Reducing the Effect of OOV Query Words by Using Morph-Based Spoken Document Retrieval

Ville T. Turunen

Adaptive Informatics Research Centre, Helsinki University of Technology, Finland
ville.t.turunen@tkk.fi

Abstract

Morph-based spoken document retrieval uses morpheme-like subword units for both language modeling and as index terms. Problems of out-of-vocabulary (OOV) words are avoided as the morph recognizer can recognize any word in speech as a sequence of subwords. The effect of previously unseen query words (i.e. words that are not in the language model training text) is analyzed for Finnish spoken document retrieval. The performance of the morph-based system is compared to a word-based approach. Language models with artificially high OOV query word rates are built and the results show that morph-based retrieval suffers significantly less from the OOV query words than word-based. Extracting alternative recognition candidates from confusion networks further improves the results, especially for morph-based retrieval.

Index Terms: Spoken Document Retrieval, Out-of-Vocabulary, Subword Indexing, Morpheme, Confusion Network

1. Introduction

As more and more audio-visual material is published and stored, there is a need for methods that can be used for finding interesting segments from large amounts of data. The goal of Spoken Document Retrieval (SDR) is to enable users to perform searches on audio material based on speech content.

Typically, SDR systems work in two phases: (i) transforming the speech into textual form with an automatic speech recognizer (ASR) and (ii) building an index based on the text. In the simplest case, the textual form consists of single best ASR output of words that can be indexed with any standard text retrieval tool. This word-based approach suffers from the limited vocabulary of the speech recognizer. Any out-of-vocabulary (OOV) word will be misrecognized and a query with a such word may fail. However, if the ASR accuracy is high enough and if the material is suitable (e.g. relatively long news stories that contain redundant information), this approach will give satisfying performance, as has been seen in the SDR track of TREC [1].

An alternative approach is to transcribe the speech using phones, e.g. in the form of phone lattices. The queries are also transformed to phones and then matched to the recognizer output. This phone-based approach is not limited by any vocabulary but the disadvantages are that the error rates of the phone recognizers are higher and that the search algorithms are more complex. Hybrid methods combine both phone and word representations [2] [3].

The retrieval method used in this work is based on morpheme-like subword units or *morphs* for short. The language model is trained on a corpus that is first segmented to morphs using an unsupervised segmentation method, *Morfessor* [4]. The recognizer now transcribes the speech as a string of morphs with word break positions marked. This approach is especially suited for morphologically rich languages. The advantage compared to word-based methods is two-fold. First, the vocabulary of the recognizer will be much smaller leading to more efficient recognition. Second, the morph model can recognize previously unseen word forms by recognizing them as sequences of shorter familiar fragments. This should help with the problem of OOV query words.

Using only the 1-best ASR result may not be sufficient if the error rates are too high. Speech recognizers can also produce multiple hypotheses that can be used to improve retrieval performance. A compact representation of the alternative recognition candidates is the *confusion network* [5]. Confusion networks have been used for speech retrieval in different ways. Mamou et al. [6] have improved English SDR performance by extracting and weighing index terms from word confusion networks. Turunen and Kurimo [7] have adapted the method for Finnish SDR by using confusion networks of morphs. Hori et al. [8] use word-phone combined confusion network for open-vocabulary spoken utterance retrieval in English.

In this paper, the effect of unseen query words (i.e. words that are not in the language model training text) is analyzed for Finnish spoken document retrieval. The performance of the morph-based system is compared to a word-based approach. Language models with artificially high OOV query word rates are built and the results show that morph-based retrieval suffers significantly less from the OOV query words than word-based. The results are further improved by extracting alternative recognition candidates from confusion networks.

The retrieval system used here is an updated version of the system in [7]. The key contribution of this work is the analysis of the effect of previously unseen query words for morph and word-based retrieval. Recognition of unseen words with the morph-based system has been previously studied in [9], but since the methods and goals in recognition and retrieval are different, further evaluations are warranted.

2. Morph-based SDR

Speech retrieval in a morphologically complex language such as Finnish, Estonian or Turkish poses challenges that do not exist in other, morphologically simpler languages, such as English. Words in these languages are formed by joining together morphemes. Thus, the number of possible word forms is large, which affects both recognition and retrieval part of the process.

2.1. Speech Recognition

N-gram language models are the standard in speech recognition. For morphologically simple languages, the models can be built using surface forms of the words. The language model is

Accepted after peer review of full paper

estimated by observing statistics for sequences of words. The vocabulary of the recognizer will consist of all the words in the training corpus or, if the vocabulary size is restricted, only the most frequent words.

For morphologically richer languages with a large number of distinct word forms, this approach becomes impractical. A vocabulary of all the words in the training set would be large but would still leave high proportion of OOV words in any independent test text. Data sparsity also becomes a problem as covering enough instances for all the different word forms would require a huge corpus.

One solution is to use suitable subword units in the language model. If the units are chosen well, a relatively small set of subwords will be able to cover the training corpus, reducing the OOV rate as well as data sparsity. Morfessor is an unsupervised, data-driven algorithm for the segmentation of words into morpheme-like units. The algorithm takes as an input a text corpus and finds an optimal set of subwords (*morph lexicon*) that is compact but can also be used to represent the training corpus in a compact way. The subword units discovered by the algorithm are called *statistical morphs* as they resemble linguistic morphemes and are found in a statistical manner. As the algorithm is completely data driven, it is easily applicable to new languages. Detailed description of Morfessor is given in [4].

Training the language model now consists of splitting the words in the training corpus with Morfessor, inserting special markers at word boundary positions and collecting statistics for morph n-grams. The word boundary markers mean that the language model will learn to model word boundaries explicitly, making possible to obtain word sequences as output.

Morph-based language modeling can provide help with the problem of OOV query words since it makes possible to recognize any word in speech, even word forms that were not included in the language model training corpus, by recognizing them as a sequence of familiar morphs. Even recognizing only part of the morphs in a word correctly may help to retrieve a relevant document. An analysis of morph-based speech recognition for Finnish, Estonian, Turkish and Egyptian Colloquial Arabic is provided by Creutz et al. [9].

2.2. Indexing and retrieval

The morph-based recognizer transcribes the speech as a sequence of morphs, with markers at word boundary positions. It is then possible to transform the output to a string of words and use it as as a basis for indexing by using any text retrieval tool. Traditionally, for inflectional languages like Finnish, a *morphological analyzer* is used to transform all inflected word forms in the text to their base form or *lemma*. There are some problems, however, e.g. the analyzers typically have a fixed lexicon and can not process a word that is not in it. This resembles the problem of OOV words in the recognizer.

In the morph-based approach, there is an alternative. As the statistical morphs resemble linguistic morphemes, they are also a suitable candidate to be used as index terms as such. Thus, the need for a morphological analyzer is avoided. Queries are segmented to morphs as well by using Morfessor and the same morph lexicon that was discovered when segmenting the language model training set. Morfessor can also segment previously unseen words, i.e. words that were not in the training set, by using the Viterbi algorithm to find the most likely component morphs. In previous work for Finnish SDR, this approach has provided results that are equal to using lemmatized words [10].

2.3. Confusion networks

Retrieval performance is decreased if a relevant term is not recognized correctly and is thus missing from the 1-best transcript. However, it is possible that the correct term was considered by the recognizer but was not the top choice. Thus, adding alternative results to the index is expected to increase recall. Including alternative candidates is especially important when we want to retrieve previously unseen words, because the probability of the top candidate morph being correct is lower. A compact representation of alternative hypotheses is the confusion network (CN) proposed by Mangu et al. [5]. A confusion network consists of a number of alignment positions, and for each position, a set of mutually exclusive terms ranked by their probability.

The indexing method for CNs is the same that was found successful in [7]. For each term in the confusion network of the document, the *term frequency* (tf) of the term is estimated. The weight for each occurrence of the term is given by $1/rank$. Thus, the most likely candidate gets the weight of 1 and the competing candidates are given less and less weight as their rank increases. The tf of a term is given by summing the weights of all occurrences of the term in the CN. Consult [7] for details.

3. Experiments

The goal of the experiments was to compare morph-based and word-based methods for Finnish spoken document retrieval in the presence of previously unseen query words.

The retrieval corpus consisted of 288 spoken news stories in Finnish read by single female speaker [11]. Each story belonged to exactly one of 17 different topics. The topic descriptions were used as queries. The spoken documents were recognized using a large vocabulary continuous speech recognition system (LVCSR) developed at Helsinki University of Technology [12]. The lattices produced by the recognizer were transformed to confusion networks with the SRI LM Toolkit [13].

For acoustic modeling, conventional Hidden Markov Models (HMMs) with Gaussian mixtures were used. Speaker independent triphone models were trained using 19.5 hours of speech from 310 speakers from the Finnish SPEECON database [14]. The 39-dimensional feature vectors consisted of 12 standard MFCC, log power and their delta and delta-delta derivatives. Cepstral mean subtraction and maximum likelihood linear transformation were applied.

The language model training text consisted of a collection of newspaper articles, books and newswire stories [15]. To compare the performance of word-based and morph-based approach, both word n-gram and morph n-gram models were trained. Also, to test the effect of unseen query words, two different versions of the corpus were used. The other corpus (called `all`) contained all the available text, while some sentences were dropped from the other (called `drop`). A few descriptive words (e.g. names of people, places, companies etc.) were selected from each query and any sentence with any of these words in any inflected form were excluded from the `drop`-corpus. Thus, in total, four different language models were trained and used for testing.

Before training the morph language models, each corpus was first segmented to morphs as explained in Section 2.1. Naturally, for the word models, this part was skipped. Instead, to limit the vocabulary size, the least frequent words were replaced with a special OOV symbol. Table 1 shows the corpus and lexicon sizes for each setup and, for the word models, the propor-

Table 1: *Language model statistics*

	all	drop
LM training set [Mwords]	158	156
Morph lexicon [kmorphs]	19	19
Word lexicon [kwords]	487	479
OOV LM training set [%]	4.7	4.7
New words in queries [%]	1.1	8.0

Table 2: *Recognizer performance statistics.*

	morph all	morph drop	word all	word drop
WER [%]	26.01	29.62	28.63	34.74
LER [%]	7.50	8.50	8.30	10.01
TER [%]	26.90	32.47	43.81	51.53
RT-factor	1.23	1.27	2.17	2.32

tion of OOV words in each LM training set. Bottom row indicates the proportion of previously unseen words in the queries.

The optimal order of the n-gram model depends on whether morphs or words are used. In this work, all models were trained using VariKN-toolkit [16] that can optimize the performance in a way that is neutral with respect to morphs vs. words. The method does not use any fixed maximum order n but instead grows the model such that those n-grams that maximize the training set likelihood are gradually added to the model. A complexity term is used to restrict the growth of the model.

The classic vector space model with cosine similarity measure was used to rank the speech documents. The confusion networks were indexed using term-frequency estimation method described in Section 2.3. The 1-best transcripts were indexed using traditional tf-idf weights. Before indexing the word-based results, a commercial morphological analyzer [17] was used to lemmatize each word. Same procedure was applied to the queries. For the morph transcripts, the analyzer was not used, but the morphs were used as index terms and the queries were segmented to morphs as well.

4. Results

The error rates of the four different recognizer setups are summarized in Table 2. Words in Finnish are relatively long and thus word error rate (WER) tends to be high and is not the best indicator of recognition performance. Therefore, letter error rate (LER) is also used. The real time factor (RT-factor) indicates the total processing time required, proportional to the length of the audio. This includes recognizing the speech, forming the confusion networks and indexing.

For retrieval, a better indicator of recognition performance is the term error rate (TER). TER compares how much the index terms between the recognized transcripts and the reference text (either segmented or lemmatized) differ. TER is calculated as the difference of term frequency histograms.

Retrieval performance statistics are shown in Table 3. Standard measures of mean average precision (MAP) and precision at 5 and 15 documents (P@5, P@15) are used. The first column indicates which of the four language models was used and the second column indicates whether 1-best transcription or confusion network (CN) was used for indexing. The average interpolated recall-precision curves in Figure 1 show that the order of the indexes is the same for almost all levels of recall.

Statistical analysis was performed by pairwise comparison

Table 3: *Retrieval performance statistics*

LM	index	MAP	P@5	P@15
morph all	1-best	0.844	0.918	0.694
	CN	0.869	0.941	0.722
morph drop	1-best	0.642	0.729	0.533
	CN	0.706	0.776	0.600
word all	1-best	0.779	0.871	0.675
	CN	0.800	0.894	0.671
word drop	1-best	0.480	0.612	0.451
	CN	0.498	0.624	0.471

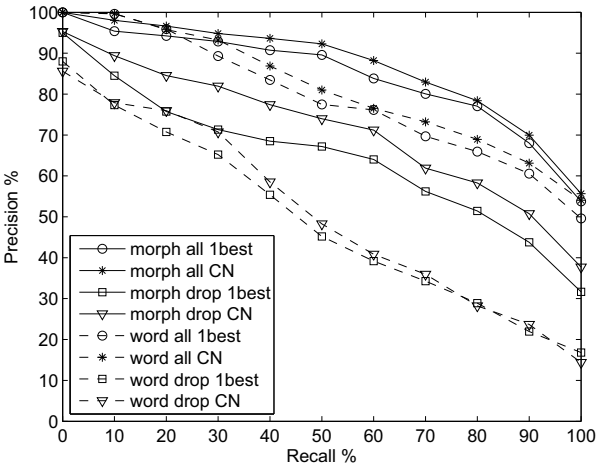

Figure 1: *Recall vs. precision*

of MAP with the t-test using MATLAB statistics toolbox. 95% confidence interval was used. The results show that in all cases, there are statistically significant improvements when using the morph index compared to the corresponding word index. Also, all CN indexes perform significantly better than the corresponding 1-best index.

5. Discussion

The recognition statistics show that the morph-based recognition is more efficient as it provides smaller error rates, but faster recognition speeds. As expected, the error rates rise for both setups when using the drop LMs but the increase for the word-based system is higher.

In retrieval, the morph-based system suffers significantly less from previously unseen query words. In the 1-best case, the word-based system has 38% relative decrease in MAP while the morph-based system only suffers 24%. Using confusion networks helps in all setups, but seems especially useful for morph-based retrieval in the presence of unseen query words. If we consider the word-based 1-best system as a baseline, switching to the morph-based system that uses confusion networks provides 47% relative increase in MAP.

The top of Table 4 shows some examples of how previously unseen query words are recognized. Some words are recognized without errors and others have a few morphs in common with the segmented queries, which will help in retrieval. In some cases, completely wrong morphs are substituted. However, it is still possible that the correct morph is in the confusion network as can be seen at the bottom of Table 4. Same instances recog-

Table 4: *At top, example recognition results for previously unseen query words. The second column shows Morfessor segmentation of the word i.e. the terms used in retrieval. In the third column, recognition results for the word at two different locations. "/" symbol separates different instances of the word. At bottom, a simplified segment of the CN for the spoken word "Iliescun" . Each alignment corresponds to an interval in time and for each alignment, competing hypotheses are given ordered by their probability. Each term is associated with an indexing weight. *DEL* denotes empty hypothesis and <w> word break. The morph "escu" is present both in the query and in the CN, but not in the 1-best result "ilja kun".*

Unseen query word	Morfessor segmentation	Morph recognition examples
Persianlahden	per si an lahden	per sia lahden / per si an lahden
Iliescun	ili **escu** n	ili a s kun / ilja kun
Namibian	na mi bi an	ami bi an / na min pi an
Jugoslavian	ju go slav i an	ju go slav ia n / juhla via n
konkurssi	kon kurssi	on kurssi / kon kur si

align 286	ilja 1	jees 0.5	lies 0.33	iris 0.25	jos 0.2
align 287	*DEL* 1	isku 0.5	kunno 0.33	s 0.25	jees 0.2
	jos 0.17	jes 0.14	**escu** 0.13		
align 288	*DEL* 1	<w> 0.5			
align 289	kun 1	n 0.5	un 0.33		

nized using the word-based system usually produced words that are similarly sounding but with different meanings.

One drawback of the word-based retrieval system used here is that "near misses" are not rewarded like in the morph-based. E.g. the unseen query word "Jugoslavian" (*Yugoslavia's*) was often recognized as two words "jugo slavian". This will not help, however, as the query word is not segmented in the word-based retrieval system.

The advantage of the morph-based system compared to other OOV capable retrieval techniques is the simplicity. There is no need for separate OOV query matching steps like phone-lattice searches. Instead, many of the existing text retrieval techniques can be used by simply using morphs as index terms. However, the disadvantage of this simple indexing is that not always are the morphs optimal as index terms. Especially the unseen words are usually recognized as small pieces that are very unlike linguistic morphemes. Also, Morfessor may segment the query word using different morphs.

One direction of future work is developing retrieval techniques that best utilize the information in the CNs, e.g. by taking in account the proximity of the morphs. Other future work includes building a new, bigger and more realistic database to verify the results and testing the methods on other languages.

6. Conclusions

The effect of previously unseen query words was analyzed for morph and word-based spoken document retrieval. In morph-based retrieval, the speech is transcribed as morpheme-like subword units and these units are used as index terms. Any word in speech can potentially be recognized as a sequence of familiar morphs thus avoiding the problem of OOV words. The results for Finnish SDR show improvements in performance compared to a word-based approach, especially if the proportion of unseen query words is high.

7. Acknowledgements

The author is grateful to his instructor Mikko Kurimo and to the rest of the AIRC Speech group for their helpful comments and for their effort in developing the recognition system. For financial support, ComMIT Graduate School in Computational Methods of Information Technology, Finnish Foundation for Technology Promotion (TES) and Emil Aaltonen Foundation are thanked.

8. References

[1] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proc. TREC-9*. National Institute of Standards and Technology NIST, 2000.

[2] T. J. Hazen and I. Bazzi, "A comparison and combination of methods for OOV word detection and word confidence scoring," in *Proc. ICASSP*, Salt Lake City, Utah, USA, 2001.

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

[4] 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.

[5] 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*, vol. 14, pp. 373–400, 2000.

[6] J. Mamou, D. Carmel, and R. Hoory, "Spoken document retrieval from call-center conversations," in *Proc. SIGIR '06*. New York, NY, USA: ACM Press, 2006, pp. 51–58.

[7] V. T. Turunen and M. Kurimo, "Indexing confusion networks for morph-based spoken document retrieval," in *Proc. SIGIR '07*. New York, NY, USA: ACM, 2007, pp. 631–638.

[8] T. Hori, I. L. Hetherington, T. J. Hazen, and J. R. Glass, "Open-vocabulary spoken utterance retrieval using confusion networks," in *Proc. ICASSP*, Honolulu, Hawaii, April 2007.

[9] M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, and A. Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages," *ACM Trans. Speech Lang. Process.*, vol. 5, no. 1, pp. 1–29, 2007.

[10] M. Kurimo and V. Turunen, "An evaluation of a spoken document retrieval baseline system in Finnish," in *Proc. Interspeech*, Jeju Island, Korea, October 2004.

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

[12] 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.

[13] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.

[14] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "SPEECON - speech databases for consumer devices: Database specification and validation," in *Proc. LREC*, 2002, pp. 329–333.

[15] CSC Tieteellinen laskenta Oy, "Finnish language text bank," http://www.csc.fi/kielipankki/.

[16] V. Siivola and B. Pellom, "Growing an n-gram model," in *Proc. Interspeech*, Lisbon, Portugal, September 2005, pp. 183–188.

[17] Lingsoft Oy, "FINTWOL," http://www.lingsoft.fi/.