An Evaluation of a Spoken Document Retrieval Baseline System in Finnish

Mikko Kurimo and Ville Turunen

Neural Networks Research Centre
Helsinki University of Technology, Finland
mikko.kurimo@hut.fi, ville.t.turunen@hut.fi

Inger Ekman

Department of Information Studies
University of Tampere, Finland
inger.ekman@uta.fi

Abstract

This paper presents a baseline spoken document retrieval system in Finnish. Due to its agglutinative structure, Finnish speech can not be adequately transcribed using the standard large vocabulary continuous speech recognition approaches. The definition of a sufficient lexicon and the training of the statistical language models are difficult, because the words appear transformed by many inflections and compounds. In this work we apply a recently developed unlimited vocabulary speech recognition system that allows the use of n-gram language models based on morpheme-like subword units discovered in an unsupervised manner. In addition to word-based indexing, we also propose an indexing based on the subword units provided directly by our speech recognizer. In an initial evaluation of newsreading in Finnish, we obtained a fairly low recognition error rate and average document retrieval precisions close to that from human reference transcripts.

1. Introduction

With the rapidly exploding amount of spoken information available in digital libraries and other digital audio archives all over the world, the interest in searching information spoken in different languages is growing fast. The evaluations of the state-of-art spoken document retrieval (SDR) systems for broadcast news data in English have shown that the accuracy of retrieval from transcripts produced by speech recognition can already be very close to that from human reference transcripts [1]. Audio indexing systems have recently been demonstrated for several other languages, too, but the majority of the world's languages are still lacking sufficiently accurate large-vocabulary continuous speech recognition (LVCSR). Even though substantial audio archives exist, the portability of LVCSR systems to new languages is restricted by the severe structural differences to English, thus the English-driven speech technology must seek for fundamentally different solutions for success.

This paper present an evaluation of a full text recognition based SDR for Finnish. As far as we know this is pioneering work, not only for Finnish, but also for the other languages of similar agglutinative word structure, such as Estonian, Hungarian, and Turkish. The main difficulty in using the standard LVCSR technology is the required lexical and language modeling. Because the words commonly consist of many inflections and compounds, training the models of sufficient coverage of the language would not only require huge corpora, but the models would also become unfeasible to process in speech recognition. Finding a suitable set of subword units that could substitute words in building the lexicon and language models (LMs) is not an easy task, either. Furthermore, for a purely phonetic transcription approach without lexicon and LMs, the problem in continuous speech is that the recognition error rate rises very high [2].

Our approach for SDR in Finnish relies on a recently developed unlimited vocabulary speech recognition system that allows the use of statistical n-gram LMs based on morpheme-like subword units discovered in an unsupervised manner from a large text corpora [3, 4]. Related LVCSR systems that have previously been presented are, for example, the one using a more heuristically motivated unit set for Finnish [5] and the ones utilizing rule-based units for Czech [6], and Turkish [7]. The indexing of the automatically transcribed text documents can utilize a traditional weighted bag-of-words approach using stopping, stemming and suitable index weights as, for example, in [8, 9]. In this paper we evaluate two systems, one that uses baseformed words as index terms and another that used directly the morphemes produced by our speech recognizer. The retrieval is evaluated by processing the test queries into index terms, respectively, and ranking the proposed documents based on their match.

2. Speech recognition in Finnish

The system utilized for transcribing the Finnish speech into text is basicly the same as described in [3], but with a few small improvements [10, 11]. In this section we briefly describe its main features and discuss their implications to SDR and differences to other (English) SDR systems such as [8, 9].

Like standard LVCSR systems, our system has not been specificly optimized for SDR accuracy, but rather

just to minimize the word and letter error rates in order to make the transcripts generally as readable as possible. In fact, most important for SDR would be to correctly recognize all content words and to avoid adding incorrect ones. For a weighted bag-of-words index, the function words frequent in all documents and the order of appearance of the words in documents are unimportant.

2.1. Acoustic modeling

The analyzed short-time features of speech were rather conventional mel-cepstral coefficients along with the total energy and the delta values. Hidden Markov models (HMMs) were trained for 25 Finnish phonemes and 16 of their long variants. Unlike the systems described in [3, 11] and most LVCSR, the phonemes in this work were modeled context-independently. The main reason for this was to get a simpler and more compact system that would be easier to train, because the SDR evaluation task did not have much training data for the speaker. Our stack decoder that allows a flexible use of different LMs [3] also restricts the use of context dependent acoustic models, in practise, to within-word contexts, which decreases the accuracy improvement obtainable by using context dependent phoneme models.

The probability density function of emitted features in each HMM state is modeled by a mixture of 10 diagonal Gaussians including a global maximum likelihood linear transformation to uncorrelate the feature vector components. Because the phoneme durations are contrastive in Finnish, the HMMs are equipped by explicit duration models [11].

2.2. Language modeling

In agglutinative languages such as Finnish, the main problem in large-vocabulary lexical and language modeling is that the conventional word-based approach does not work well enough [3]. Lexical models suffer from the vast amount of inflected word forms and n-gram LMs additionally from the virtually unlimited word order. A solution is to split the words into morpheme-like units to build the lexicon and statistical LMs. This is possible, because the set of subword units can be selected so that all the words are adequately represented and still the pronunciation of the units can be determined from simple rules. The unsupervised machine learning algorithm presented in [4] that selects such units based on a large text corpus seems to provide means to train good LMs for unlimited vocabulary, at least for Finnish [3] and Turkish [7]. In this work we utilized back-off trigram LMs with Kneser-Ney smoothing by the SRILM toolkit [12] for a data-driven set of 65K morpheme-like units. The text corpus for morpheme discovery and LM training included totally 30M words from electronic books, newspaper texts, and short news stories.

One problem with LMs of data-driven morphemes that is very relevant in SDR is the correct transcription of foreign words, especially many different names. In our sýstem the foreign words are transformed to correspond as well as possible to the Finnish pronunciation using a set of manually designed rules. However, the pronunciation of the foreign words is variable and generally quite different from Finnish. Furthermore, many foreign names that would be important for SDR occur infrequently in the Finnish text data, so the statistically formed subword units will typically represent them by splitting into short segments, which increases the changes of confusions and reduces the strength of the LMs.

Because the recognition is performed using the morpheme-like language units, the recognition result is naturally a sequence of morphemes, not words. To be able to segment the morpheme sequences into word sequences, a special symbol was introduced in LMs to model the word break points. The word break symbol is treated like any other morpheme and it has turned out that in this way we can determine the true word breaks fairly accurately, even when no silence can be heard between consecutive words. However, frequent errors are made in word breaks related to compound words, which are difficult to human listeners, as well.

3. Indexing the transcribed documents

The full text indexing approach applied in this work means that all the words in the transcribed document are used as the index terms of the document with appropriate weights. Before building the index the inflected word forms were returned to the baseforms by a commercial morphological analyzer[1]. The words that the analyzer [13] could not process were used as index terms as such. For highly inflective languages like Finnish the use of baseforms as index terms is important, because all the inflected forms usually bear the same meaning as their baseform, with respect to the topic of the document. Splitting the compound words would also be possible, but was not applied in this work, because it might have side effects such as losing some compounds that have a specific meaning of their own.

In addition to the word-based indexing, another approach was evaluated, as well. Because the speech recognizer already knows how to split the words into morpheme-like subword units, we tried to build an index directly based on those units. The motivation for this is that we wanted be to separate the “function” morphemes from the “content” morphemes that would be more directly related to the topic of the document and thus, gain higher index weights. This approach would also make the transcription and indexing process simpler and maybe help to avoid some errors, because the transformation

[1] Licensed from Lingsoft <http://www.lingsoft.fi>.

of the morpheme sequences to word sequences could be skipped as well as the morphological analyzer needed for finding the baseforms.

The index weight of each index term in a document was the standard TFIDF, that is, the term frequency in the document divided by the frequency of documents in the whole collection, where the term occurs. The index was prepared from the transcripts using the MG toolkit [14] which was also used for the retrieval experiments. In the information retrieval (IR) phase the words in the query are first normalized, and then either baseformed or split to the morphemes, depending on the index used, to get the right kind of index terms. Then the documents pointed by the index terms are ranked by the usual way summing all the connecting index weights, and finally, the required amount of the best matching documents are returned.

4. The document retrieval evaluation

4.1. The evaluation task

The motivation of the evaluation was to find out how well a Finnish baseline SDR system would perform compared to retrieval from the corresponding human reference transcripts. The speech database consisted of 288 spoken news stories. The average news story lasted one minute. The speech documents were read aloud from written stories by one single (female) speaker in a studio environment. Before reading, the stories were modified to resemble radio broadcasts. This consisted of removing or rephrasing numeral expressions, quotation and information included in braces. The news were accompanied with binary relevance judgements for 17 topics made by multiple independent judges [15].

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.

4.2. Results

The speech recognition performance statistics shown in Table 1 indicate that this transcription task is slightly more difficult than the book reading evaluation [10]. The speech is clear, but the news data includes quite many foreign names and the match with the LM training data used is generally not very good. Furthermore, the amount of acoustic training data used for the speaker was rather small, only about 2 hours of speech, which makes the acoustic modeling task more difficult. With respect to these observations, the obtained recognition results can be considered very good.

The document retrieval performance statistics shown in Table 2 are significantly better than, for example, in TREC-SDR evaluation on North American broadcast news [1]. These two evaluations are, however, not very well comparable, because, in addition to the speech recognition issues, the performance naturally depends on the difficulty of the queries with respect to the document data. Because the obtained precisions are high and even close to the ones obtained by the reference index composed of the corresponding human transcripts, both the word- and morpheme-based indexing on the speech recognition transcripts seem to have succeeded very well.

Table 1: *The speech recognition performance in the Finnish SDR evaluation task.*

Letter error rate (LER)	7.1 %
Word error rate (WER)	30.4 %
Real-time factor (RT)	2.4

The recall-precision curves for the different indexing approaches are shown in Figure 1. The difference between the word- and morpheme-based approaches varies depending on the recall level, but the reference index seems to dominate at the different recall levels leaving still some room for improvements to be obtained by improving the speech recognition.

Table 2: *Some of the key retrieval precision statistics in the Finnish SDR evaluation. P5 refers to the precision of the documents ranked to the top five of the results.*

	Morphs	Words	Ref.
R-precision (RP)	0.72	0.71	0.79
Average precision (AP)	0.79	0.78	0.84
Top-5 precision (P5)	0.91	0.92	0.93

5. Conclusions

A new approach and a baseline system for spoken document indexing are presented based on unlimited vocabulary speech recognition. This approach allows a LM-based transcription and indexing also for highly inflective and agglutinative languages such as Finnish. The baseline system is successively evaluated in a SDR task and the obtained recognition error rate is fairly low and average document retrieval precision close to the one obtained from human reference transcripts. The future work is to check how much these baseline results can be improved by more accurate speech recognition and advanced indexing methods, such as query and document expansions using suitable background texts. The creation of a larger evaluation task using broadcast news and other radio and television programs is in progress, as well. It would also be interesting to try this approach for other languages that have either lots of inflections such as Russian or lots of compound words such as German, or both, such as Hungarian, Turkish, and Estonian.

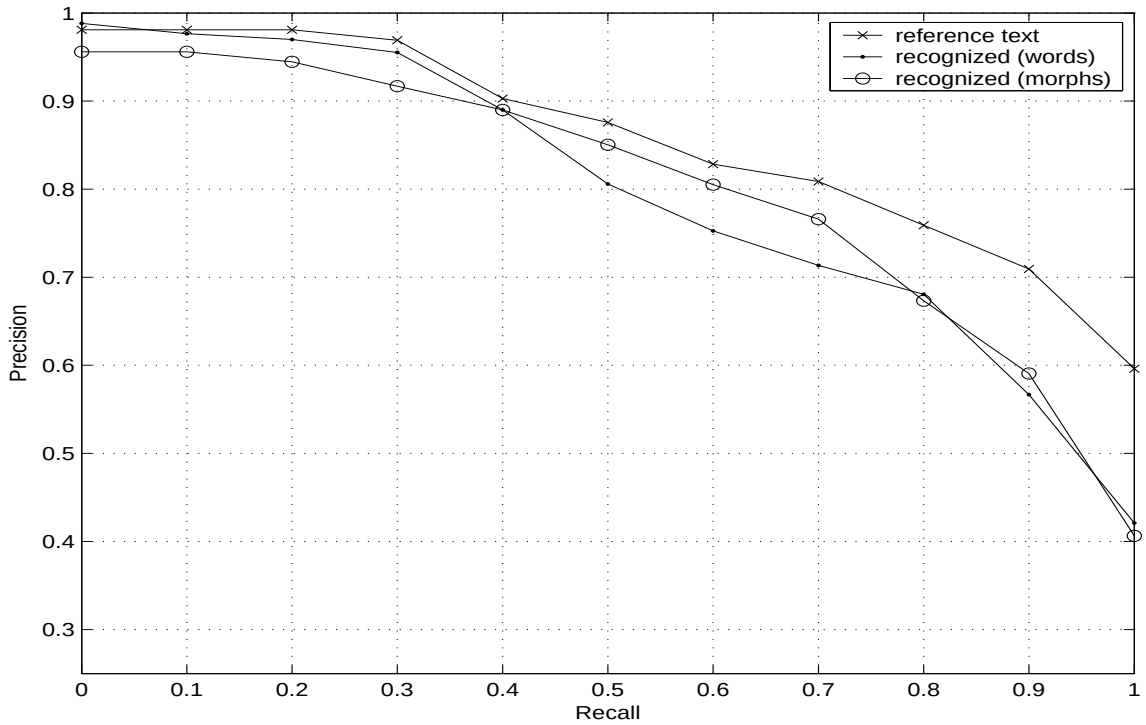

Figure 1: The IR precision at different recall levels for the automatic transcripts and the reference text.

6. Acknowledgements

The authors are grateful to the rest of the speech recognition team at the Helsinki University of Technology for help in the speech transcriptions 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*.

7. References

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

[2] I. Ekman, "Finnish speech retrieval," Master's thesis, University of Tampere, Finland, 2003, (in Finnish).

[3] V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo, "Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner," in *Proc. Eurospeech*, 2003, pp. 2293–2296.

[4] M. Creutz, "Unsupervised discovery of morphemes," in *Proc. Workshop on Morphological and Phonological Learning of ACL-02*, 2002, pp. 21–30.

[5] J. Kneissler and D. Klakow, "Speech recognition for huge vocabularies by using optimized sub-word units," in *Proc. Eurospeech*, 2001, pp. 69–72.

[6] W. Byrne, J. Hacič, P. Ircing, F. Jelinek, S. Khudanpur, P. Krbec, and J. Psutka, "On large vocabulary continuous speech recognition of highly inflectional language — Czech," in *Proc. Eurospeech*, 2001, pp. 487–489.

[7] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, and M. Creutz, "On lexicon creation for turkish LVCSR," in *Proc. Eurospeech*, 2003, pp. 1165–1168.

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

[9] B. Zhou and J. Hansen, "Speechfind: An experimental on-line spoken document retrieval system for historical audio archives," in *Proc. ICSLP*, 2002.

[10] T. Hirsimäki, M. Creutz, V. Siivola, and M. Kurimo, "Morphologically motivated language models in speech recognition," in *Proc. ICSLP*, 2004, (submitted).

[11] J. Pylkkönen and M. Kurimo, "Using phone durations in Finnish large vocabulary continuous speech recognition," in *Proc. Nordic Signal Processing Symposium (NORSIG)*, 2004, (accepted).

[12] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. ICSLP*, 2002.

[13] K. Koskenniemi, "Two-level morphology: A general computational model for word-form recognition and production," PhD thesis, University of Helsinki, 1983.

[14] I. Witten, A. Moffat, and T. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishing, 1999, 2nd edition.

[15] E. Sormunen, "A method for measuring wide range performance of Boolean queries in full-text databases, " PhD thesis, University of Tampere, 2000.