Speech Transcription and Spoken Document Retrieval in Finnish

Mikko Kurimo[1], Ville Turunen[1], and Inger Ekman[2]

[1] Helsinki University of Technology, Neural Networks Research Centre, FI-02150 Espoo, Finland
[2] Department of Information Studies, University of Tampere, Finland
Mikko.Kurimo@hut.fi
http://www.cis.hut.fi/mikkok

Abstract. This paper presents a baseline spoken document retrieval system in Finnish that is based on unlimited vocabulary continuous speech recognition. 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 the recently developed language model that enables n-gram models of 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, and a combination of the both. In an initial evaluation of newsreading in Finnish, we obtained a fairly low recognition error rate and average document retrieval precisions close to what can be obtained from human reference transcripts.

1 Introduction

The interest in searching information spoken in different languages is growing fast, because the rapid increase of spoken information available in digital libraries and other digital audio and video archives all over the world. For English data the state-of-art of spoken document retrieval (SDR) have reached the point where even the archives of spoken audio and video without manual annotation have become valuable sources of information. Some examples of such multimodal data are broadcast news, sports videos, and recordings of meetings or even telephone conversations. In some applications such as broadcast news 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 than English, 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 of such languages already exist, the portability of LVCSR systems to new languages is restricted by the

S. Bengio and H. Bourlard (Eds.): MLMI 2004, LNCS 3361, pp. 253–262, 2005.

severe structural differences of the languages. Thus, the English-driven speech technology must seek for fundamentally new solutions for success there.

This paper describes and evaluates a full text recognition based SDR system 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 also become unfeasible to process in close-realtime speech recognition. Finding a suitable set of subword units that could substitute words as building blocks of 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].

The novel Finnish SDR system relies on our research group's 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 [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]. These systems could be used for SDR, as well, given that the recognition performs sufficiently well for the rare but important content words which usually fall out of the reach of rule-based word splitting.

The indexing of the automatically transcribed text documents normally utilizes a traditional weighted bag-of-words approach with stopping, stemming and suitable index weighting as, for example, in [8, 9]. In this paper we evaluate two indexing methods, one that uses baseformed words as index terms and another that takes 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 Automatic Speech Transcripts for Finnish

The LVCSR system utilized for transcribing the Finnish speech into text is basically the same as in [3], but with a few small improvements [10]. The goal of the system development has been to make the transcripts generally as readable as possible by minimizing the average amount of word and letter errors. The SDR precision depends most on certain semantically important content words that weigh most as the index terms for the documents. Thus, it is interesting to see how well this more general LVCSR system performs in a SDR evaluation and 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].

2.1 Acoustic Modeling

The system applies context-independent hidden Markov models (HMMs) that are trained for 25 Finnish phonemes and 16 of their long variants. 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 mel-cepstral and their delta feature vector components. Because the phoneme durations are contrastive in Finnish, the HMMs are equipped by explicit duration models [10]. Most modern LVCSR systems such as the Finnish systems described in [3, 10] apply context-dependent HMMs. The main reason for deviating from this approach here, 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 practice, to within-word contexts, which somewhat decreases its benefits.

2.2 Language Modeling

The LMs in this work are back-off trigrams with Kneser-Ney smoothing trained by the SRILM toolkit [11] for a data-driven set of 65K morpheme-like units. 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]. The text corpus used in this work for morpheme discovery and LM training includes 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 the proper names. In our sstem 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.

A further problem for recognition based on subword units is that the recognition result comes as 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 LMs including the word break

symbols can then determine the word breaks 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

3.1 Word-Based Index Terms

First the obtained automatic speech transcripts must be segmented into documents which here coincides with the actual speech files. To prepare the index terms for each document, the traditional approach (in English) is to perform stopping and stemming for all the words in the transcripts. Instead of a stem, it is more convenient in Finnish to use the base form of the word that can be found by a morphological analyzer[1]. This is because the inflections may also change the root of the word so much that it would be generally difficult to define and extract unique stems. The words that the analyzer [12] 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. The initial experiments that we performed using the unprocessed inflected forms as index terms lead to very bad performance, which was no surprise. We also observed that the effect of stopping for this task was small, probably due to the applied index weighting that already strongly favours the rare words. 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 processed transcripts using the MG toolkit [13] which was also used for the retrieval experiments. In the information retrieval (IR) phase the words in the query are processed exactly like the documents to produce a list of the right kind of index terms. Each document is ranked by summing the index weights of the processed query words. Finally the ranked list is cut off at any desired level to produce the search result.

3.2 Morpheme-Based Index Terms

Whereas the first indexing method was based on word baseforming that requires several further processing steps after the speech recognition, we developed another method, as well, which is much simpler and more direct. Because the speech recognizer already knows how to split the words into morpheme-like subword units designed for obtaining better LMs, we took those units directly from the recognizer's output as the index terms for the document. Typically, these units that we call morphs perform an operation that resembles the English stemming, that is, separates a root-like morpheme that often occurs in the corpus as such from the frequently occurring prefixes and suffixes. Although the statistical

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

morphs are very rough approximations of stems, prefixes and suffixes, because they are only based on the available training corpus, they have also other qualities that make them highly plausible as index terms. This approach makes the transcription and indexing process very simple, because we can skip the word building and baseforming phases. Thus, it is also likely to avoid all the errors caused by the transformation of the morpheme sequences into word sequences and limitations of the morphological analyzer needed for finding the baseforms.

The MG toolkit [13] is applied as in word-based index, but directly on the speech recognizer's output corresponding to each document. IR is performed similarly as well, except that instead of processing queries by baseforming, the words are split into the morphs in exactly the same way as the texts used for training the recognizer's LMs.

3.3 Combined Index

Experiments were also performed to combine the word-based and morpheme-based indexes in the retrieval. As a simple way to obtain a combined index we concatenated both index term lists for each documents and then proceeded by MG to build the total index. The same concatenation approach was then utilized for processing the queries.

4 Experiments and Results

4.1 Goal and Measures for the SDR Evaluation

The purpose of the evaluation was twofold. First, we wanted to evaluate our recently developed Finnish LVCSR system using various metrics relevant to the intended application. Second, we wanted to check how well the new baseline SDR system performs compared to retrieval from human reference transcripts.

The most common measure of LVCSR performance is the word error rate (WER). In WER all word errors (substituted, added and deleted words) are counted equally significant. For applications where LVCSR is needed to understand the content or to perform some actions, it is natural that not all the words nor word errors are equally meaningful. A step towards document retrieval is to use the term error rate (TER) instead of WER. TER counts only errors that most likely affect the indexing, so it compares only the frequencies of words after stemming (word suffixes excluded) and stopping (common function words excluded). TER is defined as the difference of two index term histograms (the recognition result H and the correct transcription R) (summation t is over all resulting terms):

$$\mathrm{TER} = \sum_t |R(t) - H(t)| / \sum_t R(t) * 100\% \ . \tag{1}$$

For languages like Finnish WER is sometimes quite inaccurate measure of speech recognition performance, because the words are long and constitute of a highly variable amount of morphemes. For example, a single misrecognized

morpheme in the word "Tietä-isi-mme-kö-hän" leads to 100 % WER, whereas in English the corresponding WER for the translation "Would we really know" would be only 25 %. Because the exact extraction of the morphemes is often difficult, the phoneme or letter error rate (LER) has been used instead.

To evaluate the SDR performance we have adopted the measures used in the TREC-SDR evaluation [1]. The ranked list of relevant documents obtained for each test query is analyzed according to the precision at different recall levels and the total recall-precision curve for different systems is plotted (see Fig. 1, for example). Some key statistics can also be computed such as the average precision (AP) over all recall levels and the precision of the top R documents (RP), where R is the amount of documents relevant to the query. In practical IR work, the precision of the top ranked documents, the five best (P5), for example, is also quite relevant. Although it is obviously difficult to compose an exhaustive set of test queries and human relevance judgments, these recall and precision measures are expected to differentiate the performance of the LVCSR systems in a more meaningful manner than by using the direct transcription error rates.

4.2 Transcription and Retrieval Task

The speech data consists of 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. 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 are accompanied with binary relevance judgments for 17 topics made by multiple independent judges [14]. 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.

4.3 Results

Table 1 shows the performance of the described baseline Finnish speech recognition system. The current task seems to be more difficult than the previous book reading evaluation [3, 10]. The speech is clear and the noise level low, but there was only about two hours of suitable training data available for the speaker. Due to this lack of training data, we choose to apply context-independent phoneme models, in contrary to the earlier works [3, 10]. This reduces dramatically the amount of acoustic models to be estimated. The LM training data is the same as in the previous evaluations, but it does not necessarily match well to the spoken news that were from a different decade than the newswire texts. Given these somewhat inaccurate acoustic and language models, the obtained recognition results are not bad at all.

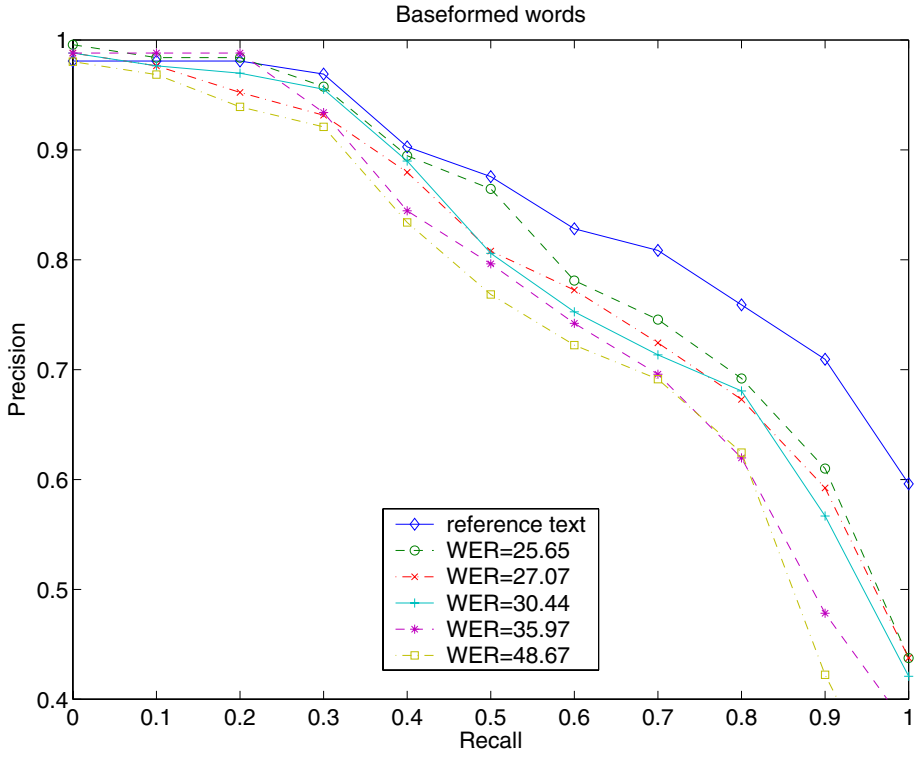

Fig. 1. The IR precision at different recall levels for the alternative ways to run the speech recognizer automatic transcripts compared to using the human reference transcripts

Table 1. The performance statistics of the speech recognizer in the transcription task. The corresponding average precision obtained from human reference transcripts was 84.1%. Beam size refers to the pruning settings of the decoder. For "Data+" we got 50 % more acoustic training data

	Beam 20	Beam 30	Beam 40	Beam 70	Data+
Real-time factor (RT)	0.7	1.3	2.4	8.1	8.1
Word error rate (WER) %	48.7	36.0	30.4	27.1	25.7
Letter error rate (LER) %	12.1	8.5	7.1	6.2	5.6
Term error rate (TER) %	44.8	31.8	26.2	22.3	20.8
Average precision (AP) %	72.9	75.9	78.0	78.0	80.4

The different transcriptions in Table 1 were obtained by changing the amount of pruning (the beam width increased from 20 to 70 hypothesis) in the decoder and finally adding 50% more acoustic training data. The results indicate that more training data and less pruning does not only decrease recognition errors,

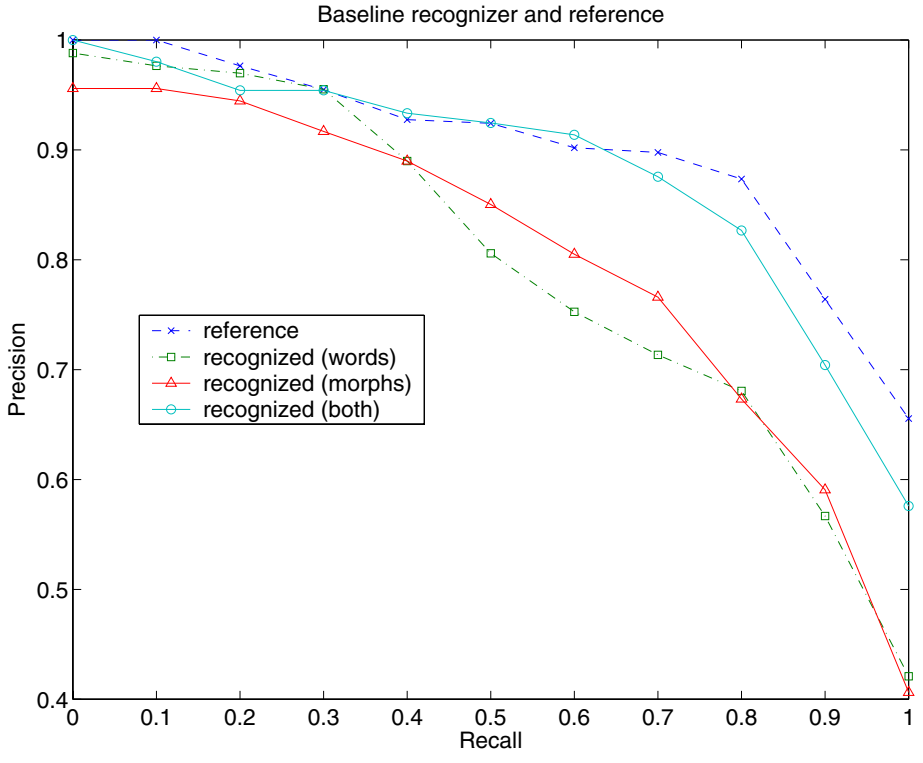

Fig. 2. The IR precision at different recall levels for the alternative ways to index the automatic transcripts. The speech recognizer is the same for all indexes (and the same as "Beam 40" in Table 1). The reference index was made as "both", but using the human reference transcripts

Table 2. Some of the key retrieval precision statistics in the SDR evaluation obtained for the alternative indexes. The speech recognizer is the same for all indexes (and the same as "Beam 40" in Table 1). The reference index was made as "Words+Morphs", but using the human reference transcripts

	Morphs	Words	Words+Morphs	Reference
R-precision (RP) %	71.6	71.2	81.3	84.9
Average precision (AP) %	79.2	78.0	87.5	89.4
Top-5 precision (P5) %	90.6	91.8	92.9	94.1

but also improves the average SDR precision. Figure 1 shows the more detailed evaluation of the SDR using the standard recall-precision curve.

Figure 2 compares the recall-precision curve of the baseline document index (baseformed words as index terms) to index prepared from the recognizer's output morphs directly. Although there are clear performance differences along the

curve, the average precisions by the two indexes are almost the same. However, as Table 2 clearly demonstrates, combining the two indexes (morphs and baseform words) seems to give the best results which are already very close to the precisions obtained from the human reference transcripts.

5 Discussion

As a more meaningful evaluation of the speech recognition performance than the standard error rate analysis, the SDR recall and precision show clear improvements obtained by increasing training data and decreasing hypothesis pruning. Table 1 indicates as well that all the improvements in speech recognition, even as measured by the term error rate, do not imply a higher retrieval precision. An example of this is the increase of the decoder's beam parameter above 40 (see Table 1). Even though the best obtained speech transcripts still fail to produce as accurate an index as the reference transcripts, the performance is so close that the baseline speech recognizer seems to be good enough for this application.

Based on these experiments in Finnish we obviously cannot state, how successful would the morpheme-based speech transcription and retrieval be in other languages. The applicability of morphemes may depend on several issues: how much information can be read from the morpheme structure, how well can it be automatically revealed by the unsupervised word-splitting algorithm, and how well does the morpheme-based recognition fit to the decoder and LMs at hand. However, it seems that because some recognized documents are better retrieved by using baseformed words and some by morphs, the combination of both indexes would maximize the recall and precision of the retrieval.

6 Conclusions

We described a new spoken document indexing and retrieval system based on unlimited vocabulary speech recognition. This approach enables the use of statistical language models in the transcription and indexing for highly inflective and agglutinative languages such as Finnish. The baseline system is successively evaluated in a recently developed Finnish SDR task. The obtained recognition error rate is fairly low and the average document retrieval precision close to the one obtained from human reference transcripts.

Future work is to check how much the 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 SDR evaluation task using broadcast news and other radio and television programs is already in progress. It will 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.

Acknowledgements

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.

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. J. Pylkkönen and M. Kurimo, "Using phone durations in Finnish large vocabulary continuous speech recognition," in *Proc. Nordic Signal Processing Symposium (NORSIG)*, 2004.
11. A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. ICSLP*, 2002.
12. K. Koskenniemi, "Two-level morphology: A general computational model for word-form recognition and production," PhD thesis, University of Helsinki, 1983.
13. I. Witten, A. Moffat, and T. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images.* Morgan Kaufmann Publishing, 1999, 2nd edition.
14. E. Sormunen, "A method for measuring wide range performance of Boolean queries in full-text databases, " PhD thesis, University of Tampere, 2000.