Using Latent Semantic Indexing for Morph-based Spoken Document Retrieval

Ville T. Turunen, Mikko Kurimo

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

Abstract

Previously, phone-based and word-based approaches have been used for spoken document retrieval. The former suffers from high error rates and the latter from limited vocabulary of the recognizer. Our method relies on unlimited vocabulary continuous speech recognizer that uses morpheme-like units discovered in an unsupervised manner. The morpheme-like units, or "morphs" for short, have been successfully used also as index terms. One problem using morphs as index terms is that the segmentation does not always separate the same stem for different inflected forms of the same word. This resembles the problem of synonyms. In this paper, we apply latent semantic indexing to morph based retrieval. The idea is to project morphs that correspond to the same word, as well as other semantically related terms, to the same dimension. The results show clear improvements in Finnish spoken document retrieval performance.

Index Terms: spoken document retrieval, latent semantic indexing, morpheme segmentation.

1. Introduction

Today, huge amount of information is generated and stored in spoken form. Therefore, spoken document retrieval (SDR), the task of finding relevant segments from recorded audio, is a significant problem. Two approaches have been commonly used for this task: in *phone-based* retrieval, the speech is transcribed at the phone level while *word-based* retrieval uses a large vocabulary speech recognizer to transcribe the speech at the word level. Another approach we presented in [1] uses a subword based method that transcribes the speech into morpheme-like units that can also be used as index terms.

Phone-based retrieval methods have been studied for example by Wechsler et al. [2]. Phoneme recognizers produce phone-level transcriptions that do not have markers for word boundaries and additional processing is required to match queries (typically word-level) to the phoneme strings. The index can be built using phone n-grams, and the queries are transformed to phonemes by using a pronunciation dictionary. Phoneme recognizers do not use any lexicon or language model, which means that they are not limited by any closed vocabulary. However, the lack of a language model also means that the error rates of phoneme recognizers are very high. Naturally, high error rates also hurt the retrieval performance.

Word-based retrieval methods have recently been the most popular and the most successful approach, particularly in the SDR-track of the Text Retrieval Conference (TREC) [3] and with others, for example [4]. The idea is to use a large vocabulary speech recognizer to transcribe the speech into words. The word level transcription can then be indexed by an information retrieval system the same way as any text document. Because a language model is used, the error rates of word-based recognizers are typically much lower than with phoneme recognizers. Additionally, retrieval is more robust with words than with phone n-grams. However, the word-based approach suffers from the limited vocabulary of the recognizer. Any word in the speech that is not in the vocabulary will always be misrecognized. Typically, only the most frequent words in the language model training corpus are selected to the vocabulary. But from retrieval point of view, the less frequent words, such as proper names, are usually the most interesting.

TREC and a lot of the other research on spoken document retrieval has focused on retrieval of English speech. Other languages have properties that make methods developed for English less useful for them. These properties affect both the recognition and retrieval phase of the process. One important property is the level of agglutination. In agglutinative languages, such as Finnish and Turkish, words are formed by joining morphemes together. Thus, the number of distinct word forms in such language is very high. For speech recognition this means that the language model can not include all the word forms without the size of the model getting too large for efficient recognition. Also the training corpus would have to be huge to cover sufficient number of instances of each word form.

In [1], we presented a base-line spoken document retrieval system for Finnish. It relies on our unlimited vocabulary speech recognition system that utilizes statistical n-gram language models based on morpheme-like subword units discovered in an unsupervised manner from large text corpora [5, 6]. The morpheme-like units used are called *morphs* for short and thus we call this approach *morph-based SDR*. This approach makes also possible to recognize previously unseen words by recognizing their component morphs. The recognizer transcribes the speech as a string of morphs, with the word boundary positions marked. Thus, both word-level and morph-level information can be used for indexing.

The high number of word forms affect also the retrieval phase as typically the user is interested in finding the documents that contain a query term in any of its inflected forms. A natural solution is to return the inflected forms in the documents to their base form before indexing. This can be done with a *morphological analyzer* [7]. The problem here is that building a morphological analyzer needs expert knowledge of the language and not all languages have one available. Also, the analyzer works on a limited vocabulary which would have to be updated from time to time for optimal performance. An alternative to using base forms is to use the morpheme-like units from the recognizer as index terms as such. This method resembles stemming in that, typically, affix morphs are separated from the word stem morphs. In [1], we found that a morph index works equally good or better as a base

form index for Finnish SDR using classical vector space model for retrieval.

A clear drawback of the morph index is that for many words in the Finnish language, there is no optimal point where to draw the line between the stem and an affix, because there are often changes in the stem when an affix is added. Thus, the inflected forms cannot always be split to morphemes in a way that would produce the same stem for all forms and that no other word would produce that stem. The problem is similar to *overstemming* and *understemming*. In practice, understemming is more common as the statistical nature of the morpheme segmentation algorithm makes that frequent word forms produce longer morphs.

The problem of different morphs that all correspond to the same base form resembles the problem with *synonyms*. Latent semantic indexing (LSI) [8] is a method that tries to find underlying latent semantic structure in the data by dimensionality reduction. It has been shown to help with the problem of synonyms by projecting words with similar meaning to the same dimension.

In this paper, we apply LSI for morph-based Finnish spoken document retrieval. The idea is to reduce the effect of mismatches between morphs in the query and morphs in the recognizer transcripts caused by varying segmentation of different inflected forms of the same word, as well as finding other semantically related morphs. For comparison, the analysis is also performed on the base formed index.

2. Latent Semantic Indexing

Latent semantic indexing is a statistical method for information retrieval that has been designed to help with the problem of synonyms (different words that have the same meaning) and polysems (words that have more than one distinct meaning) [8]. It uses singular value decomposition (SVD) to reduce the dimensionality of the term-document association matrix X. The columns of X are the document vectors and each element of a document vector correspond to an index term. The reduced dimensions are hoped to match better the "true", latent, meaning of the words by filtering any random noise.

Singular value decomposition of X is defined as:

$$\underset{t\times d}{X} = \underset{t\times m}{U}\ \underset{m\times m}{\Sigma}\ \underset{m\times d}{V^T}, \tag{1}$$

where d is the number of documents, t is the number of index terms and m is the rank of X ($m \leq \min(d,t)$). U and V are matrices of *left* and *right singular vectors* and have orthonormal columns. Σ is a diagonal matrix of *singular values* that are ordered by value.

Dimensionality reduction is achieved by taking only the k largest singular values and the corresponding singular vectors:

$$\underset{t\times d}{X_k} = \underset{t\times k}{U_k}\ \underset{k\times k}{\Sigma_k}\ \underset{k\times d}{V_k^T} \tag{2}$$

The reduced model X_k is the best rank-k approximation of X in the least squares sense. To match queries to documents, the query vector Q has to be projected to the same, k-dimensional, space:

$$\underset{1\times k}{\hat{Q}} = \underset{1\times t}{Q^T}\ \underset{t\times k}{U_k}\ \underset{k\times k}{\Sigma_k^{-1}} \tag{3}$$

The documents can now be ranked by measuring the similarity between $\hat{Q}$ and the reduced document vectors (from the rows of V_k) by some similarity measure, such as the cosine measure.

In latent semantic analysis, co-occurrence is considered evidence for semantic relatedness. Semantically related terms are then projected to the same dimension. In morph-based retrieval, morpheme segmentation of different inflected forms can cause differing stems for each form. A word that is relevant to a document often occurs in the document several times and often in different inflected forms. Thus, LSI has the potential to improve results by projecting all the stems that correspond to one base form to the same dimension, as well as other semantically related terms.

3. Speech transcription

The spoken documents were transcribed to text using our unlimited vocabulary speech recognition system that uses language models based on morpheme-like subword units found in an unsupervised manner. Below, we describe only some features relevant to the experiments. For a detailed description of the system, we refer to [5].

3.1. Data

The evaluations were performed using 270 spoken news stories in Finnish. On average, each story was about one minute long. The stories were read by a single female speaker in a studio environment [9]. Each news story was assigned to one of 17 topics by multiple independent judges [10]. The topic descriptions were used as test queries. Reference text (the error-free transcription of the speech) was also available and used as a comparison to see how much recognition errors decline the performance.

3.2. Acoustic and language modeling

Two different sets of acoustic models were used to produce transcripts with different error rates. Both sets used conventional Hidden Markov Models (HMMs) with Gaussian mixtures and mel-cepstral coefficient features along with the total energy and the delta values. Context dependent triphone models were trained for 25 Finnish phonemes and 16 of their long variants.

The first set of models were trained for just the speaker of the stories (*speaker dependent* models). For training, the news stories were randomly split to two separate sets. Models trained on one set were then used to recognize the documents in the other set and vice versa. This way, all the available speech documents could be used for evaluations.

The other set of acoustic models were the same as used in [11], that is, *speaker independent* models trained on 26 hours of speech from 207 speakers from the Finnish SPEECON database. The training data did not include any speech from the reader of the news stories.

For a highly inflectional language like Finnish, sufficiently low out-of-vocabulary (OOV) rate cannot be achieved by the traditional method of using the word forms in text as units in the language model. Even with a model with a lexicon of 500M word forms, an acceptable OOV rate was not achieved [5]. We have applied the unlimited vocabulary language modeling approach, where the language model training corpus is segmented to morpheme-like units (called morphs) using an unsupervised segmentation algorithm [6]. The language model can also learn to mark word boundary positions by introducing an additional unit to the model. For language model training, a text corpus of 30 million words from electronic books, newspaper text and short news stories was segmented using a lexicon of 26k morphs. This approach is easy to port to other languages, as the completely data-driven unsupervised unit selection algorithm is not dependent on

any given morphological rules.

Table 1 presents the word error rates (WER), the letter error rates (LER) and real-time factors (RT) obtained with the two sets of acoustic models. Not surprisingly, the speaker dependent models trained on matching data performed better than the speaker independent models.

Table 1: *Statistics of recognizer performance.*

	Spk. dependent	Spk. independent
WER (%)	20.9	34.0
LER (%)	4.6	11.2
RT-factor	4.7	9.8

4. Indexing the transcripts

The morph-based recognizer transcribes the speech as a string of morphs with the word boundary positions marked. This means we can use both the word-level and morph-level transcripts as a source for indexing. When indexing words of a inflectional language like Finnish, it is conventional to return the inflected word forms to their base form. We used a commercial rule-based morphological analyzer for the process. Another option is to use the morphs in the transcripts as index terms as such. In previous experiments [1, 12], we have noticed that these approaches achieve about equal performance, thus avoiding the need of the costly morphological analysis process. Best performance, however, was achieved when both indexes were combined.

Naturally, the query sentences have to be processed to match the index terms. In the case of the morph index, this means processing the queries with the morpheme segmentation algorithm, using the same set of morphs that are used in the language model. With the base form index, queries were returned to base form using the morphological analyzer. The reference text was also indexed for comparison, and same processes were performed to it as well.

The index terms were weighted using the traditional TF$\times$IDF (term frequency times inverse document frequency) formula. In this paper, the performance of latent semantic indexing method is compared against the traditional vector space model. In both cases, the similarity of the query against the document is measured using the cosine measure. That is, the documents with the smallest angle between their document vector and the query vector are returned first. The difference is that when using latent semantic indexing, the angle is measured in the reduced, latent, dimensions. Singular value decompositions for the term-document association matrixes were calculated using 150 latent dimensions.

5. Experiments and results

For the experiments, the 270 spoken news stories were transcribed using the two different sets of acoustic models and indexed as described above. We have a total of 12 setups to compare as the combination of three transcripts (speaker dependent, independent and reference text), two processing methods (base forming and morphs), and two similarity measures (LSI and vector space). As the correct relevance information was available, the performance of the indexes can be compared using standard measures such as the recall-precision curve and the average precision.

Figure 1 shows the performance of the morph indexes by interpolated recall-precision curves. The curve is obtained from the

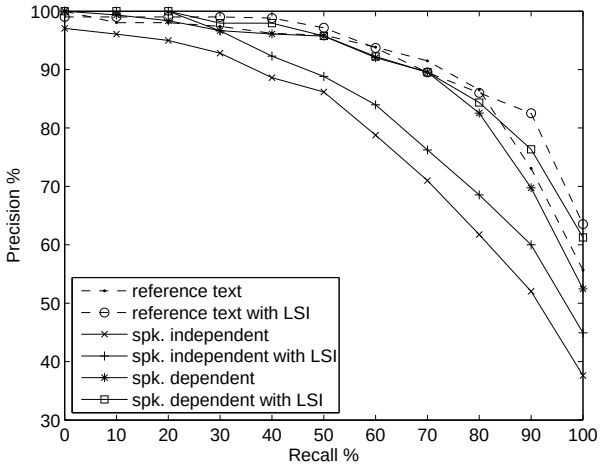

Figure 1: *Recall-precision curves for the morph indexes.*

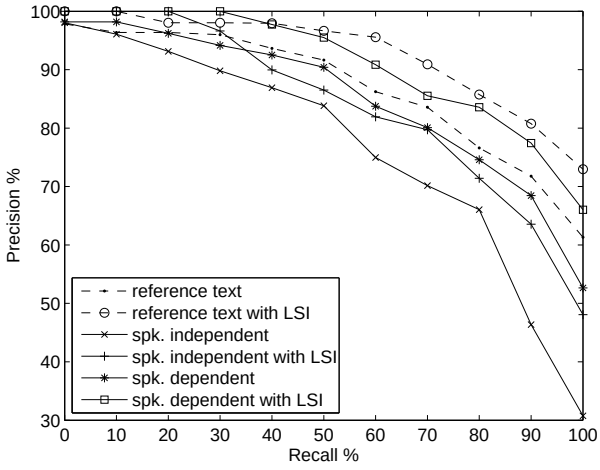

Figure 2: *Recall-precision curves for the base form indexes.*

ranked list of relevant documents by measuring the averaged interpolated precision at each standard level of recall. For both recognizer transcripts, LSI improves precision at all levels of recall. The improvements at high recall values were expected as LSI should help retrieve those documents that use different words for the same concept or have different stems from the same word and would otherwise be left unretrieved. But rather surprisingly, the improvements were just as good also at lower levels of recall actually achieving perfect retrieval for both transcripts until recall of 20%. The improvements are bigger for the speaker independent recognizer transcript, that also contains more errors. Thus, it seems that LSI is also helping to recover from some of those recognition errors.

The recall-precision curves for the base form indexes in Figure 2 show similar behavior. Relative improvements are little bigger than with morph indexes, but only because the performance without LSI was somewhat lower than with morph indexes. With LSI, both type of indexes perform about equally well, which can be seen also from the retrieval statistics in Table 2.

Table 2: *Statistics of retrieval performance. Rp is the R-precision, AP the average precision and P_5 the Top-5 precision.*

Setup		Morphs		Base forms	
		LSI	no LSI	LSI	no LSI
Ref. text	Rp %	85.6	83.6	87.4	80.2
	AP %	91.2	89.2	92.0	85.2
	P_5 %	97.6	94.1	95.2	90.5
Spk. dep.	Rp %	84.3	80.0	86.1	77.9
	AP %	90.3	88.1	90.6	83.6
	P_5 %	95.2	94.1	96.4	90.5
Spk. indep.	Rp %	76.9	69.8	75.2	68.1
	AP %	82.7	77.7	82.9	75.7
	P_5 %	94.1	89.4	91.7	87.0

However, in [12], we obtained even better improvements in SDR performance by using query expansion. In that work, a parallel text corpus was used to extract relevant terms which were then added to the query in order to help retrieve more relevant documents. Although the results in this work are not quite as good, we have shown that LSI can improve results in morph-based retrieval and with further development it should be possible to achieve the performance of query expansion methods. One possible way to improve results is to use a parallel corpus, similar to the one used in query expansion, to extract information on term similarities. The singular value decomposition could be computed on the large text corpus and then the same projection could used to fold-in the spoken documents to the latent dimensions. That way it may be possible to achieve a model that is more robust and can generalize better.

6. Conclusions

In this paper, we have applied latent semantic indexing for morph-based Finnish spoken document retrieval. The system is based on our unlimited vocabulary speech recognizer that uses a language model based on morpheme-like units found in an unsupervised manner. For retrieval, both base formed words or morphs can be used as index terms. Latent semantic indexing was applied to help reduce the mismatch between query and index terms that are caused among other things by differing segmentation of different inflected word forms. The results indicate that retrieval performance can be improved, especially with transcripts with higher error rates.

Future work includes confirming the results on larger databases and applying the methods to other similar languages. Also, future improvements to the method will be studied e.g. by using a parallel text corpus. Improvements could also be achieved by combining LSI with other methods, such as query expansion and extraction of alternative recognition results from the speech recognizer. The semantic analysis itself could be improved by replacing singular value decomposition with another factor analysis method such as *independent component analysis* (ICA).

7. Acknowledgements

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. The work was supported by the Academy of Finland in the project *New adaptive and learning methods in speech recognition*. We also thank ComMIT graduate school in Computational Methods of Information Technology for funding. 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.

8. References

[1] Mikko Kurimo and Ville 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.

[2] Martin Wechsler, Eugen Munteanu, and Peter Schäuble, "New techniques for open-vocabulary spoken document retrieval," in *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998, pp. 20–27, ACM.

[3] John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of the Recherche d'Informations Assiste par Ordinateur: ContentBased Multimedia Information Access Conference*, 2000.

[4] Steve Renals, Dave Abberley, David Kirby, and Tony Robinson, "Indexing and retrieval of broadcast news," *Speech Communication*, vol. 32, pp. 5–20, 2000.

[5] Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to Finnish," *Computer Speech and Language*, 2006.

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

[7] Kimmo Koskenniemi, *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*, Ph.D. thesis, University of Helsinki, Department of General Linguistics, 1983.

[8] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[9] Inger Ekman, "Suomenkielinen puhehaku (Finnish spoken document retrieval)," M.S. thesis, University of Tampere, Finland, 2003, (in Finnish).

[10] Eero Sormunen, *A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases*, Ph.D. thesis, University of Tampere, 2000.

[11] Janne Pylkkönen, "New pruning criteria for efficient decoding," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005)*, Lisboa, Portugal, September 2005, pp. 581–584.

[12] Mikko Kurimo and Ville Turunen, "To recover from speech recognition errors in spoken document retrieval," in *Interspeech 2005*, Lisbon, Portugal, September 2005, pp. 605–608.