CASTSEARCH - CONTEXT BASED SPOKEN DOCUMENT RETRIEVAL

Lasse Lohilahti Mølgaard, Kasper Winther Jørgensen, and Lars Kai Hansen

Informatics and Mathematical Modelling
Technical University of Denmark Richard Petersens Plads
Building 321, DK-2800 Kongens Lyngby, Denmark

ABSTRACT

The paper describes our work on the development of a system for retrieval of relevant stories from broadcast news. The system utilizes a combination of audio processing and text mining. The audio processing consists of a segmentation step that partitions the audio into speech and music. The speech is further segmented into speaker segments and then transcribed using an automatic speech recognition system, to yield text input for clustering using non-negative matrix factorization (NMF). We find semantic topics that are used to evaluate the performance for topic detection. Based on these topics we show that a novel query expansion can be performed to return more intelligent search results. We also show that the query expansion helps overcome errors of the automatic transcription.

Index Terms— Audio Retrieval, Document Clustering, Non-negative Matrix Factorization, Text Mining

1. INTRODUCTION

The rapidly increasing availability of audio data via the Internet has created a need for automatic sound indexing. However, broadcast news and other podcasts often include multiple speakers in widely different environments which makes indexing hard, combining challenges in both audio signal analysis and text segmentation.

Access to broadcast news can be aided by topic detection methods to retrieve only relevant parts of the broadcasts. Efficient indexing of such audio data will have many applications in search and information retrieval. The spoken document indexing issue has been approached in different systems notably in the 'Speechfind' project described in [1]. This project utilizes audio segmentation as well as automatic speech transcription to retrieve relevant sub-segments.

Segmentation of broadcast news to find topic boundaries can be approached at two different levels. Starting from analysis of the audio, locating parts that contain the same speaker in the same environment can indicate story boundaries and may be used to improve automatic speech recognition performance. We have investigated this approach in [2].

The speaker segments generated through the audio analysis can then be processed through an automatic speech-to-text system to generate transcripts. Utilizing these transcripts enables a top-down segmentation based on the semantic content of the news stories. The area of topic detection in text has been widely researched for the last decade, see e.g., [3] for a presentation. Though, automatically transcribed text poses additional difficulties than manually made transcripts, due to imperfect transcriptions. We utilize non-negative matrix factorization (NMF) for document topic detection. NMF has earlier shown to yield good results for this purpose, see e.g., [4, 5].

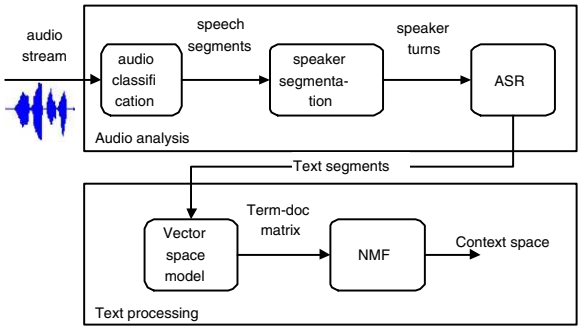


Fig. 1. The system setup. The audio stream is first processed using audio segmentation. Segments are then using an automatic speech recognition (ASR) system to produce text segments. The text is then processed using a vector representation of text and non-negative matrix factorization (NMF) to find a topic space.

2. SYSTEM OVERVIEW

The system presented here operates in two domains combining audio based processing and text based topic detection. The overall system pipeline is shown in figure 1. The audio segmentation part was described recently in [2] and will only be briefly outlined here. The text processing step will be covered in section 3.

2.1. Audio Segmentation

The basic sound representation is a feature set consisting on 12-dimensional MFCC coefficients as well as MPEG type features such as the zero-crossing rate, short-time energy, and spectral flux. The first step is to separate speech parts, excluding 'jingles'. The step is performed in a supervised classification step using a trained linear classifier. The classifier operates on 1 sec windows and detects the two classes speech and music. The audio classification step was shown to have a correct classification rate of 97.8%, see [6] for additional details and feature set evaluation.

To aid topic spotting and locate story boundaries in the news stream we use additional audio cues to segment the audio into subsegments containing only one speaker in a given environment. Broadcast news shows typically have unknown number and identities of speakers, including news anchors and guests appearing both in the radio studio and in the field. Therefore we invoke an unsupervised speaker change detection algorithm, described in detail in [2]. The algorithm is based on a 12-dimensional MFCC feature set, that is statistically summarized by vector quantization in sliding windows on both sides of a hypothesized change-point. The similarity of these

windows is then measured using the vector quantization distortion (VQD) and possible speaker changes are detected by simple thresholding. A second step is invoked for removal of false alarms inferred by the change detection algorithm. This step uses larger windows around a hypothesized change-point to yield more precise statistically models. An overall F-measure of 0.85 is found using this algorithm [6].

2.2. Automatic Speech Recognition

The sound clips produced by the segmentation step is transcribed using the Sphinx4 [7] speech recognition system. Sphinx4 is a large vocabulary speaker independent open source speech recognition system from Carnegie Mellon University. The system was set up using the pretrained acoustic and language models available on the Sphinx4 web-page [1]. Using these models Sphinx4 gives a word accuracy of $50 - 80\%$ depending on the speaking style of the speaker and the background noise level. This accuracy could be improved by training models on the material used in our experiments but this is not the focus of this work, and the performance of the model will therefore suffice.

3. TOPIC DETECTION

The result of the audio segmentation and speech recognition performed on these clips are considered as a collection of documents. The database of documents could be approached using standard indexing methods. To further increase the user friendliness we propose to use text modeling tools to spot relevant contexts of the documents. Probabilistic latent semantic analysis has shown very powerful for high-level statistical modeling of text [8]. PLSI was defined as a conditional model, which complicates generalization to news documents. We note that by modeling the joint probability the topic detection problem can be solved by non-negative matrix factorization (NMF). Xu et al. [4] have shown that NMF outperforms SVD- and eigen-decomposition clustering methods, see also [5].

3.1. Document Representation

Given the speaker documents $D = \{d_1, d_2, ..., d_m\}$ and the vocabulary set $T = \{t_1, t_2, ..., t_n\}$, the $n \times m$ term-document matrix $\mathbf{X}$ is created, where $\mathbf{X}_{i,j}$ is the count of term t_i in the speaker document d_j.

NMF factorizes the non-negative $n \times m$ matrix $\mathbf{X}$ into the non-negative $n \times r$ matrix $\mathbf{W}$ and the non-negative $r \times m$ matrix $\mathbf{H}$. This is done to minimize the objective function $J = \frac{1}{2}\|\mathbf{X} - \mathbf{WH}\|$, where $\|\cdot\|$ denotes the squared sum of the elements of the matrix.

The columns of $\mathbf{W}$ forms a r-dimensional semantic space, where each column can be interpreted as a context vocabulary of the given document corpus. Each document, columns of $\mathbf{H}$, is hence formed as a linear combination of the contexts. Usually r is chosen to be smaller than n and m.

3.2. NMF for Document Retrieval

Let N be the sum of all elements in $\mathbf{X}$. Then $\widetilde{\mathbf{X}} = \frac{\mathbf{X}}{N}$ form a frequency table approximating the joint probability of terms t and documents d

$$\widetilde{\mathbf{X}} \equiv \frac{\mathbf{X}}{N} \approx p(t, d). \tag{1}$$

[1] http://cmusphinx.sourceforge.net

Expanding on a complete set of disjoint contexts k we can write $p(t, d)$ as the mixture

$$p(t, d) = \sum_{k=1}^{K} p(t|k)p(d|k)p(k), \tag{2}$$

where $p(t|k)$ and $p(d|k)$ are identified as $\mathbf{W}$ and $\mathbf{H}$ of the NMF respectively, if columns of $\mathbf{W}$ and rows of $\mathbf{H}$ are normalized as probability distributions

$$p(t, d) = \sum_{k=1}^{K} \mathbf{W}_{t,k}\mathbf{H}_{k,d} \tag{3}$$

$$= \sum_{k=1}^{K} \frac{\mathbf{W}_{t,k}}{\alpha_k}\frac{\mathbf{H}_{k,d}}{\beta_k}\alpha_k\beta_k, \tag{4}$$

where $\alpha_k = \sum_t \mathbf{W}_{t,k}$, $\beta_k = \sum_d \mathbf{H}_{k,d}$, and $p(k) = \alpha_k\beta_k$. Thus, the normalized $\mathbf{W}$ is the probability of term t given a context k, while the normalized $\mathbf{H}$ is the probability of document d in a context k. $p(k)$ can be interpreted as the prior probability of context k.

The relevance (probability) of context k given a query string d^* is estimated as

$$p(k|d^*) = \sum_t p(k|t)p(t|d^*), \tag{5}$$

where $p(t|d^*)$ is the normalized histogram of (known) terms in the query string d^*, while $p(k|t)$ is found using Bayes theorem using the quantities estimated by the NMF step

$$p(k|t) = \frac{p(t|k)p(k)}{\sum_{k'} p(t|k')p(k')} \tag{6}$$

$$= \frac{\mathbf{W}_{t,k}p(k)}{\sum_{k'} \mathbf{W}_{t,k'}p(k')}. \tag{7}$$

The relevance (probability) of document d given a query d^* is then

$$p(d|d^*) = \sum_{k=1}^{K} p(d|k)p(k|d^*) \tag{8}$$

$$= \sum_{k=1}^{K} \mathbf{H}_{k,d}p(k|d^*). \tag{9}$$

The relevance is used for ranking in retrieval. Importantly we note that high relevance documents need not contain any of the search terms present in the query string. If the query string invokes a given subset of contexts the most central documents for these context are retrieved. Thus, the NMF based retrieval mechanism acts as a kind of association engine: "These are documents that are highly relevant for your query".

4. EVALUATION

In the following section we evaluate the use of NMF for topic detection and document retrieval.

To form a database 2099 CNN-News podcasts have been automatically transcribed and segmented into 30977 speaker documents, yielding a vocabulary of 37791 words after stop-word removal. The news show database was acquired during the period 2006-04-04 to 2006-08-09.

Based on the the database a term-document matrix was created and subjected to NMF decomposition using $K = 70$ contexts

Label	No. segments
Crisis in Lebanon	8
War in Iraq	7
Heatwave	7
Crime	5
Wildfires	1
Hurricane season	2
Other	30
Total	60

Table 1. The specific contexts used for evaluation by manual topic delineation.

... california governor arnold's *fortson agar* inspected the california mexico border by helicopter wednesday to see ...

... president bush asking california's governor for fifteen hundred more national guard troops to help patrol the mexican border but governor orville *schwartz wicker* denying the request saying...

Fig. 2. Two examples of the retrieved text for a query on 'schwarzenegger'.

producing matrices $\mathbf{W}$ and $\mathbf{H}$. The implementation of the NMF-algorithm was done using the approach of [9]. For each context the ten most probable terms in the corresponding column of $\mathbf{W}$ were extracted as keywords. Based on the keyword list each context was manually labeled with a short descriptive text string (one-two words).

For evaluation of the topic detector eight CNN-News shows were manually segmented and labeled in a subset of six contexts out of the $K = 70$ contexts identified by NMF. Segments that were not found to fall into any of six topics were labeled as 'other'. The six labels and the number of segments for each label can be seen in table 1.

4.1. NMF for Query Expansion and Segmentation

As described above the probabilistic interpretation of the NMF factorization allows query expansion by 'association'.

To illustrate the usefulness of this system let us consider a specific example. We query the system with the term 'schwarzenegger', the governor of the state of California.

The query expansion first uses eq. (5) to evaluate probabilities of the contexts given the query string. The result of the 'schwarzenegger' query produces the following three most probable contexts that were hand-labeled from the automatically generated keyword list:

- 'California Politics' $p(k|d^*) = 0.38$
- 'Mexico border' $p(k|d^*) = 0.32$
- 'Politics' $p(k|d^*) = 0.17$

Illustrating that the system indeed is able to find associate relevant topics from broadcasts in the database, consisting of data from the summer of 2006.

Traditional text indexing would return documents containing the exact term 'schwarzenegger'. This can be sufficient but using imperfect transcriptions might mean that relevant sound clips are ignored. The method presented here overcomes this problem by expanding the query onto the 'topic space', given by eq. (8). That is, documents with the highest relevance conditioned on the k topics are returned. This is useful when errors occur in automatic transcription. This is indeed the case for the 'schwarzenegger' query, where two relevant documents include wrongly transcribed versions of the word, as can be seen in figure 2. So the method compensates for transcription errors when the objective is to retrieve relevant spoken documents. These documents would have been missed by a conventional search.

To perform a more quantitative evaluation eq. (5) is used to calculate the posterior probabilities for each context given short query strings d^*. This can be used to segment a news cast into homogenous topic segments. In particular we treat a short sequence of ten words as a query string, and associate topics to these queries based on the probability $p(k|d^*)$. 'Sliding' d^* along the news cast, topic changes can be found when $\arg\max_k p(k|d^*)$ changes.

For evaluating the segmentation task we use the recall (RCL) and precision (PRC) measures defined as:

$$\text{RCL} = \frac{\textit{no. of correctly found change-points}}{\textit{no. of manually found change-points}} \quad (10)$$

$$\text{PRC} = \frac{\textit{no. of correctly found change-points}}{\textit{no. of estimated change-points}}, \quad (11)$$

where an estimated change-point is defined as correct if a manually found change-point is located within ± 5 words.

In the test we concatenated the eight manually segmented news shows and removed stop-words. Running the topic detection algorithm resulted in a recall of 0.88 with a precision of 0.44. It shows that almost every manually found change-point is found by our algorithm. On the other hand the method produces a number of false alarms. This is mostly because some of the manually found segments are quite long and include subsegments. For instance some of the segments about 'crisis in Lebanon' contain segments where the speaker mentions the US Secretary of State Condoleezza Rice's relationship to the crisis. These subsegments have a larger probability with other 'US politic' contexts, so the system will infer a change-point, i.e., infer an off-topic association induced by a single prominent name. If such events are unwanted, we probably need to go beyond mere probabilistic arguments, hence, invoke a 'loss' matrix penalizing certain association types.

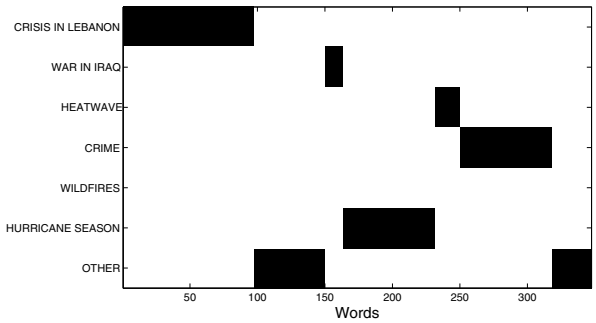
Figure 3 shows an example of the segmentation process for one of the news shows. Figure 3(a) shows the manual segmentation, while figure 3(b) shows the $p(k|d^*)$ distribution forming the basis of figure 3(c). The NMF-segmentation is consistent with the manual segmentation, with exceptions, such as a segment which is manually segmented as 'crime' but missed by the NMF-segmentation.

4.2. Topic Classification

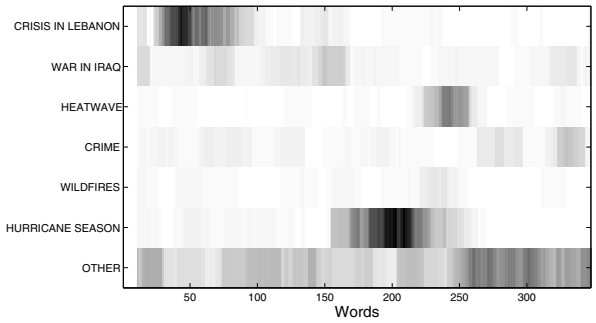
The segmentation procedure described above provides labels for all instances of d^*. As in [4, 5] we use the accuracy (AC), defined as $\text{AC} = \frac{1}{n}\sum_{i=1}^{n}\delta\big(c_m(i), c_s(i)\big)$ to quantitatively evaluate the classification performance. $\delta(x, y)$ is 1 if $x = y$, 0 otherwise. c_m and c_s denotes the manually and system labels respectively and n is the number of test strings. Using the same data as in the segmentation task we achieve an overall AC of 0.65.

The confusion matrix for the experiment is shown in table 2. The table shows that most of the errors occur when the system is classifying c1, c2, c3, and c4 as c7 ('other'). The system outputs the class 'other' when none of the six selected topic classes has the highest relevance $p(k|d^*)$.

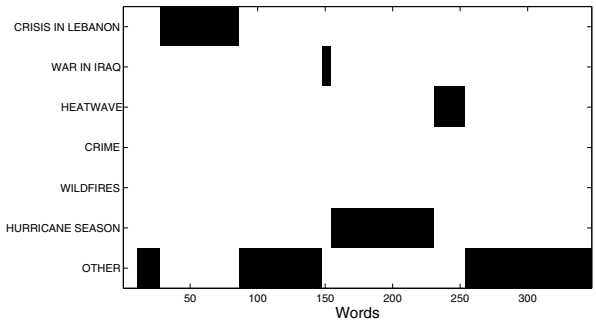
In the above classification the query string d^* is based on a sequence of ten words. If instead we use the 60 manually found segments as query string we are able to detect 53 correctly, which gives an accuracy of $\text{AC} = 0.88$.



(a) Manual segmentation.



(b) $p(k|d^*)$ for each context. Black means high probability.



(c) The segmentation based on $p(k|d^*)$.

Fig. 3. Figure 3(a) shows the manual segmentation of the news show into 7 classes. Figure 3(b) shows the distribution $p(k|d^*)$ used to do the actual segmentation shown in figure 3(c). The NMF-segmentation is in general consistent with the manual segmentation. The segment manually segmented as 'crime' is erroneously labeled 'other' by the NMF-segmentation

5. CONCLUSION

We have presented a system capable of retrieving relevant segments of audio broadcast news. The system uses audio processing to segment the audio stream into speech segments. A speech-to-text system is utilized to generate transcriptions of the speech. Furthermore a strategy for application of non-negative matrix factorization of joint probability tables allows us to retrieve relevant spoken documents in a broadcast news database. We have demonstrated that the system retrieves relevant spoken documents even though the query terms have been transcribed wrongly, hence showing that global topic models can assist interpretation of speech-to-text data. The system is fully implemented as a web demo available at: http://castsearch.imm.dtu.dk

	c1	c2	c3	c4	c5	c6	c7
c1	**370**	32	10	0	8	0	380
c2	0	**131**	1	2	0	8	52
c3	0	0	**105**	7	0	8	88
c4	0	16	2	**9**	0	0	112
c5	0	0	0	0	**13**	0	0
c6	0	0	29	0	0	**88**	0
c7	3	29	8	0	6	0	**759**

Table 2. Classification confusion matrix, where rows are manual labels and columns are estimated labels. The used classes are: (c1) crisis in Lebanon, (c2) war in Iraq, (c3) heatwave, (c4) crime, (c5) wildfires, (c6) hurricane season, and (c7) other.

ACKNOWLEDGEMENTS

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound', www.intelligentsound.org (STVF No. 26-04-0092).

6. REFERENCES

[1] J. H. L. Hansen, R. Huang, B. Zhou, M. Seadle, J. R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 712–730, september 2005.

[2] K. W. Jørgensen, L. L. Mølgaard, and L. K. Hansen, "Unsupervised speaker change detection for broadcast news segmentation," in *Proc. EUSIPCO*, 2006.

[3] J. Allan, Ed., *Topic Detection and Tracking: Event-Based Information Organization*, Kluwer Academic Publishers, Norwell, MA, 2002.

[4] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. of the 26th International ACM SIGIR*, New York, NY, USA, 2003, pp. 267–273, ACM Press.

[5] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Inf. Process. Manage.*, vol. 42, no. 2, pp. 373–386, 2006.

[6] K. W. Jørgensen and L. L. Mølgaard, "Tools for automatic audio indexing," M.S. thesis, Informatics and Mathematical Modelling, Technical University of Denmark, 2006.

[7] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Tech Report TR-2004-127, Sun Microsystems, 2004.

[8] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proc. 22nd Annual ACM Conf. on Research and Development in Information Retrieval*, Berkeley, California, August 1999, pp. 50–57.

[9] C.-J. Lin, "Projected gradient methods for non-negative matrix factorization," Tech. Rep., Department of Computer Science, National Taiwan University, 2005.