Spoken Document Retrieval from Call-Center Conversations

Jonathan Mamou, David Carmel, Ron Hoory
IBM Haifa Research Labs
Haifa 31905, Israel
{mamou,carmel,hoory}@il.ibm.com

ABSTRACT

We are interested in retrieving information from conversational speech corpora, such as call-center data. This data comprises spontaneous speech conversations with low recording quality, which makes automatic speech recognition (ASR) a highly difficult task. For typical call-center data, even state-of-the-art large vocabulary continuous speech recognition systems produce a transcript with word error rate of 30% or higher. In addition to the output transcript, advanced systems provide word confusion networks (WCNs), a compact representation of word lattices associating each word hypothesis with its posterior probability. Our work exploits the information provided by WCNs in order to improve retrieval performance. In this paper, we show that the mean average precision (MAP) is improved using WCNs compared to the raw word transcripts. Finally, we analyze the effect of increasing ASR word error rate on search effectiveness. We show that MAP is still reasonable even under extremely high error rate.

Categories and Subject Descriptors

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

General Terms

Algorithms

1. INTRODUCTION

The role of call-centers is becoming increasingly central to corporates in recent years. There are two main reasons for that. The first one is the increase of the importance of individuals to companies, and the drive of the latter to acquire and keep customers for the long haul. The second relates to the rapid pace of advances in technology that creates an ever-growing gap between users and automated systems, prompting them to require more technical assistance.

SIGIR'06, August 6–10, 2006, Seattle, Washington, USA.

Call-center is a general term for help desks, information lines and customer service centers operating via voice conversations; among the services typically provided by these centers are customer support, operator services and inbound and outbound telemarketing. A generalization of call-centers, known as *contact-centers*, provides any dialog-based support system in which a user receives a service by a professional. Some contact-centers provide additional means of conversing, e.g., web-based services, e-mail or on line chat services.

Information retrieval (IR) from call-center conversations is very useful for call-center managers and supervisors. It allows them to find calls on specific topics. It supports mining capabilities for monitoring agent conduct and customer satisfaction. Furthermore, it enables detection up-sell/cross-sell opportunities, and analysis of general trends. Agents can also benefit from having efficient information retrieval capabilities, for example by searching previous calls, which provide the solution to the caller's problem [12].

There are two main approaches for IR on speech data. The first approach converts the speech to a phonetic transcript and represents the query as a string of phonemes. The retrieval is based on searching the string of phonemes representing the query in the phonetic transcript, as described by Clements et al. [6]. The second approach converts the speech to a word transcript using large vocabulary continuous speech recognition tool and applies standard text retrieval methods over the transcript when processing queries. This approach, combined with the improvements in speech recognition accuracy and IR techniques, opens new opportunities in this field.

In past years, most of the research effort in spoken document retrieval (SDR) has been conducted in the framework of NIST SDR tracks [7]. These tracks focused on retrieval from word transcripts of 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 ASR systems depend 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. have concluded that "SDR is a solved problem" [7].

In this paper we argue that SDR is far from being solved, especially for noisy data such as call-center telephone calls. We follow the SDR approach over call-center conversations using the call transcriptions. We experiment with call-center calls, comprising conversations of speakers from a large population, that are much harder to transcribe automatically. State-of-the-art transcription systems achieve around 60%-

70% accuracy in transcription of such noisy spontaneous telephone calls. In other words, for this kind of data, approximately one out of three words is mis-recognized. In some circumstances like noisy channel, foreign accent, under trained engines etc., accuracy may fall to 50% or even less. The low accuracy of the data can have a dramatic effect on the precision and the recall of the search.

We present a novel scheme for information retrieval over noisy transcripts, that uses additional output from the transcription system in order to reduce the effect of recognition errors in the word transcripts. Our approach takes into consideration all word hypotheses provided by the transcription server as well as their posterior probabilities. We analyze the retrieval effectiveness at different word error levels and when different ranking models are used. We show that even for word error rate of about 50% our retrieval approach is able to perform reasonably well.

The paper is organized as follows. We describe the ASR engine and its output in Section 2. The 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 works. Finally, we conclude in Section 6.

2. AUTOMATIC SPEECH RECOGNITION (ASR) SYSTEM

We use the IBM research prototype ASR system, described in [16], for transcribing call-center data consisting of a single channel 6KHz speech (agent and caller are mixed). The output words are selected from a large US English vocabulary, which has a good coverage of the spoken language. The ASR engine works in speaker-independent mode, applying the same models for all speakers (agents and callers)[1]. 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 word lattices that can be considered as directed acyclic graphs. Each vertex is associated with a timestamp and each edge (u, v) is labeled with a word hypothesis and its *prior probability*, which is the probability of the signal delimited by the timestamps of the vertices u and v, given the word hypothesis. The 1-best path transcript is obtained from the word lattice using dynamic programming techniques.

Mangu et al. [11] and Hakkani-Tur et al. [9] 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 [9], 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.

[1]Some ASR systems train a specific speaker-dependent model, for each of the system agents.

The 1-best path of a WCN is obtained from the path containing the best hypotheses. As stated in [11], 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 word lattice and a WCN are given in Figure 1. An excerpt of the 1-best path automatic transcript of a call is shown in Figure 2. The corresponding manual transcript is also provided for reference. The automatic transcript clearly demonstrates the extremely noisy data to be handled by the SDR system. The WCN depicted in Figure 3 corresponds to the output of the ASR for the uppercase words of the last sentence of the automatic transcript. We can see that the different hypotheses appearing at the same offset have a certain acoustic similarity.

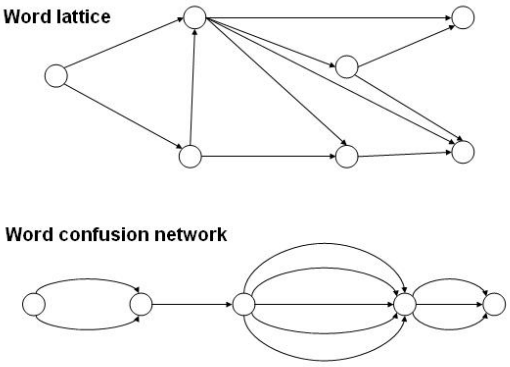

Figure 1: Typical structures of a word lattice and a WCN.

3. RETRIEVAL MODEL

The main problem with retrieving information from call-center conversations is the low accuracy of the transcription. Generally, the accuracy of a 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
- **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.

3.1 WCN indexing

Search recall can be enhanced by expanding the transcript with extra words. These words can be taken from the other

operator: thanks for calling the ibm customer service center this is john am i speaking with mark
caller: yes
operator: hey mark what's going on
caller: well i'm trying to connect to the at and t net client and i got an error it came back and gave me error one twenty it says invalid access list and i'm not sure what it means by that
operator: one twenty invalid access list
caller: uhhuh
operator: let's go and take a look at your setup real quick and see see what's going on there
caller: ok
...
caller: well it ah trying to ah oh i got something um the t and t at and t net client got A GRAPHIC ON MY SCREEN so it's connecting and it's counting the connect time so let's see it's got a little downer i think i can minimise it by clicking on it oops ah that's not it ah yeah i can minimise it ok great and i'll just try to get into lotus notes and see what happens

operator: you for calling i b m customer service center this is john to bookmark
caller: yes
operator: hey mark what's going on
caller: well i'm trying to connect to the uh a t n t net client and uh i got an error or came back and gave me an error one twenty it says invalid access list and i'm not sure what it means by that
operator: ok one twenty invalid access list
caller: ok
operator: it's gonna take a look at your setup real quick and see what see what's going on there
caller: ok
...
caller: i with to trying to install ok i've got something uhhuh at a and t a t n t net client HAVE GLASSES ON MY SCREEN fast so it's connecting and it's counting the connect time off so let's see if that'll down there at the ticket number a second notes left that's not it yeah i can minimize ok great and i'll just try to get into uh lotus notes and see what happened

Figure 2: Transcripts of a excerpt of a call. Left - a manual transcript. Right - an automatic 1-best path transcript of the same call, with 36.75% WER level .

alternatives provided by the WCN. Consequently, these alternatives may have been spoken but were not the top choice of the ASR. Such an expansion might improve recall but will probably reduce precision.

In order to improve the recall without decreasing the precision, we exploit two pieces of information provided by WCN concerning the occurrences of a term: its posterior probability and its rank among the other hypotheses. This additional information is used in order to improve the precision of the results and consequently to improve search effectiveness as measured by mean average precision (MAP) and precision at k ($P@k$).

The posterior probability reflects the confidence of the ASR in the hypothesis given the signal, thus the retrieval process will boost calls for which the query term occurs with higher probability. The rank of the hypothesis among other alternatives reflects the importance of the term relatively to other alternatives, thus, a call containing a query term that is ranked higher should be preferred over a call where the same term is ranked lower.

Let D be a call-center conversation modeled by a WCN. We denote by $Pr(t|o, D)$ the posterior probability of a term t at the offset o in the WCN of D. We denote by $rank(t|o, D)$ the rank of a term t at the offset o, where all hypotheses at offset o are sorted in decreasing order according to their posterior probabilities.

When terms are stemmed, the posterior probability of a stem t at offset o is the sum of all posterior probabilities of the terms having the same stem t at offset o. The rank of the stem is also reevaluated according to the new posterior probabilities. For example, in figure 3, the terms `graphic` and `graphics` both are stemmed to `graphic`; hence the posterior probability of the stem is $0.22+0.13 = 0.35$ and its rank is 1. The probability for the term `glass` is 0.27 and its rank is 2. It is interesting to note that the term `graphic` does not occur in the automatic 1-best path transcript although it occurs in the manual transcript (see Figure 2), thus the automatic 1-best path wrongly transcribes the speech to "have glasses on my screen" rather than "a graphics on my screen". After stemming, the term `graphic` is ranked first and appears in the automatic stemmed 1-best path.

To conclude, each occurrence of a term t in the WCN of a call D is indexed with the following information:

- the **word offset** o of t in D,
- the **confidence level** of the occurrence of t that is evaluated by its posterior probability $Pr(t|o, D)$,
- the **rank** of t among the other hypotheses at offset o, $rank(t|o, D)$.

3.2 Ranking the search results

We use the classical tf-idf vector space model for ranking the search results. Usually, term frequency is evaluated according to the number of occurrences of the term in the document. For WCNs, we readjust the term frequency of a term in the call: it is evaluated by summing the posterior probabilities of all of its occurrences over the conversation and is boosted by the rank of the term among the other hypotheses.

Consider a query Q, associated with a *boosting vector* $B = (B_1, ..., B_l)$. This vector associates a boosting factor to each rank of the different hypotheses. If the rank r is larger than l we assume $B_r = 0$. Let $occ(t, D) = (o_1, o_2, ..., o_n)$ be the sequence of all the occurrences of t in D. We define the term frequency of a term t in a conversation D, $tf(t, D)$, by the following formula:

$$tf(t, D) = \sum_{i=1}^{|occ(t,D)|} B_{rank(t|o_i,D)} \times Pr(t|o_i, D)$$

There are some interesting configurations for the boosting vector B:

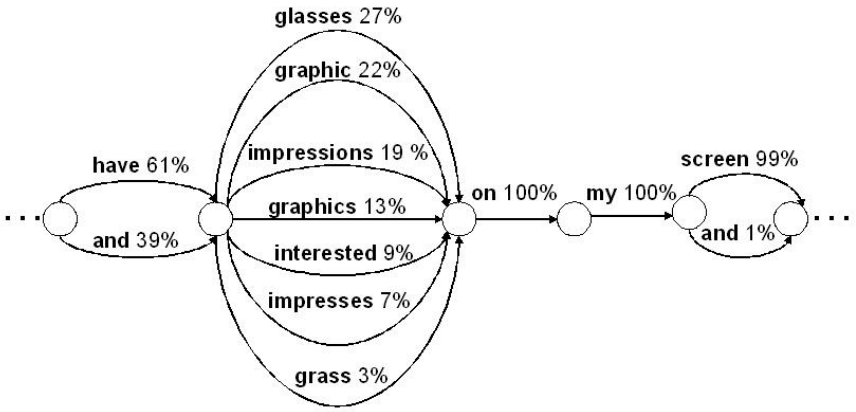

Figure 3: A fragment of the WCN of the call appeared in Figure 2. Timestamps have been omitted for clarity.

- $B = (1, 0, ..., 0)$ corresponds to the case we consider only the 1-best path transcript for retrieval, but take the confidence level of the terms into consideration.
- $B = (1, 1, ..., 1)$ corresponds to the case we consider the first l hypotheses for retrieval but ignore the rank of a term among the other alternatives.
- $B = (b_1, b_2, ..., b_l)$, where $\forall i, b_i > b_{i+1}$ corresponds to the case that the first l hypotheses are considered but higher ranked hypotheses contribute more to the document score.

Our search system also consider lexical affinities [3] during query evaluation. Lexical affinities are closely related query terms found close to each other in the document. The confidence level of a lexical affinity (t_1, t_2), is determined by the multiplication of the confidence levels of its terms: $Pr((t_1, t_2)|o, D) = Pr(t_1|o_1, D) * Pr(t_2|o_2, D)$. The rank of the lexical affinity (required for setting its boost) is determined by the maximal rank of its terms: $rank\left((t_1, t_2)|o, D\right) = \max\left(rank(t_1|o_1, D), rank(t_2|o_2, D)\right)$.

The inverse document frequency (*idf*) of a term is a measure of the relative importance of the term in the corpus. Usually, the inverse document frequency is evaluated by the ratio of the number of all the documents in the corpus to the number of documents in the corpus where the term appears. With WCNs, the occurrence of a term in a call is associated with a confidence level. We define $idf(t)$, the inverse document frequency of a term t in the corpus C by the following formula:

$$idf(t) = \log \frac{O}{O_t}$$

where O_t reflects the number of occurrences of t in the corpus and is estimated by:

$$O_t = \sum_{D \in C} \sum_{i=1}^{|occ(t,D)|} Pr(t|o_i, D)$$

and O reflects the overall number of occurrence in the corpus and is estimated by:

$$O = \sum_{t \in C} O_t$$

4. EXPERIMENTS

4.1 Experimental setup

Our corpus consists of 2236 calls made to the IBM internal customer support service during 2001. The calls deal with a large range of software and hardware problems. Manual transcripts and speaker level segmentation of the calls are also provided. The average length of a call is 9 minutes.

We divide the call-center data that we have into two sets, a training set to build the models of the ASR engine, and a test set to test the retrieval performance. When we reduce the number of calls in the training set, we increase the WER of the calls of the test set.

We use the Juru [2] search engine to index and retrieve the calls. Calls' WCNs were indexed as described in Subsection 3.1. Search results were ranked as described in Subsection 3.2. For the experiments, we have processed 120 queries that deal with typical topics of call center's discussions.

We measure precision and recall by comparing the results obtained over the automatic transcripts to the results obtained over the reference manual transcripts. Our aim is to evaluate the ability of the suggested retrieval approach to handle noisy data. Thus, the closer the automatic results to the manual results, the better the search effectiveness over the automatic transcripts. The top results returned from the manual index for a given query are considered relevant and are expected to be retrieved at the top from the automatic index. This approach for measuring search effectiveness using manual data as a reference is very common in SDR research [15, 13, 4, 5].

For evaluating retrieval effectiveness, we use the standard IR measures of precision, recall, mean average precision (MAP), precision at 10 (P@10), and precision at R (P@R) where R is the number of relevant results.

4.2 WER analysis

We use the WER in order to characterize the accuracy of the collection of calls. WER is defined as follows:

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

where N is the total number of words in the test set, 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 distribution of the error types over 1-best path transcripts extracted from automatic WCNs at different WERs. Different WER levels were achieved by providing different number of training examples to the ASR. The smaller the number of the training calls, the larger the error rate. The first row represents a limit on the lowest error rate that can be achieved by our system when all data is used for training.

The WER of the different corpora was also measured after stop-words filtering and stemming; these corpora are denoted by `stem-stpw`. Note that in our collection, around 70% of the terms are stop-words both in the manual and the automatic transcripts. This is typical for discussion transcripts where stop-words are much more frequent than in typical written text[2]. Note also the decrease in errors after stop-word filtering and stemming. The reason is that many errors relate to stop-words since shorter terms are harder to recognize. Furthermore, stemming reduces errors related to a mishmash between similar terms (e.g. `table` and `tables`).

Additionally, the table shows that while the WER increases with smaller training data, the ratio between the different error types is preserved. Around 60% of the errors are substitution errors in all collections. The rest of the errors are split differently between insertion errors and deletion errors according to the global WER. When the WER increases, DELR increases and INSR decreases. If the ASR engine is under-trained, it does not output enough words so more deletions are expected and the chance to have extra words inserted is smaller. If the ASR engine is over-trained, it outputs many alternatives and more insertion errors than deletion errors are expceted.

4.3 Lattice-based versus WCN-based retrieval

As stated in [11], the 1-best path transcript obtained from a WCN is usually more accurate than the 1-best path transcript obtained from the corresponding word lattice. However, the distribution of the different types of errors is very similar between the two 1-best path transcripts.

In order to compare the retrieval effectiveness between these two representations, we have indexed the 1-best paths of all calls as extracted from the corresponding lattices, and the 1-best paths extracted from the WCNs, and run the same 120 queries against the two indices. The results were compared to the results obtained from an index of the manual transcripts. For this experiment, the *term frequency* reflects the number of occurrences of the term in the transcript. Table 2 shows the MAP and P@10 of the results at two different WER levels.

We note that the retrieval effectiveness decreases with the increase in WER for both models. The results are very similar between 1-best paths extracted from lattices and WCNs. This can be explained by similar WER levels. However, a WCN provides much more information that can be exploited. In the following section, we show how the extra information extracted from a WCN improve search effectiveness.

[2]Using the same set of stop-words, only about 50% of the terms in the TREC-8 corpus of textual documents are marked as stop-words.

4.4 Retrieval effectiveness using WCN

In this section we test the effectiveness of the ranking model presented in Section 3.2. The following list provides a description of the different models that we compare:

- **1-best WCN TF:** the 1-best path obtained from the WCNs has been indexed. The term frequency is the number of occurrences of the term in the path.
- **all WCN TF:** all the WCN hypotheses have been indexed as explained in Section 3.1. However, the confidence level and the rank of an occurrence of a term is ignored. In other words, we index all the terms appearing in the WCN without taking into account their confidence level. The term frequency is the number of occurrences of the term in the WCN. This model is expected to improve recall while reducing precision.
- **1-best WCN CL:** the 1-best path obtained from the WCN has been indexed. Each term has been indexed with its confidence level. By ranking according to the confidence levels, the model distinguishes between terms detected with high confidence compared to low confidence detected terms. It is achieved by using the boosting vector $B = (1, 0, ..., 0)$.
- **all WCN CL:** all the hypotheses in the WCN have been indexed, with their confidence level, as explained in Section 3.1. The ranking model corresponds to the case we consider all alternatives with their confidence level but do not consider the rank of the terms compared to the other hypotheses. It is achieved by using the boosting vector $B = (1, 1, ..., 1)$.
- **all WCN CL boost:** the index is the same as for `all WCN CL`. The ranking model is based on the confidence levels as explained in Section 3.2 with the boosting vector $B = (10, 9, 8, 7, 6, 5, 4, 3, 2, 1)$, in order to boost higher ranked terms.

4.4.1 Index size

Table 3 compares the size of the different indices built at two different WER levels. While extracting all data from the WCN rather than only the 1-best paths increases the corpus size by a factor of 20, the index size of the full data is only three times larger than the index of the 1-best paths.

4.4.2 Retrieval measures

We would like to test the search effectiveness when using WCNs for indexing:

- what is the effectiveness of expanding the 1-best transcript with all the terms of the WCN?
- what is the effectiveness of using confidence level boosted by term rank for all hypotheses?

The average precision and recall of the 120 queries, at different WER levels, are respectively given in Figure 4 and Figure 5. The WER values on the x-axis are linked to the WER values for the raw transcript presented in Table 1. Both recall and precision are reduced in higher WER. It is clear that the recall is significantly improved in all error

corpus	training set	test set	WER	SUBR	DELR	INSR
raw transcript	2236	2236	32	56.5	24.5	19
stem-stpw			27	38.5	41	20.5
raw transcript	581	1655	37	58.5	24	17.5
stem-stpw			32	40	41.5	18.5
raw transcript	255	1981	40.5	60	25.5	14.5
stem-stpw			36.5	41.5	42	16.5
raw transcript	128	2108	48	58.5	31.5	10
stem-stpw			45	43	43.5	13.5
raw transcript	58	2178	53.5	61	31	8
stem-stpw			51.5	46	42	12
raw transcript	31	2205	63.5	61	35.5	5.5
stem-stpw			63.5	47.5	43	9.5

Table 1: Distribution of the error types over 1-best path extracted from WCNs at different WER levels. In each row, the first line relates to the raw transcript and the second line to the transcript after stemming and stop-words filtering.

Corpus	WER(%)	MAP	P@10
1-best WCN	32	0.83	0.96
1-best lattice	34.5	0.81	0.94
1-best WCN	40.5	0.76	0.93
1-best lattice	41	0.75	0.93

Table 2: MAP and P@10 of 1-best path transcripts extracted from word lattices vs. 1-best path transcripts extracted from WCNs at different WER levels.

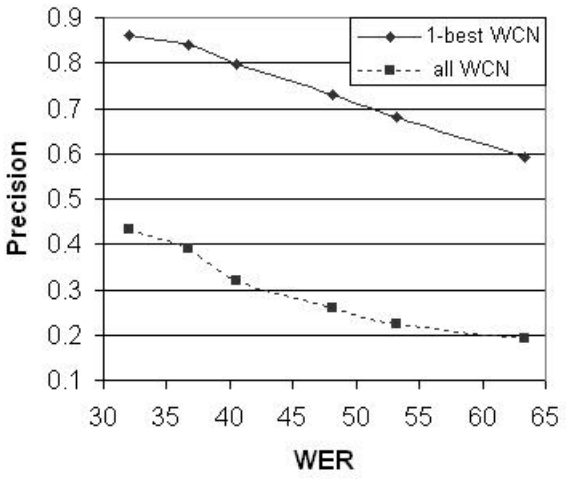

Figure 4: Precision results at different error levels.

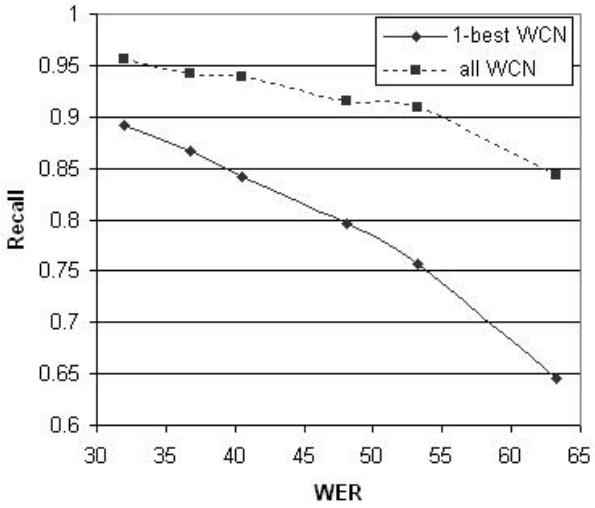

Figure 5: Recall results at different error levels.

levels, when all the hypotheses provided by the WCNs are indexed, however, precision is significantly reduced due to the added noise.

The MAP and P@R of the search results are presented in Figure 6. The graphs show that *all WCN CL boost* always outperforms the other models, especially when the WER increases. Furthermore, recall and precision are the same for the *1-best WCN CL* and *1-best WCN TF* models, however, the first outperforms the second in MAP measure due to the usage of the confidence levels. A comparison between the MAP performance of *all WCN CL* and *all WCN TF* supports this finding. This leads to the conclusion that using confidence levels improves the retrieval effectiveness.

For WER less than 42%, *1-best WCN CL* slighlty outperforms *all WCN CL* in MAP measure. However, when using the boosting factors, *all WCN CL boost* outperforms *1-best WCN CL* in MAP measure.

Despite the low precision of *all WCN* as shown in Figure 4, the ranking model based on confidence level and boosting, *all WCN CL boost*, results in the highest MAP. The conclusion is that extending the 1-best transcript with all the terms of the WCN and ranking the results according to confidence levels boosted by term rank, improves the effectiveness of the retrieval.

WER (%)	1-best			all		
	corpus	WCN TF index	WCN CL index	corpus	WCN TF index	WCN CL index
32	10.6	4.93	6.43	154	9.37	13.5
53.5	9.63	4.67	6.01	183	12.6	19.1

Table 3: Corpus and index size (in MB) at different WER levels.

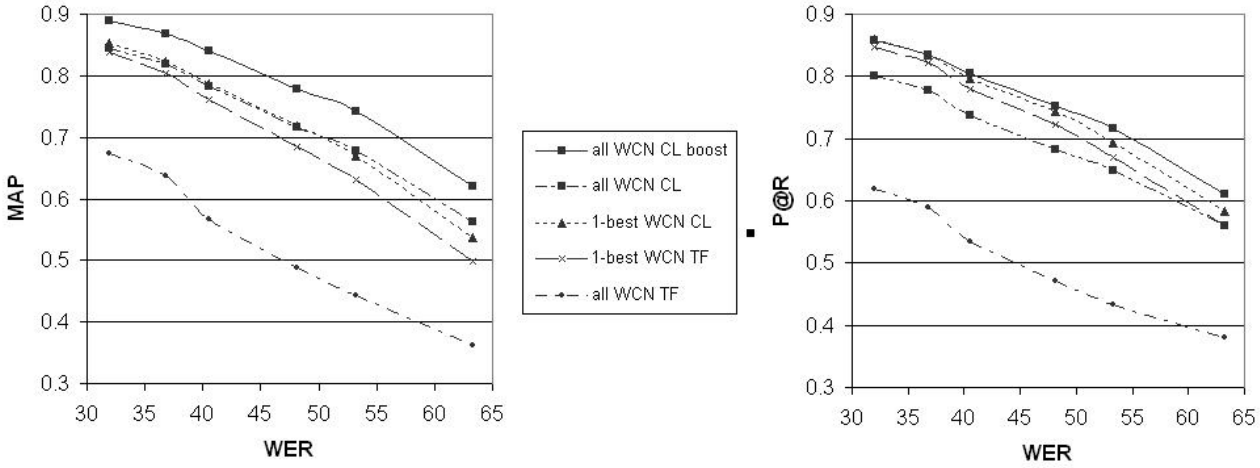

Figure 6: MAP and P@R at different error levels.

5. RELATED WORK

In the past decade, most of the research efforts on spoken data retrieval have focused on extending classical IR techniques to spoken documents. Some of these works are described by Garofolo et al. [7]. An ASR system is used to generate the transcription of the speech. This transcription is generally a 1-best path transcript. Most systems index the transcription as clean text and successfully apply a generic IR system over the text as described by Brown et al. [1] and James [10]. This strategy works well for transcripts like broadcast news stories 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. [14, 15] propose to add some terms to the transcript in order to alleviate the retrieval failures due to ASR errors. A classical way to bring new terms from an IR perspective is document expansion using a similar corpus. Their approach consists of 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.

Saraclar and Sproat in [13] show improvement in word spotting accuracy, using phonemes and word lattices, where a confidence measure of a word or a sub-word in a speech document can be derived from the lattice. Their experiments also concern telephone conversations. Similarly to our approach, they used lattices augmented with confidence measure. However, contrarily to WCNs, lattices do not provide alignment of the terms. Consequently, the offsets of the words and the sub-words are not stored in the index and proximity information between query terms cannot be exploited during retrieval. Likewise, their ranking model is only based on the confidence measure and the rank of term among the other hypotheses is not used since this piece of information is not easily retrievable from lattices. Moreover, their experiments were carried out only on a telephone conversations corpus at 40% WER and there is no analysis on retrieval effectiveness at different WER levels.

Chelba and Acero in [4, 5] 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. Furthermore, they do not test any ranking of the results based on the proximity between query terms in the speech documents. 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.

Other research has been focused on analysis of call-center data. Hakkani-Tur et al. in [17, 8] have studied the problem of spoken language understanding. More specifically, they have used WCNs for entity extraction and call classification.

Mishne et al. in [12] propose a system that assists call-center agents. The topic of the call is detected by a topic detection algorithm and then potential solutions are retrieved from a Q&A knowledge base. The system also monitors the calls by detecting off-topic segments.

6. CONCLUSIONS

This work studies how SDR can be performed efficiently over very noisy call-center data. This data comprises spontaneous speech conversations with low recording quality, which makes automatic speech recognition (ASR) a highly difficult task. For typical call-center data, even state-of-the-art large vocabulary continuous speech recognition systems produce a transcript with word error rate of 30% or higher.

Our work exploits the additional information provided by a WCN in order to improve retrieval performance. In a WCN, a compact representation of the ASR's word lattice, each word is associated with a confidence level reflecting its posterior probability given the speech signal. By taking the terms' confidence levels into consideration, and by considering the relative rank of the different hypotheses, our system is able to improve the search effectiveness.

Our experiments over true call-center conversations demonstrate the effect of increasing WER level on search effectiveness. As expected, a higher WER level hurts search results. The search recall is significantly improved by indexing all the hypotheses provided by the WCN. While the precision is decreased as expected, the MAP is improved compared to the 1-best path transcript, due to the confidence level consideration. Using our indexing scheme of WCN, the MAP is still reasonable even under extremely high error rate level, and thus effective search can be achieved.

One of problems of ASR is *out of vocabulary* (OOV) words. OOV words are missing 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. In the experiments presented in this work all query terms are in-vocabulary so this problem has been ignored. However, in real practice, OOV queries require special treatment. This is left for further research.

7. ACKNOWLEDGEMENTS

We are grateful to Olivier Siohan from the IBM T.J. Watson research center for providing the required data and for assistance on ASR topics.

8. REFERENCES

[1] 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.

[2] 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.

[3] D. Carmel, E. Farchi, Y. Petruschka, and A. Soffer. Automatic query refinement using lexical affinities with maximal information gain. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval*, pages 283–290, New York, NY, USA, 2002. ACM Press.

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

[5] 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-05)*, Ann Arbor, MI, 2005. ACL - 2005.

[6] 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.

[7] 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.

[8] D. Hakkani-Tur, F. Bechet, G. Riccardi, and G. Tur. Beyond ASR 1-best: Using word confusion networks in spoken language understanding. In *Journal of Computer Speech and Language*, 2005.

[9] 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.

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

[11] 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.

[12] G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer. Automatic analysis of call-center conversations. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 453–459, New York, NY, USA, 2005. ACM Press.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] G. Tur, D. Hakkani-Tur, and G. Riccardi. Extending boosting for call classification using word confusion networks. In *ICASSP-2004, IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.