Vocabulary Independent Spoken Term Detection

Jonathan Mamou
IBM Haifa Research Labs
Haifa 31905, Israel
mamou@il.ibm.com

Bhuvana Ramabhadran, Olivier Siohan
IBM T. J. Watson Research Center
Yorktown Heights, N.Y. 10598, USA
{bhuvana,siohan}@us.ibm.com

ABSTRACT

We are interested in retrieving information from speech data like broadcast news, telephone conversations and roundtable meetings. Today, most systems use large vocabulary continuous speech recognition tools to produce word transcripts; the transcripts are indexed and query terms are retrieved from the index. However, query terms that are not part of the recognizer's vocabulary cannot be retrieved, and the recall of the search is affected. In addition to the output word transcript, advanced systems provide also phonetic transcripts, against which query terms can be matched phonetically. Such phonetic transcripts suffer from lower accuracy and cannot be an alternative to word transcripts.

We present a *vocabulary independent* system that can handle arbitrary queries, exploiting the information provided by having both word transcripts and phonetic transcripts. A speech recognizer generates word confusion networks and phonetic lattices. The transcripts are indexed for query processing and ranking purpose. The value of the proposed method is demonstrated by the relative high performance of our system, which received the highest overall ranking for US English speech data in the recent NIST Spoken Term Detection evaluation [1].

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Speech retrieval, spoken term detection, out-of-vocabulary

1. INTRODUCTION

The rapidly increasing amount of spoken data calls for solutions to index and search this data.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

The classical approach consists of converting the speech to word transcripts using a large vocabulary continuous speech recognition (LVCSR) tool. In the past decade, most of the research efforts on spoken data retrieval have focused on extending classical IR techniques to word transcripts. Some of these works have been done in the framework of the NIST TREC Spoken Document Retrieval tracks and are described by Garofolo et al. [12]. These tracks focused on retrieval from a corpus of broadcast news stories spoken by professionals. One of the conclusions of those tracks was that the effectiveness of retrieval mostly depends on the accuracy of the transcripts. While the accuracy of automatic speech recognition (ASR) systems depends on the scenario and environment, state-of-the-art systems achieved better than 90% accuracy in transcription of such data. In 2000, Garofolo et al. concluded that "Spoken document retrieval is a solved problem" [12].

However, a significant drawback of such approaches is that search on queries containing out-of-vocabulary (OOV) terms will not return any results. OOV terms are missing words from the ASR system vocabulary and are replaced in the output transcript by alternatives that are probable, given the recognition acoustic model and the language model. It has been experimentally observed that over 10% of user queries can contain OOV terms [16], as queries often relate to named entities that typically have a poor coverage in the ASR vocabulary. The effects of OOV query terms in spoken data retrieval are discussed by Woodland et al. [28]. In many applications the OOV rate may get worse over time unless the recognizer's vocabulary is periodically updated.

Another approach consists of converting the speech to phonetic transcripts and representing the query as a sequence of phones. The retrieval is based on searching the sequence of phones representing the query in the phonetic transcripts. The main drawback of this approach is the inherent high error rate of the transcripts. Therefore, such approach cannot be an alternative to word transcripts, especially for in-vocabulary (IV) query terms that are part of the vocabulary of the ASR system.

A solution would be to combine the two different approaches presented above: we index both word transcripts and phonetic transcripts; during query processing, the information is retrieved from the word index for IV terms and from the phonetic index for OOV terms. We would like to be able to process also hybrid queries, i.e, queries that include both IV and OOV terms. Consequently, we need to merge pieces of information retrieved from word index and phonetic index. Proximity information on the occurrences

of the query terms is required for phrase search and for proximity-based ranking. In classical IR, the index stores for each occurrence of a term, its offset. Therefore, we cannot merge posting lists retrieved by phonetic index with those retrieved by word index since the offset of the occurrences retrieved from the two different indices are not comparable. The only element of comparison between phonetic and word transcripts are the timestamps. No previous work combining word and phonetic approach has been done on phrase search. We present a novel scheme for information retrieval that consists of storing, during the indexing process, for each unit of indexing (phone or word) its timestamp. We search queries by merging the information retrieved from the two different indices, word index and phonetic index, according to the timestamps of the query terms. We analyze the retrieval effectiveness of this approach on the NIST Spoken Term Detection 2006 evaluation data [1].

The paper is organized as follows. We describe the audio processing in Section 2. The indexing and retrieval methods are presented in section 3. Experimental setup and results are given in Section 4. In Section 5, we give an overview of related work. Finally, we conclude in Section 6.

2. AUTOMATIC SPEECH RECOGNITION SYSTEM

We use an ASR system for transcribing speech data. It works in speaker-independent mode. For best recognition results, a speaker-independent acoustic model and a language model are trained in advance on data with similar characteristics.

Typically, ASR generates lattices that can be considered as directed acyclic graphs. Each vertex in a lattice is associated with a timestamp and each edge (u, v) is labeled with a word or phone hypothesis and its *prior probability*, which is the probability of the signal delimited by the timestamps of the vertices u and v, given the hypothesis. The 1-best path transcript is obtained from the lattice using dynamic programming techniques.

Mangu et al. [18] and Hakkani-Tur et al. [13] propose a compact representation of a word lattice called *word confusion network* (WCN). Each edge (u, v) is labeled with a word hypothesis and its *posterior probability*, *i.e.*, the probability of the word given the signal. One of the main advantages of WCN is that it also provides an alignment for all of the words in the lattice. As explained in [13], the three main steps for building a WCN from a word lattice are as follows:

1. Compute the posterior probabilities for all edges in the word lattice.
2. Extract a path from the word lattice (which can be the 1-best, the longest or any random path), and call it the *pivot* path of the alignment.
3. Traverse the word lattice, and align all the transitions with the pivot, merging the transitions that correspond to the same word (or label) and occur in the same time interval by summing their posterior probabilities.

The 1-best path of a WCN is obtained from the path containing the best hypotheses. As stated in [18], although WCNs are more compact than word lattices, in general the 1-best path obtained from WCN has a better word accuracy than the 1-best path obtained from the corresponding word lattice.

Typical structures of a lattice and a WCN are given in Figure 1.

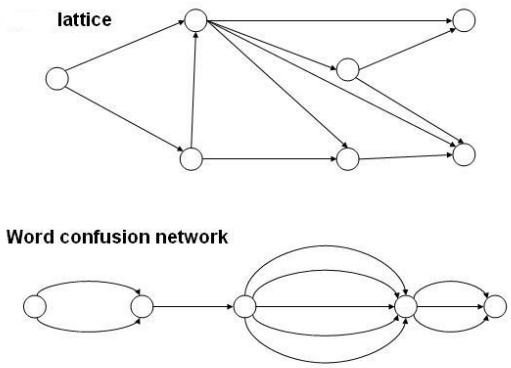

Figure 1: Typical structures of a lattice and a WCN.

3. RETRIEVAL MODEL

The main problem 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 word transcript is characterized by its word error rate (WER). There are three kinds of errors that can occur in a transcript: **substitution** of a term that is part of the speech by another term, **deletion** of a spoken term that is part of the speech and **insertion** of a term that is not part of the speech.

Substitutions and deletions reflect the fact that an occurrence of a term in the speech signal is not recognized. These misses reduce the recall of the search. Substitutions and insertions reflect the fact that a term which is not part of the speech signal appears in the transcript. These misses reduce the precision of the search.

Search recall can be enhanced by expanding the transcript with extra words. These words can be taken from the other alternatives provided by the WCN; these alternatives may have been spoken but were not the top choice of the ASR. Such an expansion tends to correct the substitutions and the deletions and consequently, might improve recall but will probably reduce precision. Using an appropriate ranking model, we can avoid the decrease in precision. Mamou et al. have presented in [17] the enhancement in the recall and the MAP by searching on WCN instead of considering only the 1-best path word transcript in the context of spoken document retrieval. We have adapted this model of IV search to term detection. In word transcripts, OOV terms are deleted or substituted. Therefore, the usage of phonetic transcripts is more desirable. However, due to their low accuracy, we have preferred to use only the 1-best path extracted from the phonetic lattices. We will show that the usage of phonetic transcripts tends to improve the recall without affecting the precision too much, using an appropriate ranking.

3.1 Spoken document detection task

As stated in the STD 2006 evaluation plan [2], the task consists in finding all the exact matches of a specific query

in a given corpus of speech data. A query is a phrase containing several words. The queries are text and not speech. Note that this task is different from the more classical task of spoken document retrieval. Manual transcripts of the speech are not provided but are used by the evaluators to find true occurrences. By definition, true occurrences of a query are found automatically by searching the manual transcripts using the following rule: the gap between adjacent words in a query must be less than 0.5 seconds in the corresponding speech. For evaluating the results, each system output occurrence is judged as correct or not according to whether it is "close" in time to a true occurrence of the query retrieved from manual transcripts; it is judged as correct if the midpoint of the system output occurrence is less than or equal to 0.5 seconds from the time span of a true occurrence of the query.

3.2 Indexing

We have used the same indexing process for WCN and phonetic transcripts. Each occurrence of a unit of indexing (word or phone) u in a transcript D is indexed with the following information:

- the **begin time** t of the occurrence of u,
- the **duration** d of the occurrence of u.

In addition, for WCN indexing, we store

- the **confidence level** of the occurrence of u at the time t that is evaluated by its posterior probability $Pr(u|t, D)$,
- the **rank** of the occurrence of u among the other hypotheses beginning at the same time t, $rank(u|t, D)$.

Note that since the task is to find exact matches of the phrase queries, we have not filtered stopwords and the corpus is not stemmed before indexing.

3.3 Search

In the following, we present our approach for accomplishing the STD task using the indices described above. The terms are extracted from the query. The vocabulary of the ASR system building word transcripts is given. Terms that are part of this vocabulary are IV terms; the other terms are OOV. For an IV query term, the posting list is extracted from the word index. For an OOV query term, the term is converted to a sequence of phones using a joint maximum entropy N-gram model [10]. For example, the term `prosody` is converted to the sequence of phones `(p, r, aa, z, ih, d, iy)`. The posting list of each phone is extracted from the phonetic index.

The next step consists of merging the different posting lists according to the timestamp of the occurrences in order to create results matching the query. First, we check that the words and phones appear in the right order according to their begin times. Second, we check that the gap in time between adjacent words and phones is "reasonable". Conforming to the requirements of the STD evaluation, the distance in time between two adjacent query terms must be less than 0.5 seconds. For OOV search, we check that the distance in time between two adjacent phones of a query term is less that 0.2 seconds; this value has been determined empirically. In such a way, we can reduce the effect of insertion errors since we allow insertions between the adjacent words and phones. Our query processing does not allow substitutions and deletions.

Example: Let us consider the phrase query `prosody research`. The term `prosody` is OOV and the term `research` is IV. The term `prosody` is converted to the sequence of phones `(p, r, aa, z, ih, d, iy)`. The posting list of each phone is extracted from the phonetic index. We merge the posting lists of the phones such that the sequence of phones appears in the right order and the gap in time between the pairs of phones $(p, r), (r, aa), (aa, z), (z, ih), (ih, d), (d, iy)$ is less than 0.2 seconds. We obtain occurrences of the term `prosody`. The posting list of `research` is extracted from the word index and we merge it with the occurrences found for `prosody` such that they appear in the right order and the distance in time between `prosody` and `research` is less than 0.5 seconds.

Note that our indexing model allows to search for different types of queries:

1. queries containing only IV terms using the word index.
2. queries containing only OOV terms using the phonetic index.
3. keyword queries containing both IV and OOV terms using the word index for IV terms and the phonetic index for OOV terms; for query processing, the different sets of matches are unified if the query terms have OR semantics and intersected if the query terms have AND semantics.
4. phrase queries containing both IV and OOV terms; for query processing, the posting lists of the IV terms retrieved from the word index are merged with the posting lists of the OOV terms retrieved from the phonetic index. The merging is possible since we have stored the timestamps for each unit of indexing (word and phone) in both indices.

The STD evaluation has focused on the fourth query type. It is the hardest task since we need to combine posting lists retrieved from phonetic and word indices.

3.4 Ranking

Since IV terms and OOV terms are retrieved from two different indices, we propose two different functions for scoring an occurrence of a term; afterward, an aggregate score is assigned to the query based on the scores of the query terms. Because the task is term detection, we do not use a document frequency criterion for ranking the occurrences.

Let us consider a query $Q = (k_0, ..., k_n)$, associated with a *boosting vector* $B = (B_1, ..., B_j)$. This vector associates a boosting factor to each rank of the different hypotheses; the boosting factors are normalized between 0 and 1. If the rank r is larger than j, we assume $B_r = 0$.

3.4.1 In vocabulary term ranking

For IV term ranking, we extend the work of Mamou et al. [17] on spoken document retrieval to term detection. We use the information provided by the word index. We define the score $score(k, t, D)$ of a keyword k occurring at a time t in the transcript D, by the following formula:

$$score(k, t, D) = B_{rank(k|t,D)} \times Pr(k|t, D)$$

Note that $0 \leq score(k, t, D) \leq 1$.

3.4.2 Out of vocabulary term ranking

For OOV term ranking, we use the information provided by the phonetic index. We give a higher rank to occurrences of OOV terms that contain phones close (in time) to each other. We define a scoring function that is related to the average gap in time between the different phones. Let us consider a keyword k converted to the sequence of phones $(p_0^k, ..., p_l^k)$. We define the normalized score $score(k, t_0^k, D)$ of a keyword $k = (p_0^k, ..., p_l^k)$, where each p_i^k occurs at time t_i^k with a duration of d_i^k in the transcript D, by the following formula:

$$score(k, t_0^k, D) = 1 - \frac{\sum_{i=1}^{l} 5 \times (t_i^k - (t_{i-1}^k + d_{i-1}^k))}{l}$$

Note that according to what we have ex-plained in Section 3.3, we have $\forall 1 \leq i \leq l, 0 < t_i^k - (t_{i-1}^k + d_{i-1}^k) < 0.2\ sec$, $0 < 5 \times (t_i^k - (t_{i-1}^k + d_{i-1}^k)) < 1$, and consequently, $0 < score(k, t_0^k, D) \leq 1$. The duration of the keyword occurrence is $t_l^k - t_0^k + d_l^k$.

Example: let us consider the sequence `(p, r, aa, z, ih, d, iy)` and two different occurrences of the sequence. For each phone, we give the begin time and the duration in second.
Occurrence 1: (`p`, 0.25, 0.01), (`r`, 0.36, 0.01), (`aa`, 0.37, 0.01), (`z`, 0.38, 0.01), (`ih`, 0.39, 0.01), (`d`, 0.4, 0.01), (`iy`, 0.52, 0.01).
Occurrence 2: (`p`, 0.45, 0.01), (`r`, 0.46, 0.01), (`aa`, 0.47, 0.01), (`z`, 0.48, 0.01), (`ih`, 0.49, 0.01), (`d`, 0.5, 0.01), (`iy`, 0.51, 0.01).
According to our formula, the score of the first occurrence is 0.83 and the score of the second occurrence is 1. In the first occurrence, there are probably some insertion or silence between the phone `p` and `r`, and between the phone `d` and `iy`. The silence can be due to the fact that the phones belongs to two different words ans therefore, it is not an occurrence of the term `prosody`.

3.4.3 Combination

The score of an occurrence of a query Q at time t_0 in the document D is determined by the multiplication of the score of each keyword k_i, where each k_i occurs at time t_i with a duration d_i in the transcript D:

$$score(Q, t_0, D) = \prod_{i=0}^{n} score(k_i, t_i, D)^{\gamma_n}$$

Note that according to what we have ex-plained in Section 3.3, we have $\forall 1 \leq i \leq n, 0 < t_i - (t_{i-1} + d_{i-1}) < 0.5\ sec$.

Our goal is to estimate for each found occurrence how likely the query appears. It is different from classical IR that aims to rank the results and not to score them. Since the probability to have a false alarm is inversely proportional to the length of the phrase query, we have boosted the score of queries by a γ_n exponent, that is related to the number of keywords in the phrase. We have determined empirically the value of $\gamma_n = 1/n$.

The begin time of the query occurrence is determined by the begin time t_0 of the first query term and the duration of the query occurrence by $t_n - t_0 + d_n$.

4. EXPERIMENTS

4.1 Experimental setup

Our corpus consists of the *evaluation set* provided by NIST for the STD 2006 evaluation [1]. It includes three different source types in US English: three hours of broadcast news (BNEWS), three hours of conversational telephony speech (CTS) and two hours of conference room meetings (CONFMTG). As shown in Section 4.2, these different collections have different accuracies. CTS and CONFMTG are spontaneous speech. For the experiments, we have processed the query set provided by NIST that includes 1100 queries. Each query is a phrase containing between one to five terms, common and rare terms, terms that are in the manual transcripts and those that are not. Testing and determination of empirical values have been achieved on another set of speech data and queries, the *development set*, also provided by NIST.

We have used the IBM research prototype ASR system, described in [26], for transcribing speech data. We have produced WCNs for the three different source types. 1-best phonetic transcripts were generated only for BNEWS and CTS, since CONFMTG phonetic transcripts have too low accuracy. We have adapted Juru [7], a full-text search library written in Java, to index the transcripts and to store the timestamps of the words and phones; search results have been retrieved as described in Section 3.

For each found occurrence of the given query, our system outputs: the **location** of the term in the audio recording (begin time and duration), the **score** indicating how likely is the occurrence of query, (as defined in Section 3.4) and a hard (binary) **decision** as to whether the detection is correct. We measure precision and recall by comparing the results obtained over the automatic transcripts (only the results having true hard decision) to the results obtained over the reference manual transcripts. Our aim is to evaluate the ability of the suggested retrieval approach to handle transcribed speech data. Thus, the closer the automatic results to the manual results is, the better the search effectiveness over the automatic transcripts will be. The results returned from the manual transcription for a given query are considered relevant and are expected to be retrieved with highest scores. This approach for measuring search effectiveness using manual data as a reference is very common in speech retrieval research [25, 22, 8, 9, 17].

Beside the recall and the precision, we use the evaluation measures defined by NIST for the 2006 STD evaluation [2]: the Actual Term-Weighted Value (ATWV) and the Maximum Term-Weighted Value (MTWV). The term-weighted value (TWV) is computed by first computing the miss and false alarm probabilities for each query separately, then using these and an (arbitrarily chosen) prior probability to compute query-specific values, and finally averaging these query-specific values over all queries q to produce an overall system value:

$$TWV(\theta) = 1 - average_q\{P_{miss}(q, \theta) + \beta \times P_{FA}(q, \theta)\}$$

where $\beta = \frac{C}{V}(Pr_q^{-1} - 1)$. θ is the detection threshold. For the evaluation, the cost/value ratio, C/V, has been determined to 0.1 and the prior probability of a query Pr_q to 10^{-4}. Therefore, $\beta = 999.9$.

Miss and false alarm probabilities for a given query q are functions of θ:

$$P_{miss}(q, \theta) = 1 - \frac{N_{correct}(q, \theta)}{N_{true}(q)}$$

$$P_{FA}(q, \theta) = \frac{N_{spurious}(q, \theta)}{N_{NT}(q)}$$

corpus	WER(%)	SUBR(%)	DELR(%)	INSR(%)
BNEWS WCN	12.7	49	42	9
CTS WCN	19.6	51	38	11
CONFMTG WCN	47.4	47	49	3

Table 1: WER and distribution of the error types over word 1-best path extracted from WCNs for the different source types.

where:

- $N_{correct}(q, \theta)$ is the number of correct detections (retrieved by the system) of the query q with a score greater than or equal to θ.
- $N_{spurious}(q, \theta)$ is the number of spurious detections of the query q with a score greater than or equal to θ.
- $N_{true}(q)$ is the number of true occurrences of the query q in the corpus.
- $N_{NT}(q)$ is the number of opportunities for incorrect detection of the query q in the corpus; it is the "Non-Target" query trials. It has been defined by the following formula: $N_{NT}(q) = T_{speech} - N_{true}(q)$. T_{speech} is the total amount of speech in the collection (in seconds).

ATWV is the "actual term-weighted value"; it is the detection value attained by the system as a result of the system output and the binary decision output for each putative occurrence. It ranges from $-\infty$ to $+1$. MTWV is the "maximum term-weighted value" over the range of all possible values of θ. It ranges from 0 to $+1$.

We have also provided the detection error tradeoff (DET) curve [19] of miss probability (P_{miss}) vs. false alarm probability (P_{FA}).

We have used the STDEval tool to extract the relevant results from the manual transcripts and to compute ATWV, MTWV and the DET curve.

We have determined empirically the following values for the boosting vector defined in Section 3.4: $B_i = \frac{1}{i}$.

4.2 WER analysis

We use the word error rate (WER) in order to characterize the accuracy of the transcripts. WER is defined as follows:

$$\frac{S + D + I}{N} \times 100$$

where N is the total number of words in the corpus, and S, I, and D are the total number of substitution, insertion, and deletion errors, respectively. The substitution error rate (SUBR) is defined by

$$\frac{S}{S + D + I} \times 100.$$

Deletion error rate (DELR) and insertion error rate (INSR) are defined in a similar manner.

Table 1 gives the WER and the distribution of the error types over 1-best path transcripts extracted from WCNs. The WER of the 1-best path phonetic transcripts is approximately two times worse than the WER of word transcripts. That is the reason why we have not retrieved from phonetic transcripts on CONFMTG speech data.

4.3 Theta threshold

We have determined empirically a detection threshold θ per source type and the hard decision of the occurrences having a score less than θ is set to `false`; false occurrences returned by the system are not considered as retrieved and therefore, are not used for computing ATWV, precision and recall.

The value of the threshold θ per source type is reported in Table 2. It is correlated to the accuracy of the transcripts. Basically, setting a threshold aims to eliminate from the retrieved occurrences, false alarms without adding misses. The higher the WER is, the higher the θ threshold should be.

BNEWS	CTS	CONFMTG
0.4	0.61	0.91

Table 2: Values of the θ threshold per source type.

4.4 Processing resource profile

We report in Table 3 the processing resource profile. Concerning the index size, note that our index is compressed using IR index compression techniques. The indexing time includes both audio processing (generation of word and phonetic transcripts) and building of the searchable indices.

Index size	0.3267 MB/HS
Indexing time	7.5627 HP/HS
Index Memory Usage	1653.4297 MB
Search speed	0.0041 sec.P/HS
Search Memory Usage	269.1250 MB

Table 3: Processing resource profile. (HS: Hours of Speech. HP: Processing Hours. sec.P: Processing seconds)

4.5 Retrieval measures

We compare our approach (`WCN_phonetic`) presented in Section 4.1 with another approach (`1-best-WCN_phonetic`). The only difference between these two approaches is that, in `1-best-WCN_phonetic`, we index only the 1-best path extracted from the WCN instead of indexing all the WCN. `WCN_phonetic` was our primary system for the evaluation and `1-best-WCN_phonetic` was one of our contrastive systems. Average precision and recall, MTWV and ATWV on the 1100 queries are given in Table 4. We provide also the DET curve for `WCN_phonetic` approach in Figure 2. The point that maximizes the TWV, the MTWV, is specified on each curve. Note that retrieval performance has been evaluated separately for each source type since the accuracy of the speech differs per source type as shown in Section 4.2.

As expected, we can see that MTWV and ATWV decrease in higher WER. The retrieval performance is improved when

	measure	BNEWS	CTS	CONFMTG
WCN_phonetic	**ATWV**	0.8485	0.7392	0.2365
	MTWV	0.8532	0.7408	0.2508
	precision	0.94	0.90	0.65
	recall	0.89	0.81	0.37
1-best-WCN_phonetic	**ATWV**	0.8279	0.7102	0.2381
	MTWV	0.8319	0.7117	0.2512
	precision	0.95	0.91	0.66
	recall	0.84	0.75	0.37

Table 4: ATWV, MTWV, precision and recall per source type.

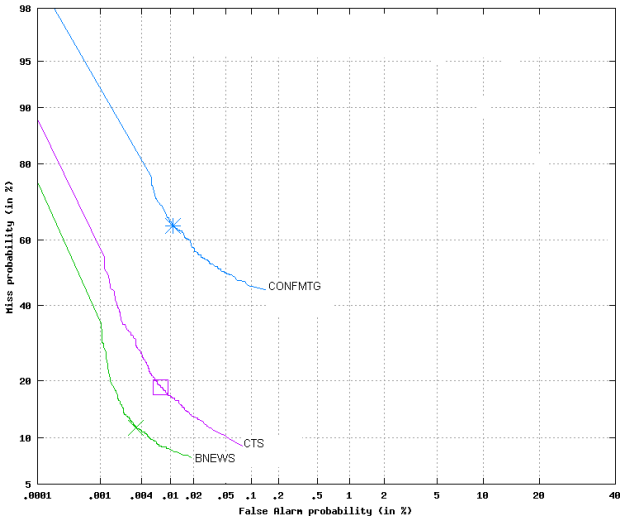

Figure 2: DET curve for WCN_phonetic approach.

using WCNs relatively to 1-best path. It is due to the fact that miss probability is improved by indexing all the hypotheses provided by the WCNs. This observation confirms the results shown by Mamou et al. [17] in the context of spoken document retrieval. The ATWV that we have obtained is close to the MTWV; we have combined our ranking model with appropriate threshold θ to eliminate results with lower score. Therefore, the effect of false alarms added by WCNs is reduced.

WCN_phonetic approach was used in the recent NIST STD evaluation and received the highest overall ranking among eleven participants. For comparison, the system that ranked at the third place, obtained an ATWV of 0.8238 for BNEWS, 0.6652 for CTS and 0.1103 for CONFMTG.

4.6 Influence of the duration of the query on the retrieval performance

We have analysed the retrieval performance according to the average duration of the occurrences in the manual transcripts. The query set was divided into three different quantiles according to the duration; we have reported in Table 5 ATWV and MTWV according to the duration. We can see that we performed better on longer queries. One of the reasons is the fact that the ASR system is more accurate on long words. Hence, it was justified to boost the score of the results with the exponent γ_n, as explained in Section 3.4.3, according to the length of the query.

quantile		0-33	33-66	66-100
BNEWS	**ATWV**	0.7655	0.8794	0.9088
	MTWV	0.7819	0.8914	0.9124
CTS	**ATWV**	0.6545	0.8308	0.8378
	MTWV	0.6551	0.8727	0.8479
CONFMTG	**ATWV**	0.1677	0.3493	0.3651
	MTWV	0.1955	0.4109	0.3880

Table 5: ATWV, MTWV according to the duration of the query occurrences per source type.

4.7 OOV vs. IV query processing

We have randomly chosen three sets of queries from the query sets provided by NIST: 50 queries containing only IV terms; 50 queries containing only OOV terms; and 50 hybrid queries containing both IV and OOV terms. The following experiment has been achieved on the BNEWS collection and IV and OOV terms has been determined according to the vocabulary of BNEWS ASR system.

We would like to compare three different approaches of retrieval: using only word index; using only phonetic index; combining word and phonetic indices. Table 6 summarizes the retrieval performance according to each approach and to each type of queries. Using a word-based approach for dealing with OOV and hybrid queries affects drastically the performance of the retrieval; precision and recall are null. Using a phone-based approach for dealing with IV queries affects also the performance of the retrieval relatively to the word-based approach.

As expected, the approach combining word and phonetic indices presented in Section 3 leads to the same retrieval performance as the word approach for IV queries and to the same retrieval performance as the phonetic approach for OOV queries. This approach always outperforms the others and it justifies the fact that we need to combine word and phonetic search.

5. RELATED WORK

In the past decade, the research efforts on spoken data retrieval have focused on extending classical IR techniques to spoken documents. Some of these works have been done in the context of the TREC Spoken Document Retrieval evaluations and are described by Garofolo et al. [12]. An LVCSR system is used to transcribe the speech into 1-best path word transcripts. The transcripts are indexed as clean text: for each occurrence, its document, its word offset and additional information are stored in the index. A generic IR system over the text is used for word spotting and search as described by Brown et al. [6] and James [14]. This strat-

index	word		phonetic		word and phonetic	
	precision	recall	precision	recall	precision	recall
IV queries	0.8	0.96	0.11	0.77	0.8	0.96
OOV queries	0	0	0.13	0.79	0.13	0.79
hybrid queries	0	0	0.15	0.71	0.89	0.83

Table 6: Comparison of word and phonetic approach on IV and OOV queries

egy works well for transcripts like broadcast news collections that have a low WER (in the range of 15%-30%) and are redundant by nature (the same piece of information is spoken several times in different manners). Moreover, the algorithms have been mostly tested over long queries stated in plain English and retrieval for such queries is more robust against speech recognition errors.

An alternative approach consists of using word lattices in order to improve the effectiveness of SDR. Singhal et al. [24, 25] propose to add some terms to the transcript in order to alleviate the retrieval failures due to ASR errors. From an IR perspective, a classical way to bring new terms is document expansion using a similar corpus. Their approach consists in using word lattices in order to determine which words returned by a document expansion algorithm should be added to the original transcript. The necessity to use a document expansion algorithm was justified by the fact that the word lattices they worked with, lack information about word probabilities. Chelba and Acero in [8, 9] propose a more compact word lattice, the *position specific posterior lattice* (PSPL). This data structure is similar to WCN and leads to a more compact index. The offset of the terms in the speech documents is also stored in the index. However, the evaluation framework is carried out on lectures that are relatively planned, in contrast to conversational speech. Their ranking model is based on the term confidence level but does not take into consideration the rank of the term among the other hypotheses. Mamou et al. [17] propose a model for spoken document retrieval using WCNs in order to improve the recall and the MAP of the search. However, in the above works, the problem of queries containing OOV terms is not addressed.

Popular approaches to deal with OOV queries are based on sub-words transcripts, where the sub-words are typically phones, syllables or word fragments (sequences of phones) [11, 20, 23]. The classical approach consists of using phonetic transcripts. The transcripts are indexed in the same manner as words in using classical text retrieval techniques; during query processing, the query is represented as a sequence of phones. The retrieval is based on searching the string of phones representing the query in the phonetic transcript.

To account for the high recognition error rates, some other systems use richer transcripts like phonetic lattices. They are attractive as they accommodate high error rate conditions as well as allow for OOV queries to be used [15, 3, 20, 23, 21, 27]. However, phonetic lattices contain many edges that overlap in time with the same phonetic label, and are difficult to index. Moreover, beside the improvement in the recall of the search, the precision is affected since phonetic lattices are often inaccurate. Consequently, phonetic approaches should be used only for OOV search; for searching queries containing also IV terms, this technique affects the performance of the retrieval in comparison to the word based approach.

Saraclar and Sproat in [22] show improvement in word spotting accuracy for both IV and OOV queries, using phonetic and word lattices, where a confidence measure of a word or a phone can be derived. They propose three different retrieval strategies: search both the word and the phonetic indices and unify the two different sets of results; search the word index for IV queries, search the phonetic index for OOV queries; search the word index and if no result is returned, search the phonetic index. However, no strategy is proposed to deal with phrase queries containing both IV and OOV terms. Amir et al. in [5, 4] propose to merge a word approach with a phonetic approach in the context of video retrieval. However, the phonetic transcript is obtained from a text to phonetic conversion of the 1-best path of the word transcript and is not based on a phonetic decoding of the speech data.

An important issue to be considered when looking at the state-of-the-art in retrieval of spoken data, is the lack of a common test set and appropriate query terms. This paper uses such a task and the STD evaluation is a good summary of the performance of different approaches on the same test conditions.

6. CONCLUSIONS

This work studies how vocabulary independent spoken term detection can be performed efficiently over different data sources. Previously, phonetic-based and word-based approaches have been used for IR on speech data. The former suffers from low accuracy and the latter from limited vocabulary of the recognition system. In this paper, we have presented a vocabulary independent model of indexing and search that combines both the approaches. The system can deal with all kinds of queries although the phrases that need to combine for the retrieval, information extracted from two different indices, a word index and a phonetic index. The scoring of OOV terms is based on the proximity (in time) between the different phones. The scoring of IV terms is based on information provided by the WCNs. We have shown an improvement in the retrieval performance when using all the WCN and not only the 1-best path and when using phonetic index for search of OOV query terms. This approach always outperforms the other approaches using only word index or phonetic index.

As a future work, we will compare our model for OOV search on phonetic transcripts with a retrieval model based on the edit distance.

7. ACKNOWLEDGEMENTS

Jonathan Mamou is grateful to David Carmel and Ron Hoory for helpful and interesting discussions.

8. REFERENCES

[1] NIST Spoken Term Detection 2006 Evaluation Website, http://www.nist.gov/speech/tests/std/.

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

[3] C. Allauzen, M. Mohri, and M. Saraclar. General indexation of weighted automata – application to spoken utterance retrieval. In *Proceedings of the HLT-NAACL 2004 Workshop on Interdiciplinary Approaches to Speech Indexing and Retrieval*, Boston, MA, USA, 2004.

[4] A. Amir, M. Berg, and H. Permuter. Mutual relevance feedback for multimodal query formulation in video retrieval. In *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 17–24, New York, NY, USA, 2005. ACM Press.

[5] A. Amir, A. Efrat, and S. Srinivasan. Advances in phonetic word spotting. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 580–582, New York, NY, USA, 2001. ACM Press.

[6] M. Brown, J. Foote, G. Jones, K. Jones, and S. Young. Open-vocabulary speech indexing for voice and video mail retrieval. In *Proceedings ACM Multimedia 96*, pages 307–316, Hong-Kong, November 1996.

[7] D. Carmel, E. Amitay, M. Herscovici, Y. S. Maarek, Y. Petruschka, and A. Soffer. Juru at TREC 10 - Experiments with Index Pruning. In *Proceedings of the Tenth Text Retrieval Conference (TREC-10)*. National Institute of Standards and Technology. NIST, 2001.

[8] C. Chelba and A. Acero. Indexing uncertainty for spoken document search. In *Interspeech 2005*, pages 61–64, Lisbon, Portugal, 2005.

[9] C. Chelba and A. Acero. Position specific posterior lattices for indexing speech. In *Proceedings of the 43rd Annual Conference of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, 2005.

[10] S. Chen. Conditional and joint models for grapheme-to-phoneme conversion. In *Eurospeech 2003*, Geneva, Switzerland, 2003.

[11] M. Clements, S. Robertson, and M. Miller. Phonetic searching applied to on-line distance learning modules. In *Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop. Proceedings of 2002 IEEE 10th*, pages 186–191, 2002.

[12] J. Garofolo, G. Auzanne, and E. Voorhees. The TREC spoken document retrieval track: A success story. In *Proceedings of the Ninth Text Retrieval Conference (TREC-9)*. National Institute of Standards and Technology. NIST, 2000.

[13] D. Hakkani-Tur and G. Riccardi. A general algorithm for word graph matrix decomposition. In *Proceedings of the IEEE Internation Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 596–599, Hong-Kong, 2003.

[14] D. James. *The application of classical information retrieval techniques to spoken documents*. PhD thesis, University of Cambridge, Downing College, 1995.

[15] D. A. James. A system for unrestricted topic retrieval from radio news broadcasts. In *Proc. ICASSP '96*, pages 279–282, Atlanta, GA, 1996.

[16] B. Logan, P. Moreno, J. V. Thong, and E. Whittaker. An experimental study of an audio indexing system for the web. In *Proceedings of ICSLP*, 1996.

[17] J. Mamou, D. Carmel, and R. Hoory. Spoken document retrieval from call-center conversations. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 51–58, New York, NY, USA, 2006. ACM Press.

[18] 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.

[19] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech '97*, pages 1895–1898, Rhodes, Greece, 1997.

[20] K. Ng and V. W. Zue. Subword-based approaches for spoken document retrieval. *Speech Commun.*, 32(3):157–186, 2000.

[21] Y. Peng and F. Seide. Fast two-stage vocabulary-independent search in spontaneous speech. In *Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP). IEEE International Conference*, volume 1, pages 481–484, 2005.

[22] M. Saraclar and R. Sproat. Lattice-based search for spoken utterance retrieval. In *HLT-NAACL 2004: Main Proceedings*, pages 129–136, Boston, Massachusetts, USA, 2004.

[23] F. Seide, P. Yu, C. Ma, and E. Chang. Vocabulary-independent search in spontaneous speech. In *ICASSP-2004, IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[24] A. Singhal, J. Choi, D. Hindle, D. Lewis, and F. Pereira. AT&T at TREC-7. In *Proceedings of the Seventh Text Retrieval Conference (TREC-7)*. National Institute of Standards and Technology. NIST, 1999.

[25] A. Singhal and F. Pereira. Document expansion for speech retrieval. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, pages 34–41, New York, NY, USA, 1999. ACM Press.

[26] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig. The IBM 2004 conversational telephony system for rich transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2005.

[27] K. Thambiratnam and S. Sridharan. Dynamic match phone-lattice searches for very fast and accurate unrestricted vocabulary keyword spotting. In *Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP). IEEE International Conference*, 2005.

[28] P. C. Woodland, S. E. Johnson, P. Jourlin, and K. S. Jones. Effects of out of vocabulary words in spoken document retrieval (poster session). In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 372–374, New York, NY, USA, 2000. ACM Press.