To recover from speech recognition errors in spoken document retrieval

Mikko Kurimo and Ville Turunen

Neural Networks Research Centre
Helsinki University of Technology, Finland
Mikko.Kurimo@hut.fi

Abstract

An important difference between the retrieval of spoken and written documents is that the indexing of the speech data is usually based on automatic speech transcripts that contain recognition errors. However, there are several ways of reducing the effect of incorrect index terms in the retrieval. This paper presents retrieval experiments with unlimited vocabulary speech recognizer that utilizes a lexicon of unsupervised morpheme-like units. Based on this recognizer, three different methods are evaluated for error recovery. First, the recognized words are expanded by adding the recognized morphemes, too. Second, the words are expanded by adding the best rival morpheme candidates that were pruned away by the recognizer. Third, the queries are expanded by the potentially relevant terms found from text documents, which were retrieved from parallel text corpora by the original queries. The best results are obtained by that latter method which significantly improves the precision compared to the original queries and brings the spoken document retrieval precision to the same level as the corresponding text document retrieval.

1. Introduction

The task of retrieving spoken documents differs from retrieving the written ones, because automatic speech recognition produces speech transcripts that contain recognition errors. While this difference can be minimized by improving the word error rate in speech recognition, there are other ways that seem to lead to better results with much less effort. This conclusion can be drawn from, for example, from the results of the TREC SDR (spoken document retrieval) track [1]. An obvious reason is that in SDR, the goal is to find the correct index terms, not to perfectly transcribe every word that is said. The error recovery methods evaluated in this paper are based on recovering the lost index terms by studying the recognition process and index term selection.

The most straight-forward approach for SDR in continuous speech is to spot pre-defined keywords directly from the speech. While such systems may be light-weight and simple to construct, the obvious limitation is the required definition of the keywords and restriction of the search to them. The one-pass full-speech recognition systems, such as [2, 3], are already much more advanced, allowing basically similar indexing and retrieval implementations than the traditional textual information retrieval systems. Despite the constant advances in recognition accuracy, recognition errors affect the obtained index terms and therefore some terms may be missed and extra index terms added. In this paper steps are taken further by expanding the words in the transcripts by the subword units provided by the recognizer and also by those that the one-pass recognizer found likely, but pruned away after finding a better hypothesis. Obviously, not all of the added index terms are likely to improve the indexing or even be correct, but it is still assumed that the effect of added noise is smaller than the positive contribution of a few correct index terms recovered. In word-based recognition and indexing, the addition of rival word candidates from n-best lists and lattices into index terms has previously been studied, e.g. in [4], but in this paper we extend this approach to morpheme based recognition and indexing.

Perhaps the most efficient method to reduce the effect of speech recognition errors in SDR is to expand the queries of the user by using query results in parallel text corpora. By adding new relevant index terms to the query, the relevant documents for the original query can be retrieved even though the asked index terms would be difficult or impossible to recognize (foreign names, e.g.). Naturally, it is important to assign proper weights to the added index terms in order to keep the effect of possible irrelevant terms small. In this work the new index terms are weighted by their relevance measured by comparing the frequency of the term in the retrieved document to its frequency in the whole corpus and only the best terms are selected. We also compare expanding based on two different text corpora.

In this paper, we present an evaluation of different term expansion methods using a Finnish news reading corpus and TREC-like human relevance judgments for the sample queries and news stories [3]. In addition to our special interest in Finnish SDR, this also serves as an evaluation of the methods in a language rich with features that are very different from English. However, some of these features, such as the high amount of inflected and compound words and agglutinative morphology within the words, exist by some degree in other languages, too, e.g. in German, Russian, Hungarian and Turkish. That is why the significance of the unsupervised morphology-learning approach developed for unlimited vocabulary language modeling [5], continuous speech recognition [6, 7] and speech indexing [3] is expected grow in future. Since the approach for lexicon and language modeling is fully data-driven, it can, in principle, be more easily and with less human resources extended to other languages than, e.g., systems based on full word vocabularies or rule-based morphology.

Even though Finnish is spoken by not much more than 5 million people, there is a high demand even for purely Finnish SDR systems within the industry. Several research projects have recently been launched to develop methods suitable for retrieving speech from radio and television broadcasts, audio books, interviews and other specific recordings, and even everyday communications between people. Especially the widespread continuous usage of mobile speech devices such as phones motivates to develop applications to store, index, retrieve, and display large quantities of data in audio form.

2. Automatic speech transcriptions

The automatic speech transcriptions are carried out by HUT's continuous speech recognition system that applies unlimited vocabulary language modeling based unsupervised morpheme-like subword units. We refer to [8, 3] for the more detailed system descriptions and remind here only some features relevant to the current application.

The news reading by a single female speaker without background noise is not, as such, most challenging continuous speech recognition task nowadays. However, there are difficulties related to the very large vocabulary and the amount of foreign names that is characteristic to the news data. Additionally, there is not much relevant acoustic training data to fully train the speaker-dependent models and neither enough relevant text data to fully train the language models. That is why we choose a light-weight version of our recognizer (monophones instead of triphones and trigrams instead of fourgrams) [3].

As an agglutinative and highly inflective language, Finnish has a particularly severe out-of-vocabulary rate problem in language modeling. Even a lexicon of 500M trained on our training data would not give an acceptable OOV rate for our test data [8]. Thus, we have applied the unlimited vocabulary language modeling approach based unsupervised morpheme-like units, which we call morphs [8] to determine a lexicon 65K of subword units. This approach can also be, in principle, easily ported to other languages, because the completely data-driven unsupervised unit selection algorithm is not dependent on any given morphological rules.

A potential problem with subword lexicon is how to segment the morph sequences into words, because word breaks cannot be directly assumed after each lexical unit and in continuous speech there are, in fact, no silences between words. This was, however, quite well solved by introducing word breaks as additional units to the language model and segmenting the output afterwards based on these tags (see Figure 1).

3. Morpheme-based indexing

In information retrieval it is conventional to use stemming to associate all the inflected word forms to the same basic form or stem which is used as an index term. In this way it is not necessary to know the exact form in which each word appears in text to make successful queries. The words in different inflected forms may have slightly different meaning, and although very simple, the stemming is still a useful semantic approximation, in practice.

In languages where the words often appear in complicated forms, due to agglutinated morphemes, inflections and compound words, the stemming is not always straight-forward, especially when recognition errors may have partly corrupted the words forms. Instead of stemming, we used a commercial rule-based morphological analyzer to find the base forms that were used as index terms. However, in morpheme-based speech recognition systems such as ours, there is an alternative way for extracting meaningful fragments from the words. Because the recognizer works by building words from the morpheme-like lexical units, a simple solution is just to use those units directly as index terms. By this way also the affixes will be used as index terms, but the frequency-based index weighting will take care of suitable weighting to emphasize enough the rare index terms, such as the actual stems. An additional benefit is that recognition errors related to the word break points can be avoided, because the correct break points are not required in this approach.

Whereas the indexing by morphs performs approximately as well as by words that are returned to the baseforms [3], a significant improvement can be obtained if both indexes are combined. In Figure 2 it is shown how the simple concatenation of words and morphs improve the retrieval performance. In fact, in the current evaluation this method almost completely recovers from the speech recognition errors, because as a result the performance increases to about the same level as the corresponding retrieval from human reference transcripts.

4. Term expansion from speech data

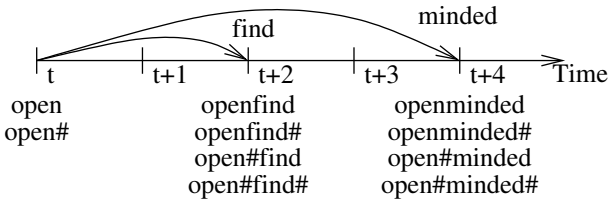

Figure 1: *Local acoustic search, stacks, and hypotheses. The two acoustically most probable fragments* find *and* minded *are used to expand the two hypotheses at frame t. The resulting hypotheses are placed in stacks according to the best ending times. The hypotheses comprise all combinations including and excluding word breaks (#).*

In HUT's continuous speech recognizer the start-synchronous stack decoder operates by storing result hypothesis in stacks corresponding to their most likely ending times [8] and then opens the stacks one at a time and expands the hypothesis by the best next morph candidates (see Fig. 1). In this process rival hypothesis are dropped whenever they fall out from the search beam.

A rather straight-forward document expansion method applied in this paper is to concatenate some of the best and most frequent morphs in the abandoned hypothesis to the speech transcription for indexing. Another slightly more elaborate method is to take into account the rejection probability of each morph, i.e. how narrow was the rejection margin, to assign a probabilistic weight to the index term.

To evaluate the document expansion by rejected morph candidates, each document is expanded by a constant amount (per document) of the best and most frequent rejected morphs. Then these expanded news documents are indexed and retrieval performance is measured as explained in Section 6. In Figure 3 the retrieval results are shown as the standard recall-precision curves for different amounts of expanded morphs (per document).

5. Term expansion from parallel text corpora

The term expansion from parallel text corpora approaches the speech recognition error recovery from the opposite side than the expansion from speech data. The main motivation is that because some of the important index terms, such as foreign names may simply be too difficult or even impossible to recognize, other related index terms are sought which are likely to appear in the same documents. An important side effect is that this is likely to provide also synonymes and other good search terms for all other query terms, as well. This effect is considered very

favorable, because it is common that some documents may be very relevant even though they do not necessarily contain the same index terms. Thus, query expansion is known to improve information from text documents as well, which can be seen in the improved reference performance in Figure 4.

A rather conventional solution to the selection of the new terms is to apply the original query to a large text corpus that is expected to contain several relevant documents. The resulting documents are then analysed and the query is expanded by index terms that are characteristic to these relevant documents. The relevance is measured by weighting the terms based on the frequency of the term in the retrieved document compared to its frequency in the whole corpus [2].

Finding the right parallel corpora for the queries is, naturally, an important practical question. Obviously, there should be enough relevant documents for each query in order to find proper expansion terms. There may also be a risk that some of the retrieved documents are not really relevant and thus, queries might get expanded by some irrelevant terms, as well. In practice, one problem is also the limited amount of suitable corpora that can be used unless the whole world-wide web is chosen (where the contexts may be too diverse).

In Figure 4 we compare two different parallel corpora, one more specific (STT) that matches well to the style and category of the spoken news articles, but from a different decade and another more general news corpus (HS) that only contains shorter articles, but includes the correct time block, as well. Preliminary experiments were first made to find the range of suitable parameters, the number of best-matching documents R and the number of best-matching terms n, and then queries were expanded and evaluated on the speech database, as usual. Only the recall-precision curves using the best parameter combinations are shown in Figure 4, but the choice of these parameters was quite robust giving about the same precisions for values +/- 5.

Naturally, the parallel text corpora could as well be used for document expansion in the same way as query expansion. In this paper we have not evaluated this, but as the TREC SDR evaluation [1] indicates, this would also be a good approach for the recovery of speech recognition errors.

6. Experiments and results

The evaluations were performed on 270 spoken news stories in Finnish. The average news story lasts one minute. The whole material is read by one single (female) speaker in a studio environment. The news are accompanied with binary relevance judgments for 17 topics made by multiple independent judges [9]. The topics are formulated as typical test queries such as: "The decisions of OPEC concerning oil price and output."

The recognized transcripts were produced by splitting the whole material into two independent sets: One for training the acoustic models of the speech recognizer and one for evaluating the recognition accuracy and the SDR performance. To be able to evaluate on the whole material we switched the roles of the sets and trained the recognizer again from the scratch [3]. The language models of the speech recognizer were trained from a corpus of 36 million words from the Finnish News Agency (newswires) and the Finnish IT center (books, newspapers, magazines).

For the query expansion experiments we applied two different parallel text corpora. STT is the same collection of newswire articles that was used to train the language models of the speech recognizer. STT contains over 100 000 short news

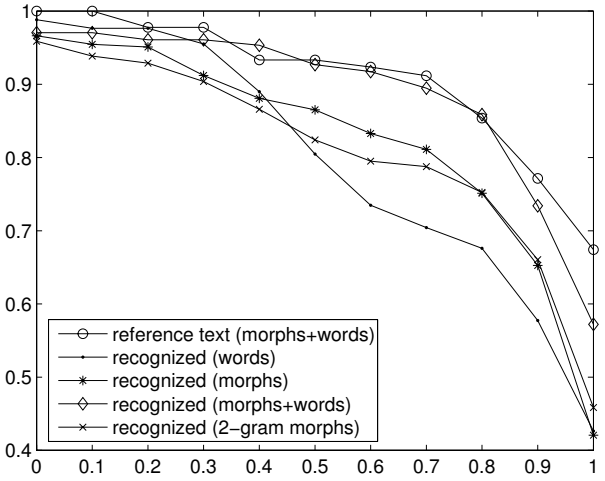

Figure 2: *Recall-precision curves in retrieval by words, single morphs, 2-gram morphs, and morphs and words combined. Reference is given by combined words and morphs from human reference transcripts.*

stories. HS is the web archive of the largest newspaper in Finland and contains 1.2 million news articles. From HS we preselected a small subcorpus using words from the titles of the spoken news as search keys to find 3728 documents. Because of the pre-selection it is expected that the expansion terms are mostly very relevant, so the results obtained probably present an upper limit of the retrieval precision rather than a representative for the whole HS corpus. Due to our limited access to HS, only a summary of about 60 words of each article was taken, which also makes the data to correspond better to the spoken news.

Figures 2, 3, and 4 represent the SDR performance by the conventional recall-precision curves. The ranked list of relevant documents is obtained for each test query and a curve is drawn to connect the averaged precisions (the vertical axis) at each of the standard recall levels (the horizontal axis). The Figure 2 indicates that combining words and morphs together improves the precisions significantly. From Figure 3 we see that expanding the morph index by the pruned morphs does not have much effect on high recall levels, but for the top ranked documents the improvements in the precision are clear. Finally in Figure 4 the query expansion in either database increases the precision both for recognized and reference texts, but reduces their difference indicating that the query expansion is indeed a good method for recovering from the recognition errors. Additionally, the HS database seems to provide better query expansions as expected in the previous paragraph.

7. Conclusions

In this paper we have presented and evaluated ways to recover from speech recognition errors in spoken document retrieval. The whole system is based on speech transcriptions obtained by a continuous speech recognizer that utilizes unlimited vocabulary language models based on unsupervised morpheme-like modeling units. The indexing of the spoken documents can be based on the language modeling units, morphs, used by the recognizer, the baseformed words built by connecting the morphs, or both which clearly obtaines the highest precisions. The evaluated expansion methods are the expansion of the documents by

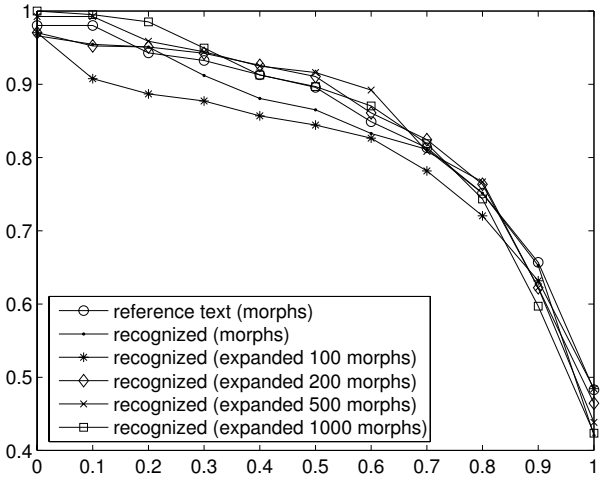

Figure 3: *Recall-precision curves in retrieval by different amounts of expanded morphs (per document). Reference is given by splitting human reference transcripts into morphs.*

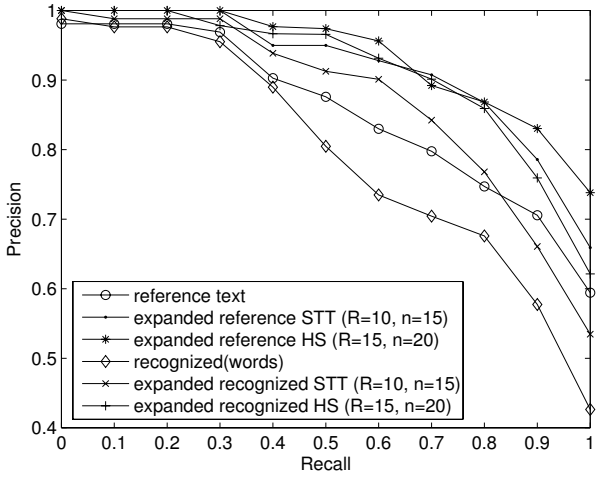

Figure 4: *Recall-precision curves in retrieval by expanding the queries. Reference is given by the same queries for human reference transcripts.*

index terms obtained from the rival morphs pruned by the recognizer and the expansion of the queries by using parallel text corpora. The results indicate that remarkable improvements can be obtained by utilizing these methods and document retrieval precision can improve close to the one obtained from human reference transcripts. Future work will be to rerun the evaluations in larger databases and also verify the portability of the results to other languages.

8. Acknowledgements

We thank Ms. Inger Ekman and the Department of Information Studies at the University of Tampere for the SDR evaluation data. The authors are grateful to the rest of the speech recognition team at the Helsinki University of Technology for help in developing the speech recognizer and the morpheme discovery, and to Mr. Nicholas Volk from University of Helsinki in expanding the numbers, abbreviations, and foreign words closer to the Finnish pronunciation for our LMs. The work was supported by the Academy of Finland in the projects *New information processing principles* and *New adaptive and learning methods in speech recognition*. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

9. References

[1] J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of Content Based Multimedia Information Access Conference*, April 12-14 2000.

[2] S. Renals, D. Abberley, D. Kirby, and T. Robinson, "Indexing and retrieval of broadcast news," *Speech Communication*, vol. 32, pp. 5–20, 2000.

[3] M. Kurimo, V. Turunen, and I. Ekman, "An evaluation of a spoken document retrieval baseline system in finnish," in *Proceedings of the International Conference on Spoken Language Processing*, 2004.

[4] M. A. Siegler, "Integration of continuous speech recognition and information retrieval for mutually optimal performance," Ph.D. dissertation, Carnegie Mellon University, 1999.

[5] M. Creutz, "Unsupervised discovery of morphemes," in *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, Philadelphia, Pennsylvania, July 2002, pp. 21–30.

[6] V. Siivola, T. Hirsimki, M. Creutz, and M. Kurimo, "Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, Sept. 2003, pp. 2293–2296.

[7] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, and M. Creutz, "On lexicon creation for Turkish LVCSR," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, Sept. 2003, pp. 1165–1168.

[8] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to finnish," *Computer Speech and Language*, 2005, (in review).

[9] E. Sormunen, "A method for measuring wide range performance of Boolean queries in full-text databases," Ph.D. dissertation, University of Tampere, 2000.