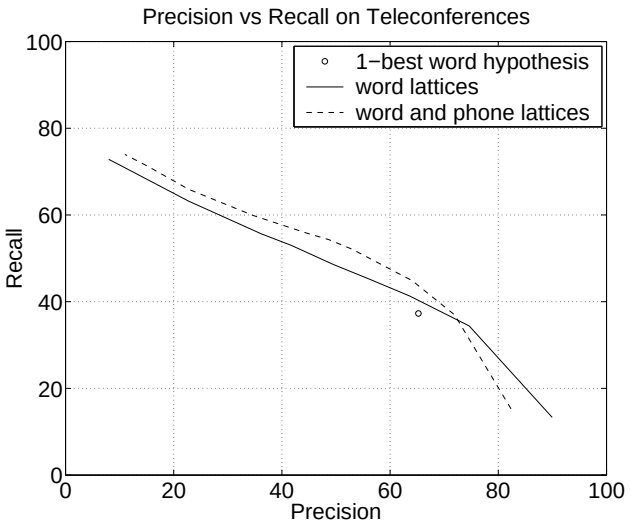

Figure 2: Comparison of word lattices and word/phone hybrid strategies for teleconferences

4.7 Effect of Minimum Pronunciation Length for Queries

When searching for words with short pronunciations in the phone index the system will produce many false alarms. One way of reducing the number of false alarms is to disallow queries with short pronunciations. In Figure 3 we show the effect of imposing a minimum pronunciation length for queries. For a query to be answered its pronunciation has to have more than *minphone* phones, otherwise no answers are returned. Best maximum F-measure result is obtained using minphone=3.

4.8 Effects of Recognition Vocabulary Size

In Figure 4 we present results for different recognition vocabulary sizes (5k, 20k, 45k) on the Switchboard corpus. The OOV rates by type are 32%, 10% and 6% respectively. The word error rates are 41.5%, 40.1% and 40.1% respectively. The precision recall curves are almost the same for 20k and 45k vocabulary sizes.

4.9 Using Word Pair Queries

So far, in all the experiments the query list consisted of single words. In order to observe the behavior of various methods when faced with longer queries we used a set of

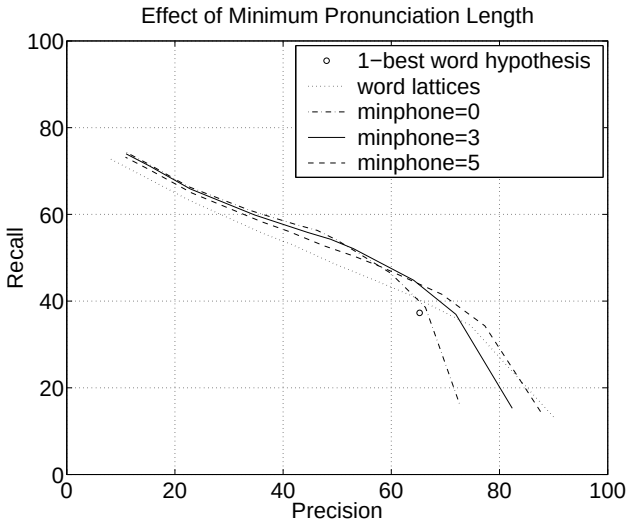

Figure 3: Effect of minimum pronunciation length using a word/phone hybrid strategy for teleconferences

word pair queries. Instead of using all the word pairs seen in the reference transcriptions, we chose the ones which were more likely to occur together than with other words. For this, we sorted the word pairs (w_1, w_2) according to their *pointwise mutual information*

$$\log \frac{p(w_1, w_2)}{p(w_1)p(w_2)}$$

and used the top pairs as queries in our experiments. Note that in these experiments only the query set is changed and the indices remain the same as before.

As it turns out, the precision of the system is very high on this type of queries. For this reason, it is more interesting to look at the operating point that achieves the maximum F-measure for each technique, which in this case coincides with the point that yields the highest recall. In Table 6 we present results on the Switchboard corpus using 1004 word pair queries. Using word lattices it is possible to increase the recall of the system by 16.4% while degrading the precision by only 2.2%. Using phone lattices we can get another 3.7% increase in recall for 1.2% loss in precision. The final system still has 95% precision.

System	Precision	Recall	maxF
Word 1-best	98.3	29.7	45.6
Word lattices	96.1	46.1	62.3
Word+Phone lattices	94.9	49.8	65.4

Table 6: Results for word pair queries on Switchboard

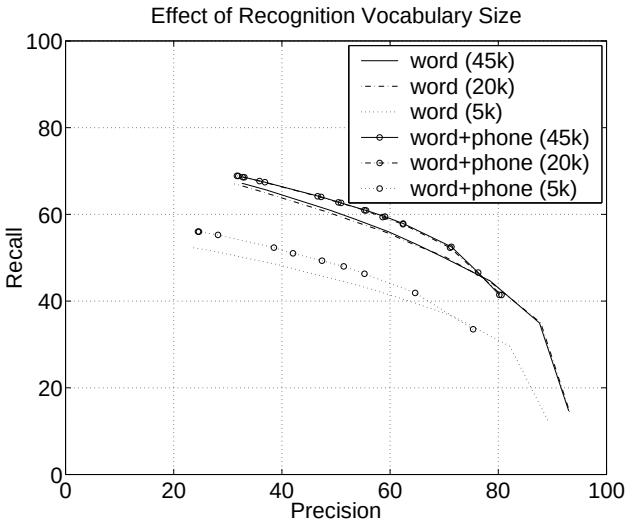

Figure 4: Comparison of various recognition vocabulary sizes for Switchboard

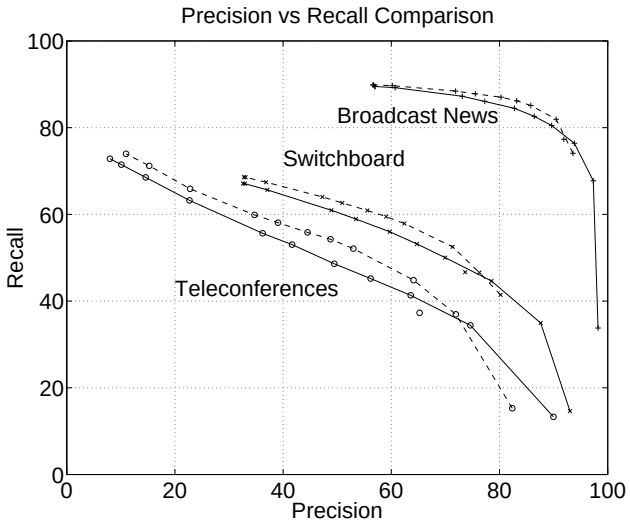

Figure 5: Precision Recall for various techniques on different tasks. The tasks are Broadcast News (+), Switchboard (x), and Teleconferences (o). The techniques are using best word hypotheses (single points), using word lattices (solid lines), and using word and phone lattices (dashed lines).

4.10 Summary of Results on Different Corpora

Finally, we make a comparison of various techniques on different tasks. In Table 7 maximum F-measure (maxF) is given. Using word lattices yields a relative gain of 3-5% in maxF over using best word hypotheses. For the final system that uses both word and phone lattices, the relative gain over the baseline increases to 8-12%.

Task	System		
	1-best	W Lats	W+P Lats
Broadcast News	84.0	84.8	86.0
Switchboard	57.1	58.4	60.5
Teleconferences	47.4	50.3	52.8

Table 7: Maximum F-measure for various systems and tasks

In Figure 5 we present the precision recall curves. The gain from using better techniques utilizing word and phone lattices increases as retrieval performance gets worse.

5 Conclusion

We proposed an indexing procedure for spoken utterance retrieval that works on ASR lattices rather than just single-best text. We demonstrated that this procedure can improve maximum F-measure by over five points compared to single-best retrieval on tasks with poor WER and low redundancy. The representation is flexible so that we can represent both word lattices, as well as phone lattices, the latter being important for improving performance when searching for phrases containing OOV words. It is important to note that spoken utterance retrieval for conversational speech has different properties than spoken document retrieval for broadcast news. Although consistent improvements were observed on a variety of tasks including Broadcast News, the procedure proposed here is most beneficial for more difficult conversational speech tasks like Switchboard and Teleconferences.

References

A. Amir, A. Efrat, and S. Srinivasan. 2001. Advances in phonetic word spotting. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pages 580–582, Atlanta, Georgia, USA.

M. G. Brown, J. T. Foote, G. J. F. Jones, K. Sparck Jones, and S. J. Young. 1996. Open-vocabulary speech indexing for voice and video mail retrieval. In *Proc. ACM Multimedia 96*, pages 307–316, Boston, November.

R. V. Cox, B. Haskell, Y. LeCun, B. Shahraray, and L. Rabiner. 1998. On the application of multimedia processing to telecommunications. *Proceedings of the IEEE*, 86(5):755–824, May.

J. Garofolo, G. Auzanne, and E. Voorhees. 2000. The TREC spoken document retrieval track: A success

story. In *Proceedings of the Recherche d'Informations Assiste par Ordinateur: Content Based Multimedia Information Access Conference*.

J. Hirschberg, M. Bacchiani, D. Hindle, P. Isenhour, A. Rosenberg, L. Stark, L. Stead, S. Whittaker, and G. Zamchick. 2001. Scanmail: Browsing and searching speech data by content. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark.

David Anthony James. 1995. *The Application of Classical Information Retrieval Techniques to Spoken Documents*. Ph.D. thesis, University of Cambridge, Downing College.

G. J. F. Jones, J. T. Foote, K. Sparck Jones, and S. J. Young. 1996. Retrieving spoken documents by combining multiple index sources. In *Proc. SIGIR 96*, pages 30–38, Zürich, August.

A. Ljolje, M. Saraclar, M. Bacchiani, M. Collins, and B. Roark. 2002. The AT&T RT-02 STT system. In *Proc. RT02 Workshop*, Vienna, Virginia.

B. Logan and JM Van Thong. 2002. Confusion-based query expansion for OOV words in spoken document retrieval. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, USA.

B. Logan, P. Moreno, and O. Deshmukh. 2002. Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio. In *Proc. HLT*.

J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava. 2000. Speech and language technologies for audio indexing and retrieval. *Proceedings of the IEEE*, 88(8):1338–1353, August.

M. Mohri, F. Pereira, and M. Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16(1):69–88.

Kenney Ng. 2000. *Subword-Based Approaches for Spoken Document Retrieval*. Ph.D. thesis, Massachusetts Institute of Technology.

NIST TREC-9 SDR Web Site. 2000. www.nist.gov/speech/tests/sdr/sdr2000/sdr2000.htm.

M. Saraclar, M. Riley, E. Bocchieri, and V. Goffin. 2002. Towards automatic closed captioning: Low latency real time broadcast news transcription. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado, USA.

Matthew A. Siegler. 1999. *Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance*. Ph.D. thesis, Carnegie Mellon University.

S. Srinivasan and D. Petkovic. 2000. Phonetic confusion matrix based spoken document retrieval. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–87.

M. Wechsler, E. Munteanu, and P. Scäuble. 1998. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 20–27, Melbourne, Australia.

Martin Wechsler. 1998. *Spoken Document Retrieval Based on Phoneme Recognition*. Ph.D. thesis, Swiss Federal Institute of Technology (ETH), Zurich.

M. Witbrock and A. Hauptmann. 1997. Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents. In *2nd ACM International Conference on Digital Libraries (DL'97)*, pages 30–35, Philadelphia, PA, July.

P.C. Woodland, S.E. Johnson, P. Jourlin, and K.Sparck Jones. 2000. Effects of out of vocabulary words in spoken document retrieval. In *Proc. SIGIR*, pages 372–374, Athens, Greece.

A. Yazgan and M. Saraclar. 2004. Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Canada.