Publication IV

Ville T. Turunen and Mikko Kurimo. Indexing confusion networks for morph-based spoken document retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, pp. 631–638, July 2007.

Indexing Confusion Networks for Morph-based Spoken Document Retrieval

Ville T. Turunen and Mikko Kurimo
Adaptive Informatics Research Centre
Helsinki University of Technology
PO Box 5400, FI-02015, TKK, Finland
{ville.t.turunen,mikko.kurimo}@tkk.fi

ABSTRACT

In this paper, we investigate methods for improving the performance of morph-based spoken document retrieval in Finnish by extracting relevant index terms from confusion networks. Our approach uses morpheme-like subword units ("morphs") for recognition and indexing. This alleviates the problem of out-of-vocabulary words, especially with inflectional languages like Finnish. Confusion networks offer a convenient representation of alternative recognition candidates by aligning mutually exclusive terms and by giving the posterior probability of each term. The rank of the competing terms and their posterior probability is used to estimate term frequency for indexing. Comparing against 1-best recognizer transcripts, we show that retrieval effectiveness is significantly improved. Finally, the effect of pruning in recognition is analyzed, showing that when recognition speed is increased, the reduction in retrieval performance due to the increase in the 1-best error rate can be compensated by using confusion networks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech recognition and synthesis*

General Terms

Algorithms, Languages, Performance

Keywords

Spoken document retrieval, Subword indexing, Morphemes, Confusion networks, Lattices

1. INTRODUCTION

As more and more spoken information is produced and archived, there is an increasing need for indexing and retrieving audio material based on the speech content. In addition to TV and radio broadcasts, increasing amount of audio and video material is distributed on the Internet, e.g. in the form of podcasts and video sharing web sites such as YouTube. Currently, these archives can only be retrieved based on human-inputted metadata rather than their actual content. As the available material becomes more diverse, the requirements for the retrieval systems increase. Furthermore, different languages pose different challenges for retrieval. Most research is done for English, but these results can not always successfully be applied to other languages with different morphology and other properties.

Content based retrieval of speech data utilizes an automatic speech recognition (ASR) system to produce a transcript of the speech for indexing. Two main approaches have commonly been used for spoken document retrieval (SDR). In phone-based retrieval, the speech is transcribed into a string of phonemes. Query words are also transformed to phoneme strings and then matched to the recognizer outputs. The second, word-based, approach uses a large vocabulary continuous speech recognition (LVCSR) system to transcribe the speech into words and then applies standard text retrieval methods to the transcripts. This has been the most successful approach in the TREC SDR tracks [3].

However, word-based methods suffer from the limited vocabulary of the speech recognizer. Any word in speech that is not in the vocabulary (*out-of-vocabulary*, OOV) will always be misrecognized and is replaced by an alternative that is deemed probable by the acoustic and language models. Phoneme recognizers are not limited to any vocabulary, but their performance is hurt by higher error rates. Typically, the vocabulary consists of the most frequent words in the language model training corpus. For retrieval this is especially problematic, since the less frequent words, such as proper names, are usually the most interesting from retrieval point of view.

The problem of limited vocabulary can be alleviated by 'backing-off' to the phoneme transcription at locations where no word of the vocabulary fits. This is the basic principle of OOV-detection proposed by Hazen and Bazzi [4]. Other methods include query and document expansion where relevant terms are extracted from a parallel text corpus. The added semantically related terms help retrieve documents with missing OOV terms [20].

Our approach is based on morpheme-like subword units learned in an unsupervised manner. We call these units *statistical morphs* for short. The recognizer transcribes the speech as a string of morphs, leaving almost no words out of vocabulary. This approach is especially useful for languages

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

with rich morphology (e.g. Finnish, Turkish), that have a high number of different inflected word forms and thus cannot use traditional language modeling based on whole words. The morph language model assigns word break positions and thus both word level and morph level information can be used for indexing. The data driven algorithm is easily applicable for other languages with similar properties.

Retrieval using the recognizer transcripts as they were error free has proved successful for low error rates [3]. State-of-the-art systems can achieve low enough error rates (better than 90% accuracy) for broadcast news material in English. But as the databases grow larger, the amount of CPU power that can be used for every fixed time of speech decreases. Also, it is more demanding to index speech recordings that have less optimal properties than noiseless non-spontaneous speech, e.g. recordings of meetings, telephone conversations, etc. Thus, it is not always possible to obtain recognizer transcripts that are accurate enough for successful retrieval.

Retrieval performance can be increased by extracting alternative hypotheses from the recognizer in addition to the most probable (1-best) candidate. A *lattice* is a graph containing a number of most probable hypotheses considered by the recognizer and can be used as a source for extracting additional terms. A more compact representation for the hypotheses is the *word confusion network* (WCN), which offers a convenient representation of competing terms along with the posterior probability for each term. Mamou et al. [11] have shown improvements of SDR performance in low accuracy conditions by indexing and weighting terms in confusion networks based on their probability and rank among competitors.

In this paper, we investigate methods for improving the performance of morph-based spoken document retrieval in Finnish by extracting relevant index terms from confusion networks. Comparing against 1-best transcripts and error-free human transcripts, we show that retrieval effectiveness is significantly improved. As far as we know, this is the first time such methods have been applied to retrieval of speech in a highly inflectional language like Finnish.

This paper is organized as follows. In section 2, we present the methods used in this work. Especially, the morph-based retrieval scheme is described in more detail as well as the generation and use of the confusion networks. Section 3 presents the experiments and the obtained results. Overview of related work is presented in Section 4. Finally, our conclusions are given in Section 5.

2. METHODS

2.1 Morph-based retrieval

Most research on speech retrieval is focused on English data, but different languages have different properties that make methods developed for one language less usable for others. One such property is the level of agglutination. Finnish is a highly agglutinative language, which means that words are formed by joining together morphemes and thus there are a high number of distinct inflected word forms. This affects both the recognition and retrieval phase of the SDR process.

2.1.1 Recognition phase

Statistical language models for speech recognition are typically built by observing co-occurrence statistics such as n-grams in a large text corpus. This works for English as a reasonably sized lexicon can cover the language well. For a highly inflective language with a huge number of distinct word forms, constructing a fixed lexicon of words becomes infeasible. Also, training an effective language model using inflected words as units becomes very hard as the amount of training data needed to cover enough instances of all the different forms grows too large. One solution is to use subword language model units instead of whole words. If the units are chosen well, the size of the lexicon and the amount of training data that are needed to cover the language are greatly reduced.

An unsupervised algorithm for finding suitable subword units has been developed by Creutz [1]. Based on the Minimum Description Length (MDL) principle, the algorithm takes in an unsegmented training corpus and finds a set of units that is compact but models the training set effectively. An n-gram model can then be built over a corpus that is segmented using these units. The units produced by the algorithm are referred to as *statistical morphs* as the algorithm chooses the units based on statistical criteria and as the boundaries between the units in segmented word forms often coincide with grammatical morpheme boundaries. Speech recognition accuracy in Finnish has been greatly improved by utilizing statistical morphs in the language model [5]. As the algorithm is completely data driven, it can be easily applied to other languages. An example transcript produced by the morph-based recognizer is shown in Figure 1.

2.1.2 Retrieval phase

A speech recognizer with morph language modeling transcribes the speech into a string of morphs with markers at word break positions. Thus, both morph-level and word-level information can be used for indexing. Typically, in retrieval of an inflectional language, a morphological analyzer is used to lemmatize each inflected word form before indexing. However, not all languages have a morphological analyzer available as building one requires special linguistic knowledge. Furthermore, in the case of speech retrieval, errors in the transcript can cause the morphological analysis to fail and produce spurious results. This happens if a morph in a word is misrecognized and the resulting word is grammatically incorrect thus confusing the morphological analyzer. Also, the word break positions are sometimes wrongly assigned, again leading to confusion. The language model should prevent most of these situations, but not all of these errors can be avoided.

Because the statistical morphs resemble grammatical morphemes, they are also an appealing candidate to be used as index terms as such. For retrieval, query terms are also segmented to morphs using the same segmentation algorithm. Thus, the need for the morphological analyzer can be avoided. This resembles stemming as it separates the affix morphs from the stems. Spoken document retrieval in Finnish using morphs as index terms produces results that are about equal compared to the lemmatized transcripts [8]. However, best results have been achieved by combining both methods.

The morph-based approach provides also alleviation for the OOV query term problem as it is now possible to recognize almost any word in speech by recognizing its component morphs. In practice, this means that the rare words, such as some proper names, get transcribed into many small morphs

```
<w> valtio ta <w> yhteis ymmärrys <w> saa ttaa
<w> purka utua <w> jo <w> anta isi <w> jäädy ttää
<w> itsenäisyys julistuksen sa <w> sadaksi <w>
päivä ksi <w> ei <w> merkitse <w> si tä <w> että
<w> riippuma ttomaksi <w tasa valla ksi <w>
julista utuneen <w> liettua <w> tinki si <w>
itsenäistymis tavoitte istaan <w> hän <w>
koro sti <w> liettua n <w> pääministeri <w>
kat si mie ra <w> p ru n ski ene <w> joka <w>
saapu i <w> lauantai na <w> kotka n <w> uusi <w>
hansa <w> seminaari n <w>
```

```
Baltia|ssa yhteis|ymmärrys, saarto
purka|utuu.  Liettua|n päätös jäädy|ttää
itsenäisyys|julistuksen|sa sadaksi
päivä|ksi ei merkitse si|tä, että
riippuma|ttomaksi tasa|valla|ksi
julista|utunut Liettua tinki|si
itsenäistymis|tavoittee|staan.  Tä|tä
koro|sti Liettua|n pääministeri
Kat|si|mie|ra P|ru|n|ski|ene, joka
saapu|i lauantai|na Kot|kaan Uusi
Hansa -seminaari|in.
```

Figure 1: Transcripts of a part of a Finnish news story about the independence struggles of Lithuania. Left, the recognized 1-best transcript of morphs with word break position marked with <w>. Right, the manual transcript, aligned by line. The morph boundaries are marked with | . Notice the recognition of the name *Kazimiera Prunskiene*. (The letter "z" is transformed to the closest Finnish pronunciation "ts".)

while the more common words are formed of bigger pieces or just one morph. This behavior is caused by the statistical nature of the segmentation algorithm. However, errors are still often made, especially with foreign names that contain foreign sounds.

A problem similar to *understemming* and *overstemming* arises from non-ideal segmentation of inflected word forms. Sometimes different inflected forms of the same base form produce different stems as the boundary is placed at a wrong place. In some cases, it is not even possible to find positions for the boundaries in all the different inflected forms to produce a stem that is not confused with the stems of other words. This problem can be alleviated by the use of query expansion [9] or latent semantic indexing [19].

2.2 Lattices and Confusion Networks

The speech recognizer takes as input the speech signal and generates morph lattices that represent a large number of alternative hypotheses in the form of directed acyclic graphs. Each node of the graph has a timestamp. Each edge is labeled with a morph hypothesis and its acoustic and language model likelihoods. The edge corresponds to the signal delimited by the timestamps of the start and end nodes.

A more compact representation of a lattice called *word confusion network* (WCN) or *sausage* has been proposed by Mangu et. al [12]. A confusion network contains a number of alignment positions and, in each position, a set of mutually exclusive word hypotheses called the *confusion set*. Each word in a confusion set is associated with its posterior probability i.e. the probability of the word given the signal at that time interval. Sentence hypotheses can be generated by freely combining hypotheses at each alignment position.

The 1-best path in a confusion network is simply obtained by picking the term with the highest posterior probability at each alignment position. The error rates are in general lower for the 1-best paths obtained from the confusion networks than for the 1-best paths of the lattice [12].

For indexing, confusion networks offer a convenient source for expanding the transcript with alternative recognition candidates. Confusion networks are more compact than lattices and they also provide alignment for all the terms in the lattice. With confusion networks, it is easy to rank locally competing terms by their posterior probability and use the information for indexing.

At general level, the algorithm for transforming a lattice to a confusion network consists of the following steps:

1. Compute the posterior probability for all edges in the lattice
2. Pruning: remove all edges with posterior probability below some threshold
3. Intra-word clustering: merge together edges corresponding to the same word instance and sum their posterior probabilities
4. Inter-word clustering: group different words which compete around the same time interval and have similar phonetic properties to form confusion sets

For a detailed description of the algorithm, see [12].

Pruning is needed to achieve better alignment of competing terms as it removes constraining low probability paths. This results in more accurate 1-best paths as explained in [12]. Removing very low probability terms can also increase retrieval performance as these terms were not likely spoken in reality. However, if the pruning threshold is too high, there is a risk of removing correct terms and thus reducing recall.

With our morph-based recognizer, the confusion networks consist of morphs instead of words. A special marker indicates word break positions. An example morph confusion network is presented in Figure 2. The network corresponds to the beginning of the transcript of Figure 1.

2.3 Indexing and ranking confusion networks

Retrieval performance is decreased if a relevant term that is spoken is misrecognized and is thus missing from the transcript. However, it is possible that the correct term was considered by the recognizer but was not the top choice. In that case, the term will appear in the confusion network. Adding these alternative hypothised terms to the index is expected to increase recall. However, as most of the candidates in the confusion network were in fact not spoken, we need to be careful so that the spurious terms do not hurt precision too much.

Following the notation in [11], let D be a document modeled by a confusion network. We use two pieces of information in the confusion network for each occurrence of a term t at position o: its posterior probability $Pr(t|o, D)$ and rank among competitors $rank(t|o, D)$. Posterior probability tells

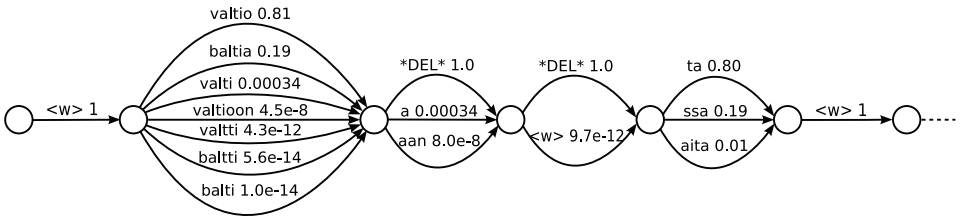

Figure 2: Beginning of the confusion network for the story of Figure 1. <w> marks word break positions and *DEL* deletions (empty hypotheses). The 1-best path is <w> valtio ta <w> (*of the state*). The correct result <w> baltia ssa <w> (*in the Baltic*) is also present in the network.

how confident the recognizer is that the term occurs in the signal at that position. Rank of the term reflects the importance of the term relative to the other alternatives. In retrieval, document with a higher probability and/or rank of a term should be preferred to one with lower values.

The classical vector space model with tf-idf weights and cosine distance relevance measure is used for ranking the search results [15]. Normally, term frequency tf is the number of times a term occurs in a document. In our case, we need to estimate a value for term frequency based on the posterior probabilities and ranks of each occurrence of the term in the confusion network of a document. We compare two methods for the estimate.

In the first method, term frequency is evaluated by summing the posterior probabilities of all of its occurrences in the confusion network. This means, that if the recognizer is confident that the term at a location is correctly recognized (posterior probability close to one), term frequency is added by (close to) one as in the case of indexing error free text documents. Less weight is given to terms with less confidence. Thus, the term frequency of a term t in a document D, $tf(t, D)$ is defined (confidence level or **CL-method**):

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

This is the same as used by Mamou et al. [11], except that we omit the boosting vector, which would assign a boosting factor to each rank of the different hypotheses. In our case, experimenting with different values of boosting did not improve the results.

Instead, we use the ranks in a different way in our second method for estimating the term frequency. Siegler [17] noted that, in the case of lattices, the probability values do not necessarily give good estimates for term frequencies. Better results were achieved by using only the ranks of locally competing terms. Motivated by this, our second method for estimating term frequency is defined by (**rank-method**):

$$tf(t, D) = \sum_{i=1}^{|occ(t,D)|} 1/rank(t|o_i, D) \quad (2)$$

This means that the highest ranked terms of each alignment of the lattice, which correspond to the 1-best result, get weights of one. The 1-best result is then expanded by competing terms, which are given less and less weight as their rank increases.

The inverse document frequency *idf* indicates the relative importance of a term in the corpus. Traditionally, *idf* is a function of the number of documents in the collection the term occurs in. In our case, we simply counted the number of confusion networks that the term occurs in at any position. In other words, term occurrence $o(t, D)$ was estimated by:

$$o(t, D) = \begin{cases} 1, & \text{if } tf(t, D) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Now, the inverse document frequency for a term t is

$$idf(t) = log(N/\sum_{D} o(t, D)), \quad (4)$$

where N is the number of documents in the collection.

In the equation for $o(t, D)$, the value of $tf(t, D)$ could also be thresholded by using a value greater than zero to eliminate the effect of terms with low estimated frequency. That resembles the method used by Siegler [17] for estimating term presence by thresholding estimated probability of occurrence. In our case, however, thresholding did not improve the results. This may be due to pruning at the confusion network calculation where the terms with very low probability are already removed.

3. EXPERIMENTS

3.1 Experimental setup

The corpus consisted of 288 spoken news stories in Finnish read by single female speaker [2]. Each story was about one minute long. The manual reference transcripts of the documents were also available. Each story belonged to exactly one of 17 different topics, assigned by multiple independent judges. The topic descriptions were used as queries.

An 'unlimited vocabulary' continuous speech recognizer [5] was used to recognize the speech into morph lattices. Lattices were transformed to confusion networks with the SRI Language Modeling Toolkit [18]. The decoder pruning parameters were varied to analyze the effect of the recognition running time and to obtain confusion networks with different error rates and sizes.

We used speaker independent acoustic models, trained on separate speech data consisting of 26 hours of speech from 207 speakers as in [14].

The language model was trained on a corpus consisting of 30 million words from electronic books, newspaper text and short news stories. Most of the text was similar to the style of the spoken news stories but from a different time period. Before training, the corpus was segmented to morpheme-like units using the unsupervised segmentation algorithm as

explained in Section 2.1. The size of the lexicon was about 26000 morphs.

The confusion networks were indexed using the estimates for term frequency and inverse document frequency from Section 2.3. The 1-best path was also extracted from the confusion networks and indexed as any text using traditional values for tf and idf based on term counts. Similarly, the retrieval experiments were also performed on the error free reference text to analyze the decline in performance due to recognition errors. Before indexing, the reference text was segmented to morphs using the same lexicon as with the language model training corpus.

As the correct relevance information was known, we could use standard IR measures provided by the `trec_eval` program [3]: mean average precision (MAP), precision at 15 documents (P@15), precision at 5 documents (P@5) and precision at R (P@R), where R is the number of relevant documents. We also plotted average interpolated recall-precision curves.

The amount of errors in the 1-best transcripts is measured by word error rate (WER) and term error rate (TER). WER is the total number of errors (substitutions, deletions and insertions) divided by the number of words in the reference text. For an agglutinative language like Finnish where the words are formed by joining together morphemes, the WER tends to be higher than e.g. English. This is because an expression that takes many words in English may be expressed in just one long word in Finnish. An error in one of the morphs in the word results in the whole word to be counted as wrong. In English, on the other hand, if one of the words in the expression is misrecognized, several correct words remain.

For retrieval, a more relevant measure of error is obtained by comparing how much the index terms (morphs in our case) differ. Term error rate is the difference of term frequency histograms between the indexes [7]:

$$TER = \frac{\sum_t |tf_{ref}(t) - tf_{rec}(t)|}{\sum_t tf_{ref}(t)}, \quad (5)$$

where $tf_{ref}(t)$ is the term frequency in the reference text and $tf_{rec}(t)$ is the term frequency in the recognized transcript.

It is also possible to compare the confusion network to the reference transcript and count the *oracle error rate* i.e. the minimum number of error counts of any path through the network. The oracle error rate indicates the upper limit for the improvements that can be obtained from the confusion network.

3.2 Results

The object of the experiments was to examine if morph-based spoken document retrieval could be improved by extracting terms from confusion networks. Retrieval effectiveness between the following indexes were compared: (1) reference text, (2) 1-best recognition result, (3) confusion network with term frequency estimated by CL-method of Equation 1, (4) confusion network with term frequency estimated by rank-method of Equation 2.

It was also examined how the performance changes with confusion networks produced by different decoder parameters to see how much the speed of recognition, the size of the lattices and the resulting 1-best error rates affect the retrieval effectiveness.

Statistical analysis of the results was performed along the lines of [6], using the MATLAB Statistics Toolbox. Performance measures were first transformed with $arcsin\sqrt{x}$ function to make them closer to normal distribution. The Lilliefors test and the Jarque-Bera test were used to test the normality assumption, which always held. Two-way Analysis of Variance (ANOVA) was performed to examine differences between the methods. 5% significance level was used in all cases.

Table 1: Recognition statistics. The lattice, confusion network (CN) and index sizes are given relative to the size of the respective 1-best transcription or index.

measure / setup	1	2	3	4
1-best WER %	47.76	40.89	38.13	37.34
1-best TER %	58.00	47.14	41.89	40.29
oracle TER %	43.12	26.72	16.87	12.50
RT-factor	0.95	2.10	4.86	7.56
lattice size ratio	24.95	60.71	204.41	526.29
CN size ratio	3.62	6.22	12.51	20.63
index size ratio	2.31	3.45	6.14	9.02

Table 1 shows WER, TER, oracle TER and real-time (RT) factors of the different recognizer runs. RT factor indicates the ratio between the time required for recognition and the length of the audio. Also presented are the resulting total sizes of uncompressed morph-lattices, confusion networks and the size of the uncompressed index using the rank-method. The CL-method produced indexes around the same size. The sizes are given relative to the respective 1-best transcription or index sizes. The size of the 1-best index was about 1100 kB for all setups.

As the level of pruning is decreased, the search space expands and the time of recognition increases as indicated by the increase in the RT factor. The resulting 1-best error rates decrease for the first three setups but stays around the same for the third and fourth. The increase in search space can also be seen in the size of the resulting lattice. The sizes of the confusion network and the index also increase but by smaller factors. The bigger lattices and confusion networks offer more potential for expanding the index with competing terms, which can be seen by the decrease in the oracle TER. On the other hand, they also require more computational power and the risk of inserting spurious terms increases.

Retrieval performance statistics for the four recognizer setups are shown in Table 2. The performance of the error free reference index is also presented. It can be seen that both expansion methods improve the performance over the 1-best index in all cases and by all measures. Also, the rank method outperforms the CL-method in all cases and by all measures except two (P@15 for setup 3 and P@R for setups 3 and 4, where the performance is in practice equal). As with the error rates, the performance of setups 3 and 4 are almost equal, with and without the expansion methods. This indicates that we have reached the level, where the pruning no longer limits the performance.

Statistical testing revealed that with all recognition setups the improvements in MAP are significant for the rank method over the 1-best method. With the CL-method, significant improvements were achieved only with the two first recognizer setups with the highest error rates. Similar results hold for P@15, with the exception of setup 4, where improvements over the 1-best case were not significant for

Table 2: Retrieval performance statistics. Statistically significant improvements over the 1-best baseline are highlighted.

setup	measure	1-best	CL	rank
1	MAP	0.716	**0.758**	**0.774**
	P@R	0.657	0.701	**0.713**
	P@5	0.847	0.871	0.871
	P@15	0.620	0.651	**0.659**
2	MAP	0.739	**0.801**	**0.833**
	P@R	0.692	**0.756**	**0.777**
	P@5	0.871	0.918	**0.929**
	P@15	0.627	**0.698**	**0.714**
3	MAP	0.765	0.817	**0.850**
	P@R	0.723	0.781	0.787
	P@5	0.871	0.894	0.941
	P@15	0.643	**0.706**	**0.702**
4	MAP	0.768	0.823	**0.852**
	P@R	0.724	0.797	0.786
	P@5	0.882	0.894	0.929
	P@15	0.651	0.706	0.710
ref. text	MAP	0.864		
	P@R	0.817	N/A	N/A
	P@5	0.929		
	P@15	0.749		

either method. For P@5, significant improvements were achieved only in setup 2 for rank-method over the 1-best. This is not surprising, because the expansion is expected to increase recall rather than precision. Thus, no improvements were expected at lower cut-off levels where the precision is already high.

Similar behavior can be seen in the interpolated averaged recall-precision curves for the different setups in Figures 3-6. For all the setups and at almost all levels of recall, the methods are again ordered from lower precision to higher: 1-best, CL-method, rank-method. The reference index performance is marked with the dashed line. As the pruning levels are decreased, both expansion methods move the performance closer to the reference. For the third and fourth setup, the performance is again almost equal.

At recall levels of 20% and lower, all indexes perform around the same level and their exact ordering is dominated by chance. For higher levels of recall, the increase in performance is bigger as expected. This indicates that the expansion helps retrieve relevant documents that were previously ranked lower and that have query terms in the confusion sets.

4. RELATED WORK

Various subword based methods have been previously used for retrieval. They usually consist of extracting sequences of phonemes from the phoneme recognizer transcripts as in [13]. Our morph-based method is different, however, as it utilizes the variable sized subword units in the language model enabling more accurate recognition of inflectional languages. These morpheme-like units also work well as index terms. More similar to our work, Logan et al. [10] utilize syllable-like units called *particles* and report improvements in retrieval of English broadcast news with high OOV ratio when combined with a word-based system. Like us, they use a data driven algorithm to find the subword units.

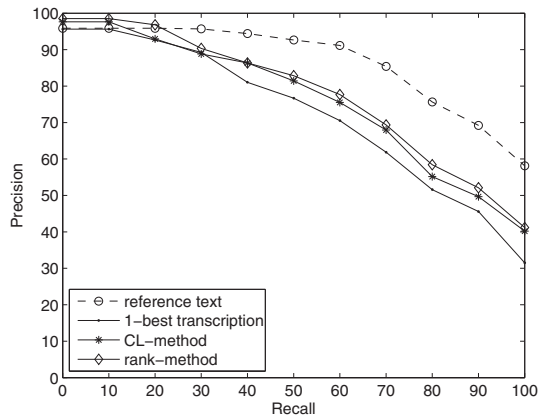

Figure 3: Recall-precision, setup 1, 1-best WER=47.76%

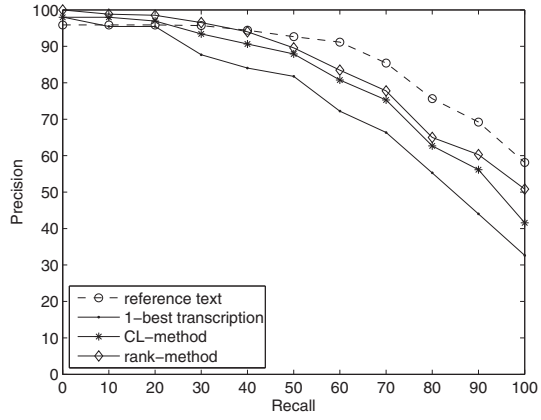

Figure 4: Recall-precision, setup 2, 1-best WER=40.89%

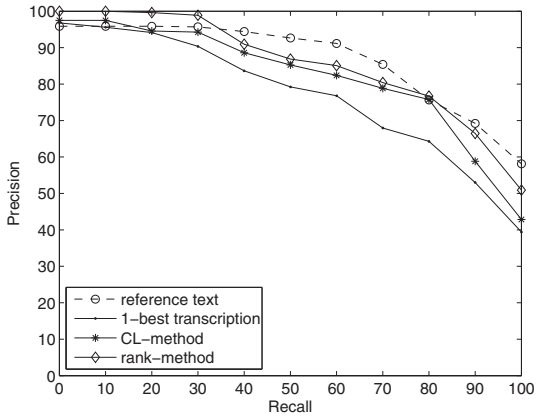

Figure 5: Recall-precision, setup 3, 1-best WER=38.13%

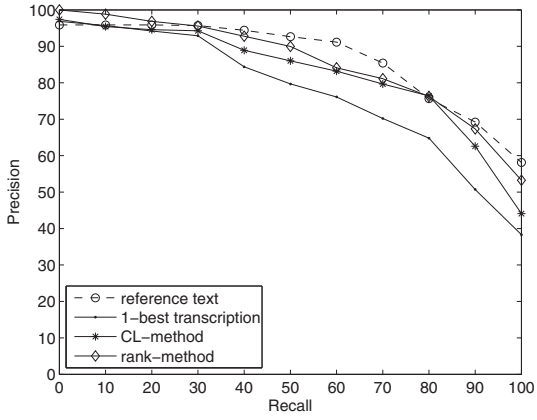

Figure 6: Recall-precision, setup 4, 1-best WER=37.34%

In our previous work [9], we presented a method for extracting alternative recognition candidates for Finnish SDR. The method is based on examining the hypothesis stack of the decoder during recognition and picking the most likely terms before they are pruned. The terms are then added to the index, unweighted. In comparison, confusion networks offer a much more flexible framework for term extraction and make possible to estimate proper values for term frequencies.

Lattices have been used for improving performance of word-based retrieval. Siegler [17] investigates methods for extracting relevant information from word lattices and n-best lists. Similarly to the confusion network method, mutually competing terms are located from the lattice and their probabilities and ranks are used for indexing, showing improvements in retrieval performance. Saraclar and Sproat [16] use phoneme and word lattices to improve word spotting accuracy in English speech. Their method uses lattice arc probabilities to derive a confidence measure for the terms in the lattice. However, as the terms in the lattice are not aligned, measures based on ranking of competing terms can not be used.

The most similar to this work is the approach by Mamou et al. [11]. They use information provided by word confusion networks to improve performance of SDR from call-center conversations in English. Compared to our work, the biggest difference is that instead of words, we use morphs for indexing, which makes our approach better suited to retrieval of inflectional languages like Finnish. Also, we had available the human relevance judgments for the speech documents in the corpus where they compared the results against the search results from the reference manual transcripts, which might introduce bias. We also changed the method for estimating the idf as the estimation used in [11] did not work well for our database. Their work also provides a good analysis on the effect of WER on retrieval, showing that confusion networks can improve the performance especially at high error levels. We also produced recognizer transcripts with different error levels. Our analysis is different in nature however, as the recognizer pruning parameters have a direct effect on the size of the lattice and thus limits the best possible improvement that the expansion can offer.

5. CONCLUSIONS

In this work, we have successfully used confusion networks of morpheme-like units to improve performance of Finnish spoken document retrieval. Confusion networks offer a convenient representation of alternative recognition candidates. Both posterior probability and rank of the locally competing terms can be used to weigh the index terms. In our experiments, discarding the probability and using only the rank to estimate the term frequency offered the best results. However, further research is needed to find the optimal way to use the information provided by the confusion networks.

Significant improvements were obtained in mean average precision and precision at 15 documents. Precision at 5 documents was not improved but was not decreased either. This shows that the estimation scheme used helps to retrieve more relevant documents but also the possibly erroneous terms that are added to the index are downweighted enough so that they do not hurt the results.

Experiments were also carried out with different recognition pruning parameters. They showed among other things

that the increase in 1-best WER, which happens when pruning is increased, can be compensated by using the confusion networks at the indexing phase. This helps indexing of large databases where fast recognition speed is essential.

Future work in using confusion networks for morph-based retrieval is still needed. That includes verifying the results with bigger databases and using different languages. Previously, it has been noted that morph-based retrieval works best when combining both morphs and lemmatized words [8]. Thus, extracting word level information from confusion networks for morphological analysis could possibly be used to improve the results. Also, other methods for estimating tf-idf values from posterior probabilities and ranks, as well as using retrieval models other than the vector space should be researched. Further improvements could be obtained by combining the confusion network approach with other methods such as query expansion and latent semantic indexing.

6. ACKNOWLEDGMENTS

We thank Inger Ekman and the Department of Information Studies at the University of Tampere for the SDR evaluation data. We are also grateful to the rest of the speech recognition team for developing the speech recognizer and the morpheme discovery. We also thank ComMIT graduate school in Computational Methods of Information Technology for funding. The work was supported by the Academy of Finland in the project *New adaptive and learning methods in speech recognition*. This publication reflects only the authors' views.

7. REFERENCES

[1] M. Creutz. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition*. Doctoral thesis, Helsinki University of Technology, 2006.

[2] I. Ekman. Suomenkielinen puhehaku (Finnish spoken document retrieval). Master's thesis, University of Tampere, Finland, 2003. (in Finnish).

[3] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. The TREC spoken document retrieval track: A success story. In *Proceedings of the Ninth Text Retrieval Conference (TREC-9)*. National Institute of Standards and Technology NIST, 2000.

[4] T. J. Hazen and I. Bazzi. A comparison and combination of methods for OOV word detection and word confidence scoring. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, Utah, USA, 2001.

[5] 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*, 2006.

[6] D. A. Hull. Using statistical testing in the evaluation of retrieval experiments. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338, New York, NY, USA, 1993. ACM Press.

[7] S. Johnson, P. Jourlin, G. Moore, K. Spärck Jones, and P. C. Woodland. The cambridge university spoken document retrieval system. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing '99*, volume 1, pages 49–52, Phoenix, AZ, 1999.

[8] M. Kurimo and V. Turunen. An evaluation of a spoken document retrieval baseline system in Finnish. In *Proceedings of the International Conference on Spoken Language Processing ICSLP 2004*, Jeju Island, Korea, October 2004.

[9] M. Kurimo and V. Turunen. To recover from speech recognition errors in spoken document retrieval. In *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005–Eurospeech)*, pages 605–608, Lisbon, Portugal, September 2005.

[10] B. Logan, P. Moreno, and O. Deshmukh. Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio. In *Proceedings of HLT-2002 Human Language Technology Conference*, 2002.

[11] J. Mamou, D. Carmel, and R. Hoory. Spoken document retrieval from call-center conversations. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 51–58, New York, NY, USA, 2006. ACM Press.

[12] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech And Language*, 14:373–400, 2000.

[13] K. Ng. *Subword-based Approaches for Spoken Document Retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.

[14] J. Pylkkönen. New pruning criteria for efficient decoding. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 581–584, Lisbon, Portugal, September 2005.

[15] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[16] M. Saraclar and R. Sproat. Lattice-based search for spoken utterance retrieval. In *HTL-NAACL: Main Proceedings*, pages 129–136, Boston, Massachusetts, USA, 2004.

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

[18] A. Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, pages 901–904, 2002.

[19] V. T. Turunen and M. Kurimo. Using latent semantic indexing for morph-based spoken document retrieval. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006–ICSLP)*, pages 341–344, Pittsburgh PA, USA, September 2006.

[20] P. C. Woodland, S. E. Johnson, P. Jourlin, and K. Spärck Jones. Effects of out of vocabulary words in spoken document retrieval. In *SIGIR '00: Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 372–374, 2000.