Publication VI

Ville T. Turunen and Mikko Kurimo. Speech retrieval from unsegmented Finnish audio using statistical morpheme-like units for segmentation, recognition, and retrieval. *ACM Transactions on Speech and Language Processing*, Vol. 8, No. 1, pp. 1–25, October 2011.

Speech Retrieval from Unsegmented Finnish Audio Using Statistical Morpheme-Like Units for Segmentation, Recognition, and Retrieval

VILLE T. TURUNEN and MIKKO KURIMO, Aalto University School of Science

This article examines the use of statistically discovered morpheme-like units for Spoken Document Retrieval (SDR). The morpheme-like units (*morphs*) are used both for language modeling in speech recognition and as index terms. Traditional word-based methods suffer from out-of-vocabulary words. If a word is not in the recognizer vocabulary, any occurrence of the word in speech will be missing from the transcripts. The problem is especially severe for languages with a high number of distinct word forms such as Finnish. With the morph language model, even previously unseen words can be recognized by identifying its component morphs. Similarly in information retrieval queries, complex word forms, even unseen ones, can be matched to data after segmenting them to morphs. Retrieval performance can be further improved by expanding the transcripts with alternative recognition results from *confusion networks*. In this article, a novel retrieval evaluation corpus consisting of unsegmented Finnish radio programs, 25 queries and corresponding human relevance assessments was constructed. Previous results on using morphs and confusion networks for Finnish SDR are confirmed and extended to the unsegmented case. As previously, using morphs or base forms as index terms yields about equal performance but combination methods, including a new one, are found to work better than either alone. Using alternative morph segmentations of the query words is found to further improve the results. Lexical similarity-based story segmentation was applied and performance using morphs, base forms, and their combinations was compared for the first time.

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

General Terms: Languages, Experimentation, Performance

Additional Key Words and Phrases: Confusion networks, lattices, morphemes, spoken document retrieval, story segmentation, subword indexing, topic segmentation

ACM Reference Format:
Turunen, V. T. and Kurimo, M. 2011. Speech retrieval from unsegmented finnish audio using statistical morpheme-like units for segmentation, recognition, and retrieval. ACM Trans. Speech Lang. Process. 8, 1, Article 1 (October 2011), 25 pages.
DOI = 10.1145/2036916.2036917 http://doi.acm.org/10.1145/2036916.2036917

1. INTRODUCTION

Spoken material is generated in large amounts in the form of radio and TV broadcasts, recordings of academic lectures, meetings, and telephone conversations. More recently, as storage costs have fallen and bandwidth increased, more and more of this type of material is stored in online archives and distributed over the Internet as videos or podcasts. While full text searching for relevant information from textual material is commonplace, searching for spoken material is typically based on human-generated

Author's address (corresponding author): V. T. Turunen, Department of Information and Computer Science, Aalto University School of Science, P.O. Box 15400, FI-00076 Aalto, Finland; email: ville.t.turunen@aalto.fi.

DOI 10.1145/2036916.2036917 http://doi.acm.org/10.1145/2036916.2036917

metadata rather than the actual spoken content. However, not all material is associated with metadata and even when it is, the user's information need might be very different from what the writer of the metadata had in mind.

To search from spoken material, the speech must first be transcribed into text. While human transcriptions would be most accurate, they are rather slow and expensive to produce and thus automatic speech recognition (ASR) is preferred. Typically, a speech search and retrieval system can be divided into three parts: (i) transforming speech to text with an ASR, (ii) indexing the ASR result, and (iii) matching the query terms to the index and returning the most relevant portions of speech.

While speech transcriptions transform the speech search problem essentially into a text search problem, there are some complications. Specifically, recognition errors and the missing structure information of the ASR transcripts degrade retrieval performance. Modern speech recognizers can achieve over 90% word accuracy on suitable speech material, that is, clean and planned speech. Accuracy this high is enough for retrieval performance that is virtually indistinguishable from retrieval from human transcripts [Garofolo et al. 2000]. However, for noisy, conversational, or accented speech, the error rates are much higher.

One source of recognition errors is the limited vocabulary of most ASRs. So called *out-of-vocabulary* (OOV) words, that is, words that are not in the recognizer vocabulary cannot be recognized correctly. Further, OOVs occurring in the speech make the recognizer lose track and also cause nearby words to be recognized incorrectly [Hacioglu et al. 2003]. OOVs are especially problematic because usually the recognizer's vocabulary is limited to only common words, but from a retrieval point-of-view, rare words such as proper names are particularly important as query terms. Phoneme recognizers are not limited by any vocabulary but try to transcribe the speech into a string of phonemes. However, recognition and retrieval performance is degraded by the higher error rates of phoneme recognizers. Other methods of dealing with the OOV problem include query expansion [Woodland et al. 2000] and the use of phone lattices [Saraçlar and Sproat 2004].

Until recent years, most research in the area has focused on English speech. English can be described as not being "morphologically rich", and words in English have a rather low number of distinct word forms. In morphologically rich languages, word forms are formed by extensive use of morphological processes such as inflection, derivation, and compounding and, as a result, have a very high number of possible word forms. For English, word-based methods for ASR and speech retrieval have proven successful. For a morphologically rich language such as Finnish, Estonian, or Turkish, the word-based approach becomes infeasible for ASR as the number of word forms in the vocabulary of the recognizer grows very large and the amount of language model-training data needed to cover enough instances for each form grows huge. For retrieval, there is the problem of matching the word forms in the query to the word forms in the collection.

In this work, a subword-based approach for speech recognition and retrieval is applied. Morpheme-like units (*morphs*) discovered in an unsupervised manner are used both as recognition units and index terms. As a result, the recognition accuracy is high but the vocabulary is small in size and the OOV rate low. The approach is especially suitable for agglutinative languages since the recognizer can construct combined words and inflected word forms by joining together a stem and one or more affix morphs. If the morph vocabulary is well constructed, then potentially any word in speech can be recognized by recognizing its component morphs. Similarly, any query word can be segmented to the most likely morph components. Example results for the morph-based system are given in Table I.

Speech retrieval performance can be improved by expanding the ASR transcripts with alternative recognition results. Even if the word or morph deemed most likely by

Table I. Example Statistical Morpheme Segmentations and Grammatical Base Form Analyses for Query Words Before and After Morph-Based ASR.
Common in-vocabulary (IV) words are usually segmented to similar units both by the statistical and grammatical analyzers and are recognized with high accuracy. Rare IV words are segmented to smaller morphs that do not necessarily resemble grammatical morphemes but may still work as index terms. Even out-of-vocabulary (OOV) words (word forms that are not in the training text at all) can be segmented and recognized at least partly by recognizing the component morphs. Here, the grammatical analyzer could not process the different forms of the names "Lavrov" and "Stubb" as the words were not in its vocabulary

	IV (common)	IV (rare)	OOV
Query word	hallitusneuvottelut	Lavrovin	Stubbiin
(English translation	government negotiations	Lavrov's	to Stubb)
Stat. morphs	hallitus neuvottelu t	la vr ovin	stu b bi in
Gramm. base form	hallitus neuvottelu	- (lavrovin)	- (stubbiin)
ASR (morph)	hallitus neuvottelu t	la vr ovin / la vr ov	stu b in / stu b b
ASR (base form)	hallitus neuvottelu	- (lavrovin) / lavrov	- (stubin) / stubb

the ASR is wrong, the correct choice might be among the considered candidates. Adding alternative candidates to the index should increase recall, but because a maximum of one of the recognition candidates can be correct at each position, appropriate weighting is needed so that precision is not degraded too much. A convenient representation of alternative candidates is the *Confusion Network* (CN) [Mangu et al. 2000]. For OOVs, using the terms in the CNs further improves the chances that the component morphs of the OOV are included in the index.

Unlike text documents, speech material is often not structured into documents of a single topic, but it is rather just a stream of words. The user will want to listen to the audio segments that are relevant to the information needed. The usage scenario consists of the user typing a query into the system and the system returning a ranked list of replay points in the audio that the system considers most relevant. The system is evaluated by comparing the returned points to the points given by human-relevance assessors. To help find the replay points relevant to a query, the audio material is first segmented into stories of a single topic. In this work, a story segmentation method based on the lexical similarity of adjacent windows in the ASR transcript (TextTiling [Hearst 1997]) is used. However, the story segments in themselves are of interest only to the extent that they help in finding the relevant replay points. The motivation is that the user will start playing back the audio at given points but will know when to stop listening.

The task in this work falls into the category of *Spoken Document Retrieval* (SDR) in an unknown boundary condition. It is related to other speech search tasks such as *Spoken Term Detection* (STD) and *Spoken Utterance Retrieval* (SUR). In STD, the goal is to find all locations where the specific word or phrase is spoken. If the words to be searched are known prior to recognition, the task is called *keyword spotting*. In SUR, the goal is to find all short segments (utterances) containing the target utterance in a collection of utterances in spoken communications. The largest difference between the tasks is that SDR is more user-centric than STD or SUR in that the goal is to find and rank all spoken segments that are relevant to a query irrespective of which words and word forms are used to describe the topic.

1.1. Properties of the Finnish language

Finnish is characterized by its highly agglutinative nature. Words are inflected, derived, and compounded by concatenation. Suffixes are used for inflection, and almost all words are inflected according to their roles in the sentence. For example, a single Finnish noun can have about 2000 different forms using a combination of different types of inflection. Suffixes are also used for word derivation. As an example, the word *kirja* (a book) has derivatives *kirjasto* (library), *kirjain* (a letter), *kirjoittaa* (to write) and many others. Finnish also has a high number of compound words. For example, the word *sanakirja* (dictionary) is a compound form of the words *sana* (a word) and *kirja* (a book).

Processing Finnish is complicated by the fact that suffixes can also cause changes in the word root by morphological processes such as *consonant gradation*. Consonants [*p t k*] can shorten, change, or disappear. For example, the genitive is formed by the suffix *n*: *koira*→*koiran* (dog→dog's). But for words *katto* (a roof), *tapa* (a habit), and *tuki* (a support), the genitives are *katon*, *tavan* and *tuen*. Further complications are caused by homography and inflectional homography. The latter means that two different base forms can have common inflectional word forms. Pirkola [2001] and Alkula [2001] study the effect of morphological properties on IR using Finnish as an example.

1.2. Contents and Contributions of the Article

This article is structured as follows. In Section 2, earlier work for speech retrieval and segmentation is reviewed. Section 3 describes the morph-based retrieval system used in this work and the experiments that were performed to evaluate the system. The results are presented in Section 4 and discussed in Section 5.

This work extends previous work in speech retrieval of morphologically rich languages. The main contributions of this paper are the following.

(1) We construct a Finnish speech retrieval collection that is much larger than the ones used before from unsegmented real-life sources.
(2) We present novel combinations of morpheme-like subword units and words for segmenting and retrieval.
(3) We expand the query with alternative morph segmentations of query words.
(4) We use subword confusion networks for unsegmented Finnish data.
(5) We provide an analysis of performance for short and long queries for Finnish SDR.

As far as we know, this is the first time a database of this size comprised of unsegmented audio streams is recognized, segmented, and indexed using morpheme-like units and morpheme-based confusion networks.

2. RELATED WORK

2.1. OOV Robust Methods for Speech Retrieval

The problem of out-of-vocabulary query words is commonly handled by using subword representations of the query and the spoken documents. Typically, the speech is recognized with a phonetic speech recognizer into a string of phonemes. Overlapping phone n-grams are extracted from the phonetic transcripts and matched to the phonetic representation of the queries. Approaches of this kind are compared in the work of Ng [2000]. Additionally, nonoverlapping, variable-length, phonetic sequences (*m-grams*) are used. These m-grams are discovered automatically from phonetic transcripts by applying an unsupervised learning algorithm.

Error rates of the 1-best phonetic transcripts tend to be higher than for word recognizers. To decrease miss rates, phone lattices can be used [Saraçlar and Sproat 2004]. However, for in-vocabulary (IV) queries, phone-based indexing produces a lower precision than word-based indexing. One solution is to use both word-based and phoneme-based approaches and combine the hypotheses [Yu and Seide 2004].

Other solutions include expanding the query with IV words. A parallel text corpus can be used to extract terms that are semantically related to the query words [Woodland et al. 2000]. Similarly, the parallel corpus can be used to expand the speech documents with OOV words. Logan and Thong [2002] take acoustic confusability and language model scores into account and expand query words into IV phrases by trying to mimic the mistakes the speech recognizer might have made. As an example, “Taliban” may be expanded to “tell a band”.

Instead of extracting phoneme strings from phonetic transcriptions, suitable subword units can be incorporated at the recognition phase. Logan et al. [2002] use syllable-like units (*particles*) that are automatically determined from phonetic representations of the words in a corpus. The particles are phoneme sequences that are determined by maximizing the leaving-one-out likelihood of a particle bigram language model. Words in the language-model training corpus are segmented into particles before training. Improvements for English speech retrieval performance were reported when combining the particle-based system with a word-based system.

The system used in this work is also based on automatically determined subword units, morphs. Unlike the methods based on phonetic transcripts, morphs are used for both recognition and retrieval. The good aspects of word and phoneme-based methods are combined, that is, IV words are recognized with very good accuracy and OOVs can be indexed by recognizing their component morphs. Morphs resemble grammatical morphemes (at least for IV words) and are thus a natural choice for index terms as well.

The morph-based approach for SDR was first introduced in Kurimo et al. [2004] where the morph-based recognizer was used and a comparison between using base forms and morphs as index terms showed about equal performance that was also close to the performance of retrieval from the reference text. Query Expansion [Kurimo and Turunen 2005] and Latent Semantic Indexing [Turunen and Kurimo 2006] were shown to increase the performance of the morph-based system in part by countering the effect of potentially suboptimal morph segmentations.

The morph-based system was compared against a word-based system that uses traditional word n-grams for language modeling with about 490,000 most frequent word forms in the lexicon in Turunen [2008]. Similarly, as in the current work, base forms were used as index terms for the word transcripts. Two different versions of the LM training corpus were prepared, one with all the available text and one where the query OOV (Q-OOV) rate was artificially increased by dropping some sentences from the corpus. The results show that the morph-based system performs significantly better than the word-based system and especially so when the Q-OOV rate is high (34% increase in MAP for retrieval from 1-best transcripts).

Studies on text retrieval [Kurimo et al. 2010] further support using morphs as index terms. Segmenting the text corpus with Morfessor and using the resulting morphs as index terms provides results that are slightly worse than using base forms, but not to a statistically significant degree.

2.2. Story Segmentation

Segmenting text and speech material into topically related units has been studied extensively. Typical methods are based on lexical or prosodic features of the data. Lexical segmentation methods rely on the fact that topic segments tend to be lexically cohesive. Segment boundaries are located, for example, based on lexical similarity between words [Kozima 1993] or cosine similarity of adjacent windows in term vector space [Hearst 1997]. Stokes [2004] and Galley et al. [2003] use *lexical chains* to determine segment boundaries. A lexical chain consists of all the repetitions of a term (possibly including synonyms) in the text. High concentration of chain beginning and endpoints is an indicator of a story boundary.

Broadcast material, especially news transcripts, has an inherent structure that can be used to assign story boundaries. Linguistic features such as cue phrases may indicate the presence of a nearby topic shift. Passonneau and Litman [1997] study the correlation of referential noun phrases, cue words, and pauses with segment boundaries. Beeferman et al. [1999] train adaptive language models to obtain a topicality feature and automatically identify cue-word features that correspond with topic shifts in labeled training text. However, on ASR transcripts, recognition errors make learning and identifying such cue words more difficult. If a segmented training corpus is available, Hidden Markov Models (HMMs) [Mulbregt et al. 1998] or aspect Hidden Markov Models [Blei and Moreno 2001] can be used for story segmenting. The segments in the training corpus are clustered, language models are estimated for each cluster, and the HMMs are trained with fixed-length word windows as observations. Topic shifts are identified by the Viterbi algorithm which finds the most likely hidden sequence of topic states.

Besides lexical information, speech carries an additional source of knowledge through its *prosody* (the patterns of stress and intonation). Shriberg et al. [2000] trained a decision tree using automatically extracted prosodic features. Especially long pauses, preboundary lengthening (slowing down toward the ends of segments), low boundary tones, and pitch resets were found to predict shifts in topic. Tür et al. [2001] and Galley et al. [2003] use decision trees and HMMs for combining lexical and prosodic information.

Methods based on the Bayesian Information Criterion (BIC) can be used to detect acoustically homogeneous regions and speaker changes in audio. While speaker change may not indicate a change in topic, the pattern of speakers throughout the audio can correspond to the topic [Renals and Ellis 2003].

Instead of trying to segment the audio before indexing, simple fixed-length sliding windowing is sometimes used [Pecina et al. 2007]. This approach was used as a baseline in this work. Sliding window-based methods will likely produce duplicate entries in the retrieval results that is, multiple windows which originate from the same story. Postretrieval merging can be used to detect and remove these duplicates [Johnson et al. 2001].

2.3. Using Lattices for Speech Retrieval

Alternative recognition hypotheses, usually in the form of lattices, have been used for speech retrieval to deal with high error rates of 1-best transcripts. Typical approaches involve estimating expected word counts and using these counts in some retrieval model. The earliest methods matched phonemic representations of query words to phone lattices for vocabulary-independent keyword spotting [James et al. 1994] or for estimating term counts for SDR [James 1995]. Word lattices and n-best lists were used in Siegler [1999] where the probability or the rank of the candidate was used to find an empirical relationship with the probability of occurrence of the term in the human transcripts. The word probabilities were used as estimates for term counts for *tf.idf*-based retrieval.

More recently, Saraçlar and Sproat [2004] used phone and word lattices for SUR by estimating expected counts of the query word in the lattices and returning those utterances that had the count above a certain threshold. Chia et al. [2010] use similar expected counts in word lattices to estimate probability models for the language modeling-based retrieval method for SDR. Yu et al. [2005] use posterior probabilities in hybrid word/phoneme lattices as estimates of term counts for searching recordings of conversational speech. A two-pass system is used where the costly phonetic match is only performed on a small subset of candidate lattices.

Instead of estimating term counts directly from lattices, the lattice can be first transformed into a simpler form, which leads to a savings in storage space. In a confusion network (CN) [Mangu et al. 2000], competing hypotheses occurring around the same time are aligned and clustered, and each hypothesis is associated with a posterior probability. For SDR, confusion networks have been used to estimate term frequencies based on the sum of posterior probabilities [Mamou et al. 2006] or rank [Turunen and Kurimo 2007] of the competing terms. Hori et al. [2007] combined word and phone confusion networks for open-vocabulary SUR. The query is represented as a sequence automaton with words, phones, or mixed word-phones and matched to the word-phone confusion networks by automata intersection.

In Chelba and Acero [2005] and Chelba et al. [2007], a comparable structure, *Position Specific Posterior Lattice* (PSPL), is used for SDR. A PSPL also has clusters of competing words with their posterior probabilities, but unlike CN, it retains the positions (i.e., the number of words since the start) of the words. Each index entry (so-called *soft hit*) represents the posterior probability that the word occurs at that particular position in the sequence of spoken words. A variation of the standard forward-backward algorithm is used to construct a PSPL: each path through the lattice is enumerated, each word hypothesis on each path is assigned the posterior probability of the entire path, and the posterior probabilities for the same word occurring at the same position are summed and the instances merged. The ranking of the spoken documents can utilize the proximity of the query words in the lattice. The n-grams present in the query are matched to the PSPLs and weighted increasingly with the n-gram order. CNs and PSPLs are compared for SDR in Pan et al. [2007] and for SUR in Kazemian et al. [2008].

Matching phrases from confusion networks is difficult due to the existence of null links. In a PSPL, time information for individual hypotheses is lost. Zhou et al. [2006] proposed a method, *Time-based Merging for Indexing* (TMI), to deal with these problems. TMI also uses a posterior probability representation of the lattice and merges hypotheses with similar time boundaries. Necessary information for phrase matching is kept, while also keeping the time information.

In this work, confusion networks are also used, but, unlike Mamou et al. [2006], the CNs consist of morphs instead of words and, unlike Turunen and Kurimo [2007], the experiments are carried out on a much bigger and more realistic corpus. Previous work in Turunen [2008] compared word-based and morph-based CNs for Finnish SDR and found that the morph-based approach works significantly better especially when the Q-OOV rate is high.

3. METHODS

In this work, methods for selecting index terms, automatic story segmentation, and extracting alternative recognition hypotheses for Finnish speech material are compared. Our focus is on unsupervised methods that do not need labeled training data. A lot of the research in the field focuses on English speech, but different languages have different properties, and thus the same methods are not always optimal.

3.1. Morph-Based Retrieval System

The retrieval system is based on the one used in our previous work, Turunen and Kurimo [2007]. A short motivation and description of the system with the emphasis on the changed parts is given in the following. This work adds the use of n-best query word segmentations, interleaving of morph and base form results, and a morph stop list.

3.1.1. Recognition Phase. Most speech recognizers use n-gram language models that are based on observing statistics for sequences of words from large text corpora. For morphologically complex languages such as Finnish, Turkish, and Estonian, morphological phenomena such as inflection and compounding cause the number of different surface forms to become very large. A vocabulary of this size would be infeasible.

Further, a huge training corpus is needed to cover enough instances of different word forms to reliably estimate n-gram statistics.

Using suitable subword units such as letters, syllables, or morphemes provides a solution to the preceding problems. The size of the vocabulary is reduced as a fewer number of subword units than words is needed to cover the language. From a retrieval point-of-view, morphemes are an attractive choice as language modeling units since they also carry a meaning. A morphological analyzer with expert crafted rules can be used to segment words into morphemes [Koskenniemi 1983]. To avoid the need for expert knowledge, data-driven automatic segmentation algorithms have also been developed.

In this work, an unsupervised data-driven morpheme segmentation algorithm Morfessor Baseline [Creutz and Lagus 2005] is used. Using a text corpus as training data, the algorithm discovers a compact set of segments that can be used to represent the training text efficiently. These units called *statistical morphs* resemble grammatical morphemes [Kurimo et al. 2008].

After the optimal morph lexicon is found, the Viterbi algorithm is used to segment each word in the LM training corpus to the most likely string of morphs. Individual letters are added to the morph lexicon with a small probability, thus ensuring that all words, even previously unseen ones, can be processed. A special word boundary morph is added between words. Any standard LM training tool can be used to estimate morph-based language models from the segmented corpus. However, it is important to use higher order n-grams than the standard 3-grams as a number of morphs span shorter segments of language than the same number of words [Hirsimäki et al. 2009]. In this work, a variable-length Kneser-Ney smoothed n-gram model is trained using a growing and pruning algorithm [Siivola and Pellom 2005].

Finnish has a straightforward almost one-to-one mapping between graphemes and phonemes. The pronunciation of any morph can therefore be easily determined as it does not change in different contexts. For speech transcription in other languages where the morphs have different pronunciation in different contexts, the different variants can be made unique by numbering then [Creutz et al. 2007]. For speech retrieval, it may be sufficient to learn the morph segmentations from phonetic representation of the training text corpus, transcribe the speech and then transform the queries into phonetic morph representations.

Previous experiments on Finnish speech transcription [Hirsimäki et al. 2006] show that statistical morphs clearly outperform syllable and word-based language modeling. Also the experiments show that statistical morphs are better than or equivalent to using grammatical morphemes as language modeling units, depending on the task. In Arısoy et al. [2008], statistical morphs, grammatical morphemes, and words were compared in a speech transcription task for three languages (Finnish, Estonian, and Turkish). For all languages, statistical and grammatical approaches compared had almost equal performance that was at least as good or better than a very large vocabulary word-based language model.

3.1.2. Retrieval Phase. As the ASR output consists of a sequence of morphs with markers of word boundary positions, there are a number of alternatives for how to construct the index. The words can be strung together and a morphological analyzer can be used to normalize the words into their base forms. It is also possible to use morphs as index terms or use different combinations of their base forms and morph indexes.

For ASR output, using morphs is potentially even more beneficial than for error-free text. If an erroneous morph is introduced, the resulting word may be normalized to a base form that is not related at all to the word that was actually said. But when using morphs as index terms, at least some of the morphs in the word will be indexed correctly. For example, the Finnish name `Eero Heinäluoma` was in one instance recognized as morphs `vir heinä <w> luoma`, which would translate roughly as "created as mistakes". The symbol `<w>` marks a word boundary. The output has some common morphs to the actually spoken words but the morphological analyzer normalized this to base forms `virhe` (mistake) and `luoda` (create). In this case, using morphs as index terms preserves more information since the terms `heinä luoma` remain in the index and would match if the name was searched, while the base forms have nothing to do with the actual words.

Morph-based indexing is especially beneficial in the case of OOV words. It is possible to recognize words that did not occur in the language model training corpus by recognizing its component morphs [Hirsimäki et al. 2009]. Further, Morfessor can segment a query word by finding the most probable morphs using the Viterbi algorithm even if the word was not in the Morfessor training corpus. In Turunen [2008], the morph-based system clearly outperformed the word-based system for Finnish SDR especially in cases where the query OOV rate was high.

The morph transcripts also include a high number of suffix morphs that are not useful for retrieval. While these terms will receive low *idf* value, removing them completely from the index is still beneficial. Since the suffixes are not tagged, a simple stop list was constructed by taking the number of most common morphs in the corpus.

Morphs and base forms have different advantages as index terms and so it is natural to combine the two methods. In this article, two different approaches were tested for this purpose. First in the *combined method*, the morph text and base form text of each segment were simply concatenated and treated as one document for indexing (same as in Turunen and Kurimo [2007]). Similarly in the combined method, the morph and base form queries were concatenated. Second in the *interleaved method*, the morph and base form indexes were created and queried separately. For each query, the ranked lists from the morph index and the base form index were interleaved, starting with the index that had the highest similarity for the top-ranking document. Documents were then put into the final ranked lists one-by-one from alternating ranked lists, removing any duplicates.

When using a morph, a base form, or the combined index, queries are naturally processed in a corresponding way. They are either returned to possible base forms with a morphological analyzer, segmented to morphs with Morfessor, or both forms are used. With Morfessor, it is possible that for some inflected forms, the stem morph of one form does not quite match the stem morph of the other form because there might be different letter transformations in the stem for different affixes or the algorithm may just find the most probable boundary at slightly different places for different forms. This problem can be alleviated by introducing n-best segmentations from Morfessor. Since the Morfessor uses Viterbi algorithm to find the most probable segmentation of a word form, it is simple to modify the algorithm to find the n-most-likely segmentations. Adding these alternative segmentations to the query will increase the likelihood of matching the morphs in the index. For example, the word `ilmastomuutoksesta` (from the climate change) may be recognized as morphs `ilmasto muuto ksesta`. If the query word is in form `ilmastomuutoksen` (of the climate change), Morfessor will segment it into morphs `ilmasto muutoksen`. In this case, the subword `muutoksen` has had high enough frequency in the training corpus so that it was not further segmented. However, the second-best segmentation is `ilmasto muuto ksen`. Adding the morph `muuto` to the query will increase the number of matching terms. In initial testing, it was found that adding the new morphs in the second-best segmentation of the words to the query improved the results, but not for third-best or further segmentations.

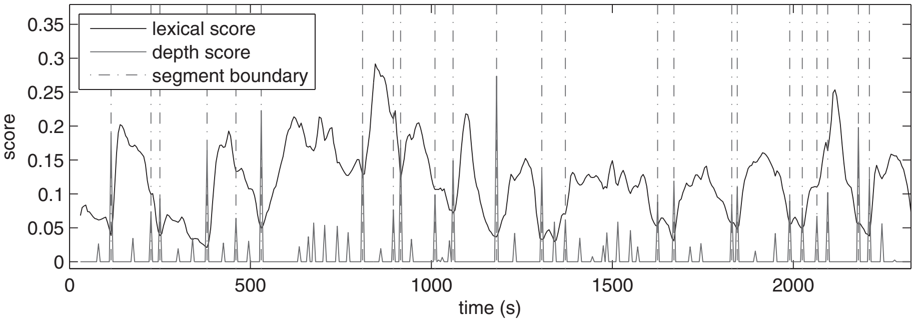

Fig. 1. The lexical scores and depth scores for a program. Vertical lines indicate the proposed segment boundaries.

3.2. Story Segmentation

Retrieval from unsegmented speech material requires finding the boundaries where the topic shifts from one story to another. The ultimate goal is to find the replay points that match the query. Perfect story segmentation of the material will help, but it is not necessary. For automatic story segmentation, it is presumably better to tune the algorithm to oversegment somewhat rather than undersegment because it will increase the chances of finding the actual replay point. Since browsing of audio material is harder than browsing of text it is important to return a replay point that is as close as possible to the actual beginning of the relevant portion.

In this work, the segmentation is performed on the ASR output using an algorithm based on TextTiling [Hearst 1997]. The lexical similarities of the preceding and following windows of speech are compared at short intervals. Story segments are hypothesized at points where there is a steep enough drop in similarity. More precisely, the *lexical score* $ls(g)$ of the adjacent windows is calculated using the cosine similarity (using the notation in Xie et al. [2008]):

$$ls(g) = cos(\mathbf{w_l}, \mathbf{w_r}) = \frac{\sum_{i=1}^{I} w_{i,l} w_{i,r}}{\sqrt{\sum_{i=1}^{I} w_{i,l}^2 \sum_{i=1}^{I} w_{i,r}^2}}, \tag{1}$$

where w_l and w_r are the document vectors of the left and right windows of point g. Each element of the document vectors w corresponds to a term, and the value is given by the frequency of the term in the window times the inverse document frequency of the term in the entire collection.

Segment boundaries are located by examining the valleys and peaks of the lexical score time trajectory. The *depth score*, $ds(v)$ of a valley v is defined as

$$ds(v) = (ls(p_l) - ls(v)) + (ls(p_r) - ls(v)), \tag{2}$$

where p_l is the peak to the left of the valley and p_r to the right of the valley. At non-valley points, the depth score is 0. If the depth score exceeds a threshold θ at some point, the point is considered a segment boundary. θ is determined based on a number of standard deviations σ from the mean μ of the lexical score: $\theta = \mu - \alpha\sigma$. α is a parameter, μ and σ are calculated for each radio program to be segmented. Example lexical and depth score curves are plotted in Figure 1.

The window sizes are typically determined by a fixed number of sentences or paragraphs, but since the ASR transcripts do not contain sentence structure, fixed-time length windows are used. Based on initial testing, 90-second windows at 5-second intervals were used. The lexical scores are smoothed using 3-point average filters before calculating depth scores.

There are some differences between the method used here and the original TextTiling method. First, the goal of TextTiling is to find subtopic structure from long documents with a single main topic. Here the goal is to segment speech material that may cover multiple widely differing topics into segments that offer the best retrieval performance. This will affect the optimal level of segmentation. Further, TextTiling calculates the lexical score between every pair of sentences using windows that have a fixed number of sentences. Because ASR transcripts have no sentence structure, the window size and step size was determined in seconds as it fits to the framework of finding replay points. Finally, the method here uses morphs as tokens, possibly making it more resilient to ASR errors and OOV words. Subwords have also been used for TextTiling for Chinese ASR transcripts by Xie et al. [2008].

The choice of the segmentation method was based on a number of factors. The algorithm was relatively simple and proven efficient. Also we wanted an algorithm that works without training data, which was not available either. This rules out a number of algorithms based on learning acoustic or lexical features from data. Since there was not any manually segmented speech data available, the segmentation performance was not evaluated directly but in terms of retrieval performance which was the desired final goal. By manually looking at some speech documents, however, it seems that the algorithm oversegments moderately.

Speaker change patterns from BIC-like methods could offer additional evidence for story segmentation. However, for this corpus, using speaker changes would most likely hot offer much benefit as usually the story change happens when there is no speaker change nearby as the anchor of the show wraps up the previous topic and then introduces the next one. With proper features and parameters, some improvement might be possible. This problem is left for future research.

When morphs were used as index terms, the segmentation was performed using morphs, and, when base forms were used as index terms, the segmentation was also performed using base forms, etc. As the real segments were not known for the speech data, it was not possible to test the success of the segmentation directly. To separate the effect of segmentation and indexing on the retrieval results, cross testing using the segments from different segmentation types was performed. For instance, morphs were used as index terms but the segmentation was achieved using the base forms.

The effectiveness of the segmentation method for retrieval was compared to using fixed-lenigth sliding windows. Initial testing with different length windows with different overlap showed that 90s windows with 15s overlap provide the best results.

The α-parameter controls the degree of segmentation. Since a manually segmented speech database was not available for optimizing α, a text retrieval database was used instead. The documents were randomly ordered and grouped into sets of about 6200 words which matches the average size of the speech documents. There were an average about 24 documents in each set. The tests were repeated 10 times with different random orderings. To simulate ASR output, all document and sentence structure was removed. The word forms in the text were either split to morphs using Morfessor, returned to base form, or the combined output of both methods was used. The resulting text was then automatically segmented using the method preceding and the segments were indexed. The starting locations of the known relevant documents were used as relevant replay points for the evaluations. The value of α was varied between 0 and 1.2, and the value with the highest resulting average GAP (see Section 3.5) was used for later testing. Note that the segmentation was not optimized with respect to matching the real segments but to the resulting retrieval performance.

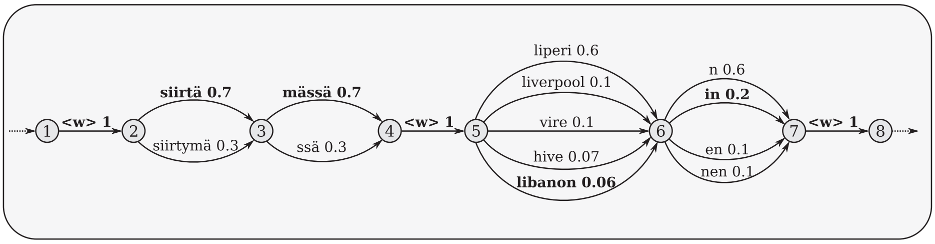

Fig. 2. A confusion network for a segment of speech. <w> marks a word break boundary. The correct morphs are in bold.

3.3. Indexing Confusion Networks

1-best ASR transcriptions will almost always have a number of erroneous words and, if the error rate is high, the errors will also have an effect on retrieval performance. However, it is possible that the correct word is among the hypotheses that the recognizer considers, and retrieval performance can be increased by utilizing these hypotheses. The recognizer outputs the candidates in a form of a word graph or *a lattice*. Lattices tend to have a high level of redundancy and connectivity which makes them unnecessarily complex for many applications. *A confusion network* (CN) is an approximation of a lattice with a strictly linear structure [Mangu et al. 2000]. A confusion network is a series of nonoverlapping *confusion sets* and each confusion set contains a number of alternative word hypotheses with their posterior probabilities. In this work, morphs are used instead of words but otherwise the structure is the same. An example of a morph confusion network is presented in Figure 2.

Confusion networks offer a convenient representation to extract competing recognition results for speech retrieval. To index the CN of a speech document, we need a method to estimate the *term frequency* (*tf*) values for the terms in the CN. In this article, the comparisons of *tf* estimation methods in Turunen and Kurimo [2007] are repeated on a new unsegmented corpus with added comparison to base form CNs.

In the first method, called the Confidence Level or *CL method*, term frequency is estimated by summing the posterior probability values of all the occurrences of the term in the confusion network [Mamou et al. 2006]. In the second method, called the *rank method*, the reciprocal rank of all the occurrences of the term are used in the sum instead of the probability [Turunen and Kurimo 2007]. Here, the highest ranking terms, that is, the 1-best recognition results, are given the full weight of one, and the competing terms are given less and less weight as their rank increases.

The *inverse document frequency* (*idf*) of a term indicates how descriptive the term is. Typically it is a function of the inverse logarithm of the number of documents the term occurs in. For estimating the *idf* for a collection of CNs, we adopt the method used in Mamou et al. [2006].

To rank the CN documents, the traditional cosine distance measure is used with the preceding *tf.idf* values for the document vectors. However, after estimating the *tf*s, many other text retrieval methods would be applicable. In Section 4, the CN methods are compared to using only the 1-best transcription. CL and rank approaches are compared in the 1-best case, also. The “1-best CL” method means weighting each term by its confidence level. The “1-best rank” method reduces to traditional indexing where *tf* is determined by the terms count.

The morph-based recognizer outputs lattices with morphs as tokens with word-break positions marked with a special symbol. As described in Section 3.1.2, the indexing can be based on morphs, base forms, or combinations of the two. To use the base forms in the case of CNs, the morph lattices have to be transformed into word lattices. This is simply

```
<top>
<num> 8 </num>
<FI-title> Libanonin kriisi </FI-title>
<FI-desc> Etsi dokumentteja joiden aiheena on Libanonin levottomuudet, rauhanturvaoperaatio or keskustelu suomalaisten rauhanturvaajien lähettämisestä. </FI-desc>
</top>
```

```
<top>
<num> 8 </num>
<EN-title> The crisis in Lebanon </EN-title>
<EN-desc> Find documents about the restlessness in Lebanon, the peacekeeping operation or the debate about sending Finnish peacekeepers. </EN-desc>
</top>
```

Fig. 3. An example of a topic description in Finnish and its translation into English.

achieved by traversing the arcs in the morph lattice, while keeping the traversed arcs in a history. When the word-break symbol is encountered, the arcs in the current history are replaced by a new arc, with the morphs joined into a word and the probabilities summed. The word lattices are transformed to confusion networks normally, but before indexing, each word is processed with a morphological analyzer [Lingsoft, Inc 2007]. Compound words are split into separate index terms. However, all parts of the word receive the same *tf*. Similarly, if the word form has multiple interpretations for the base form, all interpretations receive the same *tf*, based on the probability or the rank of the original word form in the confusion set. Any word that is not recognized by the analyzer is left as such.

3.4. Materials

3.4.1. Evaluation corpus. Currently, there are not publicly available speech retrieval evaluation corpora for Finnish, thus an effort was made to prepare one. A collection of Finnish radio broadcasts from the Finnish Broadcast Company, distributed as mp3 podcasts,[1] was gathered. 285 programs with a total length of 136 hours, spanning 3 years, were selected for the database. The selected programs included different current affairs programs about one-hour long and shorter newspaper review programs. All programs covered multiple topics and included variable types of material from planned prerecorded reports to live telephone and street interviews. The programs did not contain any music, except for short theme tunes.

The programs were associated with an RSS metadata that briefly described the contents of the program. The metadata was used to extract the programs from the Web to the database and to formulate the topic descriptions. In a real-life application, the metadata would be a good source of information for retrieval as well but, in this work, the goal of the experiments was to study the case where only the ASR transcript was available. No information about story boundaries within the programs was available.

The topic descriptions were coined manually by reading the metadata descriptions, finding topics that were covered in a number of programs and formulating the topics into sentences that were thought to correspond to real users information needs. The format of the topics was similar to the one used in TREC. An example of a topic description is given in Figure 3. The “title” field contains a description of the topic in a couple of words. The “desc” field contains a natural language description of the

[1] http://areena.yle.fi/podcast/uusimmat.

base form T query:	`libanon kriisi`
base form TD query:	`libanon kriisi libanon levottomuus rauhan turva operaatio tai keskustelu suomalainen rauhan turvaaja lähettäminen`
morph T query:	`libanon in (i n) kriisi (kri isi)`
morph TD query:	`libanon in (i n) kriisi (kri isi) libanon in (i n) levottom uudet (uude t) rauhanturva operaatio (rauhan turva) tai keskustelu (keskus telu) suomalaisten (suomalais ten) rauhanturva a jien (rauhan turvaa) lähettä misestä (lähe ttä)`

Fig. 4. The T and TD queries corresponding to the topic in Figure 3, after morphological base form analysis and morph segmentation with 2-best alternatives included in parenthesis. The start of the description ("find documents about") was removed in preprocessing.

information need. In the experiments, only the short title field was used as a query (*T query*) or the both title and description fields (*TD query*) were used (see Figure 4).

Relevance assessments were made by first looking at the metadata descriptions and finding all the programs that possibly contained relevant information to the topic. The starting times of each of the possibly relevant segments were located by listening. The location of the starting time depended on the type of the program. For news reports that covered only one topic, the starting time was at the start of the report. For interviews and debates that covered multiple topics, the start of the segment was put at the point where the topic in question was brought up. One program could contain multiple starting times relevant to the same topic.

Each marked segment was listened to by two judges who decided if it was actually relevant to the topic. At this point, a couple of the topic descriptions were changed slightly to broaden or narrow the topic and the relevance judgements were made again. Finally, a test system was used to search the database and the 50-100 (depending on similarity) highest-ranked replay points for each topic were listened in order to find any missing relevant segments. At the end, there were 25 topics and 451 relevant replay points.

The retrieval corpus is planned to be publicly available. The queries, relevance assessments, identifiers for the documents in the corpus, and the retrieval tools can be obtained by requesting them from the authors. Negotiations with the copyright holders are underway for releasing the speech data itself, its ASR transcriptions, and lattices for research use.

3.4.2. Training corpora. Speaker-independent acoustic models were trained on a separate corpus with 19.5 hours of speech from 310 speakers from the Finnish SPEECON database [Iskra et al. 2002]. Speech recognition was performed using a large vocabulary continuous speech recognizer (LVCSR) developed at the Helsinki University of Technology [Hirsimäki et al. 2009]. The recognizer uses Hidden Markov Models of decision tree-tied-triphone models using Gaussian mixtures with a gamma-probability density function to model state durations. The 39-dimensional feature vectors consisted of 12 standard MFCC and the log power plus their delta and delta-delta derivatives. Cepstral mean subtraction and maximum-likelihood linear transformation (MLLT) were applied.

The language model was trained using a 158-million word corpus of Finnish newspaper articles, books, and newswire stories [CSC Tieteellinen laskenta Oy 2007]. As described in Section 3.1, the training text was first segmented with Morfessor, a special symbol was inserted at word boundaries, and, finally, an n-gram language model was

Table II. Language Model Statistics. Here, Q-OOV rate is the proportion of query words that are not in the LM training set

LM trainig set [words]	158M
Unique word forms	4.1M
Unique morphs	19k
Q-OOV rate (T queries)	11.3%
Q-OOV rate (TD queries)	3.7%

trained with the VariKN-toolkit [Siivola et al. 2007]. The unsegmented corpus had 4.1-million unique word forms, while the segmented corpus had about 19,000 unique morpheme-like units. A summary of the statistics is given in Table II. Query OOV (Q-OOV) rates mean word forms in the query that are not in the language model training corpus at all. If a word-based model was trained, the size of the lexicon would probably have to be limited from the 4.1M word forms to much smaller size, and for a word-based LM, the Q-OOV rates would be even higher. Morfessor can split unseen words into known morphs, thus the morph versions of the queries have Q-OOV rate of 0%.

The *lattice-tool* in SRILM-toolkit [Stolcke 2002] was used to transform the lattices produced by the speech recognizer into confusion networks. None of the available IR toolkits we found supported the type of testing we wanted to perform, for instance, by allowing us to change the *tf.idf* values easily. Thus a vector-based test system was implemented for indexing and retrieval. However, whenever possible, the results were compared to the results given by corresponding methods in the LEMUR-toolkit [Ogilvie and Callan 2002] to ensure correct functioning.

For testing and optimizing the segmentation algorithm, a text retrieval corpus from Cross-Language Evaluation Forum (CLEF) [Agirre et al. 2008] was used. The corpus consisted of Finnish newspaper articles of about 55,000 documents (4.6 million words), 50 queries with about 23,000 binary relevance assessments (413 relevant).

3.5. Evaluation

The task in this experiment differs somewhat from traditional SDR. The retrieval system returns a ranked list of replay points. For evaluation, the replay points are compared to the human-assessed ground truth points. The closer the returned point is to the correct one, the better the result. The correctness of the endpoint of the segment is not evaluated. While knowing the endpoint may be useful in some cases (e.g., browsing), it is assumed here that the user will start playing at the given point and will know when to stop listening.

The retrieval effectiveness is measured by mean generalized average precision (mGAP). Originally the measure was designed to be used in the case of graded human-relevance assessments [Kekäläinen and Järvelin 2002]. For replay point retrieval, the assessment is binary but the match can be partial. The mGAP metric was suggested for use in speech retrieval evaluations in an unknown boundary condition in Liu and Oard [2006] and has been used in the CLEF speech retrieval track [Pecina et al. 2007].

Calculating GAP values involves using a relevance function R that gives full match when the returned replay point t is exactly at the same location as the ground truth point g. As t moves away from g in either direction, the value of R is decreased. That is, the bigger the distance to the relevance point, the smaller the degree of the match. At some distance w from g, R is set at 0, and any replay point further from that is considered nonrelevant.

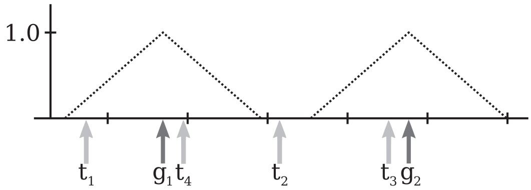

Fig. 5. Relevance functions around two ground truth points, g_1 and g_2. The system returns a ranked list of replay points (t_1, t_2, t_3, t_4). Note that t_4 will get a score of 0, because g_1 is used by t_1.

In this work, 120-second wide triangular functions are used. The function is unity at the ground truth point and decays linearly until it reaches zero when the distance to the ground truth point is plus or minus 60 seconds. Figure 5 illustrates the triangular relevance functions around two ground truth points.

For a ranked list of replay points, the precision at rank k is calculated as

$$p_k = \left(\sum_{i=1}^{k} R_i \right) \Big/ k, \tag{3}$$

where R_i is the score given by the relevance function for the replay point at rank i. The relevance function values are calculated in the order of the list, and each ground truth point is used only once. That is, any point further down the ranked list gets $R_i = 0$ even if it is within w of the used ground truth point. Using the preceding precision values, generalized average precision is calculated as

$$GAP = \left(\sum_{R_k \neq 0} p_k \right) \Big/ N, \tag{4}$$

where, N is the number of ground truth points.

Recall r_k can be calculated normally as the proportion of relevant playback points retrieved at some rank k. Using these values, it is possible to plot precision p_k as a function of recall. These functions will be heavily serrated and thus *interpolated precision* is usually used instead. Interpolated precision at a certain level of recall r is defined as the maximum precision found for any recall level greater than or equal to r.

Statistical significance testing for comparing mGAP and other precision values was performed as outlined in Hull [1993]. The values were transformed with the $\arcsin\sqrt{x}$ function to make them more normally distributed. Jarque-Bera and Lilliefors goodness-of-fit tests were performed to ensure normalness. The t-test was used pairwise to test if results differ in a statistically significant way. A significance level of 5% was used in all cases.

4. RESULTS

Table III shows retrieval results on the speech corpus, comparing morph, base form, and combined methods as well as CL or rank-based weighting for the confusion networks. Performance for each query between morph and base form indexing are compared in Figure 6 and between CN and 1-best indexing in Figure 7. Results of cross testing the segmentation and indexing term types show that segmentation using morphs only or combined morphs and base forms give about equally good results on all indexing types (Table IV). Using 2-best Morfessor segmentations of the query words offered a slight improvement of mGAP over using only 1-best morph segments (Table V).

Table III. Retrieval Statistics for Different Setups.
The best result for each metric and query type is underlined. The cases where the confusion network (CN) method was statistically significantly better than the best corresponding 1-best method are in bold. The automatic segmentation method ('seg') is also compared to the sliding window baseline ('sliding win'). Some CL combinations have been left out for brevity

method		query	GAP	P@R	P5	P15
1best rank	morph seg	TD	0.438	0.416	0.623	0.445
		T	0.318	0.335	0.452	0.344
	base form seg	TD	0.438	0.418	0.574	0.430
		T	0.347	0.345	0.518	0.346
	combined seg	TD	0.462	0.454	0.624	0.454
		T	0.379	0.370	0.558	0.390
	interleaved seg	TD	0.468	0.438	0.608	0.447
		T	0.361	0.353	0.502	0.370
	combined sliding win	TD	0.347	0.337	0.498	0.354
		T	0.262	0.263	0.399	0.279
CN rank	morph seg	TD	**0.460**	0.436	0.638	0.451
		T	**0.346**	0.353	**0.510**	**0.366**
	base form seg	TD	0.443	0.439	0.574	0.430
		T	0.354	0.356	0.501	0.352
	combined seg	TD	0.461	0.446	0.645	0.441
		T	0.376	0.365	0.572	0.376
	interleaved seg	TD	**0.492**	**0.474**	0.653	0.472
		T	0.373	**0.394**	0.525	0.390
	combined sliding win	TD	0.350	0.341	0.518	0.359
		T	0.259	0.264	0.402	0.274
1best CL	morph seg	TD	0.422	0.405	0.605	0.419
		T	0.300	0.325	0.423	0.328
	base form seg	TD	0.393	0.389	0.528	0.400
		T	0.309	0.307	0.433	0.303
	combined seg	TD	0.443	0.428	0.609	0.440
		T	0.352	0.361	0.500	0.367
CN CL	morph seg	TD	0.437	0.424	0.622	0.438
		T	0.327	0.338	0.443	0.343
	base form seg	TD	0.403	0.402	0.532	0.418
		T	0.322	0.318	0.461	0.309
	combined seg	TD	0.452	0.436	0.627	0.443
		T	0.361	0.364	0.533	0.368

The morph confusion networks had on average 3.1 morphs per confusion set. For the word confusion networks, the average is higher (5.6 words) because all the different inflections that were represented as combinations of stem and suffix morphs received their own arc. Different inflections of a word was returned to the same base form but some word forms also have multiple alternative base form interpretations. Thus, base form confusion networks had only slightly smaller than average, 5.1 unique base forms, per confusion set.

Table VI shows the mGAP results of the initial testing of the segmentation method on the text retrieval corpus. The results are optimized with respect to the α-parameter and stop list size. For α, a value of 0.5 was found to produce the best average performance. However, the results changed very little if α was in the region from 0.2 to 0.8. A stop list of 150 most common terms was found best for the morph index. The results for the base form and combined index did not improve when using a stop list. Note that separate development and test sets were not used for the text retrieval experiments, rather the text corpus was regarded as development data and the speech corpus as test data. The parameters found to work best for text retrieval were used for speech retrieval.

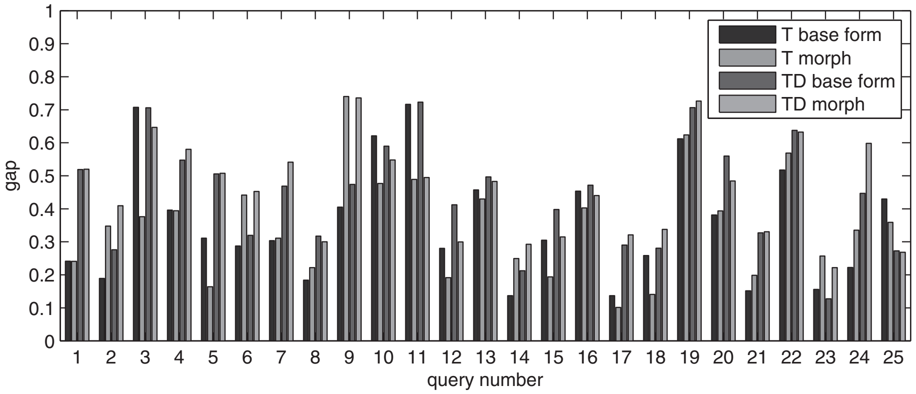

Fig. 6. GAP for each query comparing morph and base form indexing using the CN rank method.

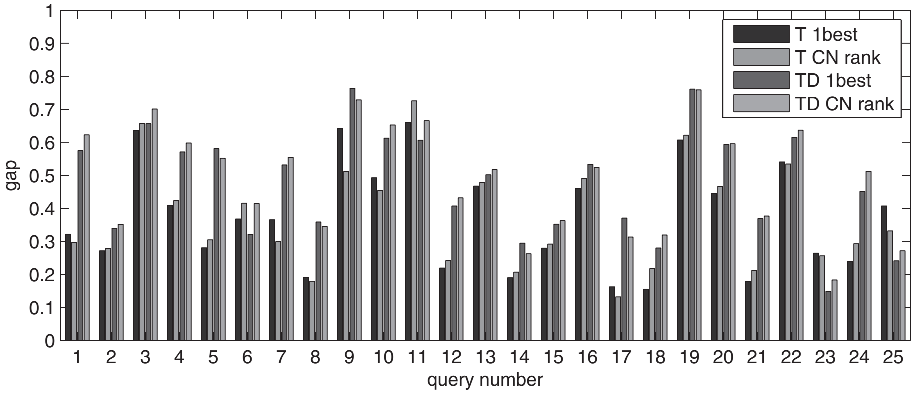

Fig. 7. GAP for each query comparing CN and 1-best indexing using the rank method and the interleaved index.

Table IV. mGAP Results for Cross Testing the Segmentation and Indexing Types Using 1-Best Indexes.
For the morph index and TD queries, the base form segmentation is significantly worse than for the morph segmentation. In other cases, there were no statistical differences between the columns (segmentation types)

		segmentation		
index	query	morph	base form	combined
morph	TD	**0.438**	0.409	**0.421**
	T	0.318	0.302	0.315
base form	TD	0.447	0.438	0.454
	T	0.355	0.347	0.359
combined	TD	0.470	0.457	0.462
	T	0.380	0.367	0.379

Table V. mGAP Results for Using 1-Best and 2-Best Morfessor Segmentations for Query Words for Morph, Combined and Interpolated 1-Best and CN Rank Indexes.
Note that here "1-best" can refer to both ASR result (rows) or Morfessor segmentation result (columns). The cases where the difference is statistically significant to the corresponding Morfessor, the 1-best case are in bold

method	query	1-best	2-best	diff
1-best morph	TD	0.430	0.438	1.80%
	T	0.304	**0.318**	4.69%
CN rank morph	TD	0.452	0.460	1.56%
	T	0.332	**0.346**	4.26%
1-best combined	TD	0.460	0.462	0.59%
	T	0.373	0.379	1.58%
CN rank combined	TD	0.454	0.461	1.47%
	T	0.370	0.376	1.69%
1-best interleaved	TD	0.466	0.468	0.51%
	T	0.351	0.361	2.78%
CN rank interleaved	TD	0.489	0.492	0.74%
	T	0.368	0.373	1.47%

Table VI. Generalized Average Precision Values for Testing the Segmentation Method with a Text Retrieval Corpus

index	query	real seg.	segmentation	diff
morph	TD	0.420	0.292	−30.7%
	T	0.367	0.244	−33.6%
base form	TD	0.401	0.313	−21.8%
	T	0.372	0.271	−27.2%
combined	TD	0.425	0.306	−27.8%
	T	0.345	0.245	−28.8%

5. DISCUSSION

5.1. Selection of Indexing Units

The speech retrieval results indicate that combing both morphs and base forms as index terms leads to the best performance. However, should a morphological analyzer not be available, using morphs only is as good as the traditional way of using base forms only. There were small (but not significant) differences between the two methods. Some query words might get segmented to morphs that do not work well as index terms due to over- or undersegmentation. If the query is short, the performance suffers, but if there are many query word forms, the other terms can compensate. This explains why morph-based indexing performs worse for T queries but not for TD queries.

The two fusion techniques, combined and interleaved, do not use any weighting between the features. But since there is a difference in performance depending on the type of query, appropriate weighting could lead to improved performance. Weighting techiques could be based on the length of the query, the existence of the query words in the LM training corpus, and the existence of the query words in the morphological analyzer lexicon. This is left as future work.

Words that have changes in the root with different inflections are problematic for morph indexing. This happened with query numbers 3 "`nokia|n vesi|kriisi`" (Water crisis in Nokia) and 18 "`ruo|an hinta`" (Price of food). Symbol "`|`" marks the morph boundary used in querying. In many documents, the word *vesi* (water) appears in the genitive form *veden*. The word *ruoka* (food) is doubly complicated because it has two alternative spellings for it's genitive, *ruoan* and *ruuan*. The morphological analyzer can return all of these forms to their base form without problems. The words are also common and reliably recognized from speech.

For rare words that are difficult to recognize correctly, the morph-based indexing can be beneficial. For example, in query number 9 "`nokia|n bo|chu|min tehtaan lopet|us`" (Shutting down of Nokia's factory in Bochum), the name *Bochum* was difficult for the recognizer and it never recognized it correctly. The word was not in the lexicon of the morphological analyzer either. But the subword `bo` was often correctly recognized. As this is a very rare sequence in Finnish, it helped in the morph-based retrieval.

Previous work shows that the problem of different roots can be alleviated by using Latent Semantic Indexing (LSI) [Turunen and Kurimo 2006]. Since different word forms of the same word often cooccur in the same document, they are conflated by LSI. LSI also helps with the usual problem of synonyms and deviation in the use of language between queries and the spoken segment. Query expansion is an another technique that has been shown to help with these issues [Kurimo and Turunen 2005] by introducing both cooccurring terms and alternative forms of root morphs to the query.

Introducing alternative Morfessor segmentations for query terms should help with the problem, and indeed there were small improvements especially for T queries. Still, for T queries, the performance level of base forms is not quite reached. It may be that having multiple alternative morph segmentations also introduces confusing query terms in some cases. For combined and interleaved indexes, alternative Morfessor segmentations offered smaller or no improvement. In these cases, having the base forms as index terms lessens the usefulness of alternative Morfessor segmentations.

5.2. Effect of Indexing Confusion Networks

When expanding the index with competing terms in the confusion network, appropriate weighting of the terms is important in order to achieve improvements over the 1-best case. This is not surprising since maximum-one term in each confusion set can be correct and giving too much weight to the incorrect ones will degrade precision. The CL approach of weighting provides minimal improvements for morph index, and for base form and combined indexes it actually degrades the results. Weighting the terms in the ASR 1-best with CL weights offers no improvements either. Weighting the terms based on the rank of the terms, however, improves the results for all index types except the combined index where there is negligible degradation.

While the speech corpus is larger than the ones used before for this language, it is still rather small as a retrieval collection. Optimal assignment of weights for the terms in the CN will most likely depend on the collection and the queries. However, as the measured improvements were statistically significant, some conclusions for using CNs to improve speech retrieval can be drawn. Confirming the results on larger more diverse corpora and in different languages is left as future work.

Also for the confusion networks, the results are somewhat different when using T or TD queries. Typically users type short queries of only a couple of words (2.4 in Spink et al. [2001]), but query expansion could be used to bring the number of query words closer to the TD case. For TD queries, the interleaved index with CN rank expansion is the best method with statistically significant improvements over all 1-best cases. The relative improvement of mGAP over the best 1-best method (also using interleaved) is 5.1%. For T queries, the combined index using only ASR 1-best gives the best results. However, if the interleaved CN rank method is used, the performance on average is not significantly worse, and for high levels of recall, there are in fact small improvements.

Confusion network improvements are largest for the morph index, which probably contributes in large part to the improvements of the interleaved index as well. Some of the effect may be explained due to the fact that the alternative results are sometimes the same word in different inflections and that for some words the stem morphs may

Table VII. Summary of the Results.
Base form indexing is considered the baseline and improvements of mGAP are given when interleaving the base form index with the morph index, with 2-best query morphs and with CNs. Improvements are always given with respect to the baseline

setup	TD		T	
baseline	0.438		0.347	
+ morphs	0.466	+6.4%	0.351	+1.2%
+ 2-best query morphs	0.468	+6.8%	0.361	+4.0%
+ CN	0.492	+12.3%	0.373	+7.5%

be different for the word in different inflections due to suboptimal morph segmentations. Having both versions of the stem in the CN will help match the query word. Combining the morph and base form versions of the CN to the same index does not seem to work as well as constructing separate indices and interleaving the results. Having different kinds of terms in the same document may lead to suboptimal *tf* or *idf* estimates.

In summary, considering the standard way of doing things, that is, using only the ASR 1-best with morphologically normalized base forms as the baseline, moving to the interleaved index with CN rank expansion, offers a 12.3% improvement of mGAP for TD queries and a 7.5% improvement for T queries (Table VII).

5.3. Effect of Story Segmentation

The TextTiling-like automatic story segmentation is able to locate relevant replay points far better than the sliding window baseline. However, experiments on the text corpus indicate that better segmentation could improve retrieval results further. On the text corpus, segmentation and indexing based on base forms gives better results than on morphs, but on the speech corpus, the methods are equally good. However, we can see that morphs work as index terms because morphs give better results than base forms for TD queries using the known segments of the text corpus for retrieval. The differences can be explained by the fact that the text corpus is error free but the ASR transcripts are not. Thus, base forms seem to work comparably better for segmenting error-free text, but not in presence of recognition errors.

The results on text retrieval support the hypothesis that it is better to oversegment for optimal replay point retrieval performance. The parameter α was tuned for best mGAP, and, as a result, there were about 50% more segment boundaries than in the original corpus. Optimizing α with respect to finding the real boundaries (e.g., in terms of the F-measure), results in a lower number of segment boundaries, but worse retrieval performance.

5.4. Further Discussion

Previous results on Finnish SDR further support the use of morphs for retrieval [Turunen 2008]. Confusion networks were found useful especially for the morph-based system in the presence of OOV query words. Considering the word-based 1-best system as a baseline, by switching to the morph-based system that uses confusion networks, MAP increases by 47%. Compared to this work, there were some differences in the setup. First, in Turunen [2008], the retrieval database was much smaller (about 300 spoken documents), and it was manually segmented to stories. Second, the *idf* metric used to index the confusion networks was different. In this work, the *idf* used previously did not offer as good results and a different formula for *idf* was adapted. Most likely this effect is caused by the difference in the size of the database. It is likely that the speech corpus, the ASR, and the models used has a large effect on the terms and probabilities that appear in the confusion network, and the optimal way of assigning weights to the terms will depend on those.

Using the language modeling approach for retrieval [Ponte and Croft 1998] would allow a more systematic way of inferring parameters than the vector space model. Similar to the work in Chia et al. [2010], the CL and rank-based estimates for term frequencies should be easily transferred to a unigram language modeling retrieval framework, but experiments need to be carried out to find out its performance. This is left for future work.

The speech retrieval corpus is relatively small compared to many text retrieval corpora. On the other hand, recall measures are more reliable as we can be reasonably sure that the relevance judgements covered the entire database. Because a large segmented speech corpus was not available, a much larger text corpus with more queries was used for development and testing. The newspaper text was mostly similar in style to the spoken documents except it lacks the interview and other spontaneous material that were present in the speech data. We believe that removing sentence and segment information made the text close enough to the actual ASR transcripts to be useful as development data. Given the added evidence from the experiments on the large text corpus, the conclusions for selecting the indexing units for retrieval and segmentations are strong. The results for selecting weights for the terms in the CNs are less reliable.

The success of the replay point retrieval is measured in terms of mean generalized average precision. While it is intuitive to give a higher score to the returned replay points the closer they are to the ground truth points, it has not yet been studied how well these values correspond to actual user experience, as far as we know. For example, the triangular windows give an equal score on both sides of the ground truth point, while in reality the perceived success might depend on which side of the ground truth point we are.

It was found that excluding the most common morphs from the index improved the results slightly but the same did not hold for the base form or combined index. In the morph index, the stop list consisted mainly of suffix morphs and function words, thus excluding them makes sense. Using the same approach for the base form and combined index, some content words did appear in the stop lists. However, common terms will receive a low *idf* value in any case and large variation in results was not observed either way. Manual stop lists would be a better approach for the base form and combined indexes, while the optimal way of gathering the morph stop list remains open.

Future work includes repeating the tests on other morphologically rich languages and testing the morph-based approach for other speech retrieval tasks such as STD. The segmentation method could be improved by obtaining a manually segmented corpus for training and testing methods that could use acoustic features as well. The confusion network-based approach should be compared to other similar approaches such as PSPLs and TMI (see Section 2). Currently, the Morfessor algorithm and the LVCSR were optimized with respect to achieving the best recognition accuracy in terms of the word-error rate (on separate but similar material). Improvements could be gained by optimizing parameters with respect to retrieval performance.

6. CONCLUSIONS

We have evaluated using statistically determined morpheme-like units for Spoken Document Retrieval from unsegmented Finnish audio. The morph-based approach alleviates the OOV problem by allowing recognition of unseen word forms by concatenation of known morphs without compromising the recognition accuracy of in-vocabulary words. For index terms, using morphs in combination with base forms was shown most effective. Story segmentation based on lexical similarity was applied and the morph-based approach was found better than base forms for segmenting ASR text but not to a significant degree. Confusion networks were used to extract alternative recognition results to the index. For morph-based indexing and retrieval based on interleaving the morph and base form results, significant improvements were achieved over using only the 1-best transcripts.

ACKNOWLEDGMENTS

We would like to thank the Adaptive Informatics Research Centre, Tivit Next Media, The Finnish Foundation for Economic and Technology Sciences–KAUTE and ComMIT–Graduate School in Computational Methods of Information Technology for partially supporting this work.

REFERENCES

Agirre, E., Nunzio, G. M. D., Ferro, N., Mandl, T., and Peters, C. 2008. CLEF 2008: Ad hoc track overview. In *Working Notes for the CLEF Workshop*.

Alkula, R. 2001. From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. *Inform. Retrieval 4*, 195–208.

Arisoy, E., Kurimo, M., Saraçlar, M., Hirsimäki, T., Pylkkönen, J., Alumäe, T., and Sak, H. 2008. Statistical language modeling for automatic speech recognition of agglutinative languages. In *Speech Recognition, Technologies and Applications*, F. Mihelic and J. Zibert, Eds., I-Tech, 193–204.

Beeferman, D., Berger, A., and Lafferty, J. 1999. Statistical models for text segmentation. *Machine Learn. 34,* 1, 177–210.

Blei, D. M. and Moreno, P. J. 2001. Topic segmentation with an aspect hidden markov model. In *Proceedings of SIGIR*. ACM, New York, NY, 343–348.

Chelba, C. and Acero, A. 2005. Position specific posterior lattices for indexing speech. In *Proceedings of the Association for Computational Linguishes (ACL'05)*. ACL, 443–450.

Chelba, C., Silva, J., and Acero, A. 2007. Soft indexing of speech content for search in spoken documents. *Comput. Speech Lang. 21,* 3, 458–478.

Chia, T. K., Sim, K. C., Li, H., and Ng, H. T. 2010. Statistical lattice-based spoken document retrieval. *ACM Trans. Inf. Syst. 28,* 1, 1–30.

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., and Stolcke, A. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Trans. Speech Lang. Process. 5,* 1, 1–29.

Creutz, M. and Lagus, K. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Tech. rep. A81, Publications in Computer and Information Science, Helsinki University of Technology. http://www.cis.hut.fi/projects/morpho/.

CSC Tieteellinen laskenta Oy. 2007. Finnish language text bank. http://www.csc.fi/kielipankki/.

Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the Association for Computational Linguishes (ACL'04)*. ACL, 562–569.

Garofolo, J. S., Auzanne, C. G. P., and Voorhees, E. M. 2000. The TREC spoken document retrieval track: A success story. In *Proceedings of the International Conference on Computer-Assisted Information Retrieval (RIAO)*. 1–20.

Hacioglu, K., Pellom, B., Ciloglu, T., Ozturk, O., Kurimo, M., and Creutz, M. 2003. On lexicon creation for turkish LVCSR. In *Proceedings of Interspeech*. 1165–1168.

Hearst, M. A. 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist. 23,* 1, 33–64.

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., and Pylkkönen, J. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech Lang. 20,* 4, 515–541.

Hirsimäki, T., Pylkkönen, J., and Kurimo, M. 2009. Importance of high-order n-gram models in morph-based speech recognition. *IEEE Trans. Audio, Speech Lang. Proces. 17,* 4, 724–732.

Hori, T., Hetherington, I. L., Hazen, T. J., and Glass, J. R. 2007. Open-vocabulary spoken utterance retrieval using confusion networks. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Hull, D. A. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of SIGIR*. ACM Press, New York, NY, 329–338.

ISKRA, D., GROSSKOPF, B., MARASEK, K., VAN DEN HEUVEL, H., DIEHL, F., AND KIESSLING, A. 2002. SPEECON–speech databases for consumer devices: Database specification and validation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 329–333.

JAMES, D. 1995. The application of classical information retrieval techniques to spoken documents. Ph.D. thesis, University of Cambridge, UK.

JAMES, D. A., YOUNG, S., AND PZ, C. 1994. A fast lattice-based approach to vocabulary independent wordspotting. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 377–380.

JOHNSON, S. E., JOURLIN, P., JONES, K. S., AND WOODLAND, P. C. 2001. Information retrieval from unsegmented broadcast news audio. *Int. J. Speech Technol. 4,* 3, 251–268.

KAZEMIAN, S., RUDZICZ, F., PENN, G., AND MUNTEANU, C. 2008. A critical assessment of spoken utterance retrieval through approximate lattice representations. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR'08)*. ACM, New York, NY, 83–88.

KEKÄLÄINEN, J. AND JÄRVELIN, K. 2002. Using graded relevance assessments in IR evaluation. *J. Amer. Soc. Inf. Sci. Technol. 53,* 13, 1120–1129.

KOSKENNIEMI, K. 1983. Two-level morphology: A general computational model for word-form recognition and production. Ph.D. thesis, Department of General Linguistics, University of Helsinki.

KOZIMA, H. 1993. Text segmentation based on similarity between words. In *Proceedings of the Association for Computational Linguistier (ACL'93)*. 286–288.

KURIMO, M., CREUTZ, M., AND VARJOKALLIO, M. 2008. Morpho Challenge evaluation using a linguistic gold standard. In *Proceedings of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF'07)*. (Revised Selected Papers). Lecture Notes in Computer Science, Vol. 5152, Springer, Berlin, 864–873.

KURIMO, M. AND TURUNEN, V. 2005. To recover from speech recognition errors in spoken document retrieval. In *Proceedings of Interspeech*. 605–608.

KURIMO, M., TURUNEN, V., AND EKMAN, I. 2004. An evaluation of a spoken document retrieval baseline system in Finnish. In *Proceedings of Interspeech*.

KURIMO, M., VIRPIOJA, S., TURUNEN, V. T., BLACKWOOD, G. W., AND BYRNE, W. 2010. Overview of Morpho Challenge 2009. In *Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF'09)*. (Revised Selected Papers). Lecture Notes in Computer Science. Springer, Berlin.

LINGSOFT, INC. 2007. FINTWOL: Finnish morphological analyser [computer software]. http://www.lingsoft.fi/.

LIU, B. AND OARD, D. W. 2006. One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. In *Proceedings of SIGIR*. ACM, New York, NY, 673–674.

LOGAN, B., MORENO, P., AND DESHMUKH, O. 2002. Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio. In *Proceedings of the Human Language Technology Conference (HLT'02)*.

LOGAN, B. AND THONG, J. V. 2002. Confusion-based query expansion for oov words in spoken document retrieval. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

MAMOU, J., CARMEL, D., AND HOORY, R. 2006. Spoken document retrieval from call-center conversations. In *Proceedings of SIGIR*. ACM Press, New York, NY, 51–58.

MANGU, L., BRILL, E., AND STOLCKE, A. 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Comput. Speech Lang. 14*, 373–400.

MULBREGT, P. V., CARP, I., GILLICK, L., LOWE, S., AND YAMRON, J. 1998. Text segmentation and topic tracking on broadcast news via a hidden markov model approach. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. 2519–2522.

NG, K. 2000. Subword-based approaches for spoken document retrieval. Ph.D. thesis, Massachusetts Institute of Technology.

OGILVIE, P. AND CALLAN, J. 2002. Experiments using the lemur toolkit. In *Proceedings of TREC*. National Institute of Standards and Technology. 103–108.

PAN, Y. C., CHANG, H. L., AND LEE, L. S. 2007. Analytical comparison between position specific posterior lattices and confusion networks based on words and subword units for spoken document indexing. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU)*. 677–682.

PASSONNEAU, R. J. AND LITMAN, D. J. 1997. Discourse segmentation by human and automated means. *Comput. Linguist. 23,* 1, 103–139.

PECINA, P., HOFFMANNOVÁ, P., JONES, G. J., ZHANG, Y., AND OARD, D. W. 2007. Overview of the CLEF-2007 Cross Language Speech Retrieval Track. In *Working Notes for the CLEF'07 Workshop*, A. Nardi and C. Peters, Eds.

PIRKOLA, A. 2001. Morphological typology of languages for IR. *J. Document. 57,* 3, 330–348.

Ponte, J. M. and Croft, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 275–281.

Renals, S. and Ellis, D. 2003. Audio information access from meeting rooms. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'03)*. 744–7.

Saraçlar, M. and Sproat, R. 2004. Lattice-based search for spoken utterance retrieval. In *Proceedings of HTL-NAACL*. 129–136.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Comm. 32,* 1-2, 127–154.

Siegler, M. A. 1999. Integration of continuous speech recognition and information retrieval for mutually optimal performance. Ph.D. thesis, Carnegie Mellon University.

Siivola, V., Creutz, M., and Kurimo, M. 2007. Morfessor and VariKN machine learning tools for speech and language technology. In *Proceedings of Interspeech*.

Siivola, V. and Pellom, B. 2005. Growing an n-gram model. In *Proceedings of Interspeech*. 183–188.

Spink, A., Building, R. I., Wolfram, D., and Saracevic, T. 2001. Searching the web: the public and their queries. *J. Amer. Soc. Inform. Science Technol. 52,* 52, 226–234.

Stokes, N. 2004. Select: a lexical cohesion based news story segmentation system. *AI Comm. 17,* 1, 3–12.

Stolcke, A. 2002. SRILM—an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. 901–904.

Tür, G., Hakkani-Tür, D., Stolcke, A., and Shriberg, E. 2001. Integrating prosodic and lexical cues for automatic topic segmentation. *Computat. Ling. 27,* 1, 31–57.

Turunen, V. T. 2008. Reducing the effect of OOV query words by using morph-based spoken document retrieval. In *Proceedings of Interspeech*. 2158–2161.

Turunen, V. T. and Kurimo, M. 2006. Using latent semantic indexing for morph-based spoken document retrieval. In *Proceedings of Interspeech*. 341–344.

Turunen, V. T. and Kurimo, M. 2007. Indexing confusion networks for morph-based spoken document retrieval. In *Proceedings of SIGIR*. ACM, New York, NY, 631–638.

Woodland, P. C., Johnson, S. E., Jourlin, P., and Spärck Jones, K. 2000. Effects of out of vocabulary words in spoken document retrieval. In *Proceedings of SIGIR*. 372–374.

Xie, L., Zeng, J., and Feng, W. 2008. Multi-Scale TextTiling for automatic story segmentation in chinese broadcast news. In *Information Retrieval Technology*, Springer, Berlin, 345–355.

Yu, P., Chen, K., Lu, L., and Seide, F. 2005. Searching the audio notebook: keyword search in recorded conversations. In *Proceedings of the Human Language Technology Conference (HLT'05)*. ACL, 947–954.

Yu, P. and Seide, F. 2004. A hybrid-word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech. In *Proceedings of Interspeech*. 293–296.

Zhou, Z., Yu, P., Chelba, C., and Seide, F. 2006. Towards spoken-document retrieval for the internet: lattice indexing for large-scale web-search architectures. In *Proceedings of HLT-NAACL*. ACL, New York, New York, 415–422.

Received June 2010; revised February 2011, May 2011; accepted July 2011