Processing and Linking Audio Events in Large Multimedia Archives: The EU *inEvent* Project

H. Bourlard[1,2], *M. Ferras*[1], *N. Pappas*[1,2], *A. Popescu-Belis*[1],
S. Renals[3], *F. McInnes*[3], *P. Bell*[3], *S. Ingram*[4], *M. Guillemot*[4]

[1]Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne, Switzerland
[3]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK
[4]Klewel SA, Martigny, Switzerland

{bourlard,ferras,andrei.popescu-belis,pappas}@idiap.ch
{s.renals,fergus.mcinnes,peter.bell}@ed.ac.uk
{sandy.ingram,mael.guillemot}@klewel.com

Abstract

In the *inEvent* EU project [1], we aim at structuring, retrieving, and sharing large archives of networked, and dynamically changing, multimedia recordings, mainly consisting of meetings, videoconferences, and lectures. More specifically, we are developing an integrated system that performs audio-visual processing of multimedia recordings, and labels them in terms of interconnected "hyper-events" (a notion inspired from hyper-texts). Each hyper-event is composed of simpler facets, including audio-video recordings and metadata, which are then easier to search, retrieve and share. In the present paper, we mainly cover the audio processing aspects of the system, including speech recognition, speaker diarization and linking (across recordings), the use of these features for hyper-event indexing and recommendation, and the search portal. We present initial results for feature extraction from lecture recordings using the TED talks.

Index Terms: Networked multimedia events; audio processing: speech recognition; speaker diarization and linking; multimedia indexing and searching; hyper-events.

1. Introduction

Databases and information management systems have been in reactive mode for the last decade, trying to keep up with novel and rapidly evolving applications, data characteristics, and data volumes. However, databases continue to extend the relational model to deal with standard "static" data management problems. Search engines derived their initial inspiration from text retrieval and have been developed around the bag-of-words model and the link structure of the web. More recently there has been intense activity on developing search engines to deal with dynamic data streams (such as Twitter and newswires), deployed within search engines such as Google and Bing.

As the amount of audio and video data found on the web has exploded, systems which allow searching for audio-visual content have been deployed, such as Google Videos or Yahoo! Video Search. Video repositories such as YouTube or Dailymotion have deployed their own search solutions. More recently, lecture repositories such as TED [2], Videolectures.Net, or Khan Academy have followed suit. However these systems are still largely based on text retrieval and rely on textual metadata, tags added by users, or text found on the same (or a linked) webpage. Any link structure for multimodal search relies solely on textual links rather than implicit links found thanks to the media content.

Over the last 10-15 years, there has been an intense research activity on semantic indexing of multimedia content using visual, audio, and text cues (see e.g. [3, 4]). There have also been some deployed audio search engines based on speech recognition, which enable content-based search and retrieval of podcasts and videos, for example Everyzing[1] and Blinkx[2].

However, most current systems do not address a number of key issues, including (1) disparate, heterogeneous, data sources capturing audio-visual data taken at different locations at different times to represent a holistic situation; (2) multimodal resources that represent social and communicative interactions (such as videoconferences, meetings and symposia); (3) dynamic, rapidly evolving multimodal streams; and (4) implicit links and connections contained within the multimodal streams, rather than easily accessible textual links and metadata.

In the present paper, we discuss our current efforts towards automatically analyzing, structuring, linking and retrieving multimedia networked objects consisting of archives of rich and complex A/V documents resulting from meetings, videoconferences, symposia, and lectures. Exploiting initial proposals presented in [5] and [6], an archive of multimedia recorded events is here represented in terms of a collection of *hyper-events*[3] accommodating all necessary attributes (either automatically extracted or manually annotated), including structural, temporal and spatial information, as well as contextual and social information. The resulting archive should be accompanied by tools that automatically reorganize events to satisfy different viewpoints and naturally incorporate new data types.

In the *inEvent* project, hyper-events are thus being used as a primary structure for organizing and accessing complex objects like multimedia recordings. This paper describes our initial steps towards the analysis of those hyper-events for feature extraction (speech recognition in Section 2 and diarization in Section 3), the use of features for linking and recommending similar hyper-events (Section 4), and the portal allowing users to access hyper-event repositories (Section 5). The system

[1]http://www.everyjoe.com/
[2]http://www.blinkx.com
[3]In reference to the hyper-texts used for static documents.

components are tested on TED lectures, on lectures recorded by Klewel, or on the AMI Corpus [7] (for formal evaluation), thus providing a promising proof of concept for *inEvent*'s vision.

2. Speech recognition

Automatic speech recognition (ASR) derives from the audio signal a text representation of the words that were spoken. Usually the input is segmented into utterances (delimited by pauses or changes of speaker) and the output consists of a transcription of each utterance. Variations on this paradigm include keyword spotting (where only selected words are transcribed) and the output of multiple hypotheses, in the form of an N-best list or a confusion network, to handle uncertainties as to what was said.

For the purposes of *inEvent*, ASR output is an important data stream both for searching *across* recordings, to find those most relevant to a given query, and for searching *within* a recording, to find time intervals in which particular words and phrases were spoken. For searching across recordings, it will often be best to derive an intermediate representation, such as a summary or a list of keywords, from the ASR output, rather than use the transcript directly. Indeed, as results in Section 4 show, using the entire transcript is less useful for indexing than the talk title or a short description. This may be more so when the transcript is derived automatically rather than manually and contains recognition errors. However, for searching within a recording, direct use of the transcript is more likely to be appropriate.

The speech recognition system should ideally be trained on data similar to the recordings to which it is to be applied. This applies both to the acoustic characteristics of the data (speaker characteristics, noise level, microphone type, etc.) and to the vocabulary and style of the spoken content.

The system currently in use in *inEvent* is a variant of the system developed for the IWSLT 2012 ASR evaluation [8, 9] and was trained primarily on recordings and transcripts of TED talks [2]. For further details of the modeling see [10].

2.1. Acoustic modeling

The recognition system adopts a hybrid modeling approach, in which HMM observation probabilities are computed using a deep neural network (DNN), as described in [10]. The current system does not incorporate MLAN features [10, 11], but it is planned to add these in future versions.

The core acoustic model training set was derived from 813 TED talks dating prior to the end of 2010. The recordings were automatically segmented, giving a total of 153 hours of speech. Each segment was matched to a portion of the manual transcriptions for the relevant talk using a lightly supervised technique described in [12]. For this purpose, we used existing acoustic models trained on multiparty meetings.

Three-state left-to-right HMMs were trained on features derived from the aligned TED data, and a re-alignment of the training segments and transcriptions was carried out, following which around 143 hours of speech remained for the final estimation of state-clustered cross-word triphone models. The resulting models contained approximately 12,000 tied states, with 16 Gaussians per state. The state tying from these (HMM-GMM) models was used in the final hybrid models, as described in [10].

The first pass of recognition uses a 7-layer hybrid DNN trained on PLP features (13-dimensional vectors with first, second and third order differential coefficients, projected to 39 dimensions using an HLDA transform). The first-pass output is used to estimate a single CMLLR transform [13] for each speaker, which is used to generate speaker-normalized features. The second pass uses a 6-layer hybrid DNN trained on speaker-normalized features from the training data.

This configuration is essentially as in the fifth row of Table 4 in [10] (baseline hybrid + SAT, giving word error rates of 18.6% and 17.6% on the dev2010 and tst2010 data sets), but with an improved language model and lattice rescoring in the final pass as described below.

2.2. Language modeling

The language models for the IWSLT 2012 evaluation were obtained by interpolating individual modified Kneser-Ney discounted LMs trained on the small in-domain corpus of TED transcripts (2.4M words) and seven larger out-of-domain sources. The out-of-domain sources were Europarl (v7), News Commentary (v7) and News Crawl data from 2007 to 2011. A random 1M sentence subset of each of News Crawl 2007-2010 was used, instead of the entire available data, for quicker processing. The total amount of out-of-domain data used was about 166M words. The vocabulary was fixed at 60,000 words, including all words found in the TED training set plus the most frequent additional words in the other sources.

The language models in the current system were obtained by interpolating the IWSLT evaluation LMs described above with the LM built for the 2009 NIST Rich Transcription evaluation (RT09), based on a range of data sources including conversational speech and meetings [14].

The system generates word lattices using a trigram model, and rescores them with a 4-gram model for the final output.

2.3. Current and future work

Work is in progress on improving the language models trained for the IWSLT 2012 evaluation. As mentioned above, the amount of data used to train the existing models was restricted because of time constraints, and it was noted that other participants in the evaluation had obtained better LMs by using more data and by refinements including domain adaptation and recurrent neural network modeling [15]. Subsequent experiments [16] have shown WER reductions of about 2% absolute due to using the NICT trigram LM [15] instead of the original UEDIN trigram LM of [9], with 4-gram and factored RNN models giving further improvements. Current work within *inEvent* is focused on applying similar techniques to obtain an improved baseline LM. This will then be taken as the starting point for topic adaptation based on generating queries from the first-pass ASR output and running web searches to retrieve relevant text [17]. It may also be helpful to use any text associated with the recording (e.g. from slides or lecture notes) for LM adaptation [18, 19].

In order to perform speaker adaptation, the ASR system requires a speaker diarization stage. In the present system this is based on the diarization module of the AMIDA system [14], applied separately to each recording. It should be possible to improve on this by performing speaker linking across recordings as described in Section 3.

Recordings of interactive meetings, as obtained for instance from a videoconferencing system, pose a particularly difficult challenge for ASR, since they typically contain more frequent changes of speaker, higher levels of noise and more disfluent speech than lecture-style recordings. Work will be required on both acoustic modeling and language modeling in order to extend the *inEvent* system successfully to data of this type.

3. Speaker diarization and linking

Speaker diarization technology structures audio data in terms of "who spoke when". In a project like *inEvent*, such information is used to enrich the semantic annotation of events to enable speaker-based search and recommendation. Speaker diarization can also drive higher-level semantic annotation by fusion with other technologies such as speech recognition, video processing and social signal processing.

Speaker diarization within the *inEvent* project must cope with specific challenges:

- A large amount of data to be processed in an appropriate time, although off-line is acceptable for a search and retrieval application.
- A large number of speakers are present in the data set, with some of them appearing in multiple recordings. Diarizing the whole data set, i.e. structuring the speaker space across all recordings, would be more than desirable, as opposed to per-recording operation of the current diarization solutions.
- The data is dynamic and the algorithms should be able to work incrementally as new data are available.
- Large variability in the recording quality and acoustic conditions, with special attention to robustness to variations of noise and room acoustics across recordings.
- Weak priors on the number of participants and interaction structure so that, ideally, a single diarization set-up works fine for different scenarios.

We have developed a diarization and linking method that is able to both uniquely identify the speakers across the data set and find the segments of each recording where each speaker is speaking. This task could be otherwise addressed by diarizing the concatenation of all recordings, but the computational cost is prohibitive given current capabilities. We opt instead for a two-stage approach, involving intra-session speaker diarization, followed by speaker linking across sessions. This system is described more in-depth in [20].

3.1. Speaker diarization

A standard speaker diarization system obtains within-recording speaker clusters using agglomerative clustering at the acoustic observation level. The speaker clusters are given a set of start and end times and a unique speaker identifier within each recording. This stage benefits from a reference model fitted to the recording conditions so that fine differences between speakers are accurately detected. It also deals with a tractable number of speakers. We use the Information Bottleneck diarization system [21] obtaining state-of-the-art performance on meeting scenarios with small computational load. This system uses information theoretic principles to find speaker clusters that are maximally informative w.r.t. a set of relevance variables, namely Gaussian mixture posterior probabilities, while keeping the cluster representation as compact as possible.

3.2. Speaker linking

A second agglomerative clustering algorithm takes as input the speaker clusters generated by the speaker diarization system, and structures the speaker space of the whole data set. The resulting speaker clusters are then given a unique speaker identifier across the data set. Speaker clusters are given a compact and robust representation obtained via Joint Factor Analysis (JFA) [22, 23]. JFA models the speaker and channel variabilities around a reference model, i.e. the Universal Background Model (UBM), obtaining speaker factor posterior distributions that are assumed to be speaker-dependent multivariate Gaussian. These objects are then linked across all recordings in the whole data set. Such speaker factor representation has been shown to be robust to across-recording variation in speaker recognition applications. The hyper-parameters of the JFA model are trained on around 50 hours of the Augmented Multiparty Interaction (AMI) meeting corpus [7] involving 130 speakers.

The clustering step takes advantage of the Gaussian properties of the objects to be clustered. The Ward algorithm [24] seeks to minimize the increase of the total within cluster variance after merging two clusters while the Lance-Williams recursion [25] enables an efficient implementation.

Amongst the similarity measures we explored, including the cosine distance of mean vectors and the symmetrized Kullback-Leibler divergence, the Hotelling t-square statistic stood out as being the most stable and performing. This measure is the multivariate version of the two-way Student-t statistic used for testing the hypothesis that the means of two Gaussian samples are different. Under the assumption that both Gaussian distributions share the same covariance matrix, this measure has the form of the Euclidean distance between spherified Gaussian distributions, therefore matching the assumptions of the Lance-Williams recursion.

It is expected that speaker clusters naturally arise during the agglomerative clustering process. In this work, we assume that speaker clusters can be simply found by thresholding the distance values in the clustering dendrogram.

Given that no labeled data is available for the Klewel and TED data sets, we evaluated the speaker diarization and linking system on a subset of the AMI corpus. These data involve meetings with 4 participants recorded using far-field microphones.

Table 1 shows the results of these experiments for two subsets involving low and high channel variability, LCV and HCV entries. For both sets the linking approach reduces the within-recording Diarization Error Rate (DER), a gain coming from further clustering speakers within the same recording.

Regarding the across-recording DER, measuring the performance of both diarization and linking stages together, similar or even lower error rates are obtained, whereas the complexity of the task has enormously increased. These numbers show that the linking stage is properly detecting speaker entities in the data set. Nonetheless, the absolute performance is dependent on the initial speaker diarization performance. The number of speakers estimated for the whole data set is close to the correct one for the low channel variability data set whereas it is significantly higher for the high channel variability data set.

System	Data set	#Spk	wr/ar DER(%)
Dia	LCV	—	24.5/
Dia+Link	LCV	58	21.7/23.6
Dia	HCV	—	27.6/
Dia+Link	HCV	86	26.8/28.0

Table 1: Speaker diarization results on the LCV and HCV data sets involving 56 speakers and 8 and 24 channels respectively. Columns 3 and 4 show the detected number of speakers and the within-recording/across-recording DER.

4. Indexing for recommendation

One of the main uses of audio features extracted from multimedia events is in information retrieval (IR) applications. These features can be complemented by features extracted from lecture or meeting metadata, such as title and speaker(s). In the *inEvent* project, we have specified two types of lecture recommendation tasks, and focused initially on the first one [26]. In the *personalized recommendation task* we aim to predict whether lectures will be interesting or not for the users [27], given their previous binary ratings, or more simply to predict the N most interesting ones (top-N task) [28]. In the *generic recommendation task*, the users' history of ratings is not available, and the goal is to predict the most similar items to a given one (non-personalized top-N recommendation). The latter task also amounts to building similarity links between hyper-events, based on all their facets.

The focus on this task was also influenced by the availability of an online repository of audiovisual recordings, the TED lectures [2], made available under a Creative Commons license. This makes possible audio, video and text processing (as in Section 2 above), along with testing recommendations against preferences expressed by users. We have recently made available the TED metadata and user profiles with ratings and comments as a public set for lecture recommendation benchmarking[4].

4.1. Recommending multimedia objects

The audio features and the metadata are used within three types of methods for personalized recommendation: (i) content-based (CB) methods using vector space similarities; (ii) collaborative filtering (CF) methods using ratings; and (iii) combined methods [26]. When using a vector space model for textual features, each TED talk d_j can be represented as a feature vector $d_j = (w_{1j}, w_{2j}, \ldots, w_{ij}, \ldots)$, where each position i corresponds to a word of the vocabulary, extracted from the textual attributes, including e.g. the title, speaker name, description, or transcript. The weights w_{ij} can be computed using various models, e.g. Boolean or TF-IDF coefficients. The talks' feature vectors can then be linked by defining a similarity measure, e.g. cosine similarity. We also investigated more sophisticated approaches, namely semantic vector spaces using LSI, LDA, Random Projections [29] and Explicit Semantic Analysis (ESA) [30].

4.2. Experiments: features and scores

Using cross-validation, we ranked the features (including metadata and audio-based ones) with respect to relevance to the personalized recommendation task with CB models. We used ground truth feature values from TED for oracle performance. Figure 1 displays the ranking of features and their combinations (see caption for acronyms), ordered by their overall relevance across several content-based models, i.e. indicating which features perform well over *all* methods. Alternatively, the optimal features found specifically for *each* method are listed in Table 2.

The results show that the human-made description of talks (DE), the title (TI), and their combinations with other features (TIDE, TIDE.RTT, and TIDE.TESP.RTT) are the most useful features for CB personalized recommendations. Knowledge of the speaker (SP) is useful too. The lowest performing features were the name of the TED event (TE) and the related themes assigned by TED experts (RTH), which presumably lack specificity. The transcript (TRA) is ranked lower than average, potentially due to the noise introduced by its large vocabulary.

[4] https://www.idiap.ch/dataset/ted

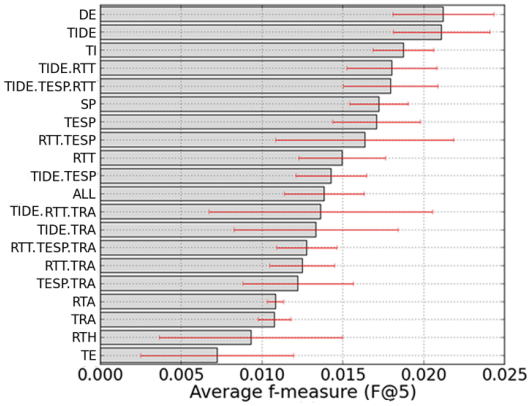

Figure 1: Ranking of features based on the decreasing average of f-measure (F@5) over all content-based recommendation methods. The atomic features are: title (TI), description (DE), related tags (RTA), related themes (RTH), transcript (TRA), speaker (SP) and TED event (TE). The combined features composed by two atomic ones are: related tags and themes (RTT), title and description (TIDE), TED event and speaker (TESP). The remaining features are combinations of the previously defined features separated by '.' symbol.

In terms of the best scores, all the semantic-based CB methods except LDA outperform significantly the TF-IDF baseline (t-statistic, $p < 0.05$): 11% improvement for LSI, 7.6% for RP and up to 64% by ESA (best method). The scores obtained appear to be low, however they are in line with previous works on top-N recommendation task (e.g. [28, 31]).

We then compared recommendation methods in a setting where users' ratings were available and hence CF methods could be used. The CF methods outperformed the CB ones, and a combined method using a neighborhood model, user/item biases and TF-IDF similarity achieved reasonable performance compared to pure CF by utilizing only the popularity bias.

The content of the TED talks as described by the metadata is important for personalized recommendations as was demonstrated in two different settings. Another promising type of information are user-generated comments or reviews as we discuss in [32]. TED data contains valuable ground-truth to evaluate quantitatively multimedia recommendations (generic and personalized) and, given that they have the same structure with hyper-events, the methods are also applicable to the *inEvent* project. In the future, we will work on improving hybrid recommender systems, especially by exploiting the rich multimodal content of the TED dataset. More advanced learning

Method	**Optimal Features**	**Performance (%)**		
		P@5	R@5	F@5
LDA	Title, desc., TED event, speaker (TIDE.TESP)	1.63	1.96	1.78
TF-IDF	Title (TI)	1.70	2.00	1.83
RP	Description (DE)	**1.83**	**2.25**	**2.01**
LSI	Title (TI)	**1.86**	**2.27**	**2.04**
ESA	Title, description (TIDE)	**2.79**	**3.46**	**3.08**

Table 2: Optimal features for content-based methods found using 5-fold cross-validation on the training set. Scores in bold are significantly higher than TF-IDF ones (t-test, $p < 0.05$).

models such as matrix factorization could improve the fusion of CB and CF information. We will also assess the variation of performance when automatic processing is used for extracting all features: e.g. ASR, speaker detection, or summarization.

5. Portal

The results of analyzing, indexing and linking hyper-events are only relevant to end-users when they are presented in an online portal offering efficient access to the event recordings.

While researchers are aware of some of the requirements of such a portal, we have also conducted, within *inEvent*, a survey with about 40 participants selected from relevant professional categories (e.g., conference organizers). The results shed light on the importance of having an online portal for managing, sharing, and replaying recorded events. The *inEvent* portal should enable contextual user interactions; provide easy means to navigate within and across hyper-events; and offer search and recommendation services to help users find their needle in a haystack. While the *inEvent* portal is currently under development, its main intended features are discussed hereafter.

5.1. Visibility and impact of hyper-events

A common goal for conference and meeting organizers is to have their multimedia content ubiquitously accessible in high quality, so as to have a high impact on the community and maximize their return on investment. Thus, it is important to ensure widespread and persistent access to event recordings, so that previous participants peruse the repository, and new ones join at any time. For maximal visibility, hyper-events should be uniquely addressable, crawlable by search engines, and accessible to third-party services for interaction, event dissemination and advertisement. Statistical graphs showing how the number of views per hyper-event evolves over time (with peak dates) serve as important impact measures.

While some public events aim at reaching the widest audience possible, some others such as internal enterprise meetings should have limited access. In these cases, we should ensure that hyper-events with restricted access can still be accessible by local search engine, though not available on the Web.

The role of social media features in spreading information and turning users from passive consumers to interactive content producers is crucial. Features such as commenting, sharing, rating, embedding videos, and tagging aim at engaging the user community and enriching hyper-events with contextual interactions. This is invaluable in increasing the accuracy of linking and recommendation using collaborative filtering approaches.

5.2. Efficient search and navigation

An important challenge to the *inEvent* portal is to offer a user-friendly and efficient way to navigate across hyper-events and within a single event in order to find segments of interest. Here, features extracted from hyper-events, along with metadata and various similarity links are crucial. Visual graphs where hyper-events are taken as nodes and the various links between them as edges are currently being explored for efficient navigation across recommended hyper-events. For instance, when a user brings a specific hyper-event in focus, it is displayed at the center of the graph along with its segments. Other related events and event segments are displayed around the central event helping end-users discover other potentially interesting events and watch them with a single click. Zoomed word clouds and/or slides appear when a user hovers the mouse over a specific

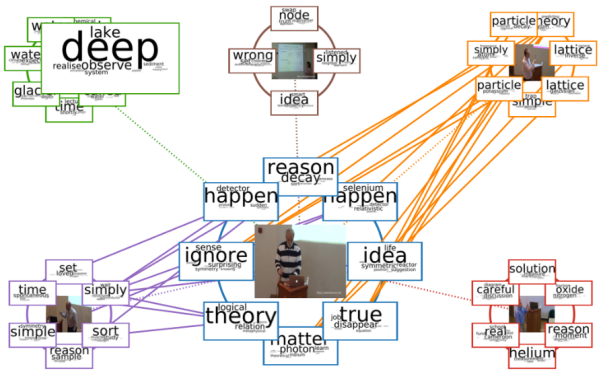

Figure 2: Visual interface for navigating hyper-events.

event, giving the user an overview of event content and helping him/her decide what to watch. (See Figure 2.)

With respect to navigating and searching within a single hyper-event, as several state-of-the-art lecture browsers do, the *inEvent* portal will respond to search requests with rich content, e.g. by highlighting the requested keywords in the transcript, slides, and media segments. Additional visual cues will emphasize the participants who uttered the keywords.

Finally, ensuring a satisfactory user experience requires a cross-browser and cross-platform multimodal player, which can render videos in different quality formats based on the available bandwidth and the target device. In the case where a hyper-event contains more than one video (e.g. the speaker and the projection screen), synchronized display should be provided.

6. Conclusions

A system for indexing multimedia lecture and meeting recordings was proposed that exploits the notion of "hyper-events" as a means to represent the multi-faceted structure of events accompanied by rich multimedia recordings and related metadata. The resulting model can integrate audio and video features, as well as social features to perform search along different axes, as well as providing generic or personalized recommendations based on the similarity of hyper-events, including their viewing profiles. The core of the model is an indexing mechanism based on automatically extracted audio features[5] such as speech-to-text outputs, together with speaker diarization and linking labels, that allows searching and recommendation within a new type of multimedia archive. The resulting system has been evaluated on several datasets – from TED, Klewel, and the AMI corpus – providing a promising proof of concept for the *inEvent* approach.

7. Acknowledgments

This work was supported by the European Union (Networked Media and Search Systems) under the *inEvent* project (Accessing Dynamic Networked Multimedia Events), contract number ICT-287872. The authors gratefully thank the EU for their financial support, and all project partners for a fruitful collaboration. More information about *inEvent* is available from the project web site `http://www.inevent-project.eu/`.

[5] Video features are also under consideration, but they are not discussed in this paper.

8. References

[1] "Accessing dynamic networked multimedia events," in *EU Project No. 287872, Network Media*. [Online]. Available: https://www.inevent-project.eu/

[2] "Riveting talks by remarkable people, free to the world." [Online]. Available: http://www.ted.com/

[3] W. Adams, G. Lyengar, M. Naphade, C. Neti, H. Nock, and J. Smith, "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP Journal on Applied Signal Processing*, vol. 2003:2, pp. 1–16, 2003.

[4] M. Larson, F. de Jong, W. Kraaij, and S. Renals, Eds., *ACM Transactions on Information Systems, Special issue on searching speech*. New York, NY: ACM Press, 2012, vol. 30, no. 3.

[5] U. Westermann and R. Jain, "Events in multimedia electronic chronicles (e-chronicles)," *Int. J. Semantic Web Inf. Syst.*, pp. 1–23, 2006.

[6] R. Jain, "Eventweb: Developing a human-centered computing system," *IEEE Computer*, vol. 41(2), pp. 42–50, 2008.

[7] J. Carletta and M. Lincoln, "Data collection," in *Multimodal Signal Processing–Human Interactions in Meetings*, S. Renals, H. Bourlard, J. Carletta, and A. Popescu-Belis, Eds. Cambridge University Press, 2012.

[8] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. International Workshop on Spoken Language Translation*, Hong Kong, Dec. 2012.

[9] E. Hasler, P. Bell, A. Ghoshal, B. Haddow, P. Koehn, F. McInnes, S. Renals, and P. Swietojanski, "The UEDIN systems for the IWSLT 2012 evaluation," in *Proc. International Workshop on Spoken Language Translation*, Hong Kong, Dec. 2012.

[10] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[11] P. Bell, M. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P. Woodland, "Transcription of multi-genre media archives using out-of-domain data," in *Proc. IEEE Workshop on Spoken Language Technology*, Miama, Florida, USA, Dec. 2012.

[12] A. Stan, P. Bell, and S. King, "A grapheme-based method for automatic alignment of speech and text data," in *Proc. IEEE Workshop on Spoken Language Technology*, Miama, Florida, USA, Dec. 2012.

[13] M. Gales, "Maximum likelihood linear transforms for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 75-98, 1998.

[14] T. Hain, L. Burget, J. Dines, P. Garner, F. Grezl, A. Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[15] H. Yamamoto, Y. Wu, C. Huang, X. Lu, P. Dixon, S. Matsuda, C. Hori, and H. Kashioka, "The NICT ASR system for IWSLT 2012," in *Proc. International Workshop on Spoken Language Translation*, Hong Kong, Dec. 2012.

[16] P. Bell, H. Yamamoto, P. Swietojanski, Y. Wu, F. McInnes, C. Hori, and S. Renals, "A lecture transcription system combining neural network acoustic and language models," in *Proc. Interspeech*, Lyon, France, Aug. 2013.

[17] I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke, and O. Cetin, "Web resources for language modeling in conversational speech recognition," *ACM Transactions on Speech and Language Processing*, vol. 5, 2007.

[18] H. Yamazaki, K. Iwano, K. Shinoda, S. Furui, and H. Yokota, "Dynamic language model adaptation using presentation slides for lecture speech recognition," in *Interspeech*, 2007, pp. 2349–2352.

[19] P. Maergner, A. Waibel, and I. Lane, "Unsupervised vocabulary selection for real-time speech recognition of lectures," in *Proc. ICASSP*, 2012, pp. 4417–4420.

[20] M. Ferras and H. Bourlard, "Speaker Diarization and Linking of Large Corpora," in *IEEE Spoken Language Technology Workshop*, 2012.

[21] D. Vijayasenan, F. Valente, and H. Bourlard, "Information Theoretic Approach to Speaker Diarization of Meeting Data," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[22] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[23] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2008.

[24] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[25] G. N. Lance and W. T. Williams, "A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems," *Computer Journal*, vol. 9, pp. 373–380, 1967.

[26] N. Pappas and A. Popescu-Belis, "Combining content with user preferences for TED lecture recommendation," in *11th Int. Workshop on Content Based Multimedia Indexing*, Veszprém, Hungary, 2013.

[27] G. Shani and A. Gunawardana, "Evaluating recommendation systems," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer, 2011.

[28] P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-n recommendation tasks," in *Proceedings of the fourth ACM conference on Recommender Systems*, ser. RecSys '10, Barcelona, Spain, 2010.

[29] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *in Knowledge Discovery and Data Mining*. ACM Press, 2001, pp. 245–250.

[30] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, ser. IJCAI'07, Hyderabad, India, 2007.

[31] R. Pan, Y. Zhou, B. Cao, N. Liu, R. Lukose, M. Scholz, and Q. Yang, "One-class collaborative filtering," in *8th Int. Conf. on Data Mining*, Pisa, Italy, 2008, pp. 502–511.

[32] N. Pappas and A. Popescu-Belis, "Sentiment analysis of user comments for one-class collaborative filtering over TED talks," in *Proceedings of the 36th ACM SIGIR Conference on Research and Development in Information Retrieval, Short Papers*, Dublin, Ireland, 2013.