Audio Information for Hyperlinking of TV Content

Petra Galuščáková and Pavel Pecina
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics,
Prague, Czech Republic
{galuscakova,pecina}@ufal.mff.cuni.cz

ABSTRACT

In this paper, we explore the use of audio information in the retrieval of multimedia content. Specifically, we focus on linking similar segments in a collection consisting of 4,000 hours of BBC TV programmes. We provide a description of our system submitted to the Hyperlinking Sub-task of the Search and Hyperlinking Task in the MediaEval 2014 Benchmark, in which it scored best. We explore three automatic transcripts and compare them to available subtitles. We confirm the relationship between retrieval performance and transcript quality. The performance of the retrieval is further improved by extending transcripts by metadata and context, by combining different transcripts, using the highest confident words of the transcripts, and by utilizing acoustic similarity.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Search process, Selection process, Query formulation*

General Terms

Experimentation, Theory

Keywords

Speech retrieval, Automatic transcripts, Acoustic similarity

1. INTRODUCTION

Acoustic information is crucial for effective retrieval of multimedia content since the audio track contains additional information which can help identify content or genre of the recording. Speech retrieval commonly utilizes automatic speech recognition (ASR) systems to mine lexical information from audio and video recordings. Automatic transcripts are then indexed and text-based information retrieval methods are applied. A wide range of acoustic attributes could

also be utilized here: applause, music, shouts, explosions, whispers, and background noise among others.

In this paper, we explore in detail the utilization of speech transcripts in speech retrieval by analyzing three ASR systems and comparing them to human subtitles. Specifically, we address the two most significant problems associated with automatic transcripts: the restricted vocabulary used in the ASR system and the lack of reliability [11]. To broaden the vocabulary, we utilize query expansion and combination of transcripts. The precision could be increased by utilizing the only most confident words of the transcripts. In addition to the lexical information acquired from the transcripts, we also use prosodic features to improve speech retrieval. Specifically, we explore the possibility of using audio music information and acoustic similarity to enhance the retrieval process.

2. HYPERLINKING TV CONTENT

The main objective of our work is to retrieve segments similar to a given query segment from the collection of television programmes i.e. to create hyperlinks between related segments of the collection [12]. Such hyperlinks can not only bring additional entertainment value to videos, and thus serve as a specific recommendation system [3], but they can also facilitate exploratory search [2, 7] and enable users to find unexplored connections within the collection.

The queries are given as segments of recordings in our video collection. As the recordings in the collection are often long (the golf tournament lasts for more than 10 hours), we aim at retrieval of particular relevant segments of videos.

Our experiments are performed using data provided in the Search and Hyperlinking Task [4] organized in the MediaEval Benchmark 2014¹. The linking of the similar segments was explored in the Hyperlinking Sub-task of this task. Some of our experiments were submitted to the task and were evaluated officially by the task organizers; the rest of our experiments were performed after the official evaluation.

2.1 Data

We utilize a collection of BBC TV programmes provided for the task. The development data consists of 1335 hours of programmes broadcast between 01.04.2008 and 11.05.2008, and the test data consists of 2686 hours of programmes broadcast between 12.05.2008 and 31.07.2008. Apart from the videos, there were manual subtitles, automatic tran-

¹<http://www.multimediaeval.org>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

Copyright is held by the owner/author(s).

SLAM'15, October 30, 2015, Brisbane, Australia.

ACM 978-1-4503-3749-6/15/10.

<http://dx.doi.org/10.1145/2802558.2814643>.

	Words Dev. Unique	Words Dev. Total	Words Test Unique	Words Test Total	WER(%)
LIMSI All	71k	13M	84k	27M	71.8
LIMSI 1-best	66k	11M	79k	22M	57.5
LIUM	95k	9M	122k	18M	65.1
NST	55k	9M	59k	17M	58.6
Subtitles	227k	13M	313k	24M	–

Table 1: Word counts of the Search and Hyperlinking development and test sets and WER of automatic transcripts calculated on 10 randomly selected BBC TV programmes.

scripts, manually-created metadata, prosodic features [5], and additional visual features also available.

The collection contains three automatic transcripts provided by LIMSI [9], TED-LIUM [13] and NST-Sheffield [10]. Lamel [9] reports Word Error Rates (WER) of the LIMSI ASR system obtained on two data sets consisting of a mixture of English Broadcast News and Broadcast Conversation as 17.3% and 20.1% respectively. The lowest WER achieved by the LIUM transcripts on the test set consisting of TED talks was 11.1% [13]. The NST system was designed for transcribing the multi-genre BBC archive [10]. The system was evaluated on BBC data consisting of radio and television drama programmes and the lowest achieved WER was 27.1%.

As our system deploys text-based retrieval, its quality depends on the quality of used transcripts. Stated WERs are calculated for different corpora and thus could not be directly compared. To compare ASR quality on our dataset, we randomly selected 10 documents from the test set and calculated the WER on these documents. We unified the transcripts (punctuation, letter case), removed additional functional and special words (e.g. $\{fw\}$) and compared them to subtitles². The comparison of calculated WERs appears in Table 1.

The confidence of each word is given in the LIUM transcripts. LIMSI transcripts also contain several word variants occurring at the same time. Similarly to the LIUM transcripts, the confidence of each word variant is given. In the NST transcripts, only the single best word with no confidence score is given.

2.2 Evaluation

We report scores using two evaluation measures: MAP-bin, and MAP-tol [1]. Both measures are adaptations of the MAP measure and they allow a certain tolerance between the beginning of a relevant segment and the beginning of a retrieved segment. According to the MAP-bin measure, each recording is segmented into 5-minute long bins. If a retrieved segment lies in the bin together with a relevant segment, it is also considered as a relevant. According to the MAP-tol measure, a retrieved segment is considered as a relevant if it begins closer than 15 seconds from the beginning of any relevant segment.

The runs submitted in the task were evaluated manually. The top 10 results were marked as relevant or irrelevant in the crowdsourcing campaign [12]. Our additional experiments were evaluated with respect to the set of officially eval-

uated segments. Therefore, runs officially submitted to the task may be advantaged compared to the additional runs.

3. EXPERIMENTS

In order to easily retrieve relevant segments, passage retrieval was applied to the recordings. We divided the documents into 60-second long segments. A new segment was created each 10 seconds – thus created segments may overlap. The text lying within the segment boundaries was then indexed. Similarly, we created a textual query from a query segment by utilizing all words in the subtitles lying within the query segment. Text-based retrieval can then be applied to the collection.

In our experiments, we used the Terrier IR Framework³ and its implementation of the Hiemstra Language Model [6] for retrieval, with its parameter set to 0.35. We used Porter stemmer and Terrier’s stopwords list. In all experiments, we post-filtered retrieved segments and removed segments which partially overlapped either with the query segment or with another higher ranked retrieved segment.

We focused on the two most significant problems of the ASR systems: out of vocabulary words and transcript reliability. The problem with out-of-vocabulary words may be partially overcome by expansion of the lexical data segments and query segments. We utilize metadata which possibly contain names and designations which are important for a particular recording but which sparsely occur in the general vocabulary. Additionally, we expand the query segment by the content surrounding the query. We also experiment with “acoustic fingerprinting”, which obtains additional information from the music contained within the query segment. Acoustic fingerprinting could be especially helpful for hyperlinking music programmes which are likely to occur in TV archives.

To address transcript reliability we compare utilization of the first, most reliable word and all word variants, and we only use words with high confidence scores. We also present experiments with a combination of different transcripts.

Finally, we describe an experiment in which we employ “acoustic similarity” to further extend retrieval based on lexical features. The experiments are schemed in Figure 1.

3.1 Query expansion

3.1.1 Metadata and context

The BBC TV program collection contains manually-created metadata such as title and description of the program. We used these data (specifically title, episode title, description, short episode synopsis, service name, and program variant) to concatenate each data segment with metadata of the cor-

²The subtitles may contain additional words such as captions and production information at the end of the program, which may increase WER, but this disadvantage is the same for all examined transcripts.

³<http://terrier.org>

Figure 1: Scheme of the performed experiments.

Transcript	Expansion	MAP-bin	MAP-tol
Subtitles	No	0.1423	0.1216
Subtitles	Yes	0.2689	0.2465
LIMS I All	No	0.1028	0.0742
LIMS I All	Yes	0.2659	0.2297
LIUM	No	0.0873	0.0604
LIUM	Yes	0.2593	0.2547
NST	No	0.1021	0.0792
NST	Yes	0.2611	0.2237

Table 2: Results for different transcripts, for employment of metadata and context of the query. The best results for each transcript are highlighted.

responding file. Similarly, we concatenated each query segment with corresponding metadata. Additionally, we expanded the context of the query segment: using 200 seconds before and after the query segment. The results for various transcripts, expansion of the data and query segments by metadata and context are shown in Table 2. Expanding the queries and segments with keywords from metadata and context significantly⁴ increased the scores achieved by subtitles and all automatic transcripts.

3.1.2 Acoustic fingerprinting

As our collection contains musical pieces either as the main object in a query segment or in the background, the identification of an artist and a song title could provide useful additional information about the query segment. We experimented with two music fingerprinting services: EchoNest⁵ and Doreso⁶, which both offer developers’ API.

The query segments were divided into 10-second long passages; a new passage was created each second. These created passages were submitted to the EchoNest and Doreso services. EchoNest service retrieved no interpret or song title for all development queries. The Doreso service retrieved the song title and artist for 4 queries out of 30 from the development set and 10 queries out of 30 from the test set. We concatenated each query segment with the corresponding title, artist and album name, but doing so caused both retrieval scores to drop.

3.2 Transcript reliability

Variants are also available for some words in the LIMS I transcripts. We compared the employment of all available words and the case when only the most confident word was

⁴Wilcoxon signed rank test [14] at the 0.05 level was used.

⁵<http://the.echonest.com>

⁶<http://developer.doreso.com>

Transcript	MAP-bin	MAP-tol
LIMS I + LIUM	0.2706	0.2421
LIMS I + NST	0.2747	0.2306
LIUM + NST	0.2687	0.2327
LIMS I + LIUM + NST	0.2736	0.2304

Table 3: Results for combination of transcripts. Metadata and context were utilized. Best results are highlighted.

used. The MAP-bin score raised from 0.2659 when all available words were used to 0.2676 when only the most confident word was used. The MAP-tol score raised from 0.2297 to 0.2320.

We also experimented with using only the words from LIMS I and LIUM transcripts with a confidence score higher than a given threshold⁷. Although this strategy increased both scores for the development set, it did not outperform fully transcribed test data. In addition to filtering of the results based on confidence we also experimented with voting [8], which should guarantee higher reliability of the indexed words as well. However, this approach also did not prove to be helpful.

Finally, various transcripts were joined into a single document – we merged all words in selected transcripts which lie within particular segment boundaries⁸. These experiments are tabulated in Table 3.

3.3 Acoustic similarity

Acoustic similarity between short audio tracks could not only help to retrieve segments containing identical acoustic tracks (e.g. signature tunes and jingles) but could also be useful for detection semantically related segments (e.g. segments containing action scenes and music). Eight prosodic features were provided in the collection (energy, loudness, voice probability, pitch, pitch direction, direction score, voice quality, and harmonics-to-noise ratio). We calculated acoustic similarity between data segments and query segments based on these features and combined it with the textual similarity between segments based on lexical information acquired from subtitles.

The prosodic features were calculated every 10ms. We took all overlapping sequences of 10 prosodic 8-dimensional feature vectors appearing up to 1 second from the beginning of the query segment and found the most similar sequences of the vectors near the beginning of each data segment.

Similarity between data and query vector sequences was calculated as the sum of differences between the corresponding vectors of the sequence. These vector differences were calculated as the sum of the absolute values of the differences between the corresponding items of the prosodic vectors. The difference of each item of the prosodic vector was normalized to have component values between 0 and 1 to ensure that all prosodic features have equal weights.

The highest acoustic similarity between segments was linearly combined with text-based similarity score calculated on the subtitles.

⁷The threshold confidence scores were tuned with the development data. The threshold on the LIUM transcripts was set to 0.7 and 0.8 for MAP-bin and MAP-tol measures, respectively, and to 0.8 and 0.6 on the LIMS I transcripts.

⁸All words in all merged segments were indexed.

Acoustic Similarity very slightly decreased the MAP-bin score from 0.2689 to 0.2687, but the MAP-tol score raised from 0.2465 to 0.2473.

4. DISCUSSION

Expanding the query with keywords from metadata and context clearly improves the quality of the retrieval. When no metadata are utilized, the performance corresponds with the WER relatively well, and the subtitles clearly outperform the automatic transcripts. However, the metadata and context produce much higher relative improvement to the automatic transcripts than to the subtitles. The improvement on the subtitles is 90% and 103% for MAP-bin and MAP-tol measures, respectively, and 197% and 322% on the LIUM transcripts. When metadata and context are utilized, the LIUM transcripts even outperform the subtitles in MAP-tol measure; the overall highest MAP-tol score is achieved in this case.

Words missing from the vocabulary are a big problem for the automatic transcripts. Number of unique words in transcripts is almost three times smaller than in subtitles. Moreover, these low frequency words are expected to be the most informative for the information retrieval. The MAP-tol score achieved on the LIUM transcript, which have the largest number of unique words among the ASR transcripts, show that expansion using metadata and context may substantially reduce this disadvantage, even though the transcripts are trained on different data type and have a relatively high WER.

The overall highest MAP-bin score is achieved using the union of the LIMS and NST transcripts; the combination even outperforms the results achieved with the subtitles. Combination of the transcripts is generally helpful (e.g. the combination of the LIUM and NST transcripts is in terms of the MAP-bin score significantly better than both LIUM and NST transcripts), but as the LIUM transcripts already achieve high MAP-tol score, this score drops when they are combined with other transcripts. Although the employment of words with higher confidence values did not improve test results, the employment of the most confident word variants of the LIMS transcripts helps slightly.

Application of the audio similarity is promising, but it needs to be further explored. Despite the computational restrictions, the MAP-tol measure increased. The employment of music information did not increase the scores, but this may be partially caused by an evaluation which did not previously pay attention to this aspect of the retrieval.

5. CONCLUSIONS

We explored several ways to use audio information in the retrieval of similar segments of television programmes. Employment of context and metadata improve retrieval quality significantly. The results were also improved by a combination of different transcripts and by utilization of only the most confident words of the transcripts instead of all available word variants. Our results were also slightly improved by using calculated audio similarity.

6. ACKNOWLEDGMENTS

This research has been supported by the project AMALACH (grant n. DF12P01OVV022 of the program NAKI of the

Ministry of Culture of the Czech Republic), the Czech Science Foundation (grant n. P103/12/G084), and the Charles University Grant Agency (grant n. 920913).

7. REFERENCES

- [1] R. Aly, M. Eskevich, R. Ordeman, and G. J. F. Jones. Adapting Binary Information Retrieval Evaluation Metrics for Segment-based Retrieval Tasks. *CoRR*, abs/1312.1913, 2013.
- [2] C. A. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. Multimodal Reranking of Content-based Recommendations for Hyperlinking Video Snippets. In *Proc. of ICMR*, pages 225–232, 2014.
- [3] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath. The YouTube Video Recommendation System. In *Proc. of RecSys*, pages 293–296, 2010.
- [4] M. Eskevich, R. Aly, D. N. Racca, R. Ordeman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking Task at MediaEval 2014. In *Proc. of MediaEval*, Barcelona, Spain, 2014.
- [5] F. Eyben, F. Wengler, F. Gross, and B. Schuller. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proc. of ACMMM*, pages 835–838, Barcelona, Spain, 2013.
- [6] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, Enschede, Netherlands, 2001.
- [7] B. Jou, H. Li, J. G. Ellis, D. Morozoff-Abegauz, and S.-F. Chang. Structured Exploration of Who, What, when, and Where in Heterogeneous Multimedia News Sources. In *Proc. of ACMMM*, pages 357–360, 2013.
- [8] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On Combining Classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3):226–239, Mar. 1998.
- [9] L. Lamel. Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data. In *Proc. of HLT*, pages 1–8, Tartu, Estonia, 2012.
- [10] P. Lanchantin, P.-J. Bell, M.-J.-F. Gales, T. Hain, X. Liu, Y. Long, J. Quinell, S. Renals, O. Saz, M.-S. Seigel, P. Swietojanski, and P.-C. Woodland. Automatic Transcription of Multi-genre Media Archives. In *Proc. of SLAM Workshop*, pages 26–31, Marseille, France, 2013.
- [11] M. A. Larson and G. J. F. Jones. *Spoken Content Retrieval: A Survey of Techniques and Technologies*, volume 5 of *Found. Trends Inf. Retr.* Now Publishers Inc., Hanover, MA, USA, 2012.
- [12] R. Ordeman, M. Eskevich, R. Aly, B. Huet, and G. J. F. Jones. Defining and Evaluating Video Hyperlinking for Navigating Multimedia Archives. In *Proc. of WWW (Companion Volume)*, pages 727–732, 2015.
- [13] A. Rousseau, P. Deléglise, and Y. Estève. Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling And More TED Talks. In *Proc. of LREC*, pages 3935–3939, Reykjavik, Iceland, 2014.
- [14] F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.