A Critical Assessment of Spoken Utterance Retrieval through Approximate Lattice Representations

Siavash Kazemian Frank Rudzicz Gerald Penn
Cosmin Munteanu
Department of Computer Science, University of Toronto
Toronto, Ontario Canada
[kazemian, frank, gpenn, mcosmin]@cs.toronto.edu

ABSTRACT

This paper compares the performance of position-specific posterior lattices (PSPL) and confusion networks applied to spoken utterance retrieval, and tests these recent proposals against several baselines in two disparate domains. These lossy methods provide compact representations that generalize the original segment lattices and provide greater recall and robustness, but have yet to be evaluated against each other in multiple WER conditions for spoken utterance retrieval. Our comparisons suggest that while PSPL and confusion networks have comparable recall, the former is slightly more precise, although its merit appears to be coupled to the assumptions of low-frequency search queries and low-WER environments.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods—*Speech indexing and retrieval*; H.5.5 [**Sound and Music Computing**]: Signal analysis, synthesis, and processing

General Terms

Speech Retrieval, Spoken Utterance Retrieval, Lossy Recognition Lattice Representation

1. INTRODUCTION

With an increase in the general availability of resources such as streaming audio and inexpensive data storage, there is an apparent shift towards multimedia data in information retrieval. To make speech-heavy multimedia data amenable to text-based search an important first step is to perform automatic speech recognition (ASR) on those data, and to hold the results in a structure appropriate for multimedia retrieval.

Spoken document retrieval (SDR) is the task of identifying a subset of spoken documents that are relevant to given keywords or phrases from a larger corpus, which is normally accomplished exclusively through automatic transcription of the data by ASR. This field has received increased attention in recent years as larger speech corpora become more readily available. The TREC SDR track [4] focused on this task in the Broadcast news domain, concluding that recognition errors in automatic transcripts do not effect the retrieval performance significantly. However, broadcast news is usually produced by trained speakers within excellent acoustic environments, allowing for especially accurate ASR transcription. There is also a lot of repetition in the content terms in Broadcast news that make retrieval robust to recognition error [1]. These conditions are not typical of more spontaneous domains such as academic lectures or phone conversations, so these domains tend to have a much higher word error rate (WER). Retrieving speech in these more complex domains also tends to be less robust to recognition error as there is far less repetition and structure [1]. Improving retrieval in these areas is an ongoing research problem.

To deal with higher WER domains, it is customary to consider the entire recognition lattice since queried terms in the source speech are more likely to appear in a wider range of hypotheses than just the top scoring path. Saraclar and Sproat showed that using lattices in domains with high WER (e.g., teleconference data at $\sim$50%) improves spoken utterance retrieval (defined below) relative to domains with low WER (e.g., broadcast news at $\sim$40%) [11]. While searching the entire lattice will typically increase the recall of retrieval, this comes at the cost of reduced precision as queried terms mistakenly appear in alternate paths for unrelated documents, particularly in difficult domains.

Due to the redundancy of information present in a recognition lattice, directly using it to index a spoken utterance can be an inefficient use of computational resources. Methods of reducing the space requirements include pruning the low probability arcs of the lattice [11], and approximating the structure of the lattice to explicitly remove redundancy. These lossy representations of the recognition lattice often imply more paths than the original lattice, and hence may improve recall, but at the further expense of precision, as has been shown in domains with high WER [5, 15]. Overall, these lossy representations have been found to be preferable to the original lattices, at least in the phrase spotting task [15].

A closely related task to SDR is Spoken Utterance Retrieval

(SUR) [11], which is the subject of this paper. In addition to being an important task in its own right, SUR is also an important part of any spoken document retrieval system. To calculate the relevance of a spoken document to a certain query, one needs to compute the relevance of each of the component utterances to that query. As an independent task, SUR identifies more specific sections of a spoken document relevant to a user's query. Given utterances labelled in this fasion, users may browse documents more easily to find the parts that are relevant to their queries.

In this paper, we compare the performance of two popular lossy methods, namely position-specific posterior lattices (PSPL) [2] and word confusion networks [6, 5] on the SUR task. These methods are also compared against a simple and significantly more compact set-of-words baseline model, and the baselines of 1-best transcripts and full lattices on domains with disparate WER levels, namely recorded lectures and broadcast news. These methods are evaluated in terms of the frequency and word length of test queries within source documents.

2. RELATED WORK

To address the problem of inaccurate recognition in speech retrieval for high WER domains, researchers have moved towards utilizing the recognition lattice or N-best lists to offset the negative effect of inaccurate recognition in speech retrieval. Siegler [14] and Saraclar and Sproat [11] show that utilizing N-best lists and recognition lattices improves retrieval performance.

More recent studies have introduced lossy methods of representing word lattices [2, 5, 15]. Chelba et al. [2] introduced Position Specific Posterior Lattice (PSPL), which only keeps the position and posterior probability of each arc in the lattice. This representation is more compact than a word lattice and can easily be utilized in an indexing scheme. Chelba et al. [2] showed that using PSPL improved SDR performance by 17-26% in comparison with the widely accepted baseline of 1-best transcripts.

Zhou et al. [15] introduced Time-based Merging for Indexing speech (TMI). They reduced the average occurrence of words from 881.7 in the original lattice to 19.5, and it was between 25 and 30% more accurate than the 1-best approach on phrase spotting. They compared the performance of TMI to PSPL and found that PSPL outperforms TMI methods in document retrieval but TMI performed better in phrase spotting by 0.1%, absolutely.

Furthermore, Hori et al. [5] used Confusion Networks (CN) [6] to represent recognition lattices for retrieval. They showed that their system outperformed 1-best transcripts and, for in-vocabulary queries, that their system was comparable to using the entire word lattice (within 0.4%).

To shed some light on which of PSPL or CN is better suited to speech retrieval, Pan et al. [3] compared the performance of spoken document retrieval on Mandarin Broadcast news data utilizing PSPL and Confusion Networks in the SDR task in terms of retrieval accuracy and index size. They found that for their data, PSPL outperformed Confusion Networks on retrieval performance but required larger disk space to be stored. Our study extends this work by comparing these two methods in the Spoken Utterance task in two domains with disparate WER levels, namely English broadcast news and recorded lecture domains. Furthermore, we compare the performance of PSPL and Confusion Networks to the widely accepted baseline of 1-best transcripts, the raw recognition lattice, and a new set-of-words baseline.

Another important issue in speech retrieval is out of vocabulary (OOV) word recognition. A general approach is to represent the spoken data and the query in sub-word units (i.e. phones or phone n-grams) [8] [13]. But Logan et al. [12] showed that using word level recognition performs better than sub-word recognition in speech retrieval for in-vocabulary queries. The authors argue in favor of mixing the two approaches for a complete retrieval system. This approach is taken by Hori et al. [5] as they combine a phone-based CN with a word-based CN system. In this work, we focus our attention on the spoken utterance retrieval (SUR) task for in-vocabulary queries. Because word-level recognition performs better than sub-word recognition in speech retrieval, the lattices used in this study are word lattices.

3. RECONSTRUCTING THE WORD LATTICE

Three methods that use compact lossy reconstructions of the lattice are described in the following subsections, namely PSPL, word confusion networks, and the set-of-words model. Each of these models is applied to whole-phrase search, the goal of which is to identify documents containing the exact query phrase $Q = q_1 ... q_n$. Differences between these methods are then explored on the task of spoken utterance retrieval in the domains of recorded lectures and broadcast news in §4.

3.1 Position-Specific Posterior Lattice (PSPL)

The **PSPL** method of Chelba et al. [2] is chiefly concerned with the positions of query words within given lattices as defined by path lengths from the start of the lattice. This method computes the probability $P(w, l|\Lambda)$ of encountering word w at a distance l from the start node of Λ, which is a measure that also lends itself to statistically ranking document relevance to a query. This method differs from the standard forward-backward algorithm [10] by partitioning the forward probability mass α_n at node n according to the lengths l of all partial paths to n beginning at the unique start of the lattice, where l is the number of arcs in those paths. That is,

$$\alpha_n[l] \doteq \sum_{\substack{\pi \,:\, end(\pi) = n, \\ length(\pi) = l}} P(\pi) \qquad (1)$$

These probabilities are computed using dynamic programming and the following rules:

$$\alpha_{start}[l] = \begin{cases} 1.0, & \text{if } l = 0 \\ 0.0, & \text{otherwise} \end{cases}$$
$$\alpha_n[l+1] = \sum_{i=1}^{q} \alpha_{s_i} \left[l + \delta\left(l_i, \epsilon\right) \right] \cdot P(e_i), \tag{2}$$

where $P(e_i)$ is the posterior probability of edge $e_i = \langle q_i, n \rangle$, computed as the weighted log sum of acoustic and language model probabilities. Given the standard backwards probability at node n, β_n, the posterior probability of word w occurring at a given position l is

$$P\left(w, l|\Lambda\right) = \sum \frac{\alpha_n[l]\beta_n}{\beta_{start}} \delta\left(w, word(n)\right) \tag{3}$$

In general, when computing the relevance score RS of spoken documents given a query sequence $Q = q_1..q_n$, the PSPL method aggregates the expected count of each subsequence of m query terms in the document according to position, as in

$$s(\Lambda, q_i..q_{i+m-1}) = \log \left[1 + \sum_s \sum_l \prod_{k=0}^{m-1} P\left(q_{i+k}, k + l|\Lambda\right) \right]$$
$$RS(\Lambda, Q) = \sum_{i=1}^{n-m+1} s(\Lambda, q_i...q_{i+m-1}) \tag{4}$$

However, in the absence of page ranking, as is the case here, one would return all documents where some offset k makes $P(q_i, i + k|\Lambda) > 0$ for each q_i in Q. Note that Chelba et al. do not consider OOV words and queries containing them always return the empty set.

3.2 Confusion Networks

Confusion networks are compact finite-state representations of multiple hypotheses through a lattice, and generally have more paths than the original, which theoretically leads to more robust retrieval [6]. The representation consists of a set of equivalence classes $L_i \in \varepsilon$ and a total order $L_j \prec L_k$ where L_i is a set of arcs in the original lattice. Figure 1 illustrates an example.

The lattice alignment algorithm is based on Mangu et al. [6] and consists of three stages. Initially, equivalence classes consist of arcs with identical associated orthographies, and start and end times,

$L_{w,t_1,t_2} = \{e \in E|Word(e) = w, start(e) = t_1, end(e) = t_2\}$,

and the partial order is the transitive closure of the arc order $\leq$ on the lattice. Each of the subsequent stages merges equivalence classes that are not mutually ordered in a best-first manner. When two classes L_1 and L_2 are merged, each is removed from ε, and the new class $L_{new} = L_1 \cup L_2$ is added, as in

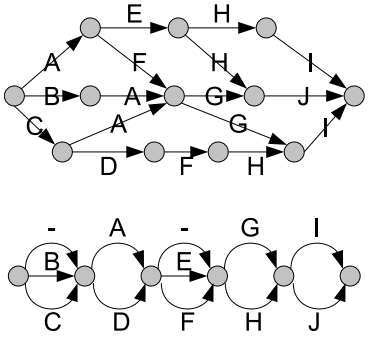

Figure 1: ***Example lattice (top) and associated confusion network (bottom) [5]. Null arcs are represented by '-'.***

$$\varepsilon := \varepsilon \cup \{L_{new}\} \setminus \{L_1, L_2\}.$$

First, *intra-word clustering* iteratively merges classes having arcs with identical orthographies, prioritized by the maximum proportional temporal overlap and posterior probability of those arcs. At the end of this stage, equivalence classes consist of overlapping instances of the same word. Finally, *inter-word clustering* iteratively merges mutually unordered classes prioritized by the following similarity score:

$$SIM(L_1, L_2) = \underset{\substack{w_1 \in Words(L_1) \\ w_2 \in Words(L_2)}}{avg} p_{L_1}(w_1) p_{L_2}(w_2) \cdot \left(1 - \frac{LD(w_1, w_2)}{||w_1|| + ||w_2||}\right),$$

where $p_{L_i}(w) = p(\{e \in L_i : Word(e) = w\})$, $LD(w_1, w_2)$ is the normalized Levenshtein edit distance between the phone string expansions of w_1 and w_2, and $||w_i||$ is the length of the phone string of w_i. Since phone strings will always have a length greater than zero, the normalized edit distance will always be less than 1, giving an overall similarity score above 0. The end of this process is a necessarily totally ordered set of equivalence classes. The confusion network can be searched by the same indexing method as for the lattice [5].

3.3 Set-of-words Baseline

Although reconstruction of the raw lattices to PSPLs is fairly fast, these methods and subsequent searching can be more time consuming on larger documents. Creating Confusion Networks in particular can be very time consuming for bigger lattices. This is not the case with the set-of-words baseline (SOW) which is the most compact representation of the lattice possible that still distinguishes all of the edge labels. This method simply keeps track of every unique word that occurs in the lattice in a list. Searching then reduces to looking for the query terms in this list; this can be performed quickly in $O(m)$ time for each word with a look-up table, where m is the size of the set.

This method ignores the position, order, and multiplicity of query terms, and retrieves all documents having all m query terms within the original lattice. Noting that CN and

PSPL differ mainly in the way they approximate the ordering information that a lattice retains, we thought it would be interesting to compare their performance against a set baseline with no ordering information at all. The performance of SOW should give us some sense of how useful this ordering information is in the SUR task.

4. EXPERIMENTS

The following experiments evaluate the three lossy methods above against the baselines of the full-lattice and 1-best hypothesis methods on spoken utterance retrieval without ranking. These are evaluated with the standard measures of precision, recall, and F-measure and compared against other studies where possible.

4.1 Data

All speech data are sampled at 16kHz, and are represented by 39-dimensional MFCC frame vectors. Acoustic models consist of context-dependent tri-phone units, and are bootstrapped with Wall Street Journal data [9]. These models are trained by sequential Viterbi forced alignment using a single maximum likelihood linear regression transform on all means and variances.

The experiments below are run on two sources of spontaneous speech, namely lecture data and broadcast news. The lecture data consist of approximately 213 minutes of audio recorded over four lectures of an introductory course on human-computer interaction made available by the University of Toronto. These data are divided into 3997 utterance segments using 200ms pause detection, with an average segment duration of 3.0 seconds. The transcripts were produced using a language model trained on WSJ and web data as described by Munteanu et al. [7]. Their ASR system had a 46% WER on these data.

The broadcast news data are a subset of the 1997 English Broadcast News Speech (HUB4) collection and consist of 48 hours of speech from various news organizations (e.g., CNN, ABC, C-SPAN). These data are manually partitioned into 52,949 segments with an average segment duration of 3.2 seconds. This domain was relatively easier to decipher, with a 28% WER. After the recognition stage, the resulting word lattices are reconstructed by the methods above.

For the lecture data, two sets of 24 queries, consisting of single- and multi-word phrases respectively, were produced by individuals familiar with the high-level themes of the corpora. Each set was further partitioned evenly according to the frequency of the key phrases in the corpus. For instance, 8 multi-word queries occurring 2 or 3 times in the corpus formed one subset, those occurring 4 or 5 times formed another, and the final subset consisted of queries occurring 6 or more times. For the broadcast news data, we picked 174 queries, with average length of 2.04 words. Like the queries on the lecture data, these queries were evenly partitioned by frequency of occurrence in the corpus.

All experiments measure the effect of using different lattice representations on spoken utterance retrieval. Here, a segment is correctly retrieved if the query term occurs exactly within that segment. The gold standard is derived from running SUR on manual transcripts of the speech data. Because the chief purpose of this study is to compare the performance of different lattice-based representations of speech, most of the discussion below is concerned with multi-word queries which show more variability. The SUR results for single-word queries are virtually identical across all lattice methods.

4.2 Results

Tables 1 and 2 show the SUR performance on lecture and broadcast news data respectively, partitioned by query word frequency. Interestingly, as we see a fairly uniform decrease in precision across the methods from 1-best to set-of-words, the recall across all methods that use some representation of the lattice is relatively stable. In our lecture data, the 1-best hypothesis is outperformed in all cases, especially for key phrases that occur more than 4 times in the source material. Furthermore, the retrieval performance of most methods improves significantly when queries occur 4 or more times in the corpus.

In broadcast news, all lossy methods have higher recall than the original word lattice, but also suffer from especially low precision, which brings their F-measure below that of the word lattice and 1-best method, the latter of which outperformed the former slightly (i.e., 87.42 F-measure to 87.17). Saraclar and Sproat [11] provide similar results of 84.0 F-measure on the word lattice, and 84.8 on the 1-best hypothesis on 3 hours of news test data.

		Lecture data (46% WER)		
		Performance by query frequency		
System		2-3	4-5	6+
1-best	**P**	100.0	100.0	99.0
	R	41.7	23.1	62.4
	F	58.8	37.6	76.5
Lattice	**P**	93.8	93.8	99.0
	R	45.8	90.0	84.1
	F	61.6	91.8	90.9
PSPL	**P**	90.6	93.8	99.0
	R	45.8	90.0	84.1
	F	60.9	91.8	90.9
ConfNet	**P**	87.5	93.8	97.3
	R	45.8	90.0	84.7
	F	60.2	91.8	90.6
SetOfWords	**P**	87.5	91.3	96.3
	R	45.8	90.0	84.7
	F	60.2	90.6	90.1

Table 1: Precision (P), recall (R) and F-measure (F) of each indexing method on lecture data and multi-word queries.

PSPL outperformed confusion networks in the SUR task in both lecture data and broadcast news, though by a small amount in F-measure (0.3% absolute in lecture data and by 0.7% absolute in broadcast news). Although confusion networks have better recall than PSPL, it comes at a significant cost in precision.

To further differentiate these methods, we looked at the false positives created by them. False positives provide an appreciation of the cost of using lattices methods, especially in the lossy case. In lossy methods, despite extra flexibility in

System		HUB-4 data (28% WER) Performance by query frequency 1	2-3	4-5	6+
1-best	**P**	98.2	99.2	94.8	97.4
	R	75.0	75.8	82.3	84.0
	F	85.1	85.9	88.2	90.2
Lattice	**P**	98.2	95.2	94.3	95.3
	R	75.0	75.8	84.7	84.5
	F	85.1	84.4	89.3	89.6
PSPL	**P**	94.6	90.9	87.5	93.8
	R	78.6	76.9	85.5	87.3
	F	85.9	83.3	86.5	90.4
ConfNet	**P**	98.2	88.2	80.6	89.2
	R	78.6	78.0	87.8	87.7
	F	87.3	82.8	84.1	88.4
SetOfWords	**P**	92.0	86.3	75.5	86.2
	R	78.6	78.0	88.4	88.4
	F	84.7	81.9	81.4	87.3

Table 2: Precision (P), recall (R) and F-measure (F) of each indexing method on HUB-4 data and multi-word queries.

finding matches with queries missed by high-WER lattices, there is the risk of retrieving segments in which the query terms occur out of sequence. We may assume that a user prefers false positives in which the query terms occur out of sequence to those in which the query terms do not occur at all, although this is not reflected in our precision and recall scores.

As shown in Table 3, a larger proportion of PSPL's false positives contain the query terms when compared to confusion networks. This is rather surprising: in addition to providing a more precise retrieval, PSPL's false positives also include more segments that have the query terms out of sequence. As expected, the majority of the false positives in set-of-words indeed contain the query terms. In fact, if we consider out-of-sequence query hits as true positive, then the set-of-words model outperforms all other methods with an F-measure of 88.7.

System	1-best	Lattice	PSPL	ConfNet	BOW
% F.P.	20.2	12.3	51.2	49.0	61.6

Table 3: Proportion of segments that contain query terms (but not in the exact sequence) among false positives in broadcast news data

Figure 2 shows the effect of WER on SUR performance. The 1-best method suffers most from a degradation of recognition quality. All the lattice methods are also negatively affected but to a lesser extent. the set-of-words baseline seems to be affected the least. This method performs rather well given its simplicity. It seems to suggest that the complexity of the other lossy methods does not provide a clear benefit for high-WER lecture data. It also suggests that keeping ordering information becomes less valuable as WER increases.

Figure 3 shows the average F-measure of each method according to the length of the given key phrase within the broadcast news domain. Here, 49 phrases consist of one

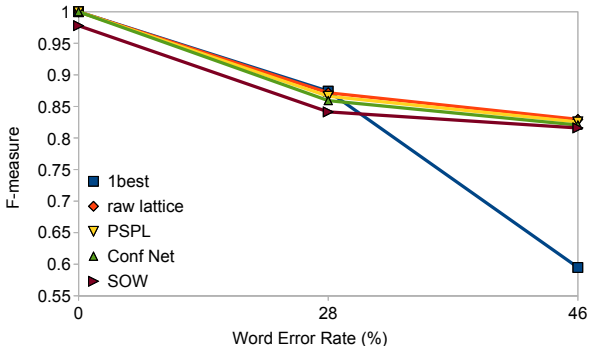

Figure 2: *F-measure vs. Word Error Rate*

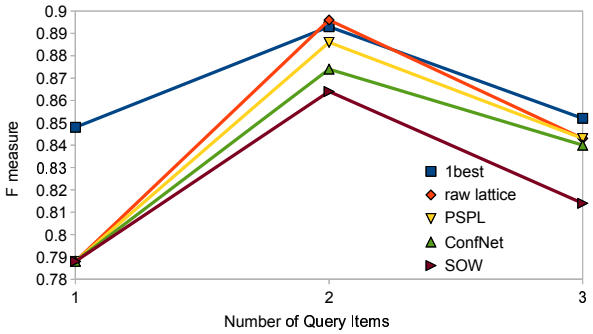

Figure 3: *Average F-measure according to query length for each indexing method.*

word, 68 had two, and 57 had three. The set-of-words method is comparable to the other lattice-based methods for 1-word and 2-word queries, but is particularly worse otherwise. This is almost exclusively due to the low precision of the SOW method, since that method has $\sim$ 82% recall on 3-word queries against $\sim$ 80% for PSPL and the lattice methods. Also, while all lattice-derived and lossy methods averaged F-measures above 80.3 across all multi-word lecture data, the 1-best method achieved only 57.6 by that score. Moreover, that score was down by 12.7 relative to 1-word queries while lattice and lossy methods lost no more than 1.58. This appears to indicate a general insufficiency of using the best path for multi-word queries, which again agrees with Saraclar and Sproat who increased the recall of two-word queries by 16.4% using word lattices over best paths on the high WER Switchboard domain [11].

Finally, we note that the set-of-words model is significantly less expensive computationally, taking only 4.3 minutes to index the resulting lattices of our 48-hour subset of HUB-4 corpus, while the lossy methods took over one hour on the same data. This may be an important factor to consider, given the similarity in recall of this method with the more complex approaches.

5. CONCLUSIONS

In our experiments, PSPL and confusion networks have comparable recall, but with the former having an edge in terms of precision which may partially be due to assumptions of low-frequency search terms and low-WER environments.

The simplistic and compact set-of-words model performs comparably to other lattice-derived methods on spoken utterance retrieval in terms of F-measure and significantly outperforms the 1-best baseline in the high-WER lecture domain, which suggests it may be a more appropriate baseline than other methods.

Future work includes a deeper examination of possible relationships between WER and segment duration and the utility of the various indexing methods. Another interesting future goal of this research is to extend the set-of-words baseline to allow for ranking and use it in the task of spoken document retrieval. Finally, another issue to investigate is the presence of false alarms produced by different lattice representations in the SUR task.

6. ACKNOWLEDGEMENTS

This research is funded by the Natural Sciences and Engineerinc Research Council of Canada and the University of Toronto.

7. REFERENCES

[1] J. Allan. Perspectives on information retrieval and speech. In *Information Retrieval Techniques for Speech Applications [this book is based on the workshop "Information Retrieval Techniques for Speech Applications", held as part of the 24th Annual International ACM SIGIR Conference on Research and Development in Infor*, pages 1–10, London, UK, 2002. Springer-Verlag.

[2] C. Chelba, J. Silva, and A. Acero. Soft indexing of speech content for search in spoken documents. *Computer Speech and Language*, 21:458–478, 2007.

[3] Y. cheng Pan, H. lin Chang, and L. shan Lee. Analytical comparison between position specific posterior lattices and confusion networks based on words and subword units for spoken document indexing. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, Kyoto, Japan, 2007.

[4] J. Garofolo, G. Auzanne, and E. Voorhees. The trec spoken document retrieval track: A success story. In *Proceedings of the Recherche d'Informations Assiste par Ordinateur: ContentBased Multimedia Information Access Conference*, April 2000.

[5] T. Hori, I. L. Hetherington, T. J. Hazen, and J. R. Glass. Open-vocabulary spoken utterance retrieval using confusion networks. In *Proceedings of the 2007 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, 2007.

[6] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer, Speech and Language*, 14(4):373–400, 2000.

[7] C. Munteanu, G. Penn, and R. Baecker. Web-based language modelling for automatic lecture transcription. In *Proceedings of the Tenth European Conference on Speech Communication and Technology - EuroSpeech / Eighth INTERSPEECH*, Antwerp, Belgium, August 2007.

[8] K. Ng. *Subword-based Approaches for Spoken Document Retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.

[9] B. Pellom and K. Hacioglu. Recent improvements in the cu sonic asr system for noisy speech: The spine task. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, April 2003.

[10] L. R. Rabiner. A tutorial on hidden markov models and selected applications inspeech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, February 1989.

[11] M. Saraclar and R. Sproat. Lattice-based search for spoken utterance retrieval. In *Proceedings of the Human Language Technologies and North American Association for Computational Linguistics (HLT-NAACL 04)*, Boston, USA, May 2004.

[12] F. Seide, P. Yu, C. Ma, and E. Chang. Word and sub-word indexing approaches for reducing the effects of oov queries on spoken audio. In *Proceedings of the second international conference on Human Language Technology Research*, San Diego, California, 2002.

[13] F. Seide, P. Yu, C. Ma, and E. Chang. Vocabulary-independent search in spontaneous speech. In *Proceedings of ICASSP*, Montreal, Canada, 2004.

[14] M. A. Siegler. *Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance*. PhD thesis, Carnegie Mellon University, 1999.

[15] Z. Zhou, P. Yu, C. Chelba, and F. Seide. Towards spoken-document retrieval for the internet: Lattice indexing for large-scale web search architectures. In *Proceedings of Human Language Technology Conference /North American chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL)*, New York City, USA, June 2006.