INDEXING SPOKEN AUDIO BY LSA AND SOMS

Mikko Kurimo
Until June 2000 at IDIAP
CP-592, Rue du Simplon 4, CH-1920 Martigny, SWITZERLAND
From June 2000 at Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, FINLAND
e-mail: Mikko.Kurimo@hut.fi

ABSTRACT

This paper presents an indexing system for spoken audio documents. The framework is indexing and retrieval of broadcast news. The proposed indexing system applies latent semantic analysis (LSA) and self-organizing maps (SOM) to map the documents into a semantic vector space and to display the semantic structures of the document collection. The SOM is also used to enhance the indexing of the documents that are difficult to decode. Relevant index terms and suitable index weights are computed by smoothing the document vectors with other documents which are close to it in the semantic space. Experimental results are provided using the test data of the TREC's spoken document retrieval track.

1 INTRODUCTION

The information retrieval (IR) from audio sources traditionally relies on handmade notes and annotations. Because huge amounts of digital audio material have recently became available and easy to access using the WWW, the demand of useful, searchable and automatically generated document indexes is rapidly increasing. Often no satisfying written transcripts are available, and thus the indexing has to be based only on the automatic annotation and segmentation of the audio. The recent advances in computational power and automatic speech recognition techniques has made it possible to decode spoken language in almost realtime. For certain very important audio sources, as broadcast news, for example, the state-of-art decoding accuracy and speed already provides means for making useful indexes for IR. A good example of this success is the THISL system [1] and the spoken document retrieval (SDR) in TREC evaluations [14, 15] organized by NIST.

For video annotations and indexing the information in audio is a very important component. Many video sources, such as broadcast news or sports, for example, have spoken commentaries which describe the essential content in a concise way. The decoded speech also provides good indexing terms (the words) for the video. The definition of suitable index terms just based on video image analysis is much less straight-forward. In addition to IR, the decoded speech flow is very useful for producing annotations of audio and video sources. It also enables users to conveniently browse the content and accurately select the interesting parts of the original data for a closer view.

The fundamental question in spoken audio indexing is how to extract the index terms from the audio stream. There are several approaches based on indexing, for example, at the phonetic level, on keyword spotting and on decoding all the speech into text. The latter one is especially interesting, because it allows to use both the best large vocabulary continuous speech recognition (LVCSR) techniques with language modeling (LM) and the latest developments in natural language processing (NLP) techniques for IR. Mainly for these reasons, the full speech decoding approach was chosen for the THISL [1] and the ASSAVID project.

The main motivations to use latent semantic indexing (LSI) [3] and self-organizing maps (SOM) [6] in spoken audio indexing are to reduce data dimensionality and noise, to eliminate the effect of recognition errors, and to exploit and visualize topics and structures derived from the data. Even the best speech recognizers frequently make recognition errors either by giving single irrelevant words or word sequences, or by deleting relevant ones. For broadcast news, which is the main application referred throughout this paper, the errors are often concentrated in some parts of the news stories, because the speech quality ranges from the well-trained newsreaders in a studio to occasional speakers interviewed in noisy conditions. In this application, another significant source of word noise is that many of the documents are very short and hardly represent their topics well in terms of the traditional word count vectors [11]. By extracting a semantic subspace and topics from the collection and indexing the documents with the help of other close-by documents, we aim at reducing the word noise and at finding new useful index terms [9].

From text databases the information is usually retrieved by browsing or querying. As shown by the WEBSOM project [5] and by several others [12, 4], the visualization of the semantic structures of the data and the topics can be very valuable. An organized representation of the contents both gives a quick understanding of the database contents to help querying and offers efficient browsing. The efficient browsing can be compared to how you search for interesting books in a conventional library. There you first select the shelf based on your interests and then start exploring the book covers from there and open the books that you find the most interesting. In this paper the SOM was selected, instead of other clustering methods, to smooth the semantic document vectors, because of its intrinsic data ordering property [6]. A document vector is smoothed by a combination of the document clusters that are close to it in the semantic space. The ordering of the topics reveals the topological structures of the input space and the hierarchies between the topics by a 2D visual display. The display itself can also be used as an efficient audio document browser like WEBSOM and others [5, 12] for text documents.

2 SYSTEM

The proposed speech indexing system starts by preprocessing the speech stream and dividing it into stories. Then a large vocabulary hybrid HMM/ANN speech recognizer [1, 2, 9] with suitable language models decodes the stories into text documents using the most probable hypothesis for the word sequences. The vector space presentation of a document is formed as the classical bag-of-words model for indexing [11], except that the word vectors are mapped to the extracted latent semantic space [3], and weighted by their importance. For smoothing the semantic document vector is then mapped into the closest clusters in this semantic space [8]. The best index terms are finally selected and the connections are weighted with respect to this mapping. For optimal retrieval results, the index terms directly derived from the actual decoded words are also used with weights that are combinations of the OKAPI frequency weight [1] of the index term and the semantic matching between the term and the document [8].

The construction of the latent semantic space starts by reducing the dimensionality of the word count vectors by random mapping [5]. By using these only almost orthogonal random word vectors, the important features of the term-document matrix for LSI still preserve quite well [10], but the semantic subspace becomes much easier to compute, even with the normal SVD. The 2D SOM of documents is trained by starting with a large smoothing neighborhood and reducing it gradually, so that the densest areas in the semantic space get higher mapping accuracy and more detailed topic specifications.

With the SOM, we can try to illustrate the index, the topics and the retrieved documents in an automatic way as shown in Figure 1. This map of the document index has 1200 units to represent the document clusters in the latent semantic space. The labels of the map units are the index terms with the best semantic match to the documents of that cluster. Thus, the chosen labels can be considered as the topics of the clusters. The coloring of the map units, known as the U-matrix [13] of the SOM, shows the relative semantical distances between the neighboring units. By imagining the color as the topography of the map, the darkest blue color shows areas where the units are semantically close together, whereas yellow and red means that the neighboring units are further away like being separated by a wall. Topic hierarchies can be formed by merging the clusters near each other and by selecting more general labels to describe the contents of these "metatopics" [7]. On the contrary, the map can be zoomed (like in Figure 1) to better see the topics in an interesting neighborhood. The results of a query can be projected on this index map by showing the location of the best-matching documents (numbers from 1 to 10 in Figure 1). The marked clusters can be further analyzed by showing all the best-matching documents there or (like in Figure 1) by showing the topics of the clusters.

3 PERFORMANCE RESULTS

This indexing method has so far been tested on 4 spoken and broadcast news collections (2 French and 2 English). Unfortunately, it is not straight-forward to test how successful the indexing is. If there is test data available with perfect transcriptions, we can estimate the word error rate (WER) of the speech decoder. However, as observed in the TREC's SDR evaluations [14, 15], WER has only a weak correlation to the actual IR performance indicators. In [9] we defined a perplexity for the documents given by the index, which corresponds to the perplexity of LMs used in speech recognition. This measure is, however, dependent on normalizations and dimensions. For directly testing the IR, on the other hand, we need a set of relevant test queries and to define what documents are relevant to those queries.

Here, the performance of the indexing method was measured using the SDR evaluation data of TREC-7 (23 test queries for 3000 stories, 100 hours of speech) [14] and TREC-8 (50 test queries for 22000 stories, 550 hours of speech) [15]. Using the provided relevance judgements it was possible to define the precisions for IR results at different recall levels and thus construct the recall-precision curves [14] for different indexes [7]. For more convenient comparisons the essential performance information for the current application can be compressed, for example, to the triple (AP, P10, RP) as in Table 1. AP is the average precision over all standard recall levels, P10 is the precision at the lowest level (10% of the relevant documents, which means just the top of the document ranking), and RP is the precision when the number of the retrieved documents equals the total number of the relevant documents (R) for the query. The Table 1 indicates that compared to the baseline

TREC-7	S1		R1	
	L	T	L	T
AP %	38.1	37.4	42.9	43.4
P10 %	38	37	43	41
RP %	63	62	64	65

TREC-8	S1		R1	
	L	T	L	T
AP %	42.3	40.0	45.4	43.8
P10 %	43	41	46	44
RP %	71	67	67	66

Table 1: IR test results for the decoded speech (S1: 35.9 % WER for TREC-7 and 32.0 % for TREC-8) and reference transcripts (R1). Precisions are shown at 10 % recall level (P10), at level R (RP), and in average (AP). IR is tested for the proposed LSI+SOM index (L) and the baseline *thislIR-0.2* system (T).

index ([1], without query expansion), the suggested LSI method gives slightly better results both for the decoded speech (S1) and for the transcribed speech (R1).

4 CONCLUSIONS AND DISCUSSIONS

This paper describes a latent semantic indexing system for spoken audio. To be able to index large databases, the normal SVD analysis is preceded by a random mapping operation that drastically reduces the dimensionality of the term-document matrix. Self-organizing map is used to smooth the document vectors and to reduce the effect of speech recognition errors. To improve the ranking of the relevant documents, probabilistic indexing weights are stored in the index.

The results of the IR tests are encouraging although the obtained improvements are mainly statistical compared to the rather straight-forward baseline IR system. Matched Pairs test applied for the APs at different test queries shows that for the larger TREC-8 database the differences between the proposed system and the baseline are significant at 95 % level [7]. The overall best systems in the TREC evaluations, however, achieved 51–57 % APs for the same tests. Whereas in this paper only the information given by the spoken audio was used, the best systems exploited external text databases to pick new index terms using query and document expansions.

The increase of computations in indexing due to the proposed LSI+SOM method is rather significant compared to the baseline system. However, the indexing is still several orders of magnitude faster than the speech decoding, which is a few times realtime. And the system is much faster as well than the conventional LSA (sparse SVD without the random mapping), which gets prohibitively slow for large databases due to its computational complexity. The complexity issues for RM, SVD, SOM and indexing are discussed more in [10, 8, 7].

ACKNOWLEDGMENTS

This work was done within European Union's ESPRIT Long Term Research Project THISL and Information Society Technologies Programme ASSAVID with financial support from the Swiss Federal Office of Education and Science. The THISL partners also provided the speech decodings of the TREC data.

REFERENCES

[1] D. Abberley, D. Kirby, S. Renals, and T. Robinson. The THISL broadcast news retrieval system. In *ESCA workshop on Accessing Information in Spoken Audio*, pp. 14–19, 1999.

[2] J. Andersen. Baseline system for hybrid speech recognition on French. COM 98-7, IDIAP, 1998.

[3] S. Deerwester, S. Dumais, G. Furdas, and K. Landauer. Indexing by latent semantic analysis. *J. Amer. Soc. Inform. Sci.*, 41:391–407, 1990.

[4] T. Hofmann. Probabilistic topic maps: Navigating through large text collections. In *Proc. Symposium on Intelligent Data Analysis (IDA'99)*, 1999.

[5] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. WEBSOM - self-organizing maps of document collections. *Neurocomputing*, 21:101–117, 1998.

[6] T. Kohonen. *Self-Organizing Maps.* Springer, Berlin, 1997. 2nd extended ed.

[7] M. Kurimo. Thematic indexing of spoken documents by using self-organizing maps. RR 00-5, IDIAP, 2000.

[8] M. Kurimo. Fast latent semantic indexing of spoken documents by using self-organizing maps. In *Proc. ICASSP*, 2000.

[9] M. Kurimo and C. Mokbel. Latent semantic indexing by self-organizing map. In *ESCA workshop on Accessing Information in Spoken Audio*, pp. 25–30, 1999.

[10] C. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In *Proc. 17th ACM Symposium on the Principles of Database Systems*, 1998.

[11] G. Salton. *The SMART Retrieval System-Experiments in Automatic Document Processing.* Prentice-Hall, 1971.

[12] S.A. Shumsky. Navigation in databases using self-organizing maps. In *Kohonen Maps*, pp. 197–206. Elsevier, 1999.

[13] A. Ultsch. Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In *Kohonen Maps*, pp. 33–45. Elsevier, 1999.

[14] E. M. Voorhees and D. K. Harman. *NIST Special Publication 500-242: TREC-7.* Department of Commerce, National Institute of Standards and Technolog, http://trec.nist.gov/pubs.html, 1999.

[15] E. M. Voorhees and D. K. Harman. *NIST Special Publication 500-244: TREC-8.* Department of Commerce, National Institute of Standards and Technolog, http://trec.nist.gov/pubs.html, 2000.

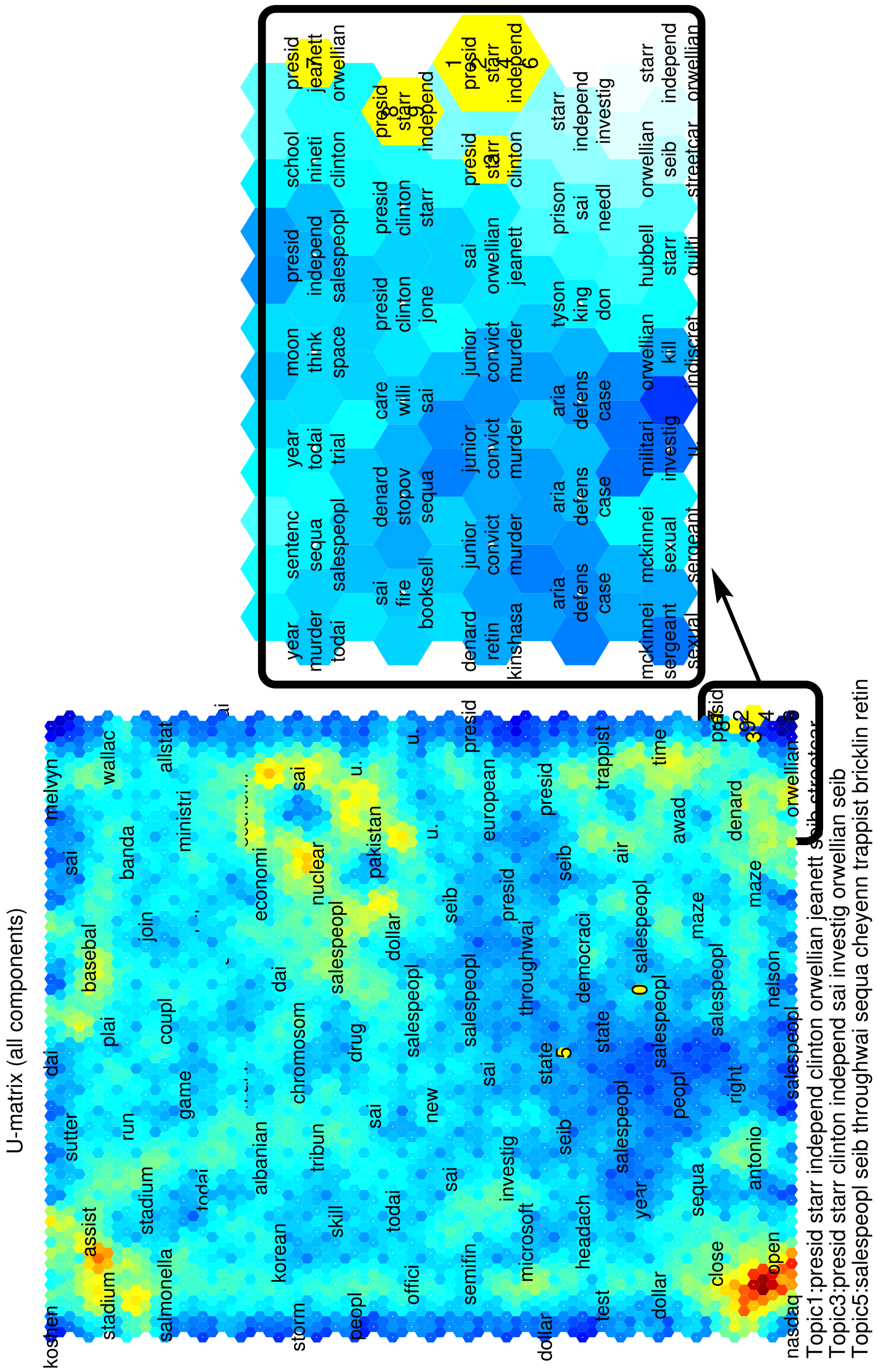

Figure 1: An U-matrix visualization of the latent semantic topics and structures found in a collection of 550 hours of North American news broadcasted during the first half of 1998 (TREC-8). The colors in this SOM show the relative semantic distance between the neighboring units. The neighbors closest together are blue and the neighbors furthest apart are red (in a gray copy the scale is from dark to light). The query for which the locations of the best-matching stories from 1 to 10 are displayed is "Lewensky" (sic.). In the zoomed display the best-matching map units are shown by magnified yellow hexagons. The labels are the word stems corresponding to the best-matching index terms (topics) for each SOM unit.