Assessing the Retrieval Effectiveness of a Speech Retrieval System by Simulating Recognition Errors

Peter Schäuble, Ulrike Glavitsch

Swiss Federal Institute of Technology (ETH)
CH-8092 Zurich, Switzerland

ABSTRACT

We show how the recognition performance of a speech recognition component in a speech retrieval system affects the retrieval effectiveness. A speech retrieval system facilitates content-based retrieval of speech documents, i.e. audio recordings containing spoken text. The speech retrieval process receives queries from users and for every query it ranks the speech documents in decreasing order of their probabilities that they are relevant to the query. The speech recognition component is an important part of a speech retrieval system, since it detects the occurrences of indexing features in the documents. Because the recognition of indexing features in continuous speech is error prone, the question arises how much an error prone recognition of indexing features affects the retrieval effectiveness. As an answer to this question and main contribution of this paper we simulated the recognition of indexing features in speech documents on standard information retrieval test collections and show the resulting retrieval accuracies.

1. Introduction

We show how the recognition performance of a speech recognition component in a speech retrieval system affects the retrieval effectiveness. A speech retrieval system facilitates content-based retrieval of speech documents, i.e. audio recordings containing spoken text [5]. The speech retrieval process receives queries from users and for every query it ranks the speech documents in decreasing order of their probabilities that they are relevant to the query. These probabilities are derived from the occurrences of indexing features that were identified in the speech documents by a speech recognition component [4]. Because the recognition of indexing features in continuous speech is error prone, the question arises how much an error prone recognition of indexing features affects the retrieval effectiveness.

The indexing features used in our speech retrieval system are phonetically motivated subword units having an intermediate specificity. The general pattern of an indexing feature is a maximum sequence of consonants enclosed by two maximum sequences of vowels at both ends. We call these indexing features VCV-features where C stands for the maximum sequence of consonants and V stands for the maximum sequence of vowels. As an example, the word INTERNATIONAL contains the VCV-features INTE, ERNA, ATIO, and IONA. The indexing vocabulary is defined to be the set of those VCV-features φ_i whose inverse document frequencies $idf(\varphi_i)$ are between a lower bound idf_{min} and an upper bound idf_{max} such that the indexing features are neither very specific nor very broad. The lower bound guarantees the suitability for indexing and the upper bound guarantees the trainability, i.e. there are enough examples to train the HMMs. Experiments on standard information retrieval test collections showed that when using an appropriate subset of only 1000 VCV-features we can achieve a retrieval effectiveness that is comparable to standard weighted retrieval which is based on a much larger indexing vocabulary [4]. In addition to the VCV-features, we have also studied indexing vocabularies that have been extended by CV- and VC-features at the word boundaries

The recognition of speech documents is carried out with standard speech recognition technology, i.e. a wordspotter [6], [11], [14] locates the occurrences of indexing features in documents. For each document in the collection we create a description vector based on the number of occurrences of each feature and use a conventional retrieval function [12] to estimate the similarity between a document and a query description.

Our indexing features consisting of VCV-features can be identified in both text and speech documents. As a consequence, the document collection may contain a mixture of text and speech documents. Furthermore, the query may also be entered as either text or speech. The controlled indexing vocabulary consisting of selected VCV-features has the advantage that the document description can be computed before the query evaluation. In particular, an access structure (e.g. an inverted file) can be constructed to allow fast query evaluation. Another important advantage of our indexing vocabulary for both text and speech retrieval is that speech retrieval can simulated by using text collections as described in the subsequent sections.

Information Retrieval on audio documents has been investigated very little. A wordspotting system for voice indexing was developed by Wilcox and Bush [14] and an information retrieval system that classifies speech messages was presented by Rose, Chang, and Lippmann [10]. Recently a project for video mail retrieval using voice was proposed by Olivetti research Limited, Cambridge University Engineering Department, and Cambridge University Computer Laboratory [7]. The effects of recognition errors on on the retrieval effectiveness has been studied in the context of OCR based Information Retrieval [1]. These results are not directly comparable because speech retrieval performance may be considerably affected by false alarms in contrast to OCR-based retrieval where false alarms can be ignored because the occur infrequently.

MEDLARS: average precision of the reference method: 0.534 (100%)

dr, fa	0	10	20	50	80	110
90%	0.539 (101%)	0.464 (87%)	0.432 (81%)	0.421 (79%)	0.428 (80%)	0.412 (77%)
80%	0.530 (99%)	0.444 (83%)	0.425 (80%)	0.403 (75%)	0.402 (75%)	0.413 (77%)
70%	0.487 (91%)	0.406 (76%)	0.400 (75%)	0.388 (73%)	0.389 (73%)	0.349 (65%)
60%	0.481 (90%)	0.400 (75%)	0.360 (67%)	0.340 (64%)	0.352 (66%)	0.335 (63%)
50%	0.431 (81%)	0.335 (63%)	0.317 (59%)	0.318 (60%)	0.319 (60%)	0.285 (53%)
40%	0.421 (79%)	0.318 (60%)	0.271 (51%)	0.284 (53%)	0.299 (56%)	0.269 (50%)

CRANFIELD: average precision of the reference method: 0.408 (100%)

dr, fa	0	10	20	50	80	110
90%	0.330 (81%)	0.315 (77%)	0.304 (75%)	0.288 (71%)	0.279 (68%)	0.264 (65%)
80%	0.324 (79%)	0.299 (73%)	0.291 (71%)	0.276 (68%)	0.265 (65%)	0.253 (62%)
70%	0.305 (75%)	0.283 (69%)	0.267 (65%)	0.265 (65%)	0.242 (59%)	0.249 (61%)
60%	0.297 (73%)	0.253 (62%)	0.234 (57%)	0.245 (60%)	0.249 (61%)	0.224 (55%)
50%	0.277 (68%)	0.236 (58%)	0.224 (55%)	0.226 (55%)	0.211 (52%)	0.206 (50%)
40%	0.259 (63%)	0.216 (53%)	0.191 (47%)	0.185 (45%)	0.191 (47%)	0.175 (43%)

CACM: average precision of the reference method: 0.257 (100%)

dr, fa	0	10	20	50	80	110
90%	0.132 (51%)	0.138 (54%)	0.155 (60%)	0.156 (61%)	0.131 (51%)	0.139 (54%)
80%	0.134 (52%)	0.133 (52%)	0.135 (53%)	0.136 (53%)	0.113 (44%)	0.116 (45%)
70%	0.124 (48%)	0.110 (43%)	0.129 (50%)	0.115 (45%)	0.104 (40%)	0.095 (37%)
60%	0.111 (43%)	0.097 (38%)	0.103 (40%)	0.096 (37%)	0.103 (40%)	0.109 (42%)
50%	0.113 (44%)	0.097 (38%)	0.084 (33%)	0.098 (38%)	0.091 (35%)	0.085 (33%)
40%	0.076 (30%)	0.078 (30%)	0.048 (19%)	0.057 (22%)	0.079 (31%)	0.069 (27%)

Table 1: Average precision values for detection rates within the range of 40% and 90% and false alarms per indexing feature (key word) per hour within the range of 0 and 140. The numbers in brackets represent the percentage of the average precision of the reference method, i.e. a standard text retrieval method.

The main contribution of this paper is the conclusion that speech retrieval is feasible to some extent even when the recognition performance is poor. A closer look reveals that recognition errors and occurrences of query features in the documents have different distributions and standard retrieval methods are quite good in distinguishing these two distributions. The next two sections describe the test setting and the results respectively. Then, some conclusions are drawn.

2. Test Setting

The experiments are performed by means of the the standard information retrieval text collections CRANFIELD, MEDLARS, and CACM [2]. The indexing vocabulary consists of the VCV-, CV-, and VC-features φ_i whose inverse document frequency

$$idf(\varphi_i) \quad := \quad \log\left(\frac{n+1}{df(\varphi_i)+1}\right)$$

is between the lower bound $idf_{min} := 1.6$ and an upper bound idf_{max} which is chosen such that the indexing vocabulary consists of exactly 1000 features. Every indexing feature φ_i and every document d_j is assigned a weight

$$a_{i,j} \quad := \quad ff(\varphi_i, d_j) * idf(\varphi_i).$$

Analogously, every indexing feature φ_i is assigned a weight

$$b_i \quad := \quad ff(\varphi_i, q) * idf(\varphi_i)$$

with respect to a given query q. As usual, the documents are presented to the user in decreasing order of the Retrieval Status Values $RSV(q, d_j)$ that are determined by the cosine measure.

$$RSV(q, d_j) \quad := \quad \frac{\sum_i a_{i,j} * b_i}{\sqrt{\sum_i a_{i,j}^2} * \sqrt{\sum_i b_i^2}}$$

The recognition errors were simulated in three steps. First, a *parser* converts a text document into a sequence of indexing features by detecting VCV-, CV-, and VC-features that belong to the indexing vocabulary. Second, the sequence of indexing features is converted into another sequence of indexing features by removing features as follows. For every feature in the input sequence it is randomly determined whether it is recognized or not. If it is recognized, the feature is included in the output sequence; otherwise, it is removed. The probability that a feature is recognized is equal to the specified *detection rate*. Third, the reduced sequence of indexing features is converted into a final sequence by adding indexing features as follows. Assume that the original document consists of k occurrences of a word. According to [8], an average

speaker needs approximately $k/170$ minutes or $\Delta t := k/1020$ hours for such a document with k word occurrences. We then add $fa * \Delta t$ occurrences of every indexing feature where fa denotes the specified *false alarms per keyword per hour.* For simplicity, every indexing feature is assumed to have the same detection rate and false alarms.

3. Results

Table 1 shows the average precision values for various detection rates and false alarm rates. The numbers in brackets represent the percentage of a average precision of the reference method. The reference method represents a standard text retrieval method which is based on words rather than on VCV-features. In our case here, the reference method uses van Rijsbergen's [13] stoplist to eliminate the high-frequency words. Furthermore, it uses the word reduction algorithm by Porter [9] to reduce different variants of a words to the same normal form. The term weights consist of simple $tf * idf$ weights and the retrieval status values are obtained by the cosine measure.

Current wordspotting systems report high detection rates and low false alarms for the recognition of entire words in speech documents [2],[3],[4]. These systems, however, are usually evaluated on small tasks: the vocabulary of the speech database is in the order of 1000 words and the number of words spotted is small. On the other hand, the task to identify 1000 VCV-features in speech documents with an unlimited vocabulary is much more difficult and the corresponding false alarms are one to two orders of magnitude higher. We therefore consider detection rates within the range of 40% and 90% and false alarms per keyword (i.e. per indexing feature) per hour within the range of 0 and 140.

4. Conclusions

We have shown that speech retrieval is feasible to some extent even when the recognition performance is poor. It should be noted that a retrieval effectiveness which is moderate because of recognition errors may well be in the range of the retrieval effectiveness of the commonly used boolean retrieval method. In the case of MEDLARS, for instance, the boolean retrieval method achieves a retrieval effectiveness which corresponds to 40 % detection rate and 140 false alarms per indexing feature per hour [3]. It seems that recognition errors and occurrences of query features in the documents have different distributions and standard retrieval methods are quite good in distinguishing these two distributions. Further investigations are needed to study the influence of the length of the queries and the length of the documents.

References

1. W. B. Croft, S. Harding, K. Taghva, and J. Borsack. An Evaluation of Information Retrieval Accuracy with Simulated OCR Output. *Journal of the ASIS*, 1994.
2. E. A. Fox. Virginia Disc One. Virginia Polytechnic Institute and State University, Department of Computer Science, 1990.
3. E. A. Fox and M. B. Koll. Practical Enhanced Boolean Retrieval: Experiences with the SMART and SIRE Systems. *Information Processing & Management*, 24(3):257–267, 1988.
4. U. Glavitsch and P. Schäuble. A System for Retrieving Speech Documents. In N. Belkin, P. Ingwersen, and A. M. Pejtersen, editors, *ACM SIGIR Conference on R&D in Information Retrieval*, pages 168–176, 1992.
5. U. Glavitsch and P. Schäuble. Speech Retrieval in a Multimedia System. In *European Signal Processing Conference (EUSIPCO)*, pages 295–298, 1992.
6. D. James and S.J. Young. A Fast Lattice-Based Approach to Vocabulary-Independent Word Spotting. In *International Conference on Acoustics, Speech, and Signal Processing*, 1994.
7. K. Sparck Jones. VMR–Video Mail retrieval Using Voice. Personal communication.
8. W. A. Lea. *Trends in Speech Recognition.* Prentice-Hall, Englewood Cliffs, NJ, 1980.
9. M. F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980.
10. R. C. Rose, E. I. Chang, and R. Lippmann. Techniques for Information Retrieval from Voice Messages. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 317–320, 1991.
11. R. C. Rose and D. B. Paul. A Hidden Markov Model Based Keyword Recognition System. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 129–132, 1990.
12. G. Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523, 1988.
13. C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.
14. L. D. Wilcox and M. A. Bush. HMM-Based Wordspotting for Voice Editing and Indexing. In *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 25–28, 1991.