Spoken Document Retrieval in a Highly Inflectional Language

Inger Ekman and Kalervo Järvelin
Department of Information Studies, University of Tampere
33014 University of Tampere
inger.ekman@uta.fi,
kalervo.jarvelin@uta.fi

Abstract

Being able to search for relevant information within a collection of documents is vital for the effective use of any kind of storage media. The main body of Spoken Document Retrieval research has concentrated on the English language, leaving other languages – and information produced in them – without the benefit of existing SDR systems. Especially the performance of vocabulary based speech recognition suffers from inflection. This also affects the usability of retrieval systems based on vocabulary-based recognition techniques. We discuss a method for rapid phoneme filtering to facilitate fast searches for words unknown by large vocabulary speech recognizers. The method is evaluated against a speech database in Finnish, which is a highly inflected language.

1 Introduction

With increasing storage capacity, providing audio information has become a common feature in many media systems. There are many sources of purely spoken documents, radio programs and voice messaging to name a few. Both education and research would benefit from effective storage and retrieval of audio material, such as lectures and interviews. Voice is also used to convey information in multimedia files.

Whereas some media file types (like pictures) are indifferent to language, speech is not. The main body of Spoken Document Retrieval (SDR) research has mainly focused on retrieval of English speech. Since English is a language with very limited inflection, current SDR has not adequately dealt with the implications inflection has on neither speech recognition nor its effects on the retrieval task.

Compared to English, Finnish is an exceptionally highly inflected language (Karlsson 1983). This does not, however, limit the value of this kind of research only to Finnish. In fact, looking at the languages spoken in Europe, English is one of the least inflectional languages (Lamel 2002). Current SDR methods developed for English do not take into account various difficulties in speech recognition and indexing that arise from inflection.

In this paper, we will examine the following questions: 1) How does inflection affect speech recognition and SDR? 2) What solutions are there to deal with inflection in SDR? 3) How to filter/retrieve spoken documents in a highly inflected language? And 4) what is the effectiveness of Spoken Document Filtering on a Finnish test database? How does it compare to text-based retrieval under the same conditions?

2 SDR methods

There are two main approaches to the SDR task: First one can transcribe the spoken material into words by means of a large vocabulary continuous speech recognizer (LVCSR). After recognizing the speech, text retrieval methods can be used to search through the produced transcripts. To enhance results, various methods of error-correction have been developed, e.g., using multiple recognizers (Jones et al. 1996; Ng 2000; Sanderson & Crestani 1998) or alternative recognitions of single words or phrases, also called n-best lists (Siegler 1999). Additionally, document expansion using a

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 44–50

text corpus has been used with some success to mitigate the effect of errors produced by the LVCSR (Singhal et al. 1998).

Because of the limited vocabulary of the recognizer, there will always be Out Of Vocabulary (OOV) words, which appear in the documents but not in their transcripts. This is especially problematic when the lost words are particularly descriptive, such as technical terms and names of people or places (Garofolo et al. 1998). Additionally OOV words tend to cause more than one error (Lamel 2002). Erroneous recognition of vocabulary words can also lead to terms disappearing from the documents.

Alternatively to recognizing words, documents can also be recognized at sub-word level as phones or phonemes. The recognition units can be single sounds (see Ng & Zue (1997) for results for various recognition categories) or sequences of several recognition units, but recognition is not restricted by a vocabulary of possible words. In these systems, retrieval is usually done by "translating" search words or phrases into their phonetic form and searching for appearances of similar-sounding slots.

The problem of OOV words does not occur with phone recognition. Phone recognition is also a magnitude faster than LVCSR. However, recognizing phones is more error prone than word-based recognition, because no higher-level information can be used to narrow down the number of possible interpretations. Another problem with phone-based recognition is the loss of word boundary information. Because of this, retrieval systems cannot use traditional indexing. The loss of word boundary information also further challenges the matching task. Phone retrieval methods therefore have to be able to deal with errors as well as managing the indexing task.

There are two major approaches to phone-recognized SDR, n-grams and word spotting. With n-grams, transcripts are split into n-length partially overlapping sequences, which are used in a similar fashion to words in the ordinary indexing approach (Wechsler & Schäuble 1995; Ng & Zue 1997, Ng & Zobel 1998). Since n-grams in effect examine words in separate small pieces it alleviates some of the problems of erroneous recognition as well as variations in inflection (to the extent that inflection occurs as suffixes). This error tolerant quality has also been used to deal with word variations e.g. when searching for historical word forms (Robertson & Willett 1992). Other solutions for phone based SDR have focused on trying to develop faster scanning techniques for the search phase instead of new indexing techniques (Brown et al 1996; Ferrieux & Peillon 1999; James 1995).

3 The Effects of Inflection on SDR

The target language can affect the suitability of SDR methods. The effects are mainly due to the morphology of a language. Morphology deals with the structure of morphemes - roots, prefixes and suffixes - and rules for their combination. Inflection occurs when inflectional affixes are added to a word stem. The stem can remain unchanged, or change depending on the affix. The rate of inflection of a language can be considered a continuous scale: At one end we find languages such as Vietnamese that have no inflection. At the other extreme are languages like Inuit that combine multiple morphemes into whole sentences, confusing the distinction between word and sentence. (Pirkola 2001.)

Morphology affects SDR by making both recognition and retrieval harder. Since morphology is concerned with the structural variation of words, it affects only vocabulary-based, not phone/phoneme-based SDR. Morphology affects recognition in two ways. First is the issue of building a recognition vocabulary. Due to inflection, the necessary vocabulary size is much greater the more inflected a language is. For example, due to inflection Finnish has a cautious estimate of 2000 different possible noun and 12000 verb forms. (Karlsson 1883.) The presence of inflected words in documents requires being able to match them with query words in different form. This feature directly affects the retrieval phase and has already called for new indexing methods for text retrieval (Alkula 2001). All forms are naturally not as frequent. In text retrieval inflection can be dealt with rather well by converting search terms to only a handful of all the possible variations (Kettunen & Airio 2006). Nevertheless, achieving necessary vocabulary coverage for LVCSR in inflected languages would demand including at least some of the most common inflected forms for each word in the dictionary. With larger vocabularies, performance usually suffers and recognition becomes slower.

Another problem imposed by inflection is that most recognizers use Language Models (LM) based on word order to help decide which words of the vocabulary are likely to occur in the speech stream. Since the function of an inflected word is indicated by its ending, word order is liberated. This aspect has been shown to affect the usability of traditional LMs for Russian that similar to Finnish has very free word order (Whittaker & Woodland 2003). Another problem with language models is that the statistical model of word order assumes a certain genre of speech. Whereas phone recognition is mostly concerned with recording quality and speaker, LM:s restrict a certain recognition system to certain styles of spoken communication: models suitable for the recognition of formal news speech are not as such usable for telephone conversations or informal interviews.

Many of the problems posed by inflection can be solved by using smaller recognition units such as morphemes (Kurimo et al. 2005). However, even morpheme recognizers are restricted to using a vocabulary of sorts, albeit more flexible in this aspect than LVCSR.

4 Using Phone Filtering in SDR for Inflected Languages

As argued in the previous chapter, the word-based approach is strongly affected by inflection and thus not as such appropriate for dealing with inflected SDR. Phone-based recognition, on the other hand, does not suffer from inflection. Moreover, the problem with inflection in the retrieval phase is often solved by the approximate-matching nature of the retrieval algorithms. Phone-based systems thus seem to provide several benefits over the LVCSR approach with inflected languages.

Further, word-based retrieval limits the possible search words to those of the recognition vocabulary. The only way to guarantee unlimited vocabulary searches seems to be to recognize speech on the sub-word level. Thus there will be a need for methods capable of dealing with phone recognized speech even in systems designed for less inflected languages.

However, while phone based recognition of speech is faster than LVCSR, retrieval by phones usually is slow compared to the speed of traditional indexing. Although there are accurate algorithms for word spotting, these are quite time-consuming. Filtering provides a faster way of processing phone-recognized transcripts, although perhaps not as accurate as more time-consuming methods.

In our view, filtering can benefit the system in three ways: First, the filtering process provides a rapid means of pre-processing and prioritizing documents for consecutive retrieval stages. Second, we hope to use filtering to provide the user with initial results. Listening to even a few speech documents will involve the user for some time. Meanwhile the system can use accurate but time-consuming matching algorithms to produce better results. Third, through immediate feedback, the system will be giving the user information based upon which they can rethink and evaluate their requests. Automatic relevance feedback based on the first few retrieved documents can be used to improve search performance.

4.1 Filtering with N-gram Signatures

The prototype retrieval system uses n-grams together with document signatures to make scanning the whole speech database as fast as possible. The underlying idea is to represent every document in the database by a standard-length bit-vector. After breaking down the phone-recognized speech document to n-grams (this is simply achieved by splitting document into partially overlapping *n* length subsequences), the content of the document is encoded in the vector. 1-bits are used to indicate that a certain n-gram exists in the document, whereas zero indicates non-existence. In the filtering phase, each request is put through the same procedure and compared to all the documents in the database with only a few bitwise operations per document signature. This method (also know as q-gram filtering) is a very fast way to scan for a certain sequence of characters compared to most known approximate string-matching algorithms (Navarro 2001).

4.2 The Finnish SDR Test Collection

To evaluate our filtering system, we used a test collection containing 288 news stories on different topics. The test collection is a subset of documents from a larger text database containing 55000 news documents and 35 test topics with relevance assessments. The documents are domestic, foreign and economic news, dating between 1988-1992

(Sormunen 2000). The 288 documents for the SDR collection were selected from among the relevant documents of 17 test topics (topics 2–18). Therefore we were able to use both the textual test topic and the relevance assessments. The documents were originally written newspaper articles, and were manipulated to resemble spoken news stories (elimination or rephrasing of numerical expressions, etc.). The stories were spoken by one single person and recorded in a studio environment. The resulting test database contains a total of 4h 47 min of speech, with individual stories about one minute each. Speech recognition used a phone recognizer developed at Tampere University of Technology. The recognizer produced a mixed transcript of phones and phone sequences with an average phone error of 42.0%.

Table 1 The Finnish test collection.

Nr of documents		288	
Total length		4h 47min	
Average length of speech documents		59.8s/ 93 words	
Average length of phone transcripts		712 phones	
Average phone error rate		42.0%	
Number of test topics		17	
Number of relevant documents/topic	Avg.	Min	Max
	16.9	4	39
Topic length (words)	Avg.	Min	Max
	13.9	3	27

Test requests were manually formed from the 17 topics, based on the textual requests by selecting informative words from a set of individually spoken words and word combinations (e.g. 'united nations' 'carl bildt'). Included were base forms of words appearing in the textual request, as well as words classified as possible to derive from the request with general information about the subject. The requests had on the average 13,9 words. On average, each topic had a recall base of 17 documents. The test collection is described in Table 1.

4.3 Evaluation of the filtering system

We tested the performance of n-gram filtering and compared different combinations of retrieval parameters. Tests were performed on n = 2, 3, 4 and 5. Additionally, signatures were formed both over whole news stories, as well as partially overlapping smaller story segments, or *windows*, of 10, 20, 50, 100, 200 and 500 phones. The total number of parameter combinations is 4*6=24.

To optimize processing time, all query words are combined before the matching phase by concatenating their transcripts. Each document is evaluated against the query and assigned a score $Sim(a,b) = |A \cap B| / \min \{|A|, |B|\}$, where A and B are the sets of n-grams of the current document and query, respectively. The windows are scored separately. The final document score is defined as the maximum of each document's window scores. Finally, all documents are ranked. For evaluation, the whole result list was inspected.

The filtering results are presented in Table 2. Comparisons to a text baseline are shown below, in Table 3. The text baseline used the parameter combination that gave the best results: 5-gram with 500 phone windowing.

Table 2 Average precision of multi-word topics (N=17). Highest values within each gram size in **bold**.

Window size	2-gram	3-gram	4-gram	5-gram
10	0.099	0.182	0.252	0.247
20	0.116	0.248	0.305	0.302
50	0.131	0.254	0.331	0.329
100	0.136	0.258	0.332	0.351
200	**0.152**	**0.281**	0.362	0.353
500	0.131	0.251	**0.384**	**0.381**
whole	0.109	0.217	0.335	0.362

Table 3 The average SDR effectiveness in relation to the baseline text search. The text baseline used 5-grams with 500 phone windowing.

Window size	2-gram	3-gram	4-gram	5-gram
10	10.9 %	20.1 %	27.8 %	27.3 %
20	12.9 %	27.4 %	33.7 %	33.4 %
50	14.4 %	28.1 %	36.6 %	36.4 %
100	15.1 %	28.5 %	36.7 %	38.8 %
200	16.8 %	31.1 %	40.0 %	39.1 %
500	14.4 %	27.7 %	42.5 %	42.1 %
whole	12.1 %	24.0 %	37.1 %	40.0 %

Table 2 shows filtering performance. At its best, filtering reaches an average precision of 0.384 for 4-grams using 500 phone windows.

Splitting the stories slightly improves performance. The ideal window size depends on *n* size. The effect is emphasized because queries are represented as one signature; multiple query words are unlikely to fit within smaller windows.

Table 3 shows relative precision for speech filtering compared to a text baseline using the same filtering methods. The results show performance as good as 42.5% of that of a text filtering.

4.4 Filtering Performance Across Recall Levels

In addition to knowing the average precisions, we are interested in how precision develops over recall levels. To save space, we will only consider the best approaches within each gram category, namely 2- and 3-grams with segment size 200 and 4- and 5-grams with 500 phone segments.

As can be seen from Figure 1, precision develops similarly for all top parameter combinations. For these approaches, one can see that precision drops in the usual way towards around or slightly above 10% precision at 100% recall.

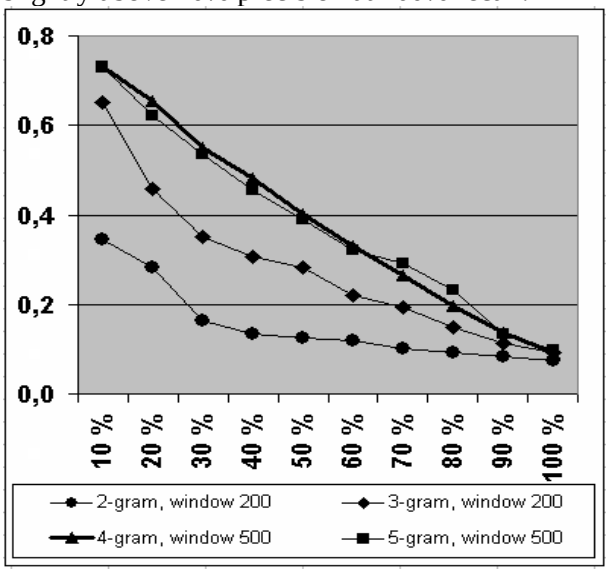

Figure 1 Precision over 10 recall levels. Precision on the y-axis, recall on the x-axis. The best approach (4-gram with window size 500) is shown in strong line.

The development of precision has implications for the use of filtering. In order to make the best of the filtering process' benefits, it is probably worth cutting off the document collection at recall levels lower than 100%. If a user would choose to inspect all the results, they will most likely also be spending more time with the system, which will allow the use of more time-consuming methods for maximal recall. However, for the typical user it is unlikely that poor precision at these levels would affect the user perception of the system, since few users can be assumed to inspect the result list this far.

More notable is the high precision obtained at the lower recall levels. The results are as high as 0.732 with 4-gram and 500 phone windowing. This means that the majority of the documents presented first in the result list are relevant, with roughly only one in four not containing relevant information. This suggests that the filtered document collection could indeed be used as a preliminary search result.

5 Conclusion and future work

The main focus of the international SDR research community has been on English. Since languages vary, it is important that research is carried out in other languages as well. The research presented in this paper aims at clarifying the effects of inflection on speech recognition and SDR and propose means to deal with SDR in highly inflected languages. One such language is Finnish, which we use as a test case to evaluate an n-gram filtering method for rapid SDR.

Morphology affects SDR on two levels: speech recognition and IR. The enormous number of words necessary for decent coverage makes LVCSR hard to implement for highly inflected languages. Less restricted word order, on the other hand, complicates the use of word order based language models. Morphology affects SDR also because request words may be in different forms than document words and thus provide less than expected evidence for retrieval – especially due to often short and sometimes variable inflectional stems.

We have examined the use of n-gram filtering for rapid scanning of a spoken document collection. The results presented in this paper suggest that this approach could be used to preprocess a database and thus shorten the retrieval phase for slower word-spotting algorithms. An average precision as high as 38.4 (42.6% relative to text filtering) was achieved with a very time-efficient parameter combination using 4-grams and a 500-phone window. Also, n-grams seem to be capable of matching words in different inflectional forms.

Comparing our results to earlier results on n-gram length for English (e.g. Ng et al. 2000) this study indicates reasonable performance on Finnish

with 4- and 5-grams, whereas results for English have shown better results for smaller 3-grams. This result may be at least in part dependent on the type of phone recognizer used. Our recognizer used phone sequences, which to some extent raises the probability of longer matches in the transcripts. However, the text baseline produced best results with 5-grams. Notably, Finnish words are longer on the average than English, so longer grams may be found beneficial in Finnish phone based SDR.

Longer n-grams naturally mean more of the descriptiveness of words is maintained. On the other hand, optimal n-gram size is affected by the recognition system used, since this dictates what kind of errors can occur. The quality of recognition also has its effects on chosen n size; the performance of long grams depends on how frequently (or consistently) recognition errors occur. Longer n-grams also increase the risk that the individual n-gram becomes over specific with regards to inflection (i.e. matches only with a certain inflectional form), or that the n-gram involves phones from several words at once. However, based on our results, this did not seem to happen. Further investigation is needed to confirm, whether this is indeed the case.

One important step towards the development of SDR methods is the availability of a suitable test database for retrieval experimentation. Unfortunately, realistic databases of several hundreds of hours of speech are as yet unavailable for Finnish. In addition to examining retrieval methods, our project also created a test database consisting of 4,7 hours of speech. This speech database, along with the spoken query words and relevance assessments, are currently being shared with other researchers and research sites, to facilitate further exploration on approaches to Finnish SDR. Promising results have been made using morph-based recognition and retrieval (Kurimo et al. 2005). Further work includes investigating the combination of LVCSR and phone retrieval to find out how approaches are optimally combined to complement each other.

6 References

Alkula, R. 2001. From plain character strings to meaningful words: Producing Better Full Text Databases for Inflectional and Compounding Languages with Morphological Analysis Software. Information Retrieval (4). 195–208.

Brown, M.; Foote, J.; Jones, G. Spärck Jones, K. & Young, S. Open-Vocabulary Speech Indexing for Voice and Video Mail Retrieval. ACM Multimedia 1996. 307–316.

Ferrieux, A. & Peillon, S. Phoneme-level Indexing for Fast and Vocabulary-Independent Voice/Voice Retrieval. Proc. ESCA ETRW Workshop on Accessing Information in Spoken Audio 1999. 60–63.

Garofolo, J.; Voorhees, E.; Auzanne, C. & Stanford, V. Spoken Document Retrieval: 1998 Evaluation and Investigation of New Metrics. ESCA ETRW workshop on Accessing Information in Spoken Audio 1998. 1–7.

James, D. The Application of Classical Information Retrieval Techniques to Spoken Documents. PhD thesis. Cambridge University. 1995.

Jones, G.; Foote, J.; Spärck Jones, K. & Young, S. Retrieving spoken documents by combining multiple index sources. ACM SIGIR 1996. ACM Press. 30–38.

Karlsson, F. Suomen kielen äänne- ja muotorakenne. [Finnish Phonetic and Morphological Structure.] WSOY. 1983.

Kettunen, K. & Airio, E. Is a morphologically complex language really that complex in full-text retrieval? In T. Salakoski et al. (Eds.): Advances in Natural Language Processing, LNAI 4139, pp. 411 - 422, 2006. Springer-Verlag Berlin Heidelberg.

Kurimo, M.; Turunen, V. & Ekman, I. Speech Transcription and Spoken Document Retrieval in Finnish. In Machine Learning for Multimodal Interaction, Revised Selected Papers of the MLMI 2004 workshop. Lecture Notes in Computer Science, Vol. 3361 pp. 253-262, Springer, 2005.

Lamel, L. Some Issues in Speech Recognizer Portability. ISCA SALTMIL SIG Workshop at LREC 2002.

Navarro, G. A guided tour to approximate string matching. ACM Computing Surveys, 33, 1 (2001). ACM Press. 31–88.

Ng, C. & Zobel, J. Speech Retrieval using Phonemes with Error Correction. ACM SIGIR 1998. ACM Press. 365–366.

Ng, C.; Wilkinson, R. & Zobel, J. Experiments in Spoken Document Retrieval using Phoneme N-grams. Speech Communication, 32, 1–2 (2000). 61–77.

Ng, K. & Zue, W. Subword Unit Representations for Spoken Document Retrieval. Eurospeech 1997. 1607–1610.

Ng, K. Information Fusion for Spoken Document Retrieval. International Conference on Acoustics Speech and Signal Processing (ICASSP) 2000.

Pirkola, A. Morphological Typology of Languages for IR. Journal of Documentation, 57, 3 (2001). 330–348.

Robertson A. & Willett, P. Searching for historical word-forms in a database of 17th-century English text using spelling-correction methods. Proc. ACM SIGIR 1992. 256–265.

Sanderson, M. & Crestani, F. Mixing and Merging for Spoken Document Retrieval. European Conference on Research and Advanced Technology for Digital Libraries 1998. 397–407.

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

Singhal, A.; Choi, J.; Hindle, D.; Lewis, D. & Pereira, F. AT&T at TREC-7. TREC-7 1998. NIST Special Publications 500-242. 239–252.

Sormunen, E. A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases. PhD thesis. University of Tampere. 2000.

Wechsler, M & Schäuble, P. Speech Retrieval Based on Automatic Indexing. Final Workshop on Multimedia Information Retrieval 1995

Whittaker, E. & Woodland, P. Language modeling for Russian and English using words and classes. Computer Speech and Language 17 (2003). 87–104.