Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM 2013)

Marseille, France, August 22-23, 2013

Session 5 : Multimedia applications and corpus

Keynote speaker: Sam Davies, BBC R&D

- **Abstract:** In this talk we will present an overview of our work on the BBC's World Service Archive. This project uses automatic speech recognition and a novel technique for topic identification & disambiguation from noisy transcripts to enable automatic semantic tagging of programmes. Of course such processing across large archives is hard to scale so we'll present some of our work towards tackling this issue. We will also present our research on attempts to classify the entire BBC's archive of radio and television programmes broadcast since 1922. This project looks to classify programmes by creating new metadata from an analysis of programme content. This will primarily focus on the work we have done in identifying semantic content from audio (including music), along with a brief overview of our work on affective indexing. Here we will briefly introduce our multimodal work on identifying the mood or emotional component of a programme, focussing on mood identification from music, speech and non-speech sounds.
- **Biography:** Sam Davies joined BBC R&D in 2007 working on a variety of projects including high frame rate television, object tracking in sporting events and image recognition. Since 2009 he has been working on the Multimedia Classification project, which has been identifying new techniques for metadata generation from audio and video content in the BBC archive. This work has resulted in prototypes which offer unique ways to analyse the semantic and affective, or emotional, content of audio, visual and text documents.

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.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

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.

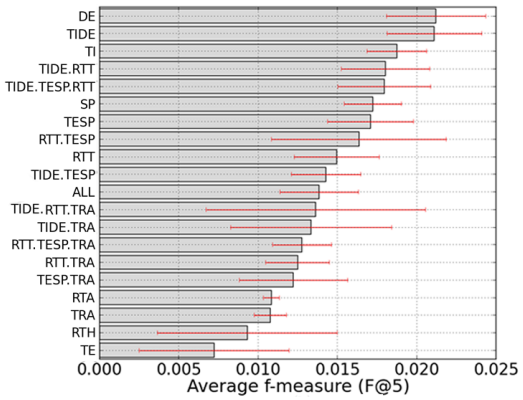


Figure 1: Ranking of features based on the decreasing average of f-measure (F@5) over all content-based recommendation methods. The atomic features are: title (TI), description (DE), related tags (RTA), related themes (RTH), transcript (TRA), speaker (SP) and TED event (TE). The combined features composed 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, $p < 0.05$).

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

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

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

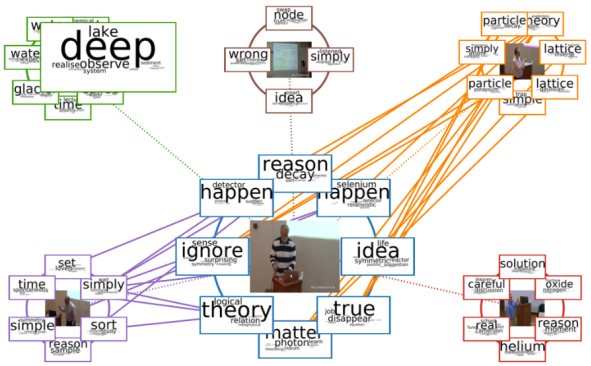


Figure 2: Visual interface for navigating hyper-events.

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.

Audio Concept Ranking for Video Event Detection on User-Generated Content

Benjamin Elizalde[1]*, Mirco Ravanelli*[2]*, Gerald Friedland*[1]

[1]International Computer Science Institute, 1947 Center Street,
Berkeley, CA 94704, USA
[2]Fondazione Bruno Kessler, via Sommarive 18, 38123 Trento, Italy

benmael@icsi.berkeley.edu, mravanelli@fbk.eu, fractor@icsi.berkeley.edu

Abstract

Video event detection on user-generated content (UGC) aims to find videos that show an observable event such as a wedding ceremony or birthday party rather than an object, such as a wedding dress, or an audio concept, such as music, speech or clapping. Different events are better described by different concepts. Therefore, proper audio concept classification enhances the search for acoustic cues in this challenge. However, audio concepts for training are typically chosen and annotated by humans and are not necessarily relevant to a specific event or the distinguishing factor for a particular event. A typical ad-hoc annotation process ignores the complex characteristics of UGC audio, such as concept ambiguities, overlap, and duration. This paper presents a methodology to rank audio concepts based on relevance to the events and contribution to the ability to discriminate. A ranking measure guides an automatic selection of concepts in order to improve audio concept classification with the goal to improve video event detection. The ranking aids to determine and select the most relevant concepts for each event, to discard meaningless concepts, and to combine ambiguous sounds to enhance a concept, thereby suggesting a focus for annotation and a better understanding of the UGC audio. Experiments show an improvement of the audio concepts mean classification accuracy per frame as well as a better-defined diagonal in the confusion matrix and a higher relevance score. In terms of accuracy, the selection of top 40 audio concepts using our methodology outperforms the highest-accuracy-based selection by a relative 17.56% and a frame-frequency-based selection by 5.74%. In terms of relevance to the events, the ranking-based selection provided the highest score.

Index Terms: event detection, audio concept, user generated content, acoustic video processing.

1. Introduction

Video event detection aims to identify videos with a semantically defined event, such as a marriage proposal. This task is implicitly multimodal because events are characterized by audio-visual cues. Multimedia detection has been explored by computer vision using different features and techniques. However, audio has been under-explored, and state-of-the-art audio-based techniques do not yet provide significant assistance to its video counterpart. Audio, however, can sometimes be more descriptive than video, especially when it comes to the descriptiveness of an event. For instance, the audio cue can quickly allow one to determine whether or not a marriage proposal was successful. Thus, there is great importance in exploring techniques to improve the use of audio for video event detection.

There have been several approaches to audio-based video event detection for UGC data. Approaches in general employ only low-level features [1] [2]. However, there are also higher-level approaches that employ audio concepts for video event detection, motivated by the idea that different events are better described by different concepts. There are techniques that automatically derive audio concepts. An example is a system [3] which extracts audio units automatically with a diarization system to create an audio concept vocabulary. A similar example is a system in [4] that defines an automatic audio concept vocabulary with a Random Forest (RF) algorithm. However, these abstract representations may or may not map to a specific humanly understandable sound, such as clapping or the buzzing of a power tool. An example of an approach with annotated audio concepts for video event detection is [5] [6]. Whether these concepts are abstract or not, they define an acoustic fingerprint that distinguishes an event from their cohorts. The relation of concepts and events can be exemplified with a language analogy as stated in [3] where concepts can be seen as words and events as ideas. The paper shows that events are defined by different distributions of concepts. Therefore, improving the classification performance of concepts enhances the detection performance of events. Following this research line, the paper [7] aims to improve audio concept classification on UGC.

Nowadays UGC videos can provide massive amounts of training data, because the videos are widely available. Ad-hoc annotations of audio concepts for video event detection on UGC videos present three main issues. One is to ignore the intrinsic characteristics of UGC, where a concept could be in the presence of background noise, be overlapped with one or more concepts, have a short duration, be unintelligible for the annotator and have acoustic ambiguities with other concepts. The second is that audio concepts for training are typically chosen and annotated by humans and are not necessarily relevant to a specific event or the distinguishing factor for a particular event. The last issue lies in the performance of the audio concept classification by the technology employed. Adding audio concept annotations alone do not help as much as in other tasks such as speech detection, where in general the more annotated speech the better the detection performance. Take for instance a set of audio concepts that can be classified with high accuracies; if the concepts are not relevant to the events, they will be of little help to discriminate between events. On the other hand, let's assume we have a relevant and unique set of an event's audio concepts, which are not classified with reliable accuracies, then the concepts would be of little help to show evidence of the event detection. Therefore, the need to define a selection procedure that addresses the issues is presented in order to maximize the usage of current au-

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

Table 1: There are audio concepts annotations of at least 10 videos from each event.

Code	Event
E001	Attempting a board trick
E002	Feeding an animal
E003	Landing a fish
E004	Wedding ceremony
E005	Working on a woodworking project
E006	Birthday party
E007	Changing a vehicle tire
E008	Flashmob gathering
E009	Getting a vehicle unstuck
E010	Grooming an animal
E011	Making a sandwich
E012	Parade
E013	Parkour
E014	Repairing an appliance
E015	Working on a sewing project

dio concept annotations and understand the UGC audio better.

This paper presents a methodology to rank audio concepts based on relevance to the events and contribution to the ability to discriminate. The ranking guides an automatic or user-based selection of concepts in order to improve audio concept classification for video event detection. The ranking aids to determine and select the most relevant concepts for each event, discard meaningless concepts, combine ambiguous sounds to enhance a concept, thereby suggesting a focus for annotations. The paper also provides an analysis on the UGC audio concept annotations.

The content of the paper is structured as follows. Section 2 presents the UGC video and the audio concept annotations for the experiments. Section 3 details the ranking methodology. Section 4 describes the audio concept classification system and the experiments. Section 5 continues with the results and expands the understanding of the UGC audio characteristics. Lastly, Section 6 states the conclusion and future work.

2. UGC Video and Annotations Sets

The video set used for the audio concept annotations is the NIST TRECVID MED 2012, which contains UGC videos. The 2012 corpus consists of 150,000 videos of about three minutes each. The audio from the videos contains environmental acoustics, overlapped sounds, and unintelligible audio among other characteristics. The annotations are based on the Event Kits subset. Table 1 contains a summary of the events.

The annotation set from SRI-Sarnoff consists of manually labeled sounds of 291 videos. The videos belong to the 15 events of the MED 2012 Event Kits dataset for a total of 11.6 hours. In total there are 28 audio concepts shown in Table 2, which attempt to describe distinctively the events.

The annotation set from CMU [8] consists of manually labeled environmental acoustics of 216 videos taken from MED 2012, totaling 5.6 hours There are at least 10 annotated videos for each of the 15 events from MED 2012. The result is a set of 42 audio concepts shown in Table 3. The main goal of the annotations was to create labels for audio segments that exist solely in the audio domain.

Table 2: List of 28 audio concepts annotated by SRI-Sarnoff in alphabetical order.

1	audio of wedding vows	15	instructional speech
2	bagpipes	16	landing after a jump
3	blowing out candles on a cake	17	laughing
4	board hitting surface	18	marching band
5	cheering	19	metallic clanking noises
6	childrens voices	20	music
7	clapping	21	noise of passing cars
8	clinking	22	power tool whine
9	conversational speech	23	rolling
10	crowd noise	24	sewing machine sound
11	dancing singing in unison in a group	25	singing
12	drums	26	someone giving a speech
13	group dancing	27	word how spoken
14	group walking	28	word tire spoken

3. Ranking Methodology

The ranking methodology is an iterative process that is divided in four steps. The first step is to calculate the relevance of the audio concepts based on how rare or common the concepts are to a specific event and to the rest of the events. The second step is to run our Audio Concept Classification system and measure the classification performance for each concept. The third step is to calculate the ranking of the audio concepts by considering the results from step one and two for each concept. Finally the fourth step consists of deciding whether a concept should be merged with another or discarded. The process iterates until the desired final quantity of concepts is reached.

3.1. Step 1: Compute relevance

The relevance of a concept to an event is expressed by the well known algorithm of Term Frequency - Inverse Document Frequency (TF-IDF) [9]. The raw frequency is the number of times a term*t* occurs in a specific document *d*. To prevent a bias with unbalanced documents, the raw frequency is divided by the maximum raw frequency of any term in the document. The TF is defined by the equation 1.

$$TF(t,d) = \frac{f(t,d)}{max\{f(w,d) : w \in d\}} \quad (1)$$

The IDF tells you whether a word is common or rare across the documents. It is the result of taking the logarithm from the division of the total number of documents by the number of documents containing the term. If the term is not in the corpus, the division will be zero, thus we add *1*. The IDF can be defined by the equation 2.

$$IDF(c,D) = log\frac{|D|}{1 + |\{d \in D : c \in D\}|} \quad (2)$$

A high TF-IDF score is reached by a high term frequency in the given document and a low document frequency of the term in the whole collection of documents; the scores therefore tend to filter out common terms. In our methodology a term

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

Table 3: List of 42 audio concepts annotated by CMU in alphabetical order.

1	anim bird	15	engine light	29	rustle
2	anim cat	16	engine quiet	30	scratch
3	anim goat	17	hammer	31	scream
4	anim horse	18	human noise	32	singing
5	applause	19	knock	33	speech
6	bang	20	laugh	34	speech not english
7	beep	21	micro blow	35	squeak
8	cheer	22	mumble	36	thud
9	child	23	music-sing	37	tone
10	clap	24	music	38	washboard
11	clatter	25	phone	39	water
12	click	26	power tool	40	whistle
13	crowd	27	processed	41	white noise
14	engine heavy	28	radio	42	wind

corresponds to a frame from an audio concept and a video event category corresponds to a document.

3.2. Step 2: Measure performance

The classification performance (CP) of the technology employed should be considered to let the system decide which concepts are more meaningful and distinguishable, along with the limitations of using a determined audio concepts set. In this paper a raw classification accuracy per frame metric is included in the ranking equation 3, but it could be substituted by any other metric that evaluates the concept CP. In this step the confusion matrix for the list of concepts is also computed in order to determine the confusability of each concept in respect to the others.

3.3. Step 3: Compute ranking

The ranking represents the relevance and classification performance for each audio concept. A higher *Rank* score means more relevance to the events and it can be represented by equation 3. The ranking is a single score for each audio concept that consists of multiplying the TF, times the IDF, times the CP across the events.

$$Rank(c, D) = (TF(c, d) \cdot IDF(c, D) \cdot CP(c)) \quad (3)$$

3.4. Step 4: Merge or discard

Once the ranking scores are computed for each audio concept comes the decision as to whether to merge or discard the lowest ranked concept. The lowest ranked concept *C-low* would be merged with the corresponding most confusable concept *C-conf* according to the confusion matrix. The cohort concepts are merged because *C-low* has low relevance and is not discriminating and distinguishable enough. The concept with higher relevance and accuracy that absorbed *C-low* will provide the name to the new resulting concept *C-merged* and will keep its corresponding annotation data. The audio classification system is run again and if the classification accuracy of *C-merged* increases, then it remains as it is. The ranking process continues the next iteration removing *C-low* from the list, but keeping its annotation data. In case the accuracy of the *C-merged* did not increase, then *C-low* is not merged and instead is discarded from the list along with its annotation data. Once again the process continues with its next iteration until the desired number of concepts is reached.

4. Experimental Setup

This section describes the classification system and details the most relevant experiments.

4.1. Audio Concept Classification System

The audio concept classification system is based on a Neural Network approach because it has demonstrated high performance on a similar task where it discriminates well between different sounds called phonemes [10] [11]. The system employs the Parallel Neural Network Trainer TNet [12] technology from Brno University of Technology. The Neural Network (NN) architecture is basic and is the first step to move on to Deep Learning, it consists of two hidden layers with 1,000 neurons each and sigmoid activation functions. The extracted acoustic features are the typical Mel-Frequency Cepstral Coefficients (MFCCs) C0-C12, with energy included, for a total of 13 dimensions. Each feature frame is computed using a 25 ms hamming window, with 10 ms frame shifts. The neural network was fed, after a mean and variance normalization step, by the specified features using a context window of nine consecutive frames. The output layer, whose softmax-based neurons dimensionality is equal to the number of audio concepts to classify. More specifically, for the training phase a stochastic gradient descent optimizing cross-entropy loss function was used. The learning rate was updated by the "newbob" algorithm: It is kept fixed at LR=0.002 as long as the single epoch increment in cross-validation frame accuracy is higher than 0.5%. For the subsequent epochs, the learning rate is being halved until the cross-validation increment of the accuracy is inferior to the stopping threshold 0.1%. The NN weights and biases are randomly initialized and updates were performed per blocks of 1024 frames.

4.2. Experiments

The objective of the experiments consists in selecting the top 40 audio concepts that provide the best trade-off between classification performance and relevance to the events. The reason for choosing 40 is that out of the 70, this is the largest number of concepts that our system was able to classify with more than one percent of accuracy. The first experiment consists of using a concept set from a selection based on highest-accuracy. The 70 concepts are fed into the audio concept classification system and then sorted to select the top 40 with highest classification accuracy. The reason for this selection is because intuitively it will lead to a high overall concepts accuracy. The second experiment uses a set based on a high frame-frequency selection. The 70 concepts annotations are analyzed and then the concepts are sorted based on the quantity of frames. The selection is motivated because concepts with more frames will most likely have longer durations or be more common, which makes them easier for the system to classify them, and more important they will have more training data available. Lastly the third experiment employs a selection based on the ranking presented in this paper.

The annotations add up to 17.2 hours and are separated into training and test. The training set contains 90% of the annotations for a total of 15.48 hours. The test set consists on the

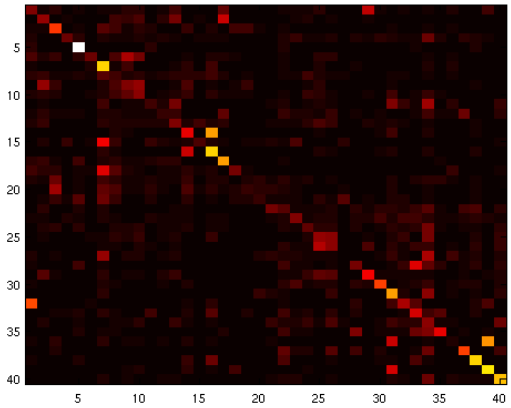


Figure 1: *The confusion matrix based on the top 40 highest-accuracy set, shows dark regions in the center of the diagonal.*

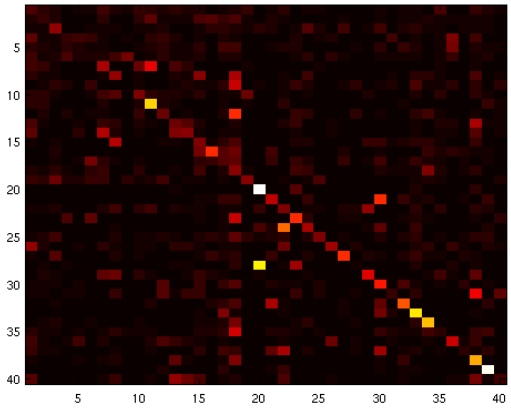


Figure 2: *The confusion matrix based on the top 40 frame-frequency set, shows dark regions on the up-left part of the diagonal.*

other 10% for a total of 1.72 hours. The classification accuracy per frame is evaluated by comparing the label from the frame's highest posterior against its corresponding label from the ground truth.

In order to provide a baseline for the three experiments against using the total number of concepts, the audio concept classification system is trained with the 70 concepts. The overall mean classification accuracy per frame is 11.5 % with a random guess of 1.42%.

5. Results and Analysis

This section presents the results from the experiments and expands the understanding of the UGC audio characteristics derived from our experiments and an analysis of both concept annotation sets.

5.1. Results

The classification performance of the three experiments is shown in the second column of Table 4. The first experiment with a highest-accuracy-based selection has an overall mean accuracy per frame of 20.38%, while the second experiment with a frame-frequency-based selection has 18.33%. The third experiment using the ranking-based set shows 21.55%. The selection of the top 40 audio concepts using our methodology outperforms the highest-accuracy-based selection by a relative 17.56% and the frame-frequency-based by a relative 5.74%.

The level of relevance to the events of the three sets of audio concepts is shown in the third column of Table 4. The score for each set is the normalized log TF-IDF, which consisted of three steps: First, the TF-IDF scores for the 40 concepts are computed as in steps one and two from Section 3. Second, the log of the TF-IDF score is computed. Finally, on the third step, a normalization is applied to the three scores, where the highest possible value of the three sets equals to one, and the other two are proportional. The lower the value, the lower the overall relevance of the set to the 15 events. The log and the normalization steps are meant to provide a more human understandable comparison. The ranking-based selection provided the highest relevance score for the 15 events, with an improvement of 17% in respect to the highest-accuracy set and 10% in respect to the frame-frequency set.

The confusion matrix is a table that allows the visualization of the accuracy performance of the system, in other words, it shows the confusability of each concept in respect to the others. Each column of the matrix represents the instances of the predicted concept, while each row represents the instances of the actual concept. A better defined diagonal means less ambiguities and higher accuracy, hence a more distinguishable and distinctive set of concepts.

The confusion matrix of the highest-accuracy set experiment is shown in Figure 1, the frame-frequency-based matrix is in Figure 2 and the one from the ranking-based set is in Figure 3.

In terms of usage of the annotated data, the 70 audio concepts comprehend about 683 minutes. The frame-frequency-based selection used 664 minutes, while the highest-accuracy selection uses 577 minutes and the ranking-based used 668 minutes. Our approach uses slightly more minutes or frames than the highest-accuracy set, which means that most of the information of the annotations is been used.

The results confirm that our methodology provided the best overall classification accuracy, the least concepts confusability and the best relevance of the audio concepts to the 15 events.

5.2. Analysis of the audio concepts

This section intends to aid the understanding of the UGC audio. The following includes an analysis of the annotation sets regarding concept overlap and duration. In addition, there is an analysis of the concepts merging and discarding step from our methodology to explain concept ambiguity.

The video events are described by a set of different sounds that occur throughout the recording therefore making it possible that one or more concepts occur at the same time, resulting in an overlap. The annotations has 38% of audio overlapping with one or more concept. The most common types of overlap are music and other audio concepts except speech 35% of the time, speech and other concepts except music 13% and speech and music 4%. The rest of the overlap types complete the total with 48%. The situation of having three or more annotated overlaps is rare and it accounts for less than 3% of the audio. It is important to mention that there could be other concepts that overlap

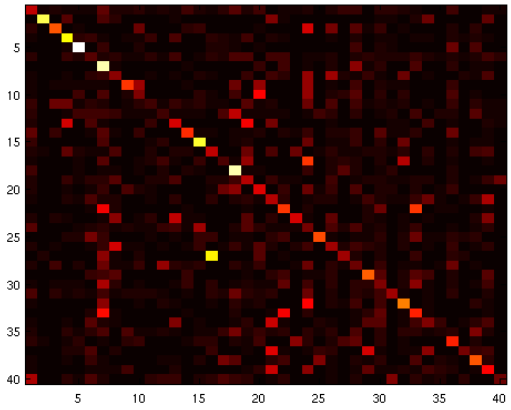


Figure 3: *The confusion matrix based on the top 40 ranked set, shows a better defined diagonal than the other two sets.*

Table 4: The ranking-based selection shows the best overall mean accuracy and the highest relevance for the top 40 audio concepts.

Selection based on	Mean accuracy	TF-IDF score
Highest-accuracy	20.38 %	0.83
Frame-frequency	18.33 %	0.90
Ranking	21.55 %	1

and are not annotated.

The nature of the concepts in the events and their duration is diverse. The annotations show that the average duration of the trials or segments is about one second. For speech trials, 38% lasts less or equal than one second and 30% lasts less or equal than two seconds, but more than one second. Examples of audio concepts with long duration are music with trials of up to 380 seconds or crowd noise of up to 260 seconds. Examples of short duration concepts are beep with trials as short as 0.8 seconds or clinking as short as 0.5 seconds. In [8] indicated that for the CMU annotations, shorter trial durations have lower accuracies, and longer duration trials have better accuracies, which was confirmed in this research with the inclusion of the SRI-Sarnoff annotations. For example, music was detected with about 90% accuracy and crowd noise with 40%, while beep and clinking resulted in less than 2% accuracy.

Discarding concepts alone intuitively suggests an improvement in accuracy. Depending on the selection process, different audio concepts will be discarded, thus affecting the overall classification in different ways. Our iterative procedure does not discard concepts for the sake of them, instead it could merge audio concepts, resulting in a more analytical usage of the annotated audio. Experiments one and two discarded the lowest 30 concepts according to their selection type. The iterative procedure from the third experiment had 30 iterations, discarded 16 concepts and merged 14. Examples of concepts discarded are: blowing out candles on a cake (SRI), clinking (SRI), dancing singing in unison in a group (SRI).

Both of the annotation sets have unique characteristics, focus and annotators. Hence, even though some of the concepts have the same or similar logical name there is no reason to assume that they should be considered as the same concept. In our methodology, merging redundant concepts from different annotation sets could sometimes make sense to the user such as cheer (CMU) and cheering (SRI) or laugh (CMU) and laughing (SRI). Nevertheless, there are other situations where is not as logical to merge sounds. Audio concepts sometimes overlap and one of them may have more "prominent" acoustic characteristics than others such as volume, pitch, duration, etc. Take for instance, the concepts group dancing (SRI) and music (SRI). The first concept is overwhelmed by music (sometimes added by the user), which has higher prominence, thus significantly decreasing the classification accuracy of the concept. The overlap information can be extracted from the annotations, but not the prominence level of the concepts involved. More complicated is when merged sounds do not have a logical semantic relation, but they could make sense from the audio concept classification system perspective. Examples are squeak (CMU) and white noise, which are broadband sounds, or thud (CMU) and click (CMU), which are impulsive sounds, or animal cat (CMU) and scream (CMU), which have similar pitch. As part of the evolution of our work we would like to include user-intervention as prior information to figure out its impact on the results of the merging process. We understand that technology and events can change and whenever this happens the iterative ranking process could be re-applied using the original set of annotations.

6. Conclusions

The research shows that the ranking methodology aids the selection of audio concepts with the best trade-off between relevance to the event and classification accuracy. The methodology discards less relevant and less accurately detected concepts and merges ambiguous sounds to enhance a concept. More important is that the ranking serves to maximize the usage of current sound concepts annotations. The improvement in classification accuracy improves the classification of concepts which provide a more reliable evidence for video event detection. The selection of top 40 audio concepts using our methodology outperforms a highest-accuracy-based selection by a relative 17.56% and a frame-frequency-based selection by 5.74%. In terms of relevance to the events, the rank-based selection provided the highest relevance score, with 17% more than the highest-accuracy-based selection and 10% more than the frame-frequency-based selection. Furthermore, the ranking suggests the audio concepts that can be enhance by more annotations and the concepts that are less relevant to the technology. Future work involves using the classification posteriors output for video event detection. The output may be used for audio segmentation or as a semantic feature, both options can feed a video event detection system.

7. Acknowledgements

Thanks to Adam Janin for his advice. Thanks to Ajay Divakaran for the SRI-Sarnoff annotations.

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusion contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

8. References

[1] R. Mertens, H. Lei, L. Gottlieb, G. Friedland, and A. Divakaran, "Acoustic Super Models for Large Scale Video Event Detection," in *ACM Multimedia*, 2011.

[2] X. Zhuang, S. Tsakalidis, S. Wu, P. Natarajan, R. Prasad, and P. Natarajan, "Compact audio representation for event detection in consumer media," in *INTERSPEECH*. ISCA, 2012.

[3] B. Elizalde, G. Friedland, H. Lei, and A. Divakaran, "There is No Data Like Less Data: Percepts for Video Concept Detection on Consumer-Produced Media," in *ACM International Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis at ACM Multimedia*, 2012.

[4] P.-S. Huang, R. Mertens, A. Divakaran, G. Friedland, and M. Hasegawa-Johnson, "How to put it into words - Using random forests to extract symbol level descriptions from audio content for concept detection," in *ICASSP*, 2012.

[5] Q. Jin, P. F. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based Video Retrieval Using Audio," in *Proceeding of the 13th Annual Conference of the International Speech Communication Association*, 2012.

[6] S. Pancoast, M. Akbacak, and M. Sanchez, "Supervised Acoustic Concept Extraction for Multimedia Event Detection," in *ACM International Workshop on Audio and Multimedia Methods for Large-Scale Video Analysis at ACM Multimedia*, 2012.

[7] A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, and B. Raj, " Audio Event Detection from Acoustic Unit Occurrence Patterns ," in *ICASSP*, 2012, pp. 489 – 492.

[8] S. Burger, Q. Jin, P. F. Schulam, and F. Metze, " Noisemes: Manual Annotation of Environmental Noise in Audio Streams ," Tech. Rep., 2012.

[9] S. Robertson, "Understanding inverse document frequency: On theoretical arguments for idf," *Journal of Documentation*, vol. 60, p. 2004, 2004.

[10] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," in *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[11] A. rahman Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[12] K. Vesely, L. Burget, and F. Grezl, "Parallel Training of Neural Networks for Speech Recognition," in *Proceeding of Interspeech*, 2010.

Segmental-GMM Approach based on Acoustic Concept Segmentation

Diego Castan[1]*, Murat Akbacak*[2]

[1]University of Zaragoza, Spain
[2]Microsoft, Sunnyvale, CA, USA

dcastan@unizar.es, murat.akbacak@ieee.org

Abstract

The amount of multimedia content is increasing day by day, and there is a need to have automatic retrieval systems with high accuracy. In addition, there is a demand for event detectors that go beyond the simple finding of objects but rather detect more abstract concepts, such as "woodworking" or a "board trick." This article presents a novelty approach for event classification that enables searching by audio concepts from the analysis of the audio track. This approach deals with the *acoustic concepts recognition* (ACR) creating a trained segmentation instead a fixed segmentation as *segmental-GMM* approach with broad concepts. Proposed approach has been evaluated on NIST 2011 TRECVID MED development set, which consists of user-generated videos from the Internet, and has shown a EER of 40%.

Index Terms: Multimedia event detection(MED), acoustic concept recognition, segmental-GMM

1. Introduction

In recent years, there has been a growing demand for high-accuracy multimedia retrieval systems due to the popularity of the video-sharing websites. For a multimedia retrieval task, video features can determine the general content of a video. However, the audio track of the video can also be critical. Consider the case of a tennis match video where a special event, like a new point, may occur. Audio analysis provide a complementary information to detect this specific event (detecting applause or cheering) that would be significantly more difficult to detect with image/video analysis. The Text Retrieval Conferences Video Retrieval Evaluation (TRECVID) addresses the problem of *Multimedia Event Detection* (MED) requiring a system that can search user-submitted quality videos for specific events [1].

Different applications for acoustic processing on multimedia videos have recently been described in the literature. These applications have been used as acoustic concept detectors in different scenarios. In [2] and [3], authors developed an SVM-based system and an HMM-based system, respectively, to classify different acoustic sounds (e.g., steps, door slams, or paper noise) in the meeting room environment. Both approaches use the CHIL-2007 database in which the acoustic concepts are isolated and recorded in a controlled environment [4].In the multimedia content analysis domain, most of the studies are concentrated on finding small events or objects rather than entire concepts. Very good summaries are provided by [5] and [6]. However, spoken concepts approaches are commonly used to detect multimedia events like [7] and [8].

Audio concept extraction approaches explored under different multimedia retrieval and content analysis projects can be grouped into two categories: (1) unsupervised and (2) supervised approaches from the perspective of modeling acoustic concepts. In the first group, one popular unsupervised approach is the Bag-of-Audio-Words (BoAW) method. In this approach, all frame-level features are clustered via vector quantization (VQ), and then VQ indices are used as features within a classifier to model audio content ([9, 10]). Other unsupervised approaches are focused on segmenting the audio track, and clustering the segments to form atomic sound units and then word-like units [11, 12], or modeling the segments with i-vectors [13] or GMM super-vectors [14] which are methods borrowed from speaker identification. In the second group of approaches, audio concept/event models are trained using annotated data [15, 16]. For example, in [15], fixed-duration segments are represented with segmental-GMM vectors where each element in the vector is a GMM score calculated from a pretrained GMM that corresponds to an annotated concept label. In [16], authors model acoustic concepts by training SVMs on 10sec audio segments which are annotated with generic concept labels (e.g., indoor vs. outdoor), and they use detected acoustic concept labels as features for multimedia event detection task. Some systems employ a combination of different approaches like in [17] where authors combine automatic speech recognition with broad-class acoustic concepts. Although the first group of approaches has the advantage of not requiring labeled acoustic event/concept data, these approaches do not present semantic labels to allow semantic searches. This is an important aspect for tasks such as multimedia event detection when the number of examples for multimedia event types becomes quite small. Therefore supervised acoustic concept detectors will be useful to tackle this problem.

This paper presents a specific study with two approaches to model five broad acoustic concepts as a MED features: segmental-GMM vectors [15] as a baseline, and a set of features based on *Acoustic Concepts Recognition* with HMMs. The broad acoustic concepts were chosen to describe sounds of different nature (people sounds, machine noises ...) and be able to model general concepts to provide a tool for retrieval information with no prior knowledge of specific acoustic events. The first part of this paper shows the classification accuracy over the isolated broad concepts. Secondly, an experiment with two extra concepts (music and speech) indicates the difficulty to provide the segmentation of a user video in general concepts. Finally, we employ an HMM-based *acoustic concept recognition* (ACR) system to segment the audio signal. The segmental information is used as features in SVM-based classification for multimedia event detection (MED) task. This approach is different from the previously mentioned supervised techniques [15, 16] in several ways. First, we do not use any fixed segmentation, but instead use recognition to extract acoustic concept segments dynamically. The second difference is that the models are not trained with specific acoustic concepts that may produce a system very constrained for a specific task.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

Abbr.	Full Name	# Train	# Test
E001	Attempting a board trick	91	32
E002	Feeding an animal	81	30
E003	Landing a fish	69	26
E004	Wedding ceremony	66	25
E005	Woodworking project	77	25

Table 1: Video event class abbreviations (Abbr.) and full names along with the number of positive samples appearing in the training and test sets

The remainder of this paper is organized as follows: the TRECVid2011 dataset and the acoustic concepts annotations are described next. Section 3 deals with the audio features and the acoustic concepts classification and recognition (segmentation and classification). The baseline of MED task using segmental-GMM vectors and the ACR system are provided in Section 4. Finally, conclusions will be presented in Section 5.

2. Data set and Annotations

2.1. TRECVid 2011

The Text Retrieval Conferences Video Retrieval Evaluation (TRECVID) [1] focuses on the problem of Multimedia Event Detection (MED) in website quality videos for hard-to-detect events (e.g., Landing a fish). The evaluation dataset consists of non-professional videos collected from the internet with high variability and short duration (a couple of minutes). Fifteen different video event categories can be found in the database with only five of those categories available for testing purposes in this study.

To develop and evaluate our proposed approach, we use three sets of data: first set (*train-1*) is for training the acoustic concept models, second set (*train-2*) is for training the MED classifiers after extracting acoustic concept indexes on this data and using them as MED features, and the third set (*test*) is for testing the system. These sets are the same used in [15] and [9] to be able to provide fair comparison to previously published works. There is a total of 2640 videos in the test set and 7881 in the training set. Table 1 shows, for each of the five video events, the numbers of positive samples in the test and training sets. Note that the categories grouped several videos. For example "feeding an animal" includes animals from different species and , therefore, different animal sounds.

2.2. Acoustic Concepts Annotations

Because the ultimate goal of the system is to perform detection of multimedia events on the videos using the recognition of acoustic concepts, it has been created an initial set of labels of acoustic concepts to be useful in discriminating the five video event classes presented in Table 1 while also being clear and understandable for the annotators.

Broad Acoustic Concepts	Abbr.
1. Crowds and audience	(CA)
2. Animal sounds	(AN)
3. Repetitive sounds	(RS)
4. Machine noise	(MN)
5. Environmental sound	(ES)
6. Music	(MU)
7. Speech	(SP)

Table 2: Broad acoustic concepts and abbreviations

The acoustic concepts are divided in five broad classes as Table 2 shows. These broad classes have been extended with Speech and Music classes because most of the videos contain speech or music as the predominant audio. In fact, some of the five acoustic concepts are overlapped with speech or music barely audible in the background. However, those segments were annotated as that acoustic concept. The following section presents the results on the classification, and recognition of the broad classes, showing how difficult is to create a well-trained model for these acoustic concepts due to the high variance of the audio.

3. Acoustic Concepts Recognition

To model the acoustic concepts, we used a HMM/GMM-based system. As it was described on the last section, to train and test these models, a subset of the National Institute of Standards and Technology (NIST) is provided for the TRECVID evaluation 2011. This set is composed of 1536 videos (47 hours approximately) averaging 1.8 minutes per file. This section is organized follows: we describe the front-end audio features used in this approach and the acoustic concepts to train the models. Also, experiments of classification and recognition are reported to show how difficult is the final goal of this task.

3.1. Front-End Audio Features

This section is a summary of the front-end audio feature extraction method used in [18]. We extract 16 MFCCs (including C0) computed in 25ms frame size with a 10ms frame step and their Δ and $\Delta\Delta$. Due to the high variability of every acoustic concept, the fact that the segments are overlapped with speech and music, and the different devices used to record the video, a normalization of these features is needed. Trying to generalize the features, a cepstral mean normalization is computed over the whole video and the mean and standard deviation are computed over 1-second windows with an overlap of 0.75 seconds. Thus, the system uses 96 features (48 for the mean and 48 for the standard deviation of the $MFCC + \Delta + \Delta\Delta$ features) every 0.25 second.

3.2. Classification System

This experiment shows how difficult the task is. The goal of this experiment is the classification of a set of cut segments in one of the broad classes. The segments are overlapped with speech and music in the background in some cases. However, the classification is done with the five broad classes (without speech and music models) keeping the seven broad classes (with speech and music models) for the recognition task. The segments are extracted from the video database generating 13520 segments of different durations. Each concept is model as one state HMM/GMM with 256 Gaussians. Table 3 shows the results using the same subset of data to train and test. As it can be seen, the task is very difficult due to the high within-class variability of each concept. The system classified 71.1% of the segments correctly.

To test the system, a 4-fold cross-validation was performed using 3 folds to train the models and 1 fold to test. Table 4 shows the confusion matrix and how the classification rate is reduced compared with Table 3, classifying a 45.9% of the segments correctly. It can be seen that the Animal Noise and the Environmental Sounds are the concepts with a higher error rate

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

	CA	AN	RS	MN	ES
CA	**0.77**	0.04	0.04	0.06	0.09
AN	0.08	**0.80**	0.04	0.02	0.06
RS	0.08	0.04	**0.75**	0.05	0.08
MN	0.11	0.04	0.09	**0.61**	0.16
ES	0.12	0.09	0.08	0.07	**0.63**

Table 3: Confusion Matrix using the same set for train and test

	CA	AN	RS	MN	ES
CA	**0.61**	0.03	0.06	0.12	0.18
AN	0.18	**0.12**	0.20	0.19	0.31
RS	0.11	0.05	**0.45**	0.18	0.21
MN	0.17	0.02	0.16	**0.40**	0.25
ES	0.24	0.07	0.15	0.21	**0.33**

Table 4: Four Folds Cross-Validation Confusion Matrix

because both classes do not have enough data to train the models.

3.3. Recognition System

In the MED task, a recognition system is needed to be able to detect and classify the acoustic concepts related with the video. Due to the fact that most of the acoustic concepts are overlapped with speech and music, two extra models are needed to identify the segments in which there is not an acoustic concept. Also, these models can be useful to describe the video in the MED task. Using the same models trained for the classification task, a segmentation is executed over the whole video where the cut-segments were extracted for the classification system in previous subsection.

In this experiment, every concept (speech and music included) is modeled by a HMM/GMM of one state. The main difference is that a segmentation is produced when there are transitions between the models in the Viterbi algorithm. Table 5 shows the recognition result per concept independently of the segment duration. As it can be seen, Crowds and Repetitive Sounds have the better results in comparison with the Animal Noise or Environmental Sound because Crowd and Repetitive sounds were trained with more data than Animal Noise or Environmental Sound. The following sections show how the multimedia events related with the acoustic concepts Animal Noise or Environmental Sound have a poor detection rate because the models are not well-trained.

4. Acoustic concepts as features for MED

4.1. Methods

The purpose of the acoustic concepts recognition is to enable a video to be modeled by the acoustic concepts present in the video. For example, the ability to identify certain properties of the audio component that correlate strongly with crowd sounds and little with environmental sounds such as water might indicate the video takes place in a setting with large number of people present away from water and is therefore more likely to belong to certain video events (i.e. parade) than others (i.e. fishing). This section shows two different approaches using acoustic concepts to detect the multimedia event.

	CA	AN	RS	MN	ES	SP	MU
CA	**0.41**	0.03	0.03	0.04	0.04	0.20	0.23
AN	0.11	**0.01**	0.01	0.05	0.07	0.52	0.20
RS	0.07	0.02	**0.35**	0.09	0.09	0.16	0.20
MN	0.14	0.10	0.10	**0.26**	0.16	0.07	0.15
ES	0.23	0.02	0.07	0.05	**0.11**	0.12	0.13

Table 5: Segmentation Confusion Matrix

The first one is described in [15] and it is known as Segmental-GMM. Training the GMM for the seven selected acoustic concepts, a score vector is generated on fixed-length segments with each element in the vector corresponding to a posterior score for a GMM. As mentioned, we refer to these score vectors as Segmental-GMM feature vectors. In our experiments the segmental GMM vectors are 7-dimensional.

The second approach is known as Acoustic Concept Recognition (ACR) in which each concept is modeled as a HMM/GMM of one state. The main difference with the Segmental-GMM approach is that the segments are not fixed-length any more, and the segmentation is based on the transitions between the HMM models following the Viterbi algorithm. The score vector is the accumulated likelihood for each model. Therefore, a video is represented by a 7xK dimensional matrix with each column corresponding a different length segments.

In order to perform classification on the multimedia event level, we need to have features that are constant length independent of the video length. These constant-length features can then be used with the SVM classifier. The original video is currently represented by a 7xK matrix and is therefore not fixed-length. In this work, we represent a video with what we refer to as a *co-occurrence matrix* in which each element represents the probability that a pair of acoustic concepts occur in the video. This process is described in [15].

We performed a verification, also referred to as *one-against-all*, experiment for each of the five video event classes. For each video event, a given file is labeled as *in-class* or *out-of-class*. For example, for E004 we would perform the binary classification into *Wedding ceremony* and non-*Wedding ceremony*. We chose to perform classifications using support vector machines (SVMs) with a linear kernel. SVMs are commonly used for similar classification experiments due their simplicity and ability to model nonlinear decision boundaries using what is referred to as the 'kernel trick.'

4.2. Results

To measure the system performance results we use Detection Error Tradeoff (DET) curves, which are commonly used to show the tradeoff between the false alarm errors and missed detections. We generated the DET-curves in this paper with plotting software available from the NIST website [19]. From these curves, we also extracted the equal error rate (EER) as the the point where the probability of false alarm (pFA) is equal to the probability of a miss (pMiss). Since TRECVid MED 2011 simulates a retrieval task from wild videos in the internet, the assumption is that high miss rates can be tolerated in favor of low false alarm probabilities. Therefore, we use a benchmark to compares the number of misses at a given false alarm rate of 6%. The percentage of misses at a given false alarm rate is computed in a similar fashion to EER.

Figure 1 shows the DET curves for every acoustic event. The blue curves represents the performance of the Segmental-GMM approach, and the red curve represents the performance of the ACR approach. As it can be seen, the systems per-

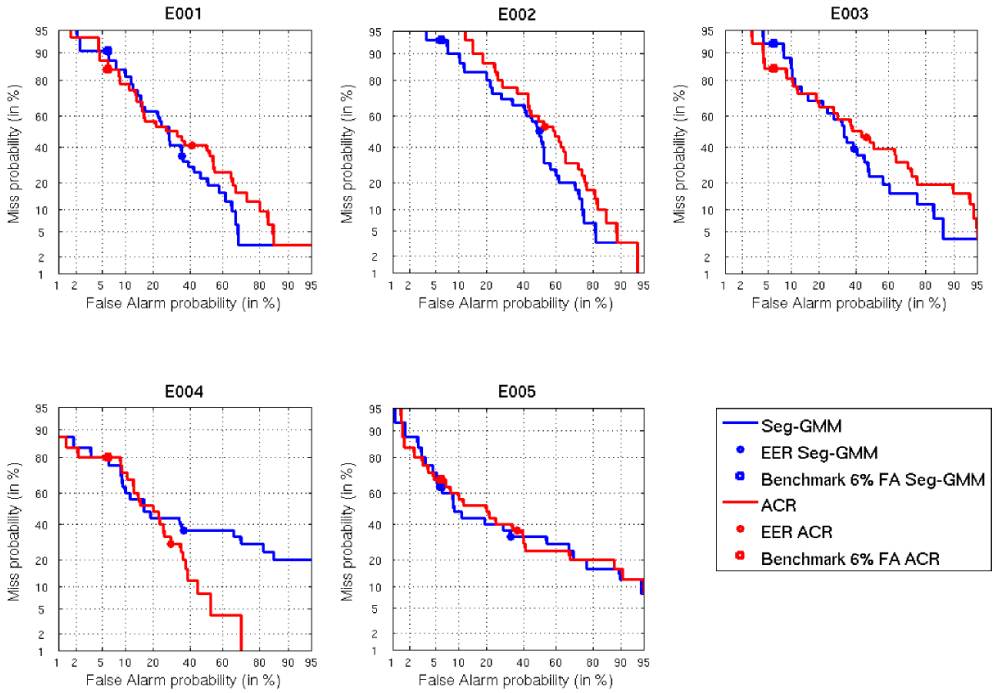


Figure 1: DET curves of Segmental-GMM approach versus ACR approach. The marks for EER and the benchmark for 6% of pFA are on the same curves

formance varies across video events. *Wedding ceremony* and *Woodworking project* show the best results while *Feeding an animal* and *Landing a fish* show the worst results. These behaviors are consistent with the previous results in section 3. It can be seen that the concepts *Animal sounds* and *Environmental sound* have the biggest error rate, and those concepts are more related with *Feeding an animal* and *Landing a fish* videos respectively. On the other hand, the concepts *Crowds and audience* and *Repetitive sounds* have the best results, and they are more related with *Wedding ceremony* and *Woodworking project* events respectively. Also, *Feeding an animal* and *Landing a fish* videos contain short bursts of sounds overlapping with a widely varying background noise, which make the detection much more difficult.

Table 6 shows the EER and the benchmark given a false alarm rate of 6% for both approaches. The EER is better using Segmental-GMM for almost all the events except for the event *Wedding ceremony*. However, the benchmark is better using ACR with the exception of E002 event where the model is poor trained and E005 where the difference between Segmental-GMM and ACR is not significant as can be seen in Figure 1.

5. Conclusions

This paper shows a comparative study between different approaches to detect multimedia events using a set of videos provided in TRECVid 2011 evaluation. These approaches are based on the analysis of the audio of the videos, and they help to improve the detection accuracy of video analysis systems. The proposed approaches create features based on the likelihood of acoustic concepts that can happen in the multimedia event.

The first set of experiments shows the accuracy to classify and recognize the acoustic concepts. The videos of the

	Segm-GMM		**ACR**	
	EER	BM-6%	EER	BM-6%
E001	**0.343**	0.906	0.406	**0.843**
E002	**0.500**	**0.933**	0.533	1.000
E003	**0.384**	0.923	0.461	**0.846**
E004	0.360	0.800	**0.280**	**0.800**
E005	**0.320**	**0.640**	0.360	0.680
Mean	**0.381**	0.840	0.408	**0.833**

Table 6: EER and Benchmark of 6% of pFA for segmental-GMM and ACR approaches

TRECVid 2011 are downloaded from different sources in internet, so the audio of these videos has a lot of variability. The acoustic features that compensate the variability of the audio are the mean and the variance of MFCCs. However, training and testing over the same set of data provide a mean error rate of 30% as it was showed in Table 3. The concepts *Animal Sounds* and *Environmental Sound* have the highest error rate for all the systems and, therefore, the events related with these concepts (as *Feeding an animal* and *Landing a fish*) have the highest detection error rates for all the event detector approaches.

We create a baseline based on the approach proposed in [15]. This baseline is known as *Segmental-GMM* and it creates a feature vector with the likelihood of the acoustic concepts from a GMM model for every acoustic concept extracted every five seconds. The novelty proposed in this paper is to create an HMM-GMM model for every acoustic concept to be able to get a segmentation based on the transitions between the models. This solution is know as ACR and it shows a little improvement over the *Segmental-GMM* as a retrieval approach.

6. Acknowledgments

We would like to thank Stephanie Pancoast for her help with the annotations and baselines.

This work has been funded by the Spanish Government and the European Union (FEDER) under the project TIN2011-28169-C05-02 when the authors were at SRI International.

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes non withstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

7. References

[1] T. multimedia event detection 2011 evaluation, "http://www.nist.gov/itl/iad/mig/med11.cmf."

[2] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[3] C. Zieger, "An hmm based system for acoustic event detection," in *Multimodal Technologies for Perception of Humans*, 2008.

[4] Mostefa, Moreau, and Choukri, "The chil audiovisual corpus for lecture and meeting analysis inside smart rooms," in *Evaluation and Language Distribution Agency*, 2008.

[5] M. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia*, 2006.

[6] C. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, 2007.

[7] S. Tsakalidis, X. Zhuang, R. Hsiao, S. Wu, P. Natarajan, and R. Prasad, "Robust event detection from spoken content in consumer domain videos," in *Interspeech 2012*, 2012.

[8] Q. Jin, P. Schulam, S. Rawat, S. Burger, D. Ding, and F. Metze, "Event-based video retrieval using audio," in *Interspeech 2012*, 2012.

[9] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Interspeech2012*, 2012.

[10] L. Li, "A novel violent videos classification scheme based on the bag of audio words features," in *International Journal of Computational Intelligence*, 2012.

[11] B. Byun, S. Kim, I.and Siniscalchi, and L. C.H., "Consumer-level multimedia event detection through unsupervised audio signal modeling," in *Interspeech 2012*, 2012.

[12] S. Chaudhuri, R. Singh, and R. Raj, "Exploiting temporal sequence structure for semantic analysis of multimedia," in *Interspeech 2012*, 2012.

[13] X. Zhuang, S. Tsakalidis, S. Wu, P. Natarajan, and R. Prasad, "Compact audio representation for event detection in consumer media," in *Interspeech 2012*, 2012.

[14] R. Mertens, H. Lei, L. Gottlieb, G. Friedland, and D. A., "Acoustic super models for large scale video event detection," in *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*. ACM, 2011.

[15] S. Pancoast, M. Akbacak, and M. Sanchez, "Supervised acoustic concept extraction for multimedia event detection," in *ACM Multimedia Workshop*, 2012.

[16] Y. Jiang, X. Zeng, G. Ye, and S. Bhattacharya, "Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching," in *NIST TRECVID 2010*, 2010.

[17] J. Van Hout, M. Akbacak, D. Castan, E. Yeh, and M. Sanchez, "Extracting spoken and acoustic concepts for multimedia event detection," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013*, 2013.

[18] D. Castan, C. Vaquero, A. Ortega, and E. Lleida, "Hierarchical audio segmentation with hmm and factor analysis in broadcast news domain," in *Interspeech2011*, 2011.

[19] N. DETware V.2., "http://www.itl.nist.gov/iad/mig/tools/."

Broadcast News Segmentation with Factor Analysis System

Diego Castan, Alfonso Ortega, Antonio Miguel and Eduardo Lleida

University of Zaragoza, Spain

[dcastan,ortega,amiguel,lleida]@unizar.es

Abstract

This paper studies a novel audio segmentation-by-classification approach based on Factor Analysis (FA) with a channel compensation matrix for each class and scoring the fixed-length segments as the log-likelihood ratio between class/no-class. The system described here is designed to segment and classify the audio files coming from broadcast programs into five different classes: speech (SP), speech with noise (SN), speech with music (SM), music (MU) or others (OT). This task was proposed in the Albayzin 2010 evaluation campaign. The article presents a final system with no special features and no hierarchical structure. Finally, the system is compared with the winning system of the evaluation (the system use specific features with hierarchical structure) achieving a significant error reduction in SP and SN. These classes represent 3/4 of the total amount of the data. Therefore, the FA segmentation system gets a reduction in the average segmentation error rate that is able to be used in a generic task.[1]

Index Terms: Audio Segmentation, Factor Analysis, Broadcast News (BN), Albayzin-2010 Evaluation

1. Introduction

Due to the increase in audio or audiovisual content, it becomes necessary to use automatic tools for different tasks such as analysis, indexation, search and retrieval. Given an audio document, the first step is audio segmentation producing a delineation of a continuous audio stream into acoustically homogeneous regions. When the audio segmentation is followed by a classification system the result is a system that is able to divide an audio file into different predefined classes chosen for a specific task.

Broadcast news (BN) domain is one of the most popular multimedia repositories because it has rich audio types and several approaches have been proposed in this scenario. For example, in the task of automatic transcriptions of BN [1] the data contain clean speech, telephone speech, music segments and speech overlapped with music and noise so the segmentation generates a boundary for every speaker change and environment/channel condition change with no explicit cues. In [2] segmentation is based on five different classes: silence, music, background sound, pure speech, and non-pure speech. The solution is based on SVM combination. In [3] the audio stream from BN domain is segmented into 5 different types including speech, commercials, environmental sound, physical violence and silence. [4] presents a review of different solutions and the acoustic features used in each one of them and also a new algorithm for computing various time-domain and frequency-domain features, for speech and music signals separately, and estimating the optimal speech/music thresholds.

[1]This work has been funded by the Spanish Government and the European Union (FEDER) under the project TIN2011-28169-C05-02

The different segmentation approaches in BN differ in either the feature extraction methods or the classifier. The features can be distinguished in *frame-based* and *segment-based* features. The frame-based features usually describe the signal within a short time period (10-30 ms), where the process is considered stationary. MFCCs or PLPs are commonly used as frame-based features like in [5] where these features are classified with an autoassociative neural network. In [6] the authors propose two pitch-density-based features and relative tonal power density to classify on BN. For segment-based feature extraction, a longer segment is taken into consideration. The length of the segment may be fixed (usually between 0.5 and 5 seconds) or variable. In [7] a content based speech discrimination algorithm is designed to exploit long-term information inherent in modulation spectrum.

Audio segmentation systems perform the segmentation in two different ways. The first one is based on detecting the boundaries and then classifying each delimited segment. We refer to them as *segmentation-and-classification* approaches. For example, in [8], an approach using a temporally weighted fuzzy C-means algorithm has been proposed. The second segmentation way is known as *segmentation-by-classification* and it consists of classifying consecutive fixed-length audio segments. The segmentation is produced directly by the classifier as a sequence of labels. This sequence is usually smoothed to improve the segmentation. An example of this procedure can be seen in [9] where the author combines different features with a GMM and a maximum entropy classifiers. The final sequence-level were smoothed with a HMM.

The different strategies outlined in the preceding paragraphs have their advantages and disadvantages described by Huang and Hansen in [10]. The most common solution to avoid the shortcomings and enjoy the benefits of each strategy is to create hierarchical systems with multiple steps where each level is designed with specific features and segmentation systems for each class. As a result, the system becomes very specific for a database and may produce segmentation errors in different domains. Recently, an audio segmentation task in BN domain was proposed in [11] in the context of the Albayzin-2010 evaluation campaign. Almost all the participants of the evaluation used hierarchical systems, including the winning system [12] based on a hierarchical architecture that used different sets of features for every level.

In this paper, we proposes a whole FA segmentation system with no-hierarchical structure where the within-class variability is compensated with a different channel matrix for each class. The remainder of the paper is organized as follows: database and metric of Albayzin 2010 evaluation is presented in Section 2. Section 3 shows the factor analysis theoretical approach based on FA. Segmentation results are presented in Section 4. Finally, the conclusions are presented in Section 5.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

2. Albayzin 2010 audio segmentation evaluation

The Albayzin evaluation campaign is an internationally open set of evaluations organized by the Spanish Network of Speech Technologies (RTTH) every 2 years. A completed description of the Albayzin 2010 evaluation can be found in [13] which describes the participant's approaches and the results of the systems . We summarize the database description and the metric of the evaluation in the next subsections.

2.1. Database

The database consists of a Catalan BN database from the public TV news channel that was recorded by the TALP Research Center from the UPC. It includes approximately 87 hours of annotated audio divided in 24 files of 4 hours long. A set of five different audio classes were defined for the evaluation with the following distribution: Clean speech: 37%; Music: 5%; Speech over music: 15%; Speech over noise: 40%; Others: 3%. The class "Others" is not evaluated in the final test. The database for the evaluation was split into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3).

2.2. Metric

The metric is defined as a relative error averaged over all acoustic classes (ACs):

$$Error = average_i \frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)},$$

where dur($miss_i$) is the total duration of all deletion errors (misses) for the *ith* AC, dur(fa_i) is the total duration of all insertion errors (false alarms) for the *ith* AC, and dur(ref_i) is the total duration of all the *ith* AC instances according to the reference file. The incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another. A forgiveness collar of 1 sec (both + and -) is not scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an AC begins/ends.

3. FA-Based audio segmentation

This study proposes a framework for automatic audio *segmentation-by-classification* system. The system deals with the problem of assigning a class label to each fixed-length clips using Factor Analysis (FA) models. The FA approach has been successfully used in speaker recognition [14] [15] [16], speaker verification [17], speaker segmentation [18] and language recognition [19]. The variability of the same class segments is known as *within-class variability*. The goal of these systems is the compensation of the *within-class variability* to reduce the mismatch between training and test. Fig. 1 illustrates the proposed framework where each block is described in the next subsections. We will discuss the feature extraction, the statistic extraction and the within-class variability compensation using FA.

3.1. Acoustic Feature Extraction and Statistics

Mel-frequency cepstral coefficients (MFCCs) [20] are used in most speech recognition tasks because the mel-scale filter bank is an approximation to human auditory system response. Therefore they work well in audio segmention task too. Typically,

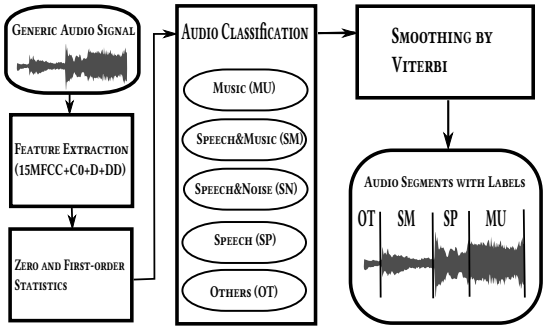


Figure 1: Block Diagram of Factor Analysis Segmentation-by-Classification System for Broadcast News Classes

MFCC features are computed at each short speech segment (e.g., 10 ms) together with their derivatives to capture the short-term speech dynamics. On this framework we extract 16 MFCCs (including C0) computed in 25 ms frame size with a 10 ms frame step, their first and second order derivatives.

The audio features are packed in clips of different lengths with 0.1 and 0.5 second clip-steps. The fixed-length clips are mapped to sufficient statistics by using a Universal Background Model (UBM) which is a class-independent GMM with C Gaussians trained with the EM-algorithm [21] on the audio feature vectors of the training data.

3.2. Theoretical Background

Data from a particular class are modeled by a GMM defined by means $m_1, m_2, ..., m_C$, weights $w_1, w_2, ..., w_C$ and covariances $\Sigma_1, \Sigma_2, ..., \Sigma_C$ where C is the number of Gaussians. We can concatenate all GMM means to one mean supervector m of $CF \times 1$ dimensions where F is the feature vector size:

$$m = [m_1^T, m_2^T, ..., m_C^T]^T. \quad (1)$$

The Factor Analysis model is the adaptation of the UBM model where the supervector of means is not fixed and it can vary from segment to segment to account for differences in the channel. These GMMs have segment and class dependent component means but fixed component weights and covariances chosen to be equal to the UBM weights and covariances. Specifically, we use a Factor Analysis model for the mean of kth component of the GMM for segment s:

$$m_k^s = t_k^{c(s)} + U_k x_s \quad (2)$$

where $c(s)$ denotes the class of segment s and $t_k^{c(s)}$ is the channel-independent-class-location vector obtained by using a single iteration of relevance-MAP adaptation from the UBM [22]. U_k is the factor loading matrix and x_s is a vector of L *segment-dependent channel factors* generated by a normal distribution $(N(0, I))$. We stack component-dependent vectors into supervectors m_s and $t^{c(s)}$ and we stack the component-dependent U_k matrices into a single tall matrix U, so that equation can be expressed more compactly as:

$$m^s = t^{c(s)} + U x_s \quad (3)$$

where U is known as the *channel matrix* and it represents the within-class variability. Note that, following the terminology in the literature, we use the terms *channel matrix* and *channel*

factors to describe the elements related with the within-class variability even if that variability is not produced by different channels (also can be produced by different speakers or different content). The columns of the U matrix are the basis spanning the subspace of the channel and the *channel factors* are the coordinates defining the position of the channel-dependent supervector in the subspace. The *channel factors* dimension (L) is smaller than CF so U matrix has low rank ($CF \times L$ dimensions). Depending on the application, the value of L is between 50 and 200 and CF can be 98304 if we have 2048 Gaussians and 48-dim feature vector (with the MFCC-UBM). The estimation of these parametres can be understood following [16].

3.3. Class/No-Class U Channel Matrices System

Most of the approaches based on FA for language recognition are implemented with a single U *channel matrix* because the nature of the within-class variability is the same for all the languages as it can be seen in [23] [24] [25] [16]. Therefore, in [26] a segmentation system was proposed with five channel-independent-class-location vectors (one vector per class) and a single compensation channel matrix U for all the classes. The paper compares the FA system with the winnner of the Albayzin-2010 evaluation and the conclusion was that the compensation matrix had a bad behavior for the *Music* class due to the different nature of the rest of the classes. However, the paper [27] shows a clear advantage when the classes are homogeneous (like SN and SP). In this scenario the channel matrix models the compensation between different speakers and different words leaving the background sound as useful information for the classification improving the segmentation.

A number of studies have focused on features to describe the distribution of sounds to be able to distinguish between speech, music or noise. Most of these approaches use a hierarchical structure where each level is specialized on the detection of an specific class with specific features for that class. The main goal of this work is the compensation of all the classes with no-specific features for each class even if the nature of the classes is not the same. We propose a ten channel-independent-class-location vectors (a class and no-class vectors for each class) and five channel matrix representing the within-class variability of each class/no-class with no hierarchical structure. Let

$$T = [t_{mu}, t_{nomu}, t_{ot}, t_{noot}, t_{sm}, t_{nosm}, t_{sn}, t_{nosn}, t_{sp}, t_{nosp}] \quad (4)$$

$$\Xi = [U_{mu-nomu}, U_{ot-noot}, U_{sm-nosm}, U_{sn-nosn}, U_{sp-nosp}] \quad (5)$$

where T represents the locations of classes and no-classes in the GMM space and Ξ the channel matrices. Our metamodel for class-segment-dependent GMM is parametrized by (T, Ξ) which describe the prior distributions of the parameters m.

This approach will be compared with the classic formulation with a single U *channel matrix* on Section 4 for the classification over the oracle segments and the final segmentation system.

3.4. Scoring

There are different scoring methods used in the state-of-the-art of speaker and language recognition. In the proposed experiments in Section 4 we use the *integration trough the channel factors distributions*. This score is a marginalization using a point estimation of the class m_s, and integrate only over the channel factors, when the statistics are centered around the point estimation m_s. The log-likelihood is defined by the equation (19) in [16] and can be understood following the Section V in the same article.

In [28], [16] and [29] the score employed to detect the speaker is the log-likelihood ratio(LLR). For a test clip χ and class c, the LLR compares the hypothesis that the clip χ belongs to the class c against the hypothesis that the clip χ does not belong to the class c. This score is shown in Formula 6 where the numerator is the likelihood of the test clip calculated with the class model and the denominator is the likelihood of the test clip calculated with UBM model. Note that the UBM model is employed as a general model to describe the not belonging hypothesis. That makes sense for speaker identification task where the hypothesized speaker represents a very small amount into the UBM. However, our problem has only four classes and, therefore, the class is highly represented by the UBM and may corrupt the LLR score.

$$LLR_{class} = log \frac{P(\chi/class)}{P(\chi/UBM)} \quad (6)$$

We propose a LLR scoring where the denominator is the likelihood of the test clip calculated with the no-class model. The compensated log-likelihood ratio (CLLR) is computed for each class/no-class as:

$$CLLR_{class} = log \frac{P(\chi/class)}{P(\chi/noclass)} \quad (7)$$

CLLR is more discriminative than LLR for a segmentation task because the hypothesized class is not presented in the denominator and, also, because the no-class model is channel compensated as the class model.

4. Experimental results

In a segmentation-by-classification system, the errors can be produced in two ways: first, a classification error due to a bad labeled frame, and a segmentation error due to a temporal mismatch between the oracle boundaries and the hypothesis boundaries. This Section shows the experiments for the evaluation data described in Section 2.1 divided into two sets. In the first set, the segments are given by the ground truth and the systems decide the class of each segment with no segmentation error to evaluate the classification accuracy of the systems. The second set of experiments shows the segmentation and the classification error and it proposes a final segmentation-by-classification system based on FA that improve the result of the winning system in the Albayzin evaluation.

4.1. Classification Experiments with Oracle Segmentation

The classification is done over the segments extracted with the ground truth to evaluate the classification accuracy over the whole segment. Most of the segments are between 5 and 20 seconds long.

We propose two sets of systems based on GMM and HMM-GMM as a baseline. Table 1 shows the results for these systems. In the first part of the table, we have tried with different number of Gaussians. The classification is based on the highest accumulated likelihood over the whole segment. Increasing the number of Gaussians improves the final result. The highest

Table 1: Classification Baseline Experiments: error per class and total error for GMM-HMM systems over the test files with perfect segmentation in %

GMM	MU	SP	SM	SN	TOTAL
32G	9.66	49.36	37.59	48.11	**36.18**
64G	10.68	45.74	36.68	45.44	**34.63**
128G	9.81	41.79	32.02	40.75	**31.09**
256G	10.43	37.61	31.85	37.67	**29.39**
512G	9.51	35.95	29.38	35.99	**27.71**
1024G	9.39	34.91	27.03	34.35	**26.42**
2048G	9.61	33.39	38.01	34.01	**26.25**
HMM-LeftToRight	MU	SP	SM	SN	**TOTAL**
1 ST - 2048G	9.61	33.39	28.01	34.01	**26.25**
2 ST - 1024G	9.48	42.75	27.45	41.26	**30.24**
4 ST - 512G	10.11	27.91	27.17	29.87	**23.77**
8 ST - 256G	8.37	31.64	26.42	32.1	**24.63**
16 ST - 128G	8.84	26.92	32.28	32.12	**25.04**
32 ST - 64G	11.33	29.81	26.64	32.48	**25.07**

number of Gaussians is 2048 because, although the final results is the best one, the MU and SM classes begin to get worse results. The next experiment of the baseline system uses 2048 Gaussians distributed in different nodes in a HMM. The second part of the Table 1 shows the results for left-to-right topologies of HMMs. These topologies increase the activity duration of each model [30], avoiding wrong transitions inside the segment and improving the results. The best baseline system (23.77% of total error) is performed using five left-to-right HMMs with four emitting states and 512 Gaussians per state where each HMM corresponds to one acoustic class.

To evaluate the strengths and weaknesses of a FA system, we assess different configurations described in Section 3. The UBM employed to compute the statistics has a fixed amount of 2048 Gaussians to be able to compare the results of the FA systems with the GMM/HMM baseline. We compute the result over the test set using the *integration trough the channel factors distributions scoring*. The experiments are calculated with a single channel matrix to compensate all the classes and different channel matrices for each class/no-class using different number of channel factors (50, 100, 150, 200 and 250).

Table 2: FA systems with a single U for all the classes and U matrix for every class/no-class over the test set with perfect segmentation in %

Single U	MU	SP	SM	SN	TOTAL
50 chnf	10.20	15.98	24.21	21.41	**17.95**
100 chnf	9.16	16.06	20.28	20.06	**16.39**
150 chnf	9.42	15.52	18.04	18.90	**15.47**
200 chnf	9.08	15.72	17.38	19.17	**15.34**
250 chnf	8.52	16.70	16.06	19.42	**15.17**
U per class	MU	SP	SM	SN	**TOTAL**
50 chnf	9.65	19.13	24.10	23.31	**19.05**
100 chnf	8.54	16.22	22.12	20.18	**16.77**
150 chnf	9.65	16.63	18.31	19.49	**16.02**
200 chnf	9.20	17.22	17.73	19.60	**15.94**
250 chnf	9.69	17.46	17.12	19.82	**16.02**

Comparing the Table 1 and the Table 2, it can be seen a significant improvement using FA as a classification system against GMM/HMMs. Using the best HMM configuration (left-to-right HMM with four states and 256 Gaussians in each state) as a reference, the worst FA system improves the total result in 4.72% (with a U matrix per class and 50 channel factors) and in 8.6% comparing with the best FA configuration (with a single U matrix and 250 channel factors).

4.2. Segmentation-by-Classification Experiments

In the last subsection, each segment was labeled with the best decision coming from the accumulated log likelihood or accumulated log likelihood ratio of the models. In this subsection, the segments are delimited with the transitions between the scores and the errors might be due to a temporal mismatch or a bad label assignment.

Table 3: Segmentation Baseline Experiments: error per class and total error for HMM systems over the test files in %

HMM-LeftToRight	MU	SP	SM	SN	TOTAL
1 ST - 2048G	35.53	59.22	65.07	58.60	**54.6**
2 ST -1024	29.96	59.26	54.79	56.82	**50.21**
4 ST - 512	26.04	49.8	45.98	50.27	**43.02**
8 ST - 256G	24.35	49.3	41.66	50.19	**41.37**
16 ST -128G	17.82	40.24	36.02	43.06	**34.28**
32 ST - 64G	17.39	39.53	33.95	41.56	**33.31**

As we did in the last subsection, GMM/HMM systems are used as the baseline. Because the segments are delimited by the scoring transitions, the scores need to be smooth using low pass filters or HMM. Table 3 shows different HMM topologies and configurations. Again, the left-to-right topology improves the result because these systems smooth the transitions between classes. The best baseline system for segmentation-by-classification (33.31% of total error) has 32 states with 64 Gaussians each state and has a left-to-right topology.

Table 4: FA segmentation-by-classification systems with a single U for all the classes and U matrix for every class/no-class in %

Win-3.0 step-0.5 100chnf					
	MU	**SP**	**SM**	**SN**	**TOTAL**
Single U	40.38	76.91	60.52	64.31	**60.53**
U per class	33.35	45.62	36.2	47.44	**40.65**

As a preliminary experiment, the first FA segmentation-by-classification system computes the statistics over a 3 second windows with 0.5 second window-steps and 100 channel factors. An increment of the channel factors or a reduction of the window-step increase the memory and the time to train the models exponentially. Experiments with a single channel matrix for all the classes and a channel matrix for each class are presented in Table 4. There is a significant improvement in the majority classes using a channel matrix for each class because the CLLR removes the information of the target class in the denominator as we pointed in Section 3.4. The bigger is the class in the data, more significant is the reduction of the error comparing with a single channel matrix for all the classes. Accordingly, the total error is reduced about 20%.

Once determined that the best configuration is the FA system with a channel matrix for each class, the window-step can be modified to get more resolution (0.1 second window-step) and the CLLR can be smoothed to avoid an over segmentation. In the experiments, a zero-phase average filter is computed to smooth the CLLR of each class and avoid a sudden change in the segment labels. Figure 2 shows the filtered-ratio scores for each class over a chunk of a test file. The ground truth is plotted in the same figure and it is represented with a square wave of

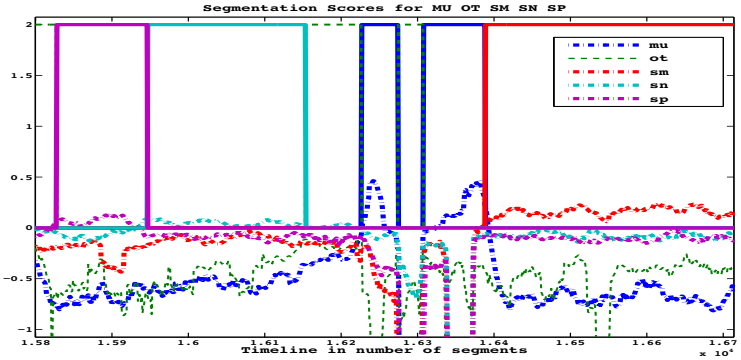


Figure 2: Scores and the ground truth of each class over a chunk of a test file

amplitude 2. The color of each score class and the corresponding ground truth is the same. The figure clearly shows that the ratio of the winning class is bigger than zero and corresponds with the ground truth class.

Due to the metric, the smallest classes have to be detected with the same accuracy as the largest classes as can be seen in section 2.2. To increase the detection of the smallest class (MU) we optimize the prior probabilities in a Viterbi algorithm checking the total result over the train files. Table 5 shows the total error over the train files. The first row shows the total error when all the classes have the same priority and it can be seen that the smallest error is obtained when MU and SN/SP have 28% and 16% of priority respectively decreasing the false alarms of the SN/SP over MU class. These priors are employed in the Viterbi over the test files and the results are shown in Table 6.

We compare the error of the system proposed in this work with the winning system of the Albayzin-2010 evaluation [12] where 15 MFCCs, the frame energy, and the derivatives are extracted. In addition, the spectral entropy and the Chroma coefficients are calculated. The mean and variance of these features are computed over 1 second interval creating a 122 dimension feature vectors. The segmentation approach chosen is HMM-based. The acoustic modeling is performed using five HMMs with three emitting states and 256 Gaussians per state. Each HMM corresponds to one acoustic class. A hierarchical organization of binary HMM detectors is used. First, audio is segmented into Music/non-Music portions. Second, the non-Music portions are further segmented into Speech-over-music/non-Speech-over-music portions. Finally, the non-Speech-over-music portions are segmented into Speech/Speech over noise.

Table 6 is divided in two parts: the first part shows the error for each class and the average error for the winning herarchical-HMM system of the evaluation (HMM-Winn). The last column shows the NIST metric used in the NIST RT Diarization evaluations [31] to compare the systems with a well-known metric. To be able to compute the NIST error with the herarchical-HMM system, we replicate the winning system according to [12] (HMM-Rep). The second part of the table shows FA segmentation-by-classification system (FA-Segm) after the Viterbi smoothing with the priors of the Table 5. The last row of the table shows the same FA system with a slight modification introducing OT segments between SN and SP to model the silence of the anchor before the coverage to avoid the false alarms. The hierarchical-HMM systems detects better the MU and SM segments than the FA systems due to the Chroma coefficients in the features. However, SN and SP classes are much better detected with the FA system decreasing the error of the classes in 2% and 9% respectively. These classes represent more that 3/4 of the total amount of the data, therefore the classification of the total time is also increased substantially. The FA systems reduces the average error in a 2% with the Albayzin metric and almost 3% with the NIST metric.

Table 5: Results over the train files to select the priors for each class in %

Prior of each class					**AVG Error**
MU	OT	SM	SN	SP	**over the train files**
0.20	0.20	0.20	0.20	0.20	15.95%
0.22	0.20	0.20	0.19	0.19	14.52%
0.24	0.20	0.20	0.18	0.18	13.75%
0.26	0.20	0.20	0.17	0.17	13.39%
0.28	0.20	0.20	0.16	0.16	**13.23%**
0.30	0.20	0.20	0.15	0.15	13.25%

Table 6: Error per class and total error for Albayzin evaluation winning system and Factor Analysis Segmentation system over the test files in %

	Error for each class					
	MU	SM	SN	SP	**TOTAL**	**NIST**
HMM-Winn	19.2	25.0	37.2	39.5	30.2	-
HMM-Rep	16.3	24.0	38.8	40.8	30.0	19.3
FA-Segm	21.7	27.6	35.4	30.5	28.8	**16.9**
FA-Segm OT	21.7	27.6	34.0	29.5	**28.2**	17.5

5. Conclusion

This paper describes a new segmentation-by-classification system based on Factor Analysis approach. The system has been applied for the segmentation of BN. The task consists of the segmentation of audio files and further classification into 5 different classes as proposed in the Albayzin 2010 evaluation. The solution we propose here compensates the within-class variability creating a channel matrix for each class and scoring the segments as the ratio between class/no-class. This approach has been compared with HMM-GMM baseline systems and with the winning system of the evaluation showing a significant improvement in both cases even if the best results in the evaluation were obtained by an HMM/GMM based hierarchical system that made use of MFCC along with Chroma features. Experimental results show that the FA approach allows a significant reduction in the classification of SP and SN and thus a reduction in the average segmentation error rate.

6. References

[1] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Workshop*, 1998.

[2] L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, Apr. 2003.

[3] T. Nwe and H. Li, "Broadcast news segmentation by audio type analysis," *Acoustics, Speech, and Signal Processing, 2005 ...*, vol. 2, pp. ii–1065, 2005.

[4] Y. Lavner and D. Ruinskiy, "A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, pp. 1–15, 2009.

[5] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, "Classification of audio signals using AANN and GMM," *Applied Soft Computing*, vol. 11, no. 1, pp. 716–723, Jan. 2011.

[6] L. Xie, Z.-H. Fu, W. Feng, and Y. Luo, "Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news," *Multimedia Systems*, vol. 17, no. 2, pp. 101–112, Sep. 2010.

[7] M. Markaki and Y. Stylianou, "Discrimination of speech from nonspeeech in broadcast news based on modulation frequency features," *Speech Communication*, vol. 53, no. 5, pp. 726–735, May 2011.

[8] N. Nguyen, M. Haque, C.-h. Kim, and J. Kim, "Audio segmentation and classification using a temporally weighted fuzzy C-means algorithm," *Advances in Neural Networks ...*, pp. 447–456, 2011.

[9] A. Misra, "Speech/Nonspeech Segmentation in Web Videos," *research.google.com*, 2012.

[10] R. Huang and J. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora," *Audio, Speech, and Language ...*, vol. 14, no. 3, pp. 907–919, 2006.

[11] T. Butko, C. N. Camprubí, and H. Schulz, "Albayzin-2010 audio segmentation evaluation: evaluation setup and results," in *FALA Evaluation*, 2010, pp. 305–308.

[12] A. G. Antolín and R. S. S. Hernández, "UPM-UC3M system for music and speech segmentation," in *Proc. II Iberian SLTech*, 2010, pp. 421–424.

[13] T. Butko and C. Nadeu, "Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, no. 1, p. 1, 2011.

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

[15] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," *Online: http://www. crim. ca/perso/patrick. kenny*, pp. 1–17, 2006.

[16] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[17] C. Vaquero, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification," in *Proc Interspeech 2010*, vol. 2010, 2010, pp. 2310–2313.

[18] C. Vaquero, A. Ortega, and E. Lleida, "Intra-session variability compensation and a hypothesis generation and selection strategy for speaker segmentation," *Acoustics, Speech and Signal ...*, pp. 3–6, 2011.

[19] N. Brummer, A. Strasheim, V. Hubeika, P. Mat\vejka, L. Burget, and O. Glembek, "Discriminative acoustic language recognition via channel-compensated GMM statistics," in *Tenth Annual Conference of the International Speech Communication Association*, 2009, pp. 2187–2190.

[20] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal ...*, no. 4, 1980.

[21] C. Bishop, *Pattern recognition and machine learning*. Springer New York, 2006, vol. 4.

[22] D. a. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[23] H. Li, B. Ma, and K. Lee, "Spoken Language Recognition: from Fundamentals to Practice," *Proceedings of IEEE*, 2013.

[24] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of Nuisance Factors for Speaker and Language Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1969–1978, Sep. 2007.

[25] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, Jan. 2008.

[26] D. Castan, A. Ortega, and E. Lleida, "Factor Analysis Segmentation and Classification in Broadcast News Domain," in *Proc. III Iberian SLTech*, 2012.

[27] D. Castan, C. Vaquero, A. Ortega, D. Martínez, and E. Lleida, "Hierarchical Audio Segmentation with HMM and Factor Analysis in Broadcast News Domain," in *Interspeech*, 2011.

[28] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with Joint Factor Analysis," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Ieee, Apr. 2009, pp. 4057–4060.

[29] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP*, vol. 1. Citeseer, 2005, pp. 637–640.

[30] J. Bilmes, "What HMMs can do," *Graphical Models*, no. 206, 2002.

[31] NIST, "The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan," pp. 1–18, 2009.

Automatic Transcription of Multi-genre Media Archives

P. Lanchantin[1], P.J. Bell[2], M.J.F. Gales[1], T. Hain[3], X. Liu[1], Y. Long[1], J. Quinnell[1]
S. Renals[2], O. Saz[3], M. S. Seigel[1], P. Swietojanski[2], P. C. Woodland[1]

[1]Cambridge University Engineering Department, Cambridge CB2 1PZ, UK
{pkl27,mjfg,xl207,yl467,jq228,mss46,pcw}@eng.cam.ac.uk
[2]Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK
{peter.bell,s.renals}@ed.ac.uk,p.swietojanski@sms.ed.ac.uk
[3]Speech and Hearing Research Group, University of Sheffield, Sheffield S1 4DP, UK
{t.hain,o.saztorralba}@dcs.shef.ac.uk

Abstract

This paper describes some recent results of our collaborative work on developing a speech recognition system for the automatic transcription or media archives from the British Broadcasting Corporation (BBC). The material includes a wide diversity of shows with their associated metadata. The latter are highly diverse in terms of completeness, reliability and accuracy. First, we investigate how to improve lightly supervised acoustic training, when timestamp information is inaccurate and when speech deviates significantly from the transcription, and how to perform evaluations when no reference transcripts are available. An automatic timestamp correction method as well as a word and segment level combination approaches between the lightly supervised transcripts and the original programme scripts are presented which yield improved metadata. Experimental results show that systems trained using the improved metadata consistently outperform those trained with only the original lightly supervised decoding hypotheses. Secondly, we show that the recognition task may benefit from systems trained on a combination of in-domain and out-of-domain data. Working with tandem HMMs, we describe Multi-level Adaptive Networks, a novel technique for incorporating information from out-of domain posterior features using deep neural network. We show that it provides a substantial reduction in WER over other systems including a PLP-based baseline, in-domain tandem features, and the best out-of-domain tandem features.

Index Terms: lightly supervised training, cross-domain adaptation, tandem, speech recognition, confidence scores, media archives

1. Introduction

The British Broadcasting Corporation (BBC) has a stated aim to open its broadcast archive to the public by 2022. Automatic transcription, metadata extraction and indexing of such material would give access to a large amount of content, indexing historic content, and enabling search based on transcriptions, speaker identity and other extracted metadata. However, technologies for this particular task are still underdeveloped. In the scope of the Natural Speech Technology EPSRC project and in collaboration with BBC Research and Development, we have begun to investigate the automatic transcription of broadcast material across different genres, using sparse or non-existent associated metadata and text resources.

This research was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). Thanks to Andrew McParland, Yves Raimond and Sam Davies of BBC R&D

Automatic transcription of arbitrary, multi-genre media content is a challenging task since the material to recognise may include broadcasts in diverse environments and drama with highly-emotional speech, overlaid background music or sound effects. Recent work on this task has for instance included automatic transcription of podcasts and other web audio [1] automatic transcription of Youtube [2, 3], the MediaEval rich speech retrieval evaluation which used blip.tv semi-professional user created content [4], and the automatic tagging of a large radio archive [5]. On the other hand, in order to train models for such large vocabulary continuous speech recognition systems, text resources and other metadata are highly desirable to provide in-domain training data. The problem is that the nature of these metadata may vary considerably over archive material in terms of completeness, reliability and precision. This partly reflects the large epoch (decades) that the data covers. A range of techniques have been proposed for this purpose such as the lightly supervised training approach [6], based on a biased language model (LM) decoding, and several methods have since been proposed along this line to improve upon this approach [7, 8, 9, 10].

In recent work described in [11, 12] which will be reviewed in this paper, we focused on two aspects related to the building of systems for automatic transcription of multi-genre media archives: lightly supervised training and evaluation using out-of-domain data. We recently proposed in [12] an approach in which phone level mismatch information is used to identify reliable regions where segment-level transcription combination can be used. Schemes for combining the imperfect original transcriptions with the confusion networks (CN) generated during the biased LM decoding can then be applied to leverage differences in the characteristics of the two forms of transcriptions. An evaluation technique based on ranking systems using imperfect reference transcripts was used to evaluate system performance. Secondly, in [11], we focused on the development of methods which can effectively combine in-domain and out-of-domain training data, using neural networks in the tandem framework [13] whereby context-dependent hidden Markov models (HMMs) with Gaussian mixture model (GMM) output distributions are trained on standard acoustic features concatenated with features derived from neural networks. A novel technique for posterior feature combination in a cross-domain setting and referred to as Multi-Level Adaptive Networks (MLAN) was then proposed. This technique has been investigated using a multi-genre broadcast corpus built from the data provided by the BBC, in terms of cross-domain speech recognition using different acoustic training data sources across different target genres.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

The new technique was evaluated in terms of a discriminatively-trained speaker-adaptive speech recognition system, comparing in-domain and out-of-domain posterior features with the features obtained using MLAN.

The rest of the paper is organised as follows. In Section 2 the available BBC datasets are presented. Section 3 presents lightly supervised approaches for the correction of timestamp positions and the proposed transcription combination schemes. Finally, Section 4 presents the multi-level adaptive network scheme for the transcription of multi-genre data followed by conclusions in Section 5.

2. Description of the BBC datasets

The stated aim of the BBC to open its broadcast archive to the public by 2022 will give access to a very large amount of data: potentially 400,000 television programmes, over 700,000 hours of video and 300,000 hours of audio. A large amount of metadata associated to these data will be available from the *Infax* cataloguing system which allows to access tags manually attributed to programmes in varying levels of detail (more than 600,000 items) some of which are already publicly available. In the scope of our collaboration with BBC research and development started in 2011, six different sets of shows with their associated metadata have been provided for the investigation and the development of methods and systems for automatic transcription of broadcast material across the full range of genres.

2.1. Diverse shows/genres

The six sets contain speech that is mostly British English with a range of regional accents and audio contents covering a broad range of genres, environments and speaking styles that we describe below.

Radio4-1day: contains 36 talk-radio programmes broadcast on the same radio channel (BBC Radio 4) over 24 hours in February 2009. The duration of programmes range from 2 minutes for weather report to 3 hours for morning news/current affair programmes to give a total duration of 18 hours. The audio material covers different genres: news, weather reports, book readings, documentaries, panel games and debates.

Archives: contains 136 radio and TV programmes some of which are publicly available on the BBC archives website (http://www.bbc.co.uk/archive). It includes 399 episodes representing 271 hours of raw audio data with 146 hours of active speech. Episodes were recorded from 1970 to 2003. As for the Radio4-1day dataset, audio material covers a broad range of genres, environments and speaking styles.

Desert Island Discs: is a radio programme broadcast on BBC Radio 4. Each week, a guest is asked to choose eight pieces of music, a book and a luxury item that they would take if they were to be castaway on a desert island, whilst discussing their lives. It includes only two speakers in each show, the presenter and the guest, and small portions of music. This set includes 180 episodes representing 108 hours of raw data with 88 hours of active speech.

Reith Lectures: are a series of annual radio lectures on significant contemporary issues, delivered by leading figures from their relevant fields. The set includes 155 episodes, covering the years from 1976 to 2010. Each lecturer had 3-6 episodes presented at different times. Each episode is composed of several regions: the lecture region given by the lecturer, a non-lecture region which contains the introduction to the lecture by a presenter and since 1988, a question and answer session after the main lecture. The duration of each episode ranges from 18-35 minutes, to give a total audio duration of 72 hours from which 71.3 hours of lecture region data were extracted.

TV-drama: includes 14 episodes of a science fiction TV-drama series broadcast in 2010. Episode durations range from 40-75 minutes, to give a total duration of 11 hours.

TV-1week: includes 169 unique shows and 333 episodes broadcast on 4 BBC TV channels during the week of May 5th, 2008 through May 11th, 2008 representing 236 hours of raw audio data. The duration of the programmes ranges from 3 minutes to 4 hours. A list of genres covered by the programmes was provided with up to 85 different categories, although programmes typically get assigned to more than one genre. This categorisation includes drama series, soap operas, different types of documentaries, live sports, broadcast news, quiz shows or animation programmes.

The available audio material contained in these sets covers different genres and a broad range of environment and speaking style. For purposes of analysis, we divided the data into three categories by broad genre:

`studio`: in which speech is controlled, recorded in studio conditions or news reports, sometimes including telephone speech from reporters or contributors;

`location`: which includes material produced on "location" including for instance parliamentary proceedings;

`drama`: TV drama series, containing dramatic, fast emotional speech, and high background noise levels, making ASR particularly challenging.

2.2. Available metadata

Metadata associated to the dataset presented in the last section varies over time, shows and media type. These can be more or less complete, accurate and reliable. In the following we first classify the metadata into three types. We then introduce the issues related to each type of metadata.

`type1`: transcriptions are produced manually and timestamps are provided (quantised to 1s) as well as speaker names and additional metadata such as indications of music or sound effects. This type of metadata is available for Radio4-1day, Desert Island Discs and the Archives dataset.

`type2`: transcriptions are not verbatim, timestamps are not provided and a number of errors which depend on the degree to which the speaker deviated from the original script. This type of metadata is typical of the Reith Lectures dataset in which scripts were used by lecturers from which they were free to deviate.

`type3`: transcriptions are derived from subtitles for hearing impaired, timestamps are provided as well as and other metadata such as an indication of music and sound effects, or indications of the way the text has been pronounced. Most of the shows include several speakers. Speaker identities are indicated by the use of several different text "colours" (which are used for subtitle display) . Timestamps were found to be unreliable due to time-lags that occur in subtitles, presumably arising from the re-speaking process for subtitle creation. This type of metadata is the one used for the TV-drama and TV-1week datasets.

These different types of metadata can be characterised in terms of completeness, accuracy and reliability. The metadata

can be more or less *complete*: the transcription can cover all the episode, or just a part of it, the timestamp information can also be available or not (e.g `type2`). The available metadata also varies over shows: some include speaker ID, sound event indications, title of music, programme genre. In terms of *accuracy*, transcriptions may include annotation of disfluencies and quantisation of the timestamps also may vary over shows (e.g 1ms for `type3` to 1s for `type1`). Finally, the *reliability* varies over the different types of metadata: `type1` include manual transcriptions and are considered to be more reliable even though they might include some variations depending on the transcriber and some episodes transcribed according to `type3` were found to have time-lag. Finally the reliability of `type2` metadata varies over episodes depending on speakers who can deviate differently from scripts.

3. Lightly Supervised Approaches

Most of the issues related to metadata described in the last section may be solved by lightly supervised approaches. In conventional lightly supervised training [6], a biased language model (LM) trained on the transcriptions (closed-captions) is used to recognise the training audio data. The recognition hypotheses are then compared to the close-captions and matching segments are filtered to be used in re-estimation of the acoustic model parameters. The entire process is carried out iteratively, until the amount of training data obtained converges. This kind of approach can first be used for the correction of timestamps when these are unreliable, imprecise or simply non-existent such as `type2` metadata. It then can be used when transcriptions are unreliable in order to select data for the training of acoustical models. We first describe our method for timestamp correction before presenting our approach for non-reliable transcription based on combined transcriptions. We finally investigate an evaluation technique based on ranking systems using imperfect reference transcription when no reference transcription is available.

3.1. Timestamp correction

Timestamps can be inaccurate due to quantisation effects (`type1`), unreliable due to time-lags that can occurs in subtitles (`type3`) or simply nonexistent (`type2`). They can however be corrected using a lightly supervised approach in the following way [14], which will also be used in section 3.2. Each show is first segmented and segments are clustered according to speakers using the CU RT-04 diarisation system [15]. Each speech segment is decoded using a two-pass[1] (P1-P2) recognition framework [16, 17] including speaker adaptation, with the decoding employing a biased language model (LM). This biased LM is initially trained on the original transcription (denoted as *origTrans* in the following) and then interpolated with a generic language model, with a 0.9/0.1 interpolation weight ratio. This results in an interpolated LM biased to the original in-domain transcripts. The vocabulary is chosen to ensure coverage of words from the original transcripts. The decoder output is then compared with the raw transcription to identify matching sequences. Non-matching word sequences from the raw transcription are force-aligned to the remaining speech segments. Finally, once realigned, the position of timestamps can be corrected.

[1] the output lattices generated in the second pass (P2 stage) when generating the 1-best hypotheses are used to generate confidence scores for both automatic transcriptions and the original transcriptions in section 3.2.

3.2. Combined transcriptions

There are two main issues with the conventional lightly supervised approaches related to `type2` metadata. As the original imperfect transcriptions deviate more from the correct ones, the constraints provided by the biased LM are increasingly less appropriate. This leads to a greater mismatch between the original transcriptions and the biased LM decoding hypotheses, which results in a reduction in the amount of usable training data after filtering is applied. Moreover, information pertaining to the mismatch between the original transcriptions and the automatic decoding outputs is normally measured at the sentence or word level. As acoustic models used in current systems are normally constructed at the phone level, the use of phone level mismatch information is preferable [9]. In [12], we proposed a method for the selection of training data using unreliable transcriptions. In this method, phone level mismatch information is used to identify reliable regions where segment-level transcription combination can be used. Schemes for combining the imperfect original transcriptions with the confusion networks (CN), generated during the biased LM decoding, can then be applied to leverage the different characteristics of the two forms of transcriptions.

3.2.1. Segment-level combination

Mismatch information at phone level is useful in order to derive combined transcriptions for the selection of training data. In order to exploit this information when the original and automatically decoded transcriptions disagree significantly, segment level phone difference rate[2] (PDR) is used to select the segments in the original transcriptions (*origTrans*) that can be combined with the automatically derived hypotheses (*aHyp*) outputs. To do so, (i) *origTrans* is first mapped into each of the *aHyp* segments using standard dynamic programming alignment, unmapped words being discarded. (ii) The mapped transcriptions are then force-aligned to obtain the phone sequences from which (iii) the PDR between the two force-aligned phone sequences can be calculated, if both exist. Finally, (iv) segment selection can be performed by selecting segments from *origTrans* which have a PDR values less than a threshold optimised on a held-out dataset. The remaining segments are then filled in to yield the transcriptions for the full training data set.

3.2.2. Word-level combination

When the mismatch between the original transcripts and the 1-best biased LM decoding hypotheses is large, the amount of training data is reduced dramatically. In this case, the hypotheses can be combined with the original transcripts by considering word level consensus networks [18], in order to limit this reduction. However, the assumption that the imperfect transcription is always present in the biased LM CN network can be too strong in cases like `type2` transcriptions in which lecturers may deviate significantly from their initial script. To handle this issue, a modified word level CN based transcription combination scheme can be used: if the word given by the original transcription is not found in the lattice, the word with the highest confidence score in the biased LM lattice is selected. To do so, (i) *origTrans* is first mapped into each of the *aHyp* segments as was carried out for the segment-level combination. (ii) Using the lattices generated in Section 3.1 to obtain the *aHyp* segments, the lattice arc posterior ratio (LAPR) presented in [19] is calculated as the confidence score (CS) for each word in *aHyp*. (iii) A "virtual" confidence score (because they are not confi-

[2] the traditional segment-level phone error rate is calculated but this is a PDR as there are no accurate transcriptions

dence scores in the usual sense) based on hard assignment is associated with each word in the mapped *origTrans*. If there are alternative word candidates in the lattices which agree with the word in *origTrans*, a score larger than the maximum value of LAPR is assigned as the confidence score (1.2), otherwise, the confidence score is set to 0.0. Finally, (iv) after confidence scores have been assigned to all words in both *aHyp* and in *origTrans*, ROVER [20] is used, taking the confidence scores into account, to do the transcript combination, yielding the final set of "best" word sequences for each segment.

3.3. Evaluation considering relative measures

Most lightly supervised training research has been focused on improving only the quality of the training transcriptions, assuming that the correct transcriptions are available for test data used in performance evaluation. However, for many practical applications accurate transcriptions that cover many diverse target domains can be impractical to manually derive for both the training and test data. Hence, alternative testing strategies that do not explicitly require correct test data transcriptions are preferred [21]. Here, we investigated the reliability of a performance rank ordering, given by the *origTrans* as an approximate reference transcription. Should such a rank ordering be consistent with that generated by the gold standard reference on the hand labelled data, it was then hoped that *origTrans* could be used for other larger sized test sets that don't have accurate transcripts

3.4. Experiments and results

To validate our proposed approach, experiments were run on the Reith Lectures dataset for which metadata are of `type2` as lecturers deviated more or less from their original prepared scripts during their speech. For the experiments, data were divided into a training set of 68 hours, a test set of 2.5 hours and two episodes of 0.8 hours of gold standard transcripts. A first comparison between *origTrans* and *aHyp* transcriptions carried out at the episode level, according to the word difference rate (WDR[3]) in the lecture regions, showed that difference rates vary strongly between speakers. The effectiveness of the segment and word level combination approaches was then validated on the gold standard transcrips, both word-level and best segment-level combined transcriptions achieving similar significant reductions in phone error rate (PER) and word error rate (WER) over the performance of the *origTrans* and *aHyp* transcriptions indicating that more accurate transcriptions could be obtained from the transcriptions combination. Given these preliminary results, we then investigated how real speech transcription systems are affected by training acoustic models using the combined training data transcriptions. Results obtained from the real transcription systems and detailed in [12] showed that both of the combination approaches investigated provide more accurate transcriptions than the original lightly supervised transcriptions, resulting in improved ML and MPE models. For MPE models, a reduction of 0.6% absolute and 1.1% absolute of WDR is obtained when using segment and word level combined transcriptions respectively, instead of *aHyp* (17.4% WDR), when added to a multi-genre broadcast dataset with accurate transcriptions. We also showed that rank ordering of the WER and WDR pairs derived from *origTrans* and from the gold standard transcript was consistent, allowing to use the *origTrans* as reference for other larger sized test sets that don't have accurate transcripts.

[3]The WDR is calculated in the same manner as the traditional word error rate, but this is a WDR as there are no accurate transcriptions

4. Multi-genre transcription using out-of-domain data

We now move our focus to a second aspect of the development of systems for the automatic transcription of media Archives which aim to effectively combine in-domain and out-of-domain training data. State-of-the-art transcription systems built for domains such as conversational telephone speech (CTS), and North American broadcast news (BN) perform with low accuracy on multi-genre data such as the BBC ones described in section 2. This is mostly due to the high mismatch in environment, speaking style, speaker and accent. Unsurprisingly, in-domain HMM-GMM systems trained on these data outperform these out-of-domain (OOD) systems, despite the fact that there is an order of magnitude less in-domain training data. For the purpose of the transcription of BBC archives, we then focused on the development of methods which can effectively combine in-domain and OOD training data using neural networks. Intensive research has been carried out recently on deep neural networks (DNNs) with promising results [22, 23]. We have used DNNs with generative pre-training to obtain posterior features used in the tandem framework [13] which is attractive for cross-domain modelling, since it allows independent adaptation of the GMM and DNN parameters. We recently proposed in [11] a novel technique called Multi-Level Adaptive Networks (MLAN) for posterior feature combination in a cross-domain setting. This technique, which will be presented below, has been investigated on a subset of the BBC dataset presented in section 2 in terms of cross-domain speech recognition using different acoustic training data sources across different target genres. It has then been evaluated in terms of a discriminatively-trained speaker adaptive speech recognition system, by comparing in-domain and out-of-domain (OOD) posterior features obtained using the proposed method.

4.1. Multi-Level Adaptive Networks

In our proposed method, DNNs are trained to model frame posterior probabilities over monophones. The structure of the DNNs is fixed following analysis of the frame error rate on held-out validation data and monophone log-posterior probabilities output from the nets are decorrelated using a single PCA transform with dimensionality reduced to 30 [13] to obtain the posterior features. These are then concatenated with the original acoustic features. Using initial OOD DNN adapted to a new domain, can be viewed as imposing a form of regularisation on the resulting net. However we observed small benefits when using deep architectures and fairly large quantities of in-domain data. We therefore proposed an alternative adaptation procedure called *Multi-level Adaptive Networks* (MLAN). In the first level of this MLAN scheme, networks trained on OOD acoustic data are used to process in-domain acoustic data to generate posterior features, which are concatenated with the original in-domain acoustic features as in the tandem framework. We would expect the OOD posterior features to enhance the discriminative abilities of the simple in-domain acoustic features. In the second level, additional DNNs are trained, using the first level tandem features as input, to minimise an in-domain objective function of log-posterior phone probabilities. The outputs from these DNNs are then used to generate the final tandem features for HMM training. Finally, by expanding the input tandem feature vector used at the second level, output from multiple networks, trained on different domains, may be included with no modification to the architecture. The main motivation for the MLAN scheme is that the new DNNs, trained discriminatively,

Feature set	1-pass (unadapted)				2-pass (adapted)			
	Studio	Location	Drama	All	Studio	Location	Drama	All
PLP	12.0	25.9	58.8	32.7	11.5	23.6	58.9	31.8
BBC tandem	11.7	23.3	54.9	30.4	11.3	22.3	54.4	29.8
AMI tandem	11.3	22.6	55.0	30.1	11.1	21.5	54.2	29.4
AMI+CTS MLAN	10.2	20.9	50.5	27.6	9.8	20.0	50.2	27.1

Table 1: Final MPE system results (WER%) on the 2.3h test set using PLP, tandem and MLAN features.

are able to learn which elements of the OOD posterior features are useful for discrimination in the new domain; whilst the direct inclusion of in-domain acoustic features in the input means that the resulting frame error rates ought never to be worse than DNNs trained purely in-domain. The additional generative pretraining carried out ensures that the new DNN does not over-fit to the in-domain data. More details (e.g DNN structure) and explanation of the method can be found in [11].

4.2. Experiments

Experiments were conducted on the Radio4-1day and the TV-drama dataset divided into the three categories by broad genres defined in Section 2.1 (`studio`, `location`, `drama`). Transcriptions were found to be reliable but timestamps were corrected according to the procedure detailed in Section 3.1, giving a total of 23 hours of transcribed and aligned speech in total. The data were divided at the show level into a training set of 20.7 hours and a test set of 2.3 hours, each containing roughly the same balance across genres. For the out-of-domain data, two diverse sets were used. The first one included 277 hours of US-English conversational telephone speech (CTS) taken from the switchboard I, switchboard II and CallHome corpora. The second set consisted of Recordings from the Augmented Multi-Party Interaction (AMI) corpus. Concerning the system architectures, development experiments were performed using a simple one-pass system and the final evaluation system was trained using MPE discriminative training [24] and had a two-pass decoding architecture.

4.2.1. Development experiments

Recognition of the test set was first performed using two OOD acoustic models trained on PLP features from the AMI and CTS training set. The results demonstrate the large acoustic mismatch between these domains and the BBC domain. The performance of tandem features was then investigated by comparing models trained purely on the BBC training set with models trained on tandem features obtained using OOD nets. It was found that OOD tandem features from AMI and CTS improved performance for all genres (with the overall WER initial value equal to 39.4% reduced by 5.6% absolute and 3.9% absolute using AMI and CTS features respectively) compared to simple PLP features supporting earlier work suggesting that posterior features are portable across domains. With respect to the broad genres, it was found that CTS and AMI OOD posteriors are both better for Studio speech by comparison with the BBC tandem results, AMI is best for Location speech and equally matched with in(domain features for Drama speech, which is the genre most mismatched to the OOD acoustic models. Performance of the MLAN was then investigated and showed substantial additional gains over standard tandem features, for both domains. The CTS posteriors which were worst-matched to the BBC domain, gain the most benefit from MLAN with a 3.6% absolute WER reduction overall (initial value 35.5%). The combination of both OOD posterior features with MLAN reduces WER still further, suggesting the second-level DNN is successfully able to exploit complementary information between AMI and CTS.

4.2.2. Final system evaluation

For the final system evaluation, the best-performing in-domain and out-of-domain tandem features, and the best MLAN features, were selected for use in training a more competitive final system. Table 1 shows the final system results on the test set with and without speaker adaptation. The HMMs were trained with MPE only on the BBC training set using STC-projected PLP features and the relevant posterior features. All the new features outperformed the baseline PLP features in both the unadapted and speaker adapted MPE systems. This supports the preliminary results from the development system and indicates that the posterior features can bring complementary information to the PLP features even when the HMMs are trained using MPE. Moreover, the overall improvement over the baseline PLP features, in both the unadapted speaker-adapted systems was dramatic, with absolute WER reductions of 5.1% and 4.7% respectively. Table 1 shows that speaker adaptation is effective in reducing the WER for all three posterior feature sets, compared with the baseline PLP features which only offers gains for the Location and Studio subsets, although for these two subsets, the gains from adaptation are larger than for the posterior features. It was then hypothesised that the posterior features are better able to capture speaker-invariant information in these subsets, whilst in the noisy drama subset, are able to model speaker-dependent structures more effectively than PLPs.

5. Conclusions and Future work

We presented our joint work on the development of a speech recognition system for multi-genre media archives from the BBC using limited text resources. We first described the different BBC datasets which were provided with their diverse audio content and metadata.We then focused on improving the transcription quality of acoustic model training data for the BBC archive task. Combination at both the word and segment level-level of the original transcriptions, with the lightly supervised transcription generated by recognising the audio using a biased language model has been presented. This provides more accurate transcriptions than the original lightly supervised transcriptions, resulting in improved models. We then presented the MLAN method for recognition of multi-genre media archives with neural network posterior features, successfully using out-of-domain data to improve performance. Results consistently show that our Multi-Level Adaptive Networks scheme results in substantial gains over over other systems including a PLP-based baseline, in-domain tandem features and the best out-of-domain tandem features. Future work will investigate further transcription combination approaches and testing schemes with imperfect transcription references. We also plan to investigate the MLAN technique in an HMM-GMM system that also incorporates speaker-adaptive training and fMPE transforms and to adapt the method for use in a hybrid DNN system. Finally the proposed approaches will be conducted on larger datasets such as Archives and TV-1week.

6. References

[1] J. Ogata and M. Goto, "Podcastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription," in *Proc. Interspeech*, 2009.

[2] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An audio indexing system for election video material," in *Proc ICASSP*, 2009, pp. 4873–7876.

[3] R. C. van Dalen, J. Yang, and M. J. F. Gales, "Generative kernels and score-spaces for classification of speech: Progress report," in *Tech. Rep. CUED/g-infeng/th.676, Cambridge University Engineering Department*, 2012.

[4] M. Larson, M. Eskevitch, R. Orderlman, C. Kofler, S. Schmiedeke, and G. J. F. Jones, "Overview of Mediaeval 2011 Rich speech retrieval task and genre tagging task," in *Working Notes Proceedings of the MediaEval 2011 Workshop*, 2011.

[5] Y. Raimond, C. Lowis, R. Hodgson, and J. Tweed, "Automatic semantic tagging of speech audio," in *Proc. WWW 2012*, 2012.

[6] L. Lamel, J. Gauvain, and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," in *Computer Speech and Language*, vol. 16, 2002, pp. 115–129.

[7] H. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. ICASSP*, vol. 1, 2004, pp. 737–740.

[8] L. Mathias, G. Yegnanarayanan, and J. Fritsch, "Discriminative Training of Acoustic Models Applied to Domains with Unreliable Transcripts," in *Proc. ICASSP*, vol. 1, 2005, pp. 109–112.

[9] B. Lecouteux, G. Linares, P. Nocera, and J. Bonastre, "Imperfect transcript driven speech recognition," in *Proc. InterSpeech'06*, 2006, pp. 1626–1629.

[10] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V. Gadde, and J. Zheng, "An Efficient Repair Procedure for Quick Transcriptions," in *Proc. ICSLP*, 2004.

[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. SLT*, 2012.

[12] Y. Long, M. J. F. Gales, P. Lanchantin, X. Liu, M. S. Seigel, and P. C. Woodland, "Improving lightly supervised training for broadcast transcriptions," in *Proc. Interspeech*, 2013.

[13] H. Hermanksy, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1630.

[14] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. Interspeech*, 2010, pp. 2222–2225.

[15] S. Tranter, M. Gales, R. Sinha, S. Umesh, and P. Woodland, "The development of the Cambridge University RT-04 diarisation system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[16] G. Evermann and P. Woodland, "Design of fast LVCSR systems," in *Proc. ASRU Workshop*, 2003.

[17] M. Gales, D. Kim, P. Woodland, H. Chan, D. Mrva, R. Sinha, and S. Tranter, "Progress in the CU-HTK Broadcast News Transcription System," in *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, 2006, pp. 1513–1525.

[18] L. Chen, L. Lamel, and J.-L. Gauvain, "Lightly supervised acoustic model training using consensus networks," in *Proc. ICASSP*, vol. 1, 2004, pp. 189–192.

[19] M. Seigel and P. Woodland, "Combining information sources for confidence estimation with CRF models," in *Proc. Interspeech*, 2011, pp. 905–908.

[20] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. ASRU Workshop*, 1997, pp. 347–352.

[21] B. Strope, D. Beeferman, A. Gruenstein, and X. Lei, "Unsupervised Testing Strategies for ASR," in *Proc. Interspeech*, Florence, Italy, 2011.

[22] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[23] A. Mohammed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[24] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, pp. 105–108.

SLIGHTLY SUPERVISED ADAPTATION OF ACOUSTIC MODELS ON CAPTIONED BBC WEATHER FORECASTS

Christian Mohr, Christian Saam, Kevin Kilgour, Jonas Gehring, Sebastian Stüker, Alex Waibel

International Center for Advanced Communication Technologies (interACT)
Institute for Anthropomatics
Karlsruhe Institute of Technology, Karlsruhe, Germany
{firstname.lastname}@kit.edu

Abstract

In this paper we investigate the exploitation of loosely transcribed audio data, in the form of captions for weather forecast recordings, in order to adapt acoustic models for automatically transcribing these kinds of forecasts. We focus on dealing with inaccurate time stamps in the captions and the fact that they often deviate from the exact spoken word sequence in the forecasts. Furthermore, different adaptation algorithms are compared when incrementally increasing the amount of adaptation material, for example, by recording new forecasts on a daily basis.

Index Terms: speech recognition, acoustic model adaptation, slightly supervised training, loose transcripts, adaptation methods

1. Introduction

Within the European Union's 7th Framework Programme's project (Bridges Across the Language Divide) (EU-BRIDGE) [1] several tasks on automatic speech recognition are defined over different data sets. The active domains are TED talks[2], a collection of public talks covering a variety of topics, academic lectures and weather bulletins. For the TED task large collections of training data are readily available which are the basis for the IWSLT ASR evaluation track [1]. The mismatch between training and testing data pertains to speaker and domain yet style is relatively consistent. The approximate transcripts of the talks are very close to verbatim. For lectures there is comparatively little training data available. Thus, general models are adapted on small data sets that often do not even have transcripts. Unsupervised adaptation must account for mismatches in speaker, domain and style. The weather bulletin data on the other hand is a new and still very small data set that has weak references in the form of captions. Again, general models must be adapted in a supervised/semi-supervised manner to account for mismatches in style, domain and speakers.

This paper investigates different approaches for acoustic model adaptation on weather forecasts when captions are available. Of special interest is the question of how to deal with imperfect transcripts and unlabeled non-speech audio as investigated by [2]. Similar to [3] we investigate the possible improvements of a system by unsupervised acoustic model training depending on the amount of training data and the reliability of transcripts. Similar to [4, 5], we made use of word level confidence scores. However, we did not exclude data from training based on the word posteriors of the transcription, as we have too little training data available as that we could afford to lose some of it. Our training conditions can be compared to [6] where new data for retraining comes from the same speaker, channel and related conversation topics. Following the implications of [7] we add low confidence score data to the training, but unlike in other work we apply word-based weighting in order to compensate for errors, as it was done by [8] for acoustic model adaptation. The assumption is that erroneous data is helpful to improve system generalization. Unlike other work, e.g. [9], we did not use a lattice-based approach. Furthermore we study the choice of a good adaptation method with increasing adaption set sizes. We assume that sufficient amounts of training data are available in order to transition from transform based techniques, such as maximum likelihood linear regression and its feature space constrained version [10], to maximum likelihood [11] or maximum a posteriori parameter re-estimation [12].

[1] http://www.eu-bridge.eu
[2] http://www.ted.com

2. The BBC Weather Data

The BBC weather data consists of audio recordings of British weather forecasts and manually generated captions. There are two different kinds of forecasts: general bulletins and regional forecasts. The captions for the general bulletins are prerecorded and therefore more accurate than the live captions for the regional weather forecasts.

The data used consists of audio of forecasts recorded between 2008 and 2012 with roughly 50 different speakers. This information is only an estimate since the tagging of speaker names is partly imprecise and inconsistent, and the airing date of the shows is not always given.

Although the speakers are well trained there are some hesitations, grammar errors or lengthy formulations in the recordings which are corrected in the captions (some examples are shown in Table 1). The captions therefore can only be regarded as loose transcripts.

Capt.	*We had some more typical summer weather*
Verb.	*We had some more **of this** typical summer weather*
Capt.	*Downpours across England and Wales*
Verb.	*Downpours **whistling** across England and Wales*

Table 1: *Two examples for differences between captions (Capt.) and the verbatim word sequences (Verb.). Words omitted in the caption are bold-faced.*

Captions are only provided for the forecast itself with time markers relative to the beginning of the forecast but without ab-

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

solute positions in the recording. Also, the recordings often contain untranscribed parts at the beginning—such as trailers and introductions by different speakers—and advertisements at the end. The length of the untranscribed parts in the audio differs, so it is not possible to simply cut it off at a specific time, in order to just obtain the portion of the data that is actually captioned.

For the test corpus described in Section 5 careful transcriptions were available in addition to the captions also covering only the forecast itself, leaving out introduction and trailers. To determine the general degree of faithfulness of the captions as a (training) reference we computed the word error rate (WER) between the verbatim references and the captions. Table 2 shows the result of this on the test data. It can be seen that the captions and the verbatim transcriptions are rather close, indicating that the speakers are indeed well trained.

	Case Sensitive	Case Insensitive
WER	7.4%	5%
# words in reference	12007	12007
Total # errors	890	600
# Substitutions	434	144
# Insertions	21	21
# Deletions	435	435

Table 2: *WER between the captions and the verbatim transcripts of the test set, and statistics on the types of errors.*

The captions' data format contains timestamps that indicate when individual captions are displayed, however these need not exactly correspond to when the respective words were spoken, because captions have to adhere to further constraints in addition to when they were spoken. E.g., they have to adhere to a certain letter rate in order to be readable, have to maintain a certain distance from scene changes and may not span several scenes. The timing information is therefore too inaccurate to be taken as timestamps in the audio.

3. Preprocessing: Finding Suitable Start and End Times

Due to the inaccuracy of the timing information we need to align the captions to the audio to be able to use them as loose transcripts. A naïve Viterbi alignment of the concatenated captions to the corresponding audio file leads to suboptimal results due to the large untranscribed parts in the audio.

To make sure that we use only audio that is properly transcribed we decode the audio data, align the resulting hypotheses to the captions and search for the first and last matching trigram. The start time of the first word of the first trigram is used as the start time of the loose transcript and the end time of last the word of the last trigram as end time. The words preceding the first trigram and following the last trigram are deleted from the transcript. This leads to some data loss but the start and end times can be iteratively refined by repeating the decoding and cutting after the model was adapted on the data obtained in the previous iteration.

Even after one iteration of model re-estimation the amount of data that is lost due to the cut-off is rather small. We tested the approach on a subset of 16 hours of audio data (parts 1-4 of the final database as described in Section 4). The acoustic model as well as the language model of the system used for decoding were trained on British general broadcast data. The baseline system is described in more detail in Section 5. On the test set described in Section 5 this system's WER was 31.9%. The baseline system was adapted on the raw recordings and then achieved a WER of 23.2%. This adapted system in turn was used to refine the start and end times of the audio it was adapted on. After cutting, the amount of audio data was reduced by approximately 37% but the text of the original captions only by around 6%. When applying this method to the final database, the reduction of audio data decreased to 35.1% and the percentage of removed words in the reference to 4.9%. So the cut off audio data consists of a small part of transcribed data plus a very large part of unwanted data.

The different results for the subset and the final database result from the small amount of data in total and from the fact that the length of introductions and trailers differs significantly. Although this heuristic for finding usable start and end times is rather simple, it is convenient for the given task, as only 4.9% of words in the reference were lost.

4. Experimental Set-Up and Data

All experiments were performed with the *Janus Recognition Toolkit* (JRTk) developed at Karlsruhe Institute of Technology and Carnegie Mellon University [13].

The training of our *Hidden Markov Model* (HMM) based acoustic model tries to maximize the likelihood of the model on the training data. In *Viterbi* training only the most probable HMM state sequence is computed and used for re-estimating the HMM's parameters. In *Expectation Maximization* (EM) training all possible alignments are taken into consideration for model estimation. Both training techniques work iteratively and require an initial set of model weights which are improved over several iterations of model re-estimation. Adaptation can be done by performing one iteration of model parameter estimation on new adaptation data using an existing set of models that was trained on different, out-of-domain data. As an alternative *maximum-a-posteriori* (MAP) estimation using the models of an existing speech recognizer as seed models for the ML estimation of the model parameters was investigated—again on the adaptation data. Various weighting factors τ to control the influence of the seed model were evaluated. We denote the MAP weights as *(Weight of the seed model* $\cdot 100$ | *Weight of the adaptation data* $\cdot 100$*)*.

From past experience these approaches are known to outperform *maximum likelihood linear regression* (MLLR) adaptation of acoustic models when the training data exceeds roughly 1.5hrs. The amount of available adaptation data suggested MLLR adaptation to be inferior, thus it was omitted.

Since the captions are not verbatim transcripts we expected the Viterbi as well as the EM training to suffer from transcription errors. The EM algorithm should not be affected as badly as the Viterbi approach, since all possible HMM state sequences are considered and not only the most likely one. To overcome the problem of transcription errors we tried altering the transcripts by introducing successions of filler states between words, that are intended to be aligned to feature vectors from words missing from the transcript. As a final alternative we tested two kinds of unsupervised adaptation on transcripts of the adaptation data that are in fact hypotheses produced by the unadapted speech recognition system. The statistics accumulated in training over these transcripts are either weighted by the confidence value of the respective hypothesis word or the weights are set to 1.0 for all words.

We split up the data and adapted the general system de-

scribed in Section 5 with the different algorithms on different growing subsets of the database. Periodically new packages of data were made available. Our final database consists of the 6 parts described in Table 3.

Part Number	# files	Comment	Duration / hours (net duration)
1	50	bulletins part 1	3.87 (2.43)
2	50	bulletins part 2	4.04 (2.48)
3	50	bulletins part 3	3.89 (2.44)
4	51	bulletins part 4	3.88 (2.49)
5	103	bulletins part 5	7.46 (4.86)
6	54	regional forecasts	1.07 (1.00)
Σ			24.21 (15.7)

Table 3: *Overview of the parts the final database consists of. The size of the general bulletins files varies between 180 and 410 seconds.*

Part 6 (regional forecasts with live captions) contains captions considered to be less verbatim even than the material in the rest of the database. These captions are produced on the fly during live airings and the results depend on the ability of the captioner to keep up with the speaking rate of the presenter.

Since not all parts of the data were available when the experiments began, we tested the general viability of some adaptation approaches only on initially available subsets of the final training data. We tested only the most promising techniques on the larger databases.

5. Results

For all tests a semi-continuous system was used as baseline system to be adapted on the given adaptation data.

As front-end we used *mel-frequency cepstral coefficients* (MFCC) with 13 cepstral coefficients. The mean and variance of the cepstral coefficients were normalized on a per-utterance basis. 15 adjacent frames were combined into one single feature vector. The resulting feature vectors were then reduced to 42 dimensions using *linear discriminant analysis* (LDA).

The acoustic model is a context dependent quinphone system with three states per phoneme, and a left-to-right topology without skip states. It uses 24,000 distributions over 8,000 codebooks. The model was trained using *incremental splitting of Gaussians* (MAS) training, followed by *semi-tied covariance* (STC) [14] training using one global transformation matrix, and one iteration of Viterbi training. The acoustic models have up to 128 mixture components per model and a total of $591k$ Gaussian components. All models use *vocal tract length normalization* (VTLN)[15].

The system was trained on about 200 hours of carefully transcriped British general Broadcast data.

A baseline 4gram case sensitive language model with modified Kneser-Ney smoothing was built for 36 sources with a total word count of 2,935.6 million and a lexicon size of $128k$ words. This was done using the SRI Language Modeling Toolkit [16]. The language models built from the text sources were interpolated using interpolation weights estimated on a tuning set resulting in a language model with $59,293k$ 2grams, $153,979k$ 3grams and $344,073k$ 4grams. For decoding, a pronunciation dictionary was used containing $142k$ entries.

A second, smaller 4-gram language model was trained on the references of the acoustic model training data containing $61,738$ words increasing the lexicon size to $129k$ words. This was interpolated with the baseline language module to produce an adapted language model. Adding pronunciations and variants for the new words in the lexicon to the pronunciation dictionary increased its size to $144k$ entries.

5.1. Test Set

The test set contains 54 minutes of general weather bulletins, the captions for which were manually corrected to be verbatim transcripts. Correct start and end times were also manually determined.

5.2. First Adaptation Tests

First adaptation tests were done on a subset of the final database originally containing 16 hours of audio data and 10.6 hours after recalculation of start and end times as described in Section 3. Of all 6 parts of the final database the first tests were only done on the first 4. Table 4 shows a comparison of the results of the adaptations via one iteration of the Viterbi or the EM algorithm, and the Viterbi-based MAP estimation. Viterbi re-estimation using the original start and end times was used as an additional baseline.

To limit time and memory consumption a segmentation of the audio files using a partial Viterbi-Alignment was performed instead of aligning over whole audio files.

System	WER
Baseline	31.9%
Viterbi 1 iteration	20.9%
Viterbi 2 iterations	26.5%
EM 1 iteration	21.5%
EM 2 iterations	32.1%
MAP 20/80	20.7%
MAP 40/60	20.5%
MAP 60/40	21.0%
MAP 80/20	21.6%

Table 4: *First adaptation results on a subset of the final database.*

It can be seen that the EM re-estimation achieves worse results than the Viterbi re-estimation. These results however are not comparable since our EM training fails for a considerable amount of the training data (approximately 31%). This may be due to the implementation being optimized under the assumption of accurate transcripts and although a pruning technique is applied the EM training exceeds the memory limit for long utterances. Tuning the pruning parameter of the EM algorithm might alleviate this problem.

After two iterations of Viterbi re-estimation the systems performance degrades since the adaptation over-fits to the adaptation data.

5.3. Results on the Iteratively Growing Database

Viterbi estimation and Viterbi MAP estimation were tested in multiple configurations trained on different parts of the database. Results are shown in Table 5 and Figure 1.

It can be seen that Viterbi MAP adaptation outperforms the Viterbi ML re-estimation for all sizes of the database but the difference in performance decreases the larger the amount of training data is. Figure 2 shows the corresponding results of the tests using the adapted language model. Here the difference in

database parts	Viterbi WER	MAP 20/80	MAP 40/60	MAP 60/40	MAP 80/20
1	26.1%	25.1%	23.4%	**22.9**%	24.0%
1+2	23.4%	23.4%	22.6%	**22.0**%	22.8%
1-3	21.9%	21.6%	**21.2**%	21.5%	22.0%
1-4	20.8%	**20.3**%	20.6%	20.7%	21.6%
1-5	20.4%	**20.1**%	20.3%	20.6%	21.3%
1-6	20.1%	**19.8**%	20.0%	20.5%	21.3%
only 6	50.9%	33.7%	31.4%	**30.7**%	31.0%

Table 5: *WERs of adapted systems for different numbers of parts of the final database. The best performance for each size of the database is bold.*

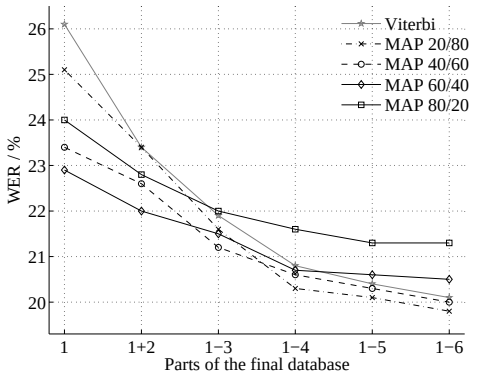


Figure 1: *Word Error rates for different adaptation methods on the test set, plotted over increasing amounts of available adaptation data* .

performance for larger amounts of training data is significantly higher and the performance of the Viterbi ML re-estimation seems to stagnate. Using the adapted language model with the unadapted acoustic model, the resulting WER is 21.5%.

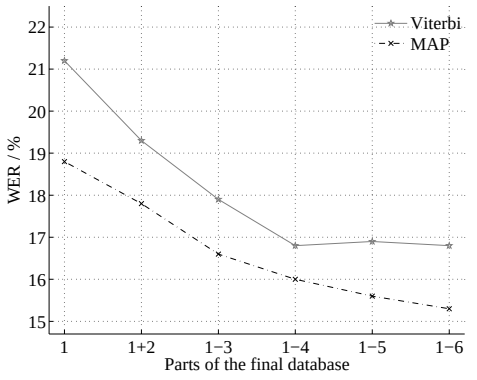


Figure 2: *Word Error rates for different adaptation methods on the test set, plotted over increasing amounts of available adaptation data with adapted language model.*

We took part in an internal EU-BRIDGE evaluation campaign on the Weather Bulletin Task using the presented Viterbi MAP re-estimation method. The initial system training mentioned in Section 5 was redone with the adaptation data also being used during the basic system training. Instead of MFCC features we used *deep bottle neck features* (DBNFs) [17] which have been shown to significantly outperform MFCC features. We also performed fMLLR and MLLR adaptation in a second decoding pass. This resulted in a single 2nd pass system with a WER of 12.4%. Adapting this system, which already saw the Weather Bulletin data during traing, still resulted in a reduced WER of 12.0% for the Viterbi ML re-estimation and 11.9% for the MAP re-estimation.

5.4. Comparison to Unsupervised Training

Table 6 compares the best results from using the captions as training transcriptions with training in an unsupervised manner. One can see that the unsupervised training performs significantly worse. This suggests that the quality of the training references, while not verbatim grade, is still good enough and that they are much more informative than recognition hypotheses. However, when time is not an issue repetitive unsupervised adaptation may yield similar results. The Viterbi ML re-estimation using the original start and end times mentioned in Section 3 was redone using the final database, improving the performance from 23.2% to 21.8%.

system	WER
Unadapted system	31,9%
Viterbi on original start and end times	21.8%
Viterbi on modified start and end times	20.1%
MAP 20/80 on modified start and end times	19.8%
Unsupervised	28.4%
Unsupervised weighted	27.9%

Table 6: *WER of the best adapted system to baseline experiments. Unsupervised adaptations are Viterbi ML re-estimations on the hypotheses from the decoding with the baseline system. In weighted unsupervised training the confidence of a word is used as a weight for the training patterns during the accumulation of the sufficient statistics during training.*

6. Conclusion

We investigated methods for using captions as loose transcripts for adapting acoustic models for automatic speech recognition to weather forecast audio data. Considerable gains can be made by determining the correct start and end times of the captions. This is necessary since the original time segments of the captions only match imprecisely to the corresponding parts in the audio. It turned out that similar to supervised adaptation methods Viterbi ML estimation is outperformed by MAP estimation but for increasing amounts of adaptation material results converge. By using an adapted language model the effect of convergence is decreased.

We showed that the proposed method leads to a WER that is 8.1% abs. lower than when using unsupervised adaptation methods, letting the WER drop from 27.9% to 19.8%. Refining start and end times for incomplete transcriptions by a simple heuristic that searches for matching trigrams of words in the alignment of hypotheses from the decoded audio files to the transcriptions improves the WER by 1.7% abs.

Using the proposed method in combination with language model adaptation and deep BNF features led to a WER of 11.9% in the EU-BRIDGE evaluation campaign on the Weather Bulletin task.

At a level of 5% WER divergence of the available transcripts from verbatim references supervised training is still much more effective than replacing the reference with automatically generated transcripts. A major drawback of the proposed method is the need to decode all of the adaptation material. Depending on the task this might not be feasible due to the time intensity of the approach.

If the divergence is higher, the investigation of the appropriate adaption method would have to be redone and data selection methods might become necessary.

7. Acknowledgements

The work leading to these results has received funding from the European Union under grant agreement no. 287658. '*Research Group 3-01*' received financial support by the '*Concept for the Future*' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative.

8. References

[1] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the iwslt 2012 evaluation campaign," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2012*, Hong Kong, December 6-7 2012.

[2] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, Jan. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S088523080190186X

[3] ——, "Unsupervised acoustic model training," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2002, pp. I–877–I–880.

[4] C. Gollan, S. Hahn, R. Schlüter, and H. Ney, "An improved method for unsupervised training of LVCSR systems," *Interspeech, Antwerp, Belgium*, pp. 2101–2104, 2007. [Online]. Available: http://www-i6.informatik.rwth-aachen.de/publications/download/366/An%20Improved%20Method%20for%20Unsupervised%20Training%20of%20LVCSR%20Systems.pdf

[5] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer using TV broadcasts," in *Proc. of ICSLP*, vol. 98, 1998, pp. 2207–2210. [Online]. Available: http://reference.kfupm.edu.sa/content/u/n/unsupervised_training_of_a_speech_recogn_110317.pdf

[6] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 301–305. [Online]. Available: http://reference.kfupm.edu.sa/content/u/t/utilizing_untranscribed_training_data_to_14022.pdf

[7] H. Li, T. Zhang, and L. Ma, "Confirmation based self-learning algorithm in LVCSR's semi-supervised incremental learning," *Procedia Engineering*, vol. 29, no. 0, pp. 754–759, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187770581200046X

[8] C. Gollan and M. Bacchiani, "Confidence scores for acoustic model adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, Apr. 2008, pp. 4289–4292.

[9] T. Fraga-Silva, J.-L. Gauvain, and L. Lamel, "Lattice-based unsupervised acoustic model training," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 4656–4659.

[10] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, Apr. 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230898900432

[11] S. Stüker, "Acoustic modelling for under-resourced languages," Ph.D. dissertation, PhD thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, 2009. 125, 2009.

[12] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[13] H. Soltau, F. Metze, C. Fuegen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *ASRU*, 2001.

[14] M. Gales, "Semi-tied covariance matrices for hidden markov models," Cambridge University, Engineering Department, Tech. Rep., February 1998.

[15] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *ICASSP*, Munich, Germany, April 1997.

[16] A. Stolcke, "Srilm—an extensible language modeling toolkit," in *ICSLP*, 2002.

[17] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stackedauto-encoders," in *Acoustics, Speech and Signal Processing (ICASSP),2013 IEEE International Conference on*. IEEE, 2013.

A Framework for Integrating Heterogeneous Sporadic Knowledge Sources into Automatic Speech Recognition

Stefan Ziegler, Guillaume Gravier

CNRS-IRISA, Campus de Beaulieu, 35042 Rennes, France

firstname.lastname@irisa.fr

Abstract

Heterogeneous knowledge sources that model speech only at certain time frames are difficult to incorporate into speech recognition, given standard multimodal fusion techniques. In this work, we present a new framework for the integration of this sporadic knowledge into standard HMM-based ASR. In a first step, each knowledge source is mapped onto a logarithmic score by using a sigmoid transfer function. Theses scores are then combined with the standard acoustic models by weighted linear combination. Speech recognition experiments with broad phonetic knowledge sources on a broadcast news transcription task show improved recognition results, given knowledge that provides complementary information for the ASR system.

Index Terms: multimodal fusion, landmark-driven ASR, event-based speech recognition

1. Introduction

Multimedia data in the form of broadcasts, podcasts as well as audio-visual content present difficult challenges for state-of-the art hidden Markov model (HMM) based automatic speech recognition (ASR), since ASR systems are still sensitive towards unseen speaking styles and changes in acoustic conditions. To improve acoustic modeling of HMM-based ASR, many studies advocate the incorporation of complementary knowledge sources into standard ASR to achieve improved recognition accuracy or robustness. Examples of such complementary knowledge sources are phonetic models, that aim at exploiting different features and modeling techniques motivated by phonological studies, to build reliable and sometimes highly specialized detectors for phonetic classes [1, 2, 3, 4]. Another example is audio-visual ASR, where, if available, the visual modality is added to the existing acoustic information, to benefit from the fact that acoustically similar speech classes might correspond to very different visual counterparts (visemes), that are reliable to detect [5]. While it has often been argued that it is desirable for each knowledge source to rely on individual features and modeling techniques, the common architecture of state-of-the-art ASR has become a bottleneck for seamlessly integrating heterogeneous knowledge into speech recognition. Consequently, external knowledge sources often rely on rather homogeneous standard modeling techniques, like frame-based Gaussian mixture models, that are integrated with conventional feature or decision fusion techniques inside the given architecture of HMM-based ASR.

In this paper, we present a new framework for integrating heterogeneous sporadic knowledge sources into HMM-based ASR, with the term sporadic referring to the fact that each knowledge is only defined at certain time frames, often referred to as *events* (e.g., [1, 6]) or *landmarks* (e.g., [7, 8]). Indeed, many acoustic or visual cues for phonetic events or visemes are naturally modeled as a sequence of discrete events, rather than continous values, which makes their integration into ASR very difficult, given common multimodal fusion techniques. In our framework, integration of theses knowledge sources into standard HMM-based ASR is performed in two steps: First, we map each knowledge source onto a logarithmic score, using a sigmoid transfer function. This allows the integration of knowledge sources of different scaling, that appear asynchronously and do model arbitrary phonemic classes. In a second step, the obtained scores are combined with the acoustic scores of standard HMM-based ASR using weighted linear combination. These modified acoustic scores are integrated into the Viterbi decoding of the first pass of a large vocabulary ASR system.

In audio-visual ASR, continuous visual knowledge is often integrated into ASR via feature-fusion, i.e., concatenating audio and visual features to train refined acoustic models [9]. This approach is also used for the integration of a burst onset landmark detector in [2]. Decision fusion at the frame level using GMMs and HMMs by weighted linear combination of log-likelihood scores is used for integration of phonetic information in [10] and for visual information in [11]. Phonetic knowledge is also integrated into ASR during the rescoring step of multi-pass ASR [3, 7]. Landmark-based phonetic models have been used inside alternative probabilistic ASR frameworks [12] and in [1] statistical-post processing of sporadic phonetic landmarks resulted in improved detection accuracy.

In the following section we will present our framework in detail, before presenting speech recognition experiments using broad phonetic knowledge sources. The paper will conclude with an outlook on future work.

2. Integration of sporadic knowledge into ASR

Given a speech utterance with t frames, we consider a sporadic knowledge source k to be a function $x_k(t)$, with $x_k(t)$ being defined only for n_k frames $\mathcal{T}_{x_k} = \{t_1, \ldots, t_{n_k}\}$. Each source is the result of an external system specialized in detecting a given set of phonemes $\mathcal{S}_k$, which is a subset of the complete set of phonemes (including non-speech symbols) $\mathcal{P}$, with $\mathcal{S}_k \subset \mathcal{P}$. To integrate this knowledge into triphone-based ASR systems, the phonemes in $\mathcal{S}_k$ have to be mapped to the corresponding states $\mathcal{I}_k$, which is equally a subset of the complete search space $\mathcal{I}$ (see Figure 2). While the range of $x_k(t)$ is arbitrary for each source k, for example one source could provide a probability from 0 to 1, while another source might correspond to a score in the range from $-\infty$ to $+\infty$ or $-\infty$ to 0, we assume a clear correlation between $x_k(t)$ and $\mathcal{S}_k$. Assuming positive correlation, low values for $x_k(t)$ are supposed to signal poor confidence in

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

the presence of $\mathcal{S}_k$ at t, while high values have a very low error rate, with a more or less sharp transition in-between.

To illustrate such external knowledge sources, we use the example of integrating phonetic landmark detectors into HMM-based ASR. Landmark detection usually consists of two steps (see for example [13]). First, the system detects potential locations for speech events (landmarks), before acoustic cues in vicinity of these landmarks are evaluated to estimate the probability of one or several phonetic classes for each landmark. For example, vowels can be detected by local maxima in the first formant frequency and evaluation of additional features around this landmark can specify the type of vowel. An additional detector might provide landmarks signaling the presence of plosives, by detecting abrupt changes in the signal and studying several cues, like voice onset time or energy of the burst around this point (see for example [14]). It is obvious, while the detection of vowels and plosives can be highly specialized for each phonetic class, both classes are only defined at very specific locations $\mathcal{T}_{x_k}$. Furthermore the landmarks for vowels and plosives will be attached with a confidence estimate $x_k(t)$ that cannot be compared with each other, since each class uses different classification algorithms and features.

With $x(t)$ not being defined for most t, sporadic knowledge can avoid to model parts of speech with high uncertainty about the acoustic content, which is a major advantage compared to HMM-based acoustic modeling. While heterogeneity, i.e., the fact that the ranges of each $x_k(t)$ are very different from each other, could be overcome by normalization, the sporadic nature of knowledge sources makes common fusion at the feature or decision level not feasible any more, since k knowledge sources cannot be mapped onto a k-dimensional vector at each frame t (see Figure 3).

In the following, we present a general framework for the integration of k knowledge sources into the Viterbi decoding of a HMM-based ASR system. Given k sources $x_k(t)$, two steps are necessary from raw knowledge to knowledge-driven ASR. First, we map each source x_k onto a log-likelihood score $\log s_k$, given a sigmoid transfer function, which parameters are estimated using cross-entropy as the objective function. In the second step, these knowledge sources are integrated into the ASR system using a weighted linear combination of the obtained scores $\log s_k$ and the acoustic scores of the ASR system.

2.1. Weighted linear combination of k knowledge sources

Given k knowledge sources, our goal is to modify the acoustic score $s(i,t)$ for state $i \in \mathcal{I}$ at frame t according to weighted linear combination of the log-likelihoods of k knowledge sources $\log s_k(i,t)$ and the unweighted log-likelihood of the acoustic model $\log s_{asr}(i,t)$, given the weights w_k:

$$\log s(i,t) = \log s_{asr}(i,t) + \sum_k w_k \log s_k(i,t) \qquad (1)$$

With $\log s_k(i,t) \geq 0$ and $w_k \geq 0$, each source k enhances states $i \in \mathcal{I}_k$ that are associated with the phonemes in set $\mathcal{S}_k$ (see Figure 2). Evidently, $\log s_k(i,t) = 0$ for all states $i \notin \mathcal{I}_k$ and for all frames t for which the source k is not defined with $t \notin \mathcal{T}_{x_k}$. All states $i \in \mathcal{I}_k$ share the same likelihood-score $\log s_k(i,t)$, to which we will refer to as $\log s_k(t)$.

The next section describes how to map $x_k(t)$ onto $\log s_k(t)$ for each source, before we discuss determining w_k.

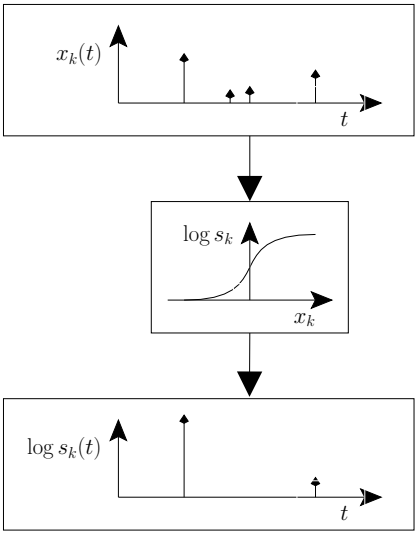


Figure 1: Mapping a sporadic knowledge source $x_k(t)$ onto $\log s_k(t)$.

2.2. Mapping of detection functions onto knowledge scores

Intuitively, $\log s_k(t)$ should maximize the scores added to the correct path, i.e., the scores added to frames t where the correct phoneme actually is a member of $\mathcal{S}_k$, but minimize the error it will introduce into the system by enhancing the wrong path. Therefore, our mapping function should result in $\log s_k(t) = 0$ for low values of $x_k(t)$, but grow according to the confidence that higher values of x_k will correctly indicate $\mathcal{S}_k$. This desired behavior can be obtained by a sigmoid function with:

$$\log s_k\,(t) = \frac{\gamma_k}{1 + \exp\left(-\alpha_k \cdot x_k\,(t) + \beta_k\right)},\ \forall t \in \mathcal{T}_{x_k} \qquad (2)$$

α_k determines the steepness of the slope of the sigmoid, β_k shifts the sigmoid to its optimal working point and γ_k is a scaling factor. For example, if a knowledge source k provides a very reliable knowledge above a certain score β_k, γ_k will be a high value reflecting the confidence in the correctness of $\log s_k(t)$ and a high α_k changes the transfer function from a smooth transition to a step-function-like behavior. Equation 2 maps noisy, unreliable values onto values very close to zero and rounding those values to a limited precision results in $\log s_k(t) = 0$. Since $\log s_k(t) = 0$ for all $t \notin \mathcal{T}_{x_k}$, $\log s_k(t)$ is effectively a sparse vector and we refer to its non-zero frames as $\mathcal{T}_{s_k}$.

To find the optimal α_k, β_k and γ_k, we maximize the cross-entropy $c_{ce}(t)$ between $\log s_k(t)$ and the correct solution $y_k(t)$ at each frame:

$$c_{ce}\,(t) = y_k(t)\frac{\log p_k(t)}{N_{k,1}} + (1 - y_k(t))\,\frac{\log\left(1 - p_k(t)\right)}{N_{k,0}} \qquad (3)$$

$y_k(t)$ is a binary vector with $y_k(t) = 1$ if $\mathcal{S}_k$ is correct at frame t and $y_k(t) = 0$ if not. $y_k(t)$ is derived from the forced alignment of the correct utterance using our baseline ASR system. $p_k(t)$ reflects the probability that knowledge source k is present at frame t. Since some knowledge sources might have a skewed distribution, we normalize $p_k(t)$ by the number of frames $N_{k,1}$ that are in $\mathcal{T}_{x_k}$ for which $y_k(t) = 1$ and respectively $N_{k,0}$ for which $y_k(t) = 0$.

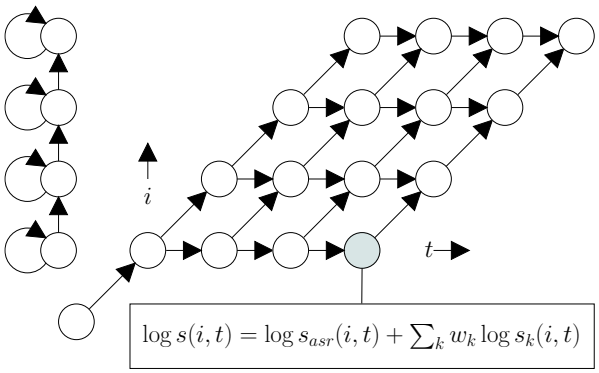


Figure 2: Integration of knowledge into the speech decoding. Arrows correspond to the transition probabilities, while the nodes represent the acoustic scores $\log s(i,t)$. The modified computation of $\log s(i,t)$ is displayed for one node highlighted in grey.

Given the log-likelihood scores of two complementary classes s_k and $\overline{s}_k$, we use the softmax function to estimate $p_k(t)$ according to:

$$p_k(t) = \frac{\exp(\log s_k(t))}{\exp(\log s_k(t)) + \exp(\log \overline{s}_k(t))} \quad (4)$$

As a consequence of the facts that all knowledge sources might model only a subset of $\mathcal{P}$ and sporadic knowledge results in asynchronous landmarks, there is no score $\log \overline{s}_k(t)$ estimating the *absence* of knowledge source k at frame t. Consequently, this *anti-score* $\log \overline{s}_k(t)$ always equals 0:

$$\log \overline{s}_k(t) = 0, \ \forall t \quad (5)$$

The final optimization problem consists in finding the parameters α_k, β_k and γ_k that maximize $c_{ce}(t)$ for all frames of the training data:

$$F_{ce,k}(\alpha_k, \beta_k, \gamma_k; x_k, y_k) = \sum_{t \in \mathcal{T}_{x_k}} c_{ce}(t) \quad (6)$$

2.3. Estimation of the combination weights

While the optimized knowledge sources $\log s_k(t)$ might achieve low error rates according to Equation 6, it has yet to be determined if this source represents *complementary* knowledge to the acoustic models of the ASR system. Therefore, we use discriminative training to determine the weight w_k for each source k, that adjusts the contribution of source k to the overall acoustic score according to Equation 1.

Estimating the weights w_k of a linear combination of log-likelihoods is a well studied problem and several discrimination criteria have been proposed in the literature [15, 11, 10]. In this paper we use the frame-based maximum mutual information (MMI) between correct alignment and n competing hypothesis according to:

$$c_{mmi}(t) = \log s(u(t),t)) - \log \sum_n \exp(\log s(\hat{u}_n(t))) \quad (7)$$

$u(t)$ is the state sequence obtained by force aligning the correct solution of an utterance, while $\hat{u}_n(t)$ corresponds to the alignment of the n-th hypothesis contained in the n-best output of the ASR system. By maximizing the MMI, the correct hypothesis will become more likely, while at the same time the competing hypothesis that do not correspond to the correct path at frame t will become less likely. In this work, we use only the best hypothesis as a competing alternative to the correct path, so that $n = 1$, which turns the MMI criterion into corrective training (see [15]). The optimization problem consists then in finding the weights w_k that maximize $c_{mmi}(t)$ over all frames in $\mathcal{T} = \bigcap_k \mathcal{T}_{s_k}$:

$$F_{mmi}(w; u, \hat{u}, \log s) = \sum_{t \in \mathcal{T}} c_{mmi}(t) \quad (8)$$

3. Experiments

The corpus used in the experiments corresponds to radio broadcast news in the French language from the ESTER2 campaign [16]. The ESTER2 dataset contains broadcast shows with speech in studio environments (RFI), but also difficult tasks like debates (Inter) or speech with strong accents (radio TVME and radio Africa 1). Since we need the correct hypothesis to generate the correct state sequences $u(t)$ and the aligned n-best recognition hypothesis $\hat{u}_n(t)$, we discard every sentence containing out-of-vocabulary words during training and testing. During testing, this allows us to assure that finding the correct path by modifying the acoustic scores during the decoding is not prevented by missing vocabulary. Additionally, we discard all telephone speech from the dataset. The estimation of the parameters α_k, β_k, γ_k and w_k are conducted on the ESTER2 development set, using only broadcasts shorter than 20 minutes, while final speech recognition experiments are conducted on the full ESTER2 test set. The speech recognizer used in this paper is a two-pass system, trained on the ESTER1 and ESTER2 training data. The first pass uses word-internal triphones with 32 Gaussians per state and a trigram language model. The second pass relies on 4-grams and cross-word triphone models. In this paper, we integrate knowledge only in the first pass of our ASR system to generate improved word graphs for rescoring.

3.1. Phonetic knowledge sources and baseline ASR system

In the experiments, we use broad phonetic classes (BPCs) as knowledge sources, obtained from the Gaussian mixture models of a Mel-frequency cepstral coefficients based monophone GMM classifier. We derive 6 detection functions $x_k(t)$ for the BPCs vowels, nasals, approximants, fricatives, plosives and a non-speech class. Each BPC at frame t is first scored with the maximum score among all phonemes of this BPC, before we perform normalization at each frame t by taking the logarithmic sum of exponentials for each source k to obtain 6 continuous detection functions. After smoothing we convert these 6 functions into $k = 6$ sporadic knowledge sources $x_k(t)$ by simple picking the local maxima for each detection function (see Figure 3). Since the monophone models were trained on the same training data like our acoustic models, it is unlikely that they actually will provide complementary information to the ASR system. To experiment with more informative knowledge sources, we additionally create oracle knowledge by adding a bias to the correct BPC at each frame t before performing normalization. We refer to this knowledge sources as BPC-oracle-$bias$, with $bias$ being the scalar added to the correct BPC. While this knowledge does not represent homogeneous knowledge in the sense that it incorporates different modeling and training frameworks, we discuss the influence of multiplicative and additive scaling of

knowledge	WER [dev]	WER [test]
baseline	28.0	31.8
BPC-0	28.0	31.8
BPC-oracle-2	27.7	31.6
BPC-oracle-3	27.4	31.3
BPC-oracle-4	26.8	31.0

Table 1: Word error rates of 4 different broad phonetic knowledge sources and the baseline ASR system on the ESTER2 development and test set.

each $x_k(t)$ in section 3.5.

3.2. Optimization

Given k knowledge sources $x_k(t)$, we have to optimize two objective functions to obtain the parameters α_k, β_k and γ_k for each knowledge source individually and the weights w_k jointly. We use L-BFGS-B minimization implemented in pythons scipy library for both objective functions, with the constraints $\alpha_k > 0$ and $\gamma_k \geq 0$ for Equation 6 and $w_k \geq 0$ for Equation 8. The gradients of the objective functions are in both cases calculated using the symbolic differentiation implemented in the Theano package [17].

For both objective functions, we could achieve fast convergence by carefully choosing initial values for both optimization problems. The scaling factor α_k should be proportional to the variance of $x_k(t)$, while the median of $x_k(t)$ is a good starting point for β_k. For Equation 8, we started with the same value w_k for all knowledge sources k, by choosing the uniform weight which maximized Equation 8. This lead to Equation 6 needing about 20 iterations to converge, while Equation 8 converged already after very few iterations. Though we maximized the weights w_k globally, instead of using gradient descent, we did not observe problems concerning convergence or overfitting.

3.3. Speech Recognition Experiments

After optimizing the mapping from $x_k(t)$ to $\log s_k(t)$ for all sources k and estimating the weights w_k on the development set, speech recognition experiments were performed for *BPC-0*, *BPC-oracle-2*, *BPC-oracle-3* and *BPC-oracle-4*. Table 1 shows the word-error-rates (WER) on the ESTER2 development and test-set along with the WER of the baseline. As expected, *BPC-0* did not provide any new information for the ASR system and obtained $w_k = 0$ for all BPCs except for the non-speech class. Consequently this led to no improvement in WER. For the oracle BPCs, the WER decreases with increasing the bias of the knowledge source. For all cases, the improvement on the development set is higher than on the test set, as often observed in discriminative training.

3.4. Evaluation of knowledge sources

Table 2 displays two criteria evaluating the quality of $x_k(t)$ and $\log s_k(t)$ for the BPCs of three different experiments. AUC (area under the curve) is a performance measurement derived from the ROC curve (receiver operator characteristic) and equal to the probability that a classifier will rank a randomly selected true BPC higher than a randomly selected false BPC. We use the AUC to give an indication of the quality of the raw knowledge source $x_k(t)$. Additionally, for every knowledge source k, we calculate a misclassification cost (MI), related to the mutual information criterion MMI in Equation 7, by calculating the average score added at each frame $\mathcal{T}_{s_k}$, with weighting every correct frame by 1 and every incorrect frame by -1. This results in a negative value if a knowledge source introduces more errors into the decoding than it enhances the correct path.

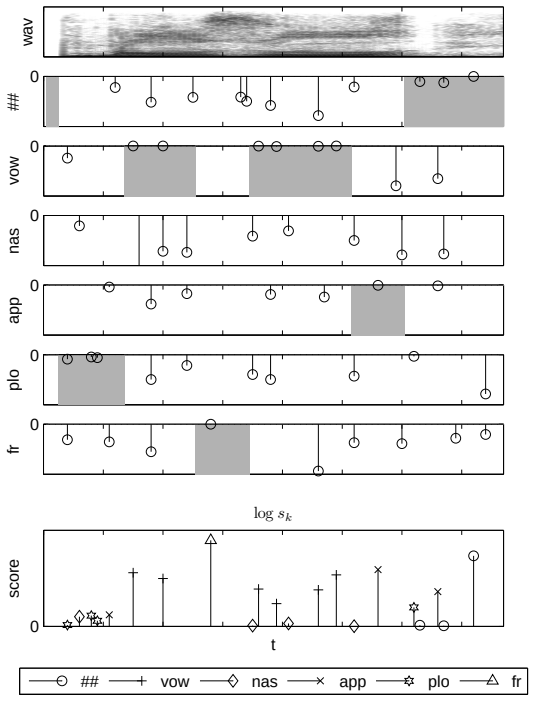


Figure 3: Spectrogram of the French word *Bonjour*, uttered at the beginning of a broadcast show, followed by six sporadic broad phonetic knowledge sources $x_k(t)$ (*BPC-oracle2*) including non-speech (##) and the obtained log likelihoods $\log s_k(t)$ at the bottom. All $x_k(t)$ are normalized, so that 0 represents the maximum value. The correct sequence of BPCs is marked in grey.

$$MI(k, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} (2y_k(t) - 1) \log s_k(t) \qquad (9)$$

$|\mathcal{T}|$ corresponds to the cardinality of the frames $\mathcal{T}$ used to calculate $MI(k, \mathcal{T})$. Both measures are shown on all available frames $\mathcal{T}_{x_k}$ for AUC and $\mathcal{T}_{s_k}$ for MI. Additionally, they are calculated only on those frames $\mathcal{T}_k^*$ where the correct BPC of the true alignment $u(t)$ differs from the BPC in $\hat{u}_n(t)$.

Since the acoustic score of the standard ASR system is not modified (see Equation 1), we expect an improvement of the WER only if a knowledge source is able to correctly enhance most of the frames that are not already correctly aligned in the best recognition hypothesis. Indeed, it can be seen that *BPC-0*, while performing relatively well on all frames $\mathcal{T}$, has a below random AUC, with $AUC < 0.5$, for all BPCs except silence for $\mathcal{T}^*$. For those BPCs MI is negative, which means these knowledge sources make it less likely for the decoder to find the correct path at frames $\mathcal{T}^*$. Consequently, discriminative training resulted in $w_k = 0$ for all BPCs except silence, to prevent the ASR system from degrading. In general, evaluating the errors of a knowledge source without taking the output of the speech recognizer into account might be misleading. Only

BPCs		$\mathcal{T}$	##	vow	nas	plo	fri	app
BPC 0	AUC	$\mathcal{T}_k$	0.84	0.90	0.95	0.93	0.96	0.83
		$\mathcal{T}_k^*$	0.41	0.43	0.37	0.35	0.34	0.46
	MI	$\mathcal{T}_k$	0.9	0.6	2.1	1.7	2.5	0.8
		$\mathcal{T}_k^*$	0	-0.1	-0.5	-0.5	-0.8	-0.1
BPC oracle 2	AUC	$\mathcal{T}_k$	0.94	0.96	0.98	0.98	0.99	0.93
		$\mathcal{T}_k^*$	0.67	0.63	0.59	0.61	0.53	0.70
	MI	$\mathcal{T}_k$	1.9	1.0	3.3	3.0	3.1	1.9
		$\mathcal{T}_k^*$	0.7	0.4	0.5	0.6	-0.2	0.6
BPC oracle 4	AUC	$\mathcal{T}_k$	0.98	0.99	0.99	0.99	0.99	0.98
		$\mathcal{T}_k^*$	0.87	0.81	0.80	0.83	0.73	0.88
	MI	$\mathcal{T}_k$	3.0	1.7	4.6	4.2	3.6	3.2
		$\mathcal{T}_k^*$	1.9	1.3	2.1	2.1	0.8	2.0

Table 2: AUC for $x_k(t)$ and MI for $\log s_k(t)$ given different knowledge sources and their broad phonetic classes silence and non-speech (##), vowels, nasals, plosives, fricatives and approximants. $\mathcal{T}_k$ corresponds either to $\mathcal{T}_{x_k}$ for AUC or $\mathcal{T}_{s_k}$ for MI.

when knowledge sources $x_k(t)$ achieve above random AUC on $\mathcal{T}^*$, MI tends to turn positive and the source contributes to improving the WER, as it is the case for *BPC-oracle-2* and *BPC-oracle-4*.

3.5. Heterogeneous knowledge

The previous broad phonetic knowledge sources were obtained using homogeneous monophone GMM classifier and thus did not represent a collection of heterogeneous knowledge sources. Assuming heterogeneous knowledge will change $x_k(t)$ into $x'_k(t)$ by multiplicative and additive scaling with $x'_k(t) = a_k x_k(t) + b_k$, it is evident that given our proposed sigmoid transfer function, this scaling can be reversed by estimating the corresponding α_k and β_k. To avoid the problem of finding an individual initialization for α_k and β_k to optimize objective function 6 for each knowledge source, we recommend to perform a simple normalization, for example mean and variance normalization for each knowledge source $x_k(t)$. All of our experiments showed, that given proper initialization for α_k and β_k, $\log s_k(t)$ and consequently $MI(k, \mathcal{T})$ was similar for different multiplicative and additive scaling factors.

One advantage of our presented framework is the fact that it is able to deal with selected knowledge sources, that may not cover the complete set of phonemes $\mathcal{P}$. This allows to design individual detectors for each phonemic group $\mathcal{S}_k$, without forcing to model the whole set $\mathcal{P}$. Table 3 shows the same speech recognition experiments as in section 3.3, but with the reduced set of BPCs vowels, nasals and plosives. It can be seen that the WER increases compared to using the complete range of BPCs and the overall impact of the provided knowledge sources is reduced. This is expected, since the broader the external knowledge sources become, the less impact they will have onto the speech decoding, even if a knowledge source inserts only few errors into the decoding.

4. Future Work

Our presented framework showed promising results given different kinds of broad phonetic knowledge sources. Before concluding the paper we want to point out several directions for future research.

Knowledge sources: Our experiments showed, while the integration of rather broad speech landmarks into HMM-based ASR improves the recognition, these landmarks need to be accurate to be effective. Obviously, efforts have to be made to research on existing and new knowledge sources that provide sufficiently accurate landmarks. Furthermore, it is desirable to experiment with additional feature systems like distinctive features, or visual features like visemes.

knowledge	WER [dev]	WER [test]
baseline	28.0	31.8
BPC-oracle-2 (vow-nas-pl)	27.6	31.8
BPC-oracle-3 (vow-nas-pl)	27.5	31.7
BPC-oracle-4 (vow-nas-pl)	27.3	31.5

Table 3: Word error rates of 3 different broad phonetic knowledge sources, using only the BPCs vowels, nasals and plosives each time.

Objective functions: While the sigmoid transfer function in connection with the cross-entropy criterion in Equation 6, as well as the MMI criterion for discriminative training provided good results, one might consider additional transfer functions and training criteria.

State dependent weights and context dependency: One disadvantage of the presented approach is the fact, that it does not include state or phoneme-dependent weights $w_{i,k}$ for Equation 1. Enhancing states that are not in $\mathcal{I}_k$ for a knowledge source k might help to reduce the error introduced into the decoding, since this might take into account common phonetic confusions, like it is the case for vowels and approximants. Additionally, the speech recognition system could be modified to accommodate for a weight w_{asr} that scales $\log s_{asr}(i, t)$ in Equation 1 to improve the discriminative training criterion.

Given phonetic landmarks, as employed in this paper, the probability of a speech class $\mathcal{S}_k$ at t depends on the context, i.e., its preceding and subsequent landmarks. To address this context dependency, landmarks $x_k(t)$ could be rescored by additional models, that are trained on landmark sequences, like it has been proposed in [1].

Integration into multi-pass ASR: In the current implementation we only implemented knowledge-driven ASR in the first pass of our speech recognizer. To fully benefit from heterogeneous knowledge sources, integration into rescoring steps of multi-pass ASR systems is desirable.

5. Conclusions

The presented framework focused on the integration of heterogeneous and sporadic knowledge sources into HMM-based ASR. It allows the use of individual training and detection algorithms for each knowledge source, that can be developed independently from each other. Furthermore, it accounts for event or landmark based models of speech and does not require the re-training of existing acoustic models. We used a transfer function to map each knowledge source onto a logarithmic score, before the obtained values were combined with the acoustic scores by weighted linear combination.

While the knowledge sources that improved the WER in this paper corresponded to oracle knowledge, we conclude from our experiments that landmarks which achieve an above random detection performance on frames where the ASR-system aligns the wrong path are likely to improve the recognition performance of HMM-based ASR systems.

6. References

[1] A. Jansen and P. Niyogi, "Point process models for event-based speech recognition," *Speech Communication*, vol. 51, no. 12, pp. 1155–1168, 2009.

[2] C.-Y. Lin and H.-C. Wang, "Burst onset landmark detection and its application to speech recognition," *Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1253–1264, 2011.

[3] S.M. Siniscalchi and C.H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," *Speech Communication*, vol. 51, no. 11, pp. 1139–1153, 2009.

[4] S. Ziegler, B. Ludusan, and G. Gravier, "Towards a new speech event detection approach for landmark-based speech recognition," in *2012 IEEE Workshop on Spoken Language Technology*, 2012.

[5] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in Visual and Audio-Visual Speech Processing*, vol. 22, pp. 23, 2004.

[6] A. Juneja, *Speech recognition based on phonetic features and acoustic landmarks*, Ph.D. thesis, University of Maryland, College Park, 2004.

[7] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang, "Landmark-based speech recognition: report of the 2004 Johns Hopkins summer workshop," in *Proc. of ICASSP'05*, 2005, pp. 213–216.

[8] G. Gravier and D. Moraru, "Towards phonetically-driven hidden Markov models: can we incorporate phonetic landmarks in HMM-based ASR?," in *Proc. of NOLISP'07*, 2007, pp. 161–168.

[9] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual lvcsr," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*, 2001, pp. 165–168.

[10] F. Metze, "Articulatory features for 'meeting' speech recognition," in *Proc. of INTERSPEECH-2006*, 2006.

[11] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and mce based hmm stream weight estimation for audio-visual asr," in *Proc. of ICASSP'02*, 2002.

[12] J. R. Glass, "A probabilistic framework for segment-based speech recognition," *Comput. Speech Lang.*, vol. 17, pp. 137–152, 2003.

[13] Amit Juneja and Carol Espy-Wilson, "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 123, pp. 1154–1168, 2008.

[14] Sharlene A Liu, "Landmark detection for distinctive feature-based speech recognition," *J. Acoust. Soc. Am.*, vol. 100, pp. 3417–3430, 1996.

[15] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Communication*, vol. 34, no. 3, pp. 287–310, 2001.

[16] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French broadcasts," in *Proc. of INTERSPEECH-2009*, 2009, pp. 1149–1152.

[17] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010, Oral Presentation.

The first official REPERE evaluation

Olivier Galibert, Juliette Kahn

Laboratoire national de métrologie et d'essais, Trappes, France

firstname.name@lne.fr

Abstract

The REPERE Challenge aims to support research on people recognition in multimodal conditions. Following a 2012 dry-run [1], the first official evaluation of systems has been conducted at the beginning of 2013. To both help system development and assess the technology progress a specific corpus is developed. It current totals at 30 hours of video with multimodal annotations. The systems have to answer the following questions: Who is speaking? Who is present in the video? What names are cited? What names are displayed? The challenge is to combine the various informations coming from the speech and the images.

Index Terms: REPERE, multimodality, evaluation, fusion, person recognition

1. Introduction

Finding people on video is a major issue when various informations come from television and from the Internet. The challenge is to understand how to use the information about people that comes from the speech and the image and combine them so as to determine who is speaking and who is present in the video.

Some evaluation campaigns [2] or [3] worked on people multimodal recognition on English databases.

Started in 2011, the REPERE Challenge aims to support the development of automatic systems for people recognition in a multimodal context. Funded by the French research agency (ANR) and the French defense procurement agency (DGA), this project has started in March 2011 and ends in March 2014.

To assess the systems' progress, the first of two international campaigns has been organized at the beginning of 2013 by the Evaluation and Language resources Distribution Agency (ELDA) and the Laboratoire national de métrologie et d'essais (LNE). The second official campaign is open to external consortia who want to participate in this challenge and will take place at the beginning of 2014.

People who are interested in the REPERE Challenge and decide to participate to the second official campaign will have access to the REPERE Corpus and to the metrics tools.

This paper presents the protocol used to estimate the systems progress and the results of the evaluation. Section 2 describes the different tasks that form the REPERE Challenge. Section 4 presents the data used to assess the systems. Section 3 is dedicated to the metrics description. Section 5 presents an overview of the evaluation results. Section 6, concludes this paper.

2. Questions and tasks

2.1. Main tasks

The first tasks in the REPERE Challenge is the identify every person who is visible and/or is speaking in the video. The goal is to combine the idiosyncratic information that comes from the speech and the video frames to answer those questions. These tasks are conducted in supervised (a-priori models of voice and face allowed) and unsupervised modes (a-priori models of voice and face not allowed).

The secondary tasks are to determine the people who are cited in the video. The people can be cited in speech. For example, a speaker can mention another person or he can name his interlocutor. In addition, the names of the people may be displayed on the video frames as show in Figure 2. Those two tasks are conducted in unsupervised mode.

2.2. Sub-tasks

Answering the four previous questions requires to combine multiple technologies. The following sub-tasks which may be useful are assessed in the REPERE Challenge:

- Speaker diarization
- Speech transcription
- Head detection and segmentation
- Overlaid words text detection and segmentation
- Optical Character Recognition (OCR)

During the 2013 REPERE Evaluation campaign, only the Speaker diarization and Speech transcription tasks had system outputs submitted.

3. Metrics

3.1. EGER

The main evaluation metric is the *Estimated Global Error Rate* (EGER). This metric is based on a comparison between the person names in the references and in the system outputs. EGER is a solution to take in account the fact that the systems have found the correct number of people.

For each annotated frame, i, the list of the names of speaking and/or visible persons is built for the reference on one side and for the hypothesis on the other side. Both lists are compared by associating the names one-on-one, each name being associated at most once.

An association between two identical names is considered correct. An association between persons with two different names is a confusion noted C_i. Each person with no association in the hypothesis is a false alarm FA_i, and in the reference a miss, M_i.

An uniform cost of 1 is associated to every error type. Among all possible association sets the one with the lowest cost is selected. Adding up all these costs gives us the total error count, which is divided by the number of expected names (i.e. sum of the size of the reference lists) to get the error rate.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

For N annotated frames, EGER is defined as :

$$EGER = \frac{\sum_{i=0}^{i=N} C_i + FA_i + M_i}{\sum_{i=0}^{i=N} P_i} \quad (1)$$

where P_i is the number of named people in the i frame.

This metric, with adapted list building methodologies, is used for three tasks:

- Who is speaking or is present in the video frame ?
- Who is speaking ?
- Who is present in the video frame ?

We also created two variants of the metric. One variant takes the persons the annotators (and systems) were not capable of naming into account. The other builds the lists per-show instead of per keyframe, measuring the capability of the systems as input to a full-show search task.

3.2. SER : What names are cited?

The expected answer to the *what names are cited?* question takes the form of a list of temporal segments to which an identity is associated. Obviously, anonymous identities do not exist in that task. We decided to use the *Slot Error Rate* as a metric. The list reference temporal segments to find is built from the audio and the annotated transcriptions through a forced alignment procedure. The hypothesis and reference intervals lists are then compared, and an error enumeration is built:

- I: For every interval of the hypothesis without an intersection with the reference we count an *Insertion* error, with a cost of 1
- D: For every interval of the reference without an intersection with the hypothesis we count an *Deletion* error, with a cost of 1
- T: For an (hypothesis, reference) interval pair in intersection where the identity is different we count a *Type* error, with a cost of 0.5
- F: For an (hypothesis, reference) interval pair in intersection where the frontiers are different by more than 500ms, we count a *Frontier* error, with a cost of 0.5

Note that a pair can end up counting as both a type and a frontier error. The SER is them computed by cumulating the error costs and dividing by the number of intervals in the reference. In other words, noting R the number of intervals in the reference:

$$SER = \frac{I + D + 0.5 \times (T + F)}{R}$$

3.3. DER

The speaker segmentation task requires to extract the speech from the recordings and split it into speaker-attributed segments. Some segments have overlapping speech and must be associated to all pertinent speakers. The naming of the speakers does not need to be related to their real name, abstract labels are plenty. Two conditions are evaluated: one where each show is considered independant, and one called *cross show* where speakers coming back from one show to another should be labelled identically.

The standard metric for the task is the *Diarization Error Rate* (DER). The metric counts the time in error and divides it by the total reference speech time. The time in error is divided in three categories:

- False alarm, where the hypothesis puts a speaker but nobody actually talks
- Miss, where the reference indicates the presence of a speaker but not the hypothesis
- Confusion, where reference and hypothesis disagree on who the speaker is

The speaker labels being abstract, establishing the confusion time requires some effort. It is done through a *mapping*, where speakers in the reference are associated 1:1 with the hypothesis speakers. Some may remain unassociated. Among all possible mappings the one that gives the best (smallest) DER is the one chosen for the evaluation. A 250ms tolerance on the reference speaker segment boundaries is taken into account to reduce the impact of the intrinsic ambiguousness of their setup.

3.4. WER : Speech transcription

For the speech transcription task, the systems have to transcribe every word spoken in a show. Segments where speech from multiple people overlap are ignored in the evaluation. The usual ASR metric, the *Word Error Rate*, is similar to the OCR one: a levenshtein distance between the words of the reference and the hypothesis. A normalisation process is used:

- Punctuation removal and downcasing.
- Substitution of dashes by spaces.
- Separation of the words at the apostrophe (l'autre becomes l' autre) except for a small number of exceptions (aujourd'hui).

Homophones are handled on a case-by-case basism through normalization tables and by putting alternatives directly in the reference in some cases.

4. The REPERE Corpus

4.1. Sources

The January 2013 corpus represented 24 hours of training data, 3 hours of development data and 3 hours of evaluation data and is described in Table 1.

The videos are selected from two French TV channels, BFM TV and LCP, for which ELDA has obtained distribution agreements. The shows are varied:

Top Questions is extracts from parliamentary "Questions to the government" sessions, featuring essentially prepared speech.

Ca vous regarde, *Pile et Face* and *Entre les lignes* are variants of the debate setup with a mix of prepared and spontaneous but relatively policed speech.

LCP Info and *BFM Story* are modern format information shows, with a small number of studio presenters, lots of on-scene presenters, interviews with complex and dynamic picture composition.

Culture et vous, previously named *Planète Showbiz*, is a celebrity news show with a voice over, lots of unnamed known people shown and essentially spontaneous speech.

These video were selected to showcase a variety of situation in both the audio and video domains. A first criteria has been to reach a fair share between prepared and spontaneous speech. A second one was to ensure a variety of filming conditions (luminosity, head size, camera angles...). For instance, the sizes of the heads the annotators would spontaneously segment varied from 146 pixels^2 to 96,720 pixels^2 for an image resolution of 720x576. Some example frames are given Figure 1.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

Show	Train	Dev	Test
BFM Story	7:57:49	1:00:50	0:59:48
Culture et Vous	2:09:28	0:15:00	0:15:03
Ça vous regarde	2:00:05	0:15:39	0:15:01
Entre les lignes	1:59:43	0:15:00	0:15:02
Pile et Face	2:01:26	0:15:04	0:15:01
LCP Info	4:07:09	0:30:08	0:29:56
Top Questions	3:57:41	0:30:02	0:27:01
Total	24:13:23	3:01:46	2:56:55

Table 1: TV shows currently present in the corpus

Figure 1: Some example frames from the video corpus

4.2. Annotations

Two kinds of annotations are produced in the REPERE corpus : audio annotation with rich speech transcription and video annotation with head and embedded text annotation.

4.2.1. Speech annotations

Speech annotation are produced in *trs* format using the Transcriber software [4]. The annotation guidelines are the ones created in the ESTER2 [5] project for rich speech transcription. The following elements are annotated :

- Speaker turn segmentation.
- Speaker naming.
- Rich speech transcription tasks gather segmentation, transcription and discourse annotation (hesitations, disfluences...)
- The annotation of named-entities of type "person" in the speech transcription with a normalized label for each identity.

4.2.2. Visual annotations

In complement to the audio annotation, the video annotation has necessitated the creation of specific annotation guidelines[1]. The VIPER-GT video annotation tool has been selected for its ability to segment objects with complex shapes and to enable specific annotation schemes. The video annotations consist in the six following tasks:

- Head segmentation: all the heads that have an area larger than 1000 pixels[2] are isolated. Heads are delimited by polygons that best fit the outlines. Figure 2 is an example of head segmentation. It is worth noting that it is head segmentation and not face segmentation. Sideways poses are annotated too.
- Head description: each segmented head may have physical attributes (glasses, headdress, moustache, beard, piercing or other). The head orientation is also indicated: face, sideways, back. The orientation choice is based on the visible eyes count. Finally, the fact that some objects hide a part of the segmented head is indicated, specifying the object's type.
- People identification: The name of the people is indicated. Only well-known people and the people named in the video are annotated. Unknown people have are identified with a unique numerical ID.
- Embedded text segmentation and transcription: the transcription of the segmented text is a direct transcript of what appears in the video. All characters are reproduced with preservation of capital letters, word wrap, line break, etc. Targeted texts are segmented with rectangles that fit best the outlines (see figure 2). Also whether a text is part of an identification cartouche is also annotated.
- Named-entities (type "person") annotation in transcripts of embedded texts
- The annotation of appearance and disappearance timestamps: the aim is to identify the segments where the annotated object (head or text) is present.

Figure 2: Segmentation example

4.2.3. Global annotations

Beyond the parallel annotation of audio and visual content, the corpus creation pays special attention to the multimodal annotation consistency. A people names database ensures the coherence of given names in audio and visual annotations. Moreover, unknown people IDs are harmonized when the same person appears both in audio and video annotations.

In addition two per-person annotation are provided for both video and audio: the gender of the person, and its role in the show under a 5-class taxonomy.

4.3. First evaluation corpus

Table 2 summaries the annotations done on the 30 hours of corpus created for that run, and the number of persons that can be found through audio or visual clues.

[1]Guidelines are available for participants on the REPERE website. They will be distributed with the REPERE corpus at the end of the project.

		Train	Dev	Test
Visual	Heads seen	13188	1534	2081
	Words seen	120384	14811	15844
Speech	Segments	12833	1602	1514
	Words	275276	34662	36489
Persons	Seen known	725	146	141
	Speaking known	556	122	126
	To find	811	172	162
	Seen unknown	1907	238	160
	Speaking unknown	1108	163	179
	Names on screen	729	138	160
	Names cited	870	190	161
Clues modalities	Name appears	504	83	83
	Name cited	544	116	101
	Never named	178	39	36
	Not speaking	255	50	36
	Not seen	86	26	21
	Speaking and seen	470	96	105

Table 2: Some number about the REPERE first evaluation corpus

We can see that in the test corpus 51% of the people to find have their name appearing on screen and 62% are introduced in the speech. In practice the OCR is much more reliable than the speech recognition for proper names, making these 51% is primary information source for the global system. Interestingly, 22% of the persons are never named, limiting the reachable level for unsupervised systems.

A number of persons appear only in one modality. In the test 22% are only visible, which is a little lower than in the rest of the corpus, and 13% are only heard.

5. Evaluation results

5.1. Participants

Three consortium participated to the evaluations. SODA is a combination of the LIUM (Computer technology lab of the Université du Maine, France) and the Idiap Research Institute. QCOMPERE is made of the LIMSI (Computer technology lab for mechanics and engineering sciences), the INRIA research centre Grenoble (Rhône-Alpes, France), the LIG (Computer technology lab of Grenoble, France), YACAST, Vocapia Research, the GREYC (Research group for computer science, image, automatic and instrumentation of Caen, France) and the Karlsruhe Institute of Technology. Finally the PERCOL consortium is composed of the Laboratoire d'Informatique Fondamentale de Marseille (LIF), the Université d'Avignon et des Pays de Vaucluse (UAPV), the Laboratoire d'Informatique Fondamentale de Lille (LIFL) and France Télécom.

5.2. Main supervised task results

The main supervised task is to find who is present and who talks in the videos by any (automatic) means necessary. The anonymized primary results for each consortium are presented in table 3 using the three EGER variants of section 3.1.

We can see that the results are quite close, with around a third of the identities incorrect. Evaluating the task as finding who is present in a given show degrades the results a little but not by much, with interestingly a different loss for different systems.

Declining per media the results for the speaker identification task are presented in table 4.

Partner	Main EGER	With unnamed	Full-show
A	32.9	43.0	34.7
B	27.9	38.0	32.8
C	29.6	37.5	35.0

Table 3: Main supervised task results

Partner	Main EGER	With unnamed	Full-show
A	22.8	23.1	25.5
B	17.6	18.0	21.7
C	17.7	18.5	21.1

Table 4: Speaker identification task results

Unsuprisingly, the results are much better for the speech side of the multimedia problem. Not only speech technologies are more mature but the task is much simpler, speech overlap being rare compared to the presence of multiple persons in the same image. That particularly shows in the results taking into account the unnamed people: it's much easier to detect whether someone is present in the speech and cluster his interventions than detecting persons in the image and clustering their apparitions.

This is confirmed by the person presence in the picture results presented in table 5.

Partner	Main EGER	With unnamed	Full-show
A	41.5	54.2	42.0
B	36.7	50.0	41.5
C	39.8	48.2	45.9

Table 5: Visible person identification task results

The results are as expected much worse than on the audio side, with the unnamed persons being particularly problematic. Image processing is the achille's heel of these integrated systems.

5.3. Main unsupervised task results

The unsupervised variant of the main task still requires the system to identify the persons speaking and present on the screen, but precludes the use of a-priori trained biometric models. The names are to be found in the signal, either pronounced or written on the screen. The results are presented in table 6.

Partner	Main EGER	With unnamed	Full-show
A	39.5	48.2	36.1
B	37.2	45.2	43.2
C	44.2	49.9	50.8

Table 6: Main unsupervised task results

The loss due to the lack of pre-trained biometric models is around 10% absolute, which isn't bad. Especially since 22% of the persons are never named, putting a hard limit to the minimum possible error rate.

We decline the results per media in tables 7 for the speakers and 8 for the persons present on screen.

The system behaviour is similar than for the supervised task, with a higher loss in the speech case showing that acoustic

Partner	Main EGER	With unnamed	Full-show
A	31.8	32.0	25.5
B	26.3	26.9	36.6
C	40.1	42.8	44.1

Table 7: Unsupervised speaker identification task results

Partner	Main EGER	With unnamed	Full-show
A	46.1	57.3	44.4
B	46.4	55.5	48.3
C	47.8	53.9	56.1

Table 8: Unsupervised visible person identification task results

biometric models are currently more efficient than visual biometric models.

5.4. Monomodal task results

The two monomodal tasks aim at measuring the quality of biometric models by asking of the participant to only use them for the identification and avoiding any fusion process. Hence the name *monomodal*, since only the speech signal modality (without ASR) is used for speaker identification, and only the images (without OCR information) is used for visible person recognition. The results are given in tables 9 for speaker identification and 10 for visible person identification.

Partner	Main EGER	With unnamed	Full-show
A	48.3	48.3	54.0
B	44.2	45.2	43.5
C	37.3	37.2	41.0

Table 9: Monomodal speaker identification task results

Partner	Main EGER	With unnamed	Full-show
B	62.2	62.6	65.9

Table 10: Monomodal visible person identification task results

The speaker identification results go from a 36% to a 49% error rate, which shows a good use of what models were pretrained. The visible person identification is worse as expected.

5.5. Speaker diarization

The speaker diarization task consists in detecting the speech segments in the audio and associating them abstract speaker labels, where the same label is used for multiple interventions of the same speaker. Two conditions were evaluated, one where labels are local to an individual show, and the cross-show one where the same label must be used for a speaker recurring in multiple shows. The results are given table 11.

Partner	DER-ind	DER-cross
A	13.70	33.09
B	13.35	16.05
C	11.10	14.20

Table 11: Speaker diarization task results

We can see that the individual show results are quite good, and at the state of the art for this kind of data. Interestingly, with one exception the cross-show results are very close to the individual-show ones. Since not taking the cross-show condition into account would have given error rates in the 60+% the problem really had to be tackled, and it has been done rather succesfully. These good results have made the cross-show diarization in combination with the OCR of names (not evaluated this year) the backbone of the information fusion efforts of the participants.

5.6. Speech transcription

The speech transcription performance is roughly state-of-the-art, as shown in table 12.

Partner	WER
A	28.03
B	16.43
C	15.18

Table 12: Speech transcription task results

The participants did not consider the speech transcription as a reliable primary information source, given how easy it is for an ASR system to make errors on proper nouns. They seem to plan to work on it more for the next evaluation.

The per-show results, table 13, confirm our expectations on the relative shows difficulties.

	A	B	C
Culture et Vous	54.53	34.56	37.87
Ça vous regarde	36.10	21.75	21.14
Entre les lignes	27.83	17.77	14.92
LCP Info	20.76	11.26	10.10
BFM Story	26.69	15.11	13.03
Pile et Face	27.81	16.27	14.34
Top Question	18.33	10.22	9.26

Table 13: Per-show speech transcription task results

6. Conclusions and Perspectives

The REPERE project focuses on identifying speakers and visible persons in multimodal conditions.

Specific metrics has been implemented. Evaluation tools are made available to interested persons to participate in the next evaluation.

30 hours of data have been created for that evaluation. The annotations are rich and useful for both training systems and evaluating their results. The corpus will double in size for the second evaluation, with the amount put aside for the test still to be decided.

The first evaluation has shown that reasonably good results are possible but a large margin of progress exists, especially on the image side. The influence of the types of programs will be discussed soon.

The sub-tasks will be redefined for the next campaign to better meet the developers needs of modular analysis (specially for video treatment)

7. Acknowledgments

This work was funded by the the ANR/DGA Repere project.

8. References

[1] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P. Joly, "A presentation of the repere challenge," in *CBMI*, P. Lambert, Ed. IEEE, 2012, pp. 1–6.

[2] A. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, 2006, pp. 321–330.

[3] J. Ortega-Garcia, J. Fierrez, F. Alonso-Fernandez, J. Galbally, M. Freire, J. Gonzalez-Rodriguez, C. Garcia-Mateo, J. Alba-Castro, E. Gonzalez-Agulla, E. Otero-Muras *et al.*, "The multiscenario multienvironment biosecure multimodal database (bmdb)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1097–1111, 2010.

[4] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," in *Speech Communication special issue on Speech Annotation and Corpus Tools*, vol. 33, January 2000.

[5] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news," in *European Conference on Speech Communication and Technology*, 2005, pp. 1149–1152.

QCompere @ REPERE 2013

Hervé Bredin[1], *Johann Poignant*[2], *Guillaume Fortier*[3], *Makarand Tapaswi*[4], *Viet Bac Le*[5], *Anindya Roy*[1], *Claude Barras*[1], *Sophie Rosset*[1], *Achintya Sarkar*[1], *Qian Yang*[4], *Hua Gao*[4], *Alexis Mignon*[6], *Jakob Verbeek*[3], *Laurent Besacier*[2], *Georges Quénot*[2], *Hazim Kemal Ekenel*[4], *Rainer Stiefelhagen*[4]

[1]LIMSI-CNRS, Université Paris-Sud, BP 133, 91403 Orsay, France
[2]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS-LIG UMR 5217, F-38041 Grenoble, France
[3]INRIA Rhone-Alpes, 655 Avenue de l'Europe, F-38330 Montbonnot, France
[4]KIT, Karlsruhe Institute of Technology, Karlsruhe, Germany
[5]Vocapia Research, 28 rue Jean Rostand, Parc Orsay Université, 91400 Orsay, France
[6] Université de Caen / GREYC UMR 6072, F-14050 Caen Cedex, France

Abstract

We describe QCompere consortium submissions to the REPERE 2013 evaluation campaign. The REPERE challenge aims at gathering four communities (face recognition, speaker identification, optical character recognition and named entity detection) towards the same goal: multimodal person recognition in TV broadcast. First, four mono-modal components are introduced (one for each foregoing community) constituting the elementary building blocks of our various submissions. Then, depending on the target modality (speaker or face recognition) and on the task (supervised or unsupervised recognition), four different fusion techniques are introduced: they can be summarized as propagation-, classifier-, rule- or graph-based approaches. Finally, their performance is evaluated on REPERE 2013 test set and their advantages and limitations are discussed.
Index Terms: speaker identification, face recognition, named entity detection, video optical character recognition, multimodal fusion

1. Introduction

The REPERE challenge[1] aims at gathering four communities (face recognition, speaker identification, optical character recognition and named entity detection) towards the same goal: multimodal person recognition in TV broadcast. It takes the form of an annual evaluation campaign and debriefing workshop. In this paper we describe the submissions of the QCompere consortium to the 2013 REPERE evaluation campaign [1]

Given TV broadcast videos such as news or talk-shows, the main objective of the REPERE challenge is to answer two questions: *who speaks when?* and *who appears when?*. We distinguish two subtasks: either supervised (when prior identity models are allowed) or unsupervised recognition (when prior models are forbidden and person names must be automatically extracted from the test videos themselves). Speaker and face recognition both rely on a priori models of each person to be recognized: they fall in the supervised recognition category. Our mono-modal (audio or visual) person recognition modules are introduced in Section 2. However, other sources of information are available in TV broadcast and can be used to achieve unsupervised person recognition; such as named entities detected in automatic speech transcription, and block titles usually written on screen to introduce reporters and interviewees. Our efforts in this direction are described in Section 3.

The main contributions of the QCompere consortium lie in the way these modules are combined into a multimodal person identification framework. In Section 4, we propose a classifier-based late fusion approach and another one modeling person recognition as a shortest path problem in a multimodal probability graph. QCompere runs submitted to the 2013 campaign are evaluated and compared in Section 5. Finally, Section 6 concludes the paper.

2. Supervised Person Recognition

In this section, we only describe the mono-modal supervised person recognition approaches (speaker and face).

2.1. Speaker Recognition

Speaker diarization (SD) is the process of partitioning the audio stream into homogeneous clusters without prior knowledge of the speaker voices and serves as a pre-processing step for the speaker identification module. Two SD systems were developed, respectively by LIMSI and KIT.

LIMSI's SD system relies on two steps: agglomerative clustering based on the BIC criterion to provide pure clusters followed by a second clustering stage using cross-likelihood ratio (CLR) as distance between the clusters [2]. Additionally, since the corpus contains several shows for each recorded program, the same identifier has to be associated to a given speaker across all the shows. Following previous experiments on cross-show speaker diarization, a first, local clustering stage is followed by a CLR clustering across all the shows; this hybrid approach was found to provide a good performance while being computationally acceptable for a corpus lasting a few hours [3].

KIT's SD system contains the following components. Audio segmentation first discriminates speech from non-speech segments. It is implemented using a HMM segmenter with 4 GMMs for speech, silence, noise and music. Speaker turn detection [4] is then applied on the segments longer than $5s$. A first-pass BIC clustering groups the segments from the same speaker together. Viterbi re-segmentation refines the segment

This work was partly realized as part of the Quaero Program and the QCompere project, respectively funded by OSEO (French State agency for innovation) and ANR (French national research agency). Corresponding author: bredin@limsi.fr

[1]http://www.defi-repere.fr

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

boundaries. The speaker models are trained on the clustering results. The features are 20-dimensional MFCC plus their first derivatives. Feature warping is applied to compensate channel effects. GMMs with 64 Gaussians are used to model the speakers. A second-pass BIC clustering and Viterbi re-segmentation further refines the segment boundaries and clustering results. Finally, the post processing merges the adjacent segments from the same speakers which are separated by the silence shorter than $0.5s$.

Unsupervised speaker diarization is followed by a cluster-wise speaker identification. We implemented two systems [5]. Our baseline system follows the standard Gaussian Mixture Model-Universal Background Model (GMM-UBM) paradigm, and the GSV-SVM system uses the super-vector made of the concatenation of the UBM-adapted GMM means to train one Support Vector Machine classifier per speaker. For both systems, each cluster is scored against all gender-matching speaker models, and the best scoring model is chosen if its score is higher than the decision threshold. Three data sources were used for training models for 648 speakers in our experiments: the REPERE training set, the ETAPE training and development data[2] and additional French politicians data extracted from French radios.

2.2. Face Recognition

The supervised face recognition process is divided into face detection and tracking stage, and the recognition stage.

Face tracking is performed using particle filtering approach [6], initialized from face detections. The first frame of each shot, and every subsequent fifth frame is scanned and face tracks are initialized from frontal, half-profile and full-profile face detections. Tracking is performed in an online fashion, i.e., using the state of the previous frame to infer the location and head pose of the faces in the current frame.

The face recognition uses a frontal face descriptor. First a detector locates nine landmarks on the face, around the eyes, the nose and the mouth. We use a tree-structured constellation model [7] that computes Histogram of Gradient (HoG) features [8] for detection of these facial landmarks. Once the landmarks are detected, faces are aligned using an affine transformation, and a second HoG descriptor is computed around each of the nine facial landmarks. The descriptor quantizes local image gradients into 10 orientation bins, and computes a gradient orientation histogram for each cell in a 7×7 spatial grid over image region around the landmark. The final descriptor concatenates the local gradient orientation histograms to form a $9 \times 10 \times 7 \times 7 = 4410$ dimensional feature vector per face (9 landmarks $\times$ 10 orientation bins $\times$ a grid of 7×7 spatial bins). For each track, we compute a mean HoG descriptor from all the frontal face detections found along the track. A database is automatically generated using a training set of annotated faces for learning the face recognition models. A Support Vector Machine (SVM) classifier is trained for each person, using one-versus-rest approach.

For the test set, we score the mean descriptor of each track using the learned models. The best scoring model is chosen and the face is tagged with the corresponding name, provided its score is higher than the decision threshold. The initial face recognition stage is followed by an unsupervised face clustering stage that is used to extend the labels to faces not named in the previous step. For each track, the mean HoG descriptor is projected on to a 200 dimensional descriptor using Logistic Discriminant Metric Learning approach (LDML) [9]. The learned face metric is used in a nearest neighbor classifier to assign names to tracks that were unlabeled so far; but only if the ratio of distances to the first and the second neighbor is sufficiently small.

[2] `http://www.afcp-parole.org/etape.html`

3. Person Name Detection

Speaker and face recognition both rely on a priori models of each person to be recognized: they fall in the "supervised recognition" category. However, other sources of information are available in TV broadcast and can be used to achieve unsupervised person recognition.

3.1. Written Name Detection

In order to detect the names written on the screen used to introduce a person, a detection and transcription system is needed. For this task we used LOOV [10] (LIG Overlaid OCR in Video). This system has been previously evaluated on another broadcast news corpus with low-resolution videos. We obtained a character error rate (CER) of 4.6% for any type of text and of 2.6% for names written on the screen to introduce a person.
From the transcriptions, we use a simple technique for detecting the spatial positions of title blocks. This technique compares each transcript with a list of famous names (list extracted from Wikipedia, 175k names). Whenever a transcription corresponds to a famous name, its spatial position is added to a list. The repeating positions in this list provide the spatial positions of title blocks used to introduce a person. However, the detected text boxes do not always contain a name. A simple filtering based on some linguistic rules allows to filter false positives. Transcription errors are corrected using our Wikipedia list when the edit distance is small. The use of LOOV pipelined with our written names detection technique provides an F1-measure of 97.5% (see Table 1). The few remaining errors are due to transcription or filtering errors.

3.2. Spoken Name Detection

The aim of this task is to detect all person names spoken during a TV program and link each instance of a spoken name to the identity of a real person in terms of a normalized identifier (in the form `Firstname_LASTNAME`). In the first step, the acoustic data is processed by a Speech-To-Text (STT) module. Second, the transcripts produced by the STT module are processed by a Named Entity Recognizer (NER) to detect person names. Note that the name in its spoken form may include only *part* of the name (first, middle or last name, eg. "Hollande" instead of "François Hollande"), an acronym or even a nickname. From these incomplete forms, the correct full name has to be guessed. This necessitates a post-processing step applied to the output of the ASR-NER modules. 6427, 1555 and 1947 spoken names were present in the training set, the development set and the test set, respectively.

A state-of-the-art off-the-shelf STT system for French [11] was used to transcribe the audio data. No task-specific adapta-

Modalities	Precision	Recall	F1-measure
written names	99.4%	95.7%	97.5%

Table 1: Quality of written names extraction for names written in title blocks

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

Post-processing	dev	test
None	61.1	60.0
Approach A	51.9	53.4
Approach B	**49.3**	**52.2**

Table 2: Spoken name detection performance in terms of SER (%).

tion was made for the REPERE evaluation (i.e., the REPERE training dataset was not used to adapt the acoustic models or the language models). The system obtained a word error rate of 16.43% (on around 36k words) during the first evaluation campaign of the REPERE challenge. In the NER module, two independent CRF models were trained on data annotated for the Quaero project: (1) a model to detect the mention of person with at least a first or a last name, and (2) a model to detect the different part of a person mention (e.g. first name or last name). These models use the same features as in [12]. In the final post-processing module to complete or correct the output of NER, two distinct approaches were studied.

Approach A used information from the *NER output itself*. Each name `N` in the output corresponding to one audio document (or TV show) was first checked if it is *full* (i.e. in the form `Firstname LASTNAME`). If not, `N` was searched inside the output. If `N` was found as *part* of another name `M` which itself is *full*, as its first, middle or last name, then each instance of `N` was replaced by `M`. For example, if the NER output contained both `MONTEBOURG` and `Arnaud_MONTEBOURG`, then each instance of the former was replaced by the latter. After this step, all remaining names which were still not full were searched in the Wikipedia. If a corresponding full name is found, it was used to replace the original name. All names remaining which were not full were discarded.

Approach B used information from the *groundtruth training data*. A Lookup Table (LUT) is created where each row contained (1) a name as it appears in the groundtruth training data, and (2) the corresponding name as it appears in the output of the ASR-NER system. When evaluating on dev or test, the LUT was used to translate each NER output to its corresponding correct form. Note that this method works only if the name occurred in the training data.

The task was evaluated by using the Slot Error Rate (SER) defined as: $\text{SER} = [\text{I} + \text{D} + 0.5 \times (\text{T} + \text{F})]/\text{R}$ where I is the Insertion error, D the Deletion error, T the Type error (i.e. a name was detected at the correct position but not the right name), F the Frontier error (i.e. the correct name was detected but not at the right time point) and R is the number of reference intervals. Table 2 shows the results obtained by the two approaches in terms of the SER. In the table, "None" refers to the case where the output of ASR-NER was directly used for evaluation. Note that the post-processing steps reduced the SER by about 10% absolute and 16.7% relative. Also, Approach B performed about 1% absolute better than A. This shows the role of training data in the performance of the system. It was also found that (1) combining A and B did not improve the scores more than B alone, and (2) about 70% of the deletion errors were a result of the ASR module.

4. Multimodal Fusion

In this section, we describe the runs submitted to the main multimodal tasks (supervised and unsupervised)

4.1. Propagation-based fusion

Unsupervised speakers recognition: This method is based on our previous work [13] (method M3). Speaker diarization and overlaid names recognition are run independently from each other. Speaker diarization is tuned to achieve the best diarization performance (i.e. minimize the diarization error rate, DER). The mapping between written names and speaker clusters is based on the following observations:

- when only one name is written on screen, any co-occurring speech turn is very likely (95% precision according to the train set) to be uttered by this person;
- the speaker diarization system can produce over-segmented speaker clusters, i.e. split speech turns from one speaker into two or more clusters.

Therefore, this method proceeds in two steps. First, speech turns with exactly one co-occurring name are tagged. Then, each remaining unnamed speech turn is tagged cluster-wise using an approach similar to the classical *Term-Frequency Inverse Document Frequency* (TF-IDF). We made two slight updates to this method: we reduce the temporal scope of each written names to the more co-occurring speech turn, this can correct the time offset between audio and written names segmentation. We also add the information of pronounced names: we name each remaining unnamed speech turn with closest pronounced names; this increases the number of speech turns named by our method.

Unsupervised faces recognition: As already stated, when one or more names are written on the screen, there is a very high probability that the name of one of the appearing face corresponds to the name written on screen. Therefore we use the information provided by written names during the face clustering process.

Before clustering, we associate each written name n to the co-occurring face. At this stage, a face can have several names if several names are written on the screen at the same time. Then, regular agglomerative clustering (based on face similarity) is performed with the constraint that merging two clusters s without at least one name n in common is forbidden.
For example, two clusters s_1 and s_2 **can** be merged into a new one s_{new} in the following case (the list of associated names is shown between brackets):

- $s_1(\emptyset) \cup s_2(\emptyset) \Rightarrow s_{new}(\emptyset)$
- $s_1(n_1) \cup s_2(\emptyset) \Rightarrow s_{new}(n_1)$
- $s_1(n_1, n_2) \cup s_2(\emptyset) \Rightarrow s_{new}(n_1, n_2)$
- $s_1(n_1, n_2) \cup s_2(n_1) \Rightarrow s_{new}(n_1)$

Below are examples where the two clusters **cannot** be merged:

- $s_1(n_1) \cup s_2(n_2) \Rightarrow$ Forbidden
- $s_1(n_1, n_3) \cup s_2(n_2) \Rightarrow$ Forbidden

The clustering is stopped according to the optimal threshold on the training set (minimizing the EGER, see Section 5.1).

4.2. Classifier-based Fusion

Speaker identification: Once all monomodal components have been run on a video, their outputs can be combined to improve the overall person recognition performance. Figure 1 draws up their list, along with two slightly modified versions of OCR: extended to the whole speech turns ($\texttt{OCR}^{+}$) or speaker diarization clusters ($\texttt{OCR}^{*}$).

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

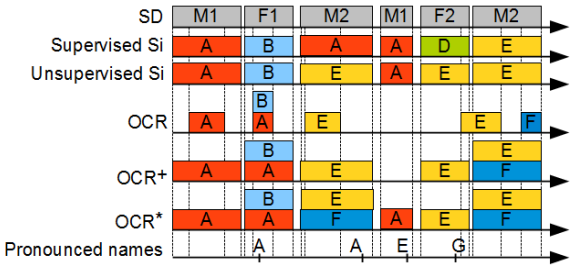


Figure 1: Several annotation timelines

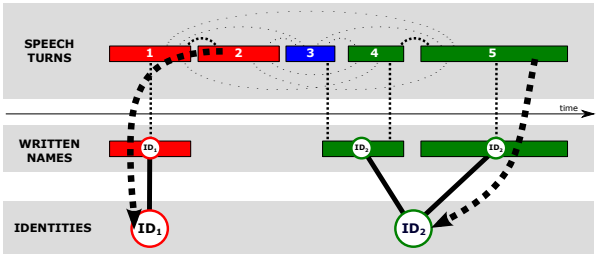


Figure 2: Multimodal probability graph for unsupervised speaker recognition, and two maximum probability paths.

Since each modality relies on its own temporal segmentation, the first step consists in aligning the various timelines onto the finest common segmentation. The final decision is taken at this segmentation granularity. For each resulting segment $\mathcal{S}$, a list of possible identities is built based on the output of all modalities. For each hypothesis identity $\mathcal{P}$, a set of features is extracted:

- Does the name of $\mathcal{P}$ appear in OCR? in OCR^+? in OCR^*?
- Duration of appearance of names in OCR^+, in OCR^* and their ratio.
- Speaker recognition scores for identity $\mathcal{P}$ provided by GSV-SVM SID and their difference to best competing scores.
- Is $\mathcal{P}$ the name proposed by the unsupervised speaker recognition system?
- Is $\mathcal{P}$ the most likely identity according to GSV-SVM SID?
- Has $\mathcal{P}$'s name been pronounced by the previous or the next speaker.

Based on these features, we trained a Multilayer Perceptron classifiers using Weka[3] to answer the following question: *"is $\mathcal{P}$ speaking for the duration of $\mathcal{S}$?"* Since these features can be either boolean or (unbounded) float, several classifiers insensitive to numerical types were used. The identity with the highest score is selected for the speaker identification task.

4.3. Rules-based Fusion

Supervised face recognition:

Several sources of information are exploited for multimodal and supervised face identification. They are combined using a set of simple rules, ordered by priority:

1. mono-modal face recognition for anchor persons;
2. names written on the screen;
3. unsupervised face recognition;
4. mono-modal face recognition for non-anchor persons;
5. multi-modal speaker recognition.

4.4. Graph-based Fusion

Alongside classifier-based approaches, the QCompere consortium also submitted a few contrastive runs based on a graphical representation of the person identification problem. For each video, a multimodal probability graph is built as illustrated in Figure 2. Each person utterance (e.g. a speech turn, a face track or a written name) is added as a vertex to this graph. For each target of supervised recognition systems (speaker identification or face recognition) and for each name found by name detection systems (written or spoken name detection), an identity vertex is added containing the normalized identifier of the person (e.g. `Nicolas_SARKOZY` or `Francois_HOLLANDE`).

Two vertices i and j are connected by an edge weighted by the probability p_{ij} that they correspond to the same person. This probability is obtained differently depending on the vertices it connects:

Intra-modal edges connect vertices of the same modality (speech turns-to-speech turns, or face tracks-to-face tracks). Probabilities are derived from the similarity scores d (BIC criterion for speech turns, learned metric for face tracks) using Bayes' theorem: $p(\mathcal{H} \mid d) = 1/(1+r)$ with $r = \frac{p(d|\overline{\mathcal{H}})}{p(d|\mathcal{H})} \frac{p(\overline{\mathcal{H}})}{p(\mathcal{H})}$ where $\mathcal{H}$ is the hypothesis that connected vertices are from the same person, and $\frac{p(d|\overline{\mathcal{H}})}{p(d|\mathcal{H})}$ and $\frac{p(\overline{\mathcal{H}})}{p(\mathcal{H})}$ are estimated using the annotated training set.

Cross-modal edges connect co-occurring vertices with two different modalities (e.g. a speech turn and a co-occurring written name) with a fixed probability estimated using the training set. For instance, two co-occurring speech turn and written name have more than 97% chance to correspond to the same person.

Identity edges connect detected names (written or spoken) to the corresponding identity with probability $p = 1$. They also connect speech turns and face tracks to target models (from supervised recognition system) with a probability derived from the identification scores.

Finally, person identification is achieved by looking for the maximum probability path between every speech turn (or face track) and all available identities. The probability of the path is simply defined as the product of the probability of its edges. It is straightforward to show that this maximum probability path problem can be modeled as a shortest path problem in the dual graph where edges are weighted by $-\log p_{ij}$ instead of p_{ij}. In Figure 2, speech turn #2 (resp. # 5) is given the identity `ID1` (resp. `ID2`).

The same framework can be used for speaker or face recognition; and for both supervised and unsupervised recognition. However, for the latter, one must remove *identity edges* coming from mono-modal speaker identification and face recognition system introduced in Section 2. Furthermore, one does not have to use all available edges to achieve the best performance. We only report on the best combination in Section 5.

The supervised run contains speech turn-to-identity (s-to-i), speech turn-to-written name (s-to-w) and w-to-i edges for speaker recognition, augmented with speech turn-to-face tracks

[3] `http://www.cs.waikato.ac.nz/ml/weka`

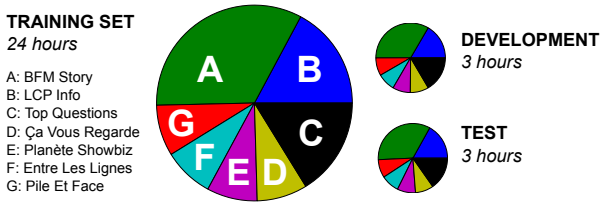


Figure 3: Training, development and test sets each contain 7 different types of shows (A to G).

(s-to-f) and f-to-w edges for face recognition. The unsupervised run contains s-to-w, h-to-w, w-to-i, s-to-h and h-to-h edges for both speaker and face recognition.

5. REPERE Evaluation Campaign 2013

5.1. Corpora & Metrics

Figure 3 provides a graphical overview of the REPERE video corpus 2013 [14] (training, development and test sets). Overall, it contains 188 videos (30 hours) recorded from 7 different shows broadcast by the French TV channels *BFM TV* and *LCP*.

While the audio annotation is dense (who speaks when?), the visual annotation (whose head appears when?) is only provided from one video frame every 10 seconds on average. [14] provides a more detailed description of the corpus and the associated annotation process.

Though the whole test set is processed, evaluation is only performed on the annotated frames $\mathcal{F}$. For each frame f, let us denote #total(f) the number of persons in the reference. The hypothesis proposed by an automatic system can make three types of errors: false alarms (#fa) when it contains more persons than there actually are in the reference; missed detections (#miss) when it contains less persons than there actually are in the reference; confusions (#conf) when the detected identity is wrong. For evaluation purposes, and because unknown people cannot – by definition – be recognized in any way, they are excluded from the scoring. The Estimated Global Error Rate (EGER) is defined by:

$$\text{EGER} = \frac{\sum_{f \in \mathcal{F}} \#\text{conf}(f) + \#\text{fa}(f) + \#\text{miss}(f)}{\sum_{f \in \mathcal{F}} \#\text{total}(f)}$$

5.2. Experimental protocol

For our experiments, the training set was split in two balanced subsets *train A* and *train B*. Target models (for speaker identification and face recognition of Section 2) are obtained using *train A*. *train B* is used to train classifiers and graph probabilities introduced in Section 4. The *development* set allows to tune various fusion parameters, and the final evaluation is done on the *test* set.

5.3. Supervised Recognition

Table 3 summarizes the performance achieved by our submissions to the supervised recognition task. Looking at the monomodal tasks, the speaker recognition system performs significantly better than the face recognition system (44.2% vs. 61.1% in EGER), probably due to more important variability factors in the image: face size, orientation, exposition, etc. The classifier-based fusion is very effective and reduces the speaker EGER to 17.8%, a 60% relative reduction compared to the mono-modal performance. The improvement brought by the rule-based fusion for faces is also important with a 39% relative reduction of errors, from 61.1% to 37.3%. The graph-based fusion is less effective but still reduces the EGER by about 20% relative compared to the mono-modal systems.

	Approach	EGER (%)
speaker	mono-modal speaker recognition	44.2
	classifier-based fusion	17.8
	graph-based fusion	35.3
head	mono-modal face recognition	61.1
	rule-based fusion	37.3
	graph-based fusion	48.1

Table 3: Performance of the QCompere submissions to the supervised person recognition tasks.

5.4. Unsupervised Recognition

The performance achieved by our submissions to the unsupervised recognition tasks are presented in Table 4; they are of course worse than the performance of a supervised multi-modal fusion, roughly 8.8% absolute above them. But they are also significantly better than the mono-modal identification scores, with 26.2% EGER for speakers and 46.2% for heads; this had already been shown for speakers after the REPERE dry-run evaluation [13]. Interestingly, the performance of unsupervised graph-based fusion for speakers and faces is almost similar to the performance observed for the supervised case (38.1% vs. 35.3% for speakers and 50.3% vs. 48.1% for faces), showing that there is room for improvement for this approach with a better integration of the person identification scores.

	Approach	EGER (%)
spk.	propagation-based fusion	26.2
	graph-based fusion	38.1
head	propagation-based fusion	46.2
	graph-based fusion	50.3

Table 4: Performance of the QCompere submissions to the unsupervised person recognition tasks.

6. Conclusion and Future Work

In this paper, we described, evaluated and discussed QCompere consortium submissions to the REPERE 2013 evaluation campaign. As expected, we showed that speaker identification and face recognition can be greatly improved when combined with name detection through video optical character recognition and automatic speech transcription available in TV broadcast. Moreover, it should be highlighted that the unsupervised person recognition approaches that we proposed perform much better than state-of-the-art supervised mono-modal ones (for both speaker and face identification). However, results show a strong performance discrepancy in favor of speaker recognition for all three participating consortia [1] as well as for QCompere various approaches. Therefore, for next year evaluation (scheduled in January 2014), a strong effort should be focused on face recognition.

7. References

[1] O. Galibert and J. Kahn, "The First Official REPERE Evaluation," in *First Workshop on Speech, Language and Audio for Multimedia (SLAM 2013)*, August 2013.

[2] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multi-Stage Speaker Diarization of Broadcast News," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1505–1512, 2006.

[3] V.-A. Tran, V. B. Le, C. Barras, and L. Lamel, "Comparing Multi-Stage Approaches for Cross-Show Speaker Diarization," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011).*, Florence, Italy, August 2011, pp. 1053–1056.

[4] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel, "Speaker segmentation and clustering in meetings," 2004.

[5] V.-B. Le, C. Barras, and M. Ferràs, "On the use of GSV-SVM for Speaker Diarization and Tracking," in *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 146–150.

[6] M. Bäuml, K. Bernardin, M. Fischer, H. K. Ekenel, and R. Stiefelhagen, "Multi-Pose Face Recognition for Person Retrieval in Camera Networks," in *International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2010.

[7] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! My name is... Buffy" Automatic Naming of Characters in TV Video," in *British Machine Vision Conference*, 2006.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[9] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *ICCV*, 2009.

[10] J. Poignant, L. Besacier, G. Quénot, and F. Thollard, "From Text Detection in Videos to Person Identification," in *International Conference on Multimedia & Expo (ICME)*, 2012.

[11] L. Lamel, S. Courcinous, J. Despres, J.-L. Gauvain, Y. Josse, K. Kilgour, F. Kraft, V. B. Le, H. Ney, M. Nußbaum-Thom, I. Oparin, T. Schlippe, R. Schlüter, T. Schultz, T. F. da Silva, S. Stüker, M. Sundermeyer, B. Vieru, N. T. Vu, A. Waibel, and C. Woehrling, "Speech Recognition for Machine Translation in Quaero," in *IWSLT*, San Francisco, CA, USA, 2011.

[12] M. Dinarelli and S. Rosset, "Models Cascade for Tree-Structured Named Entity Detection," in *IJCNLP'11, 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 2011.

[13] J. Poignant, H. Bredin, V. Le, L.Besacier, C.Barras, and G.Qunot, "Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast," in *Interspeech 2012*, 2012.

[14] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The REPERE Corpus: a Multimodal Corpus for Person Recognition," in *International Conference on Language Resources and Evaluation (LREC)*, 2012.

PERCOLI: a person identification system for the 2013 REPERE challenge

Benoit Favre[1], *Geraldine Damnati*[4], *Frederic Bechet*[1], *Meriem Bendris*[1], *Delphine Charlet*[4], *Remi Auguste*[2], *Stephane Ayache*[1], *Benjamin Bigot*[3], *Alexandre Delteil*[4], *Richard Dufour*[3], *Corinne Fredouille*[3], *Georges Linares*[3], *Jean Martinet*[2], *Gregory Senay*[3], *Pierre Tirilly*[2]

[1]Aix Marseille Université, LIF-CNRS ; [2]Université de Lille, LIFL
[3]Université d'Avignon, LIA; [4]Orange Labs, France

Abstract

The goal of the PERCOL project is to participate to the REPERE multimodal challenge by building a consortium combining different scientific fields (audio, text and video) in order to perform person recognition in video documents. The two main scientific issues addressed by the challenge are firstly multimodal fusion algorithms for automatic person recognition in video broadcast ; and secondly the improvement of information extraction from speech and images thanks to a combine decoding using both modalities to reduce decoding ambiguities. This paper describes the system PERCOLI that participated to the REPERE 2013 challenge and presents the results obtained on the main person recognition tasks.

Index Terms : multimodal fusion, person identification, video processing.

1. Introduction

The *Repere* challenge consists in identifying persons in video shows using cues from spoken content (speaker identity and words), and video content (faces and overlaid text) [1]. Systems participating in the challenge must generate a list of segments with person names according to the presence of said persons in the visual and audio modalities, using both biometric models and context analysis. The challenge provides a set of videos manually annotated with speaker segmentation, speech transcription, overlaid text transcription and face outline. All image-related annotations are sampled every 10 seconds on so-called key-frames.

Most visual indexing methods are based on face detection and recognition. Those methods require large databases of facial models trained to recognize each person who could appear in a video. However, the variability of face appearance in TV content (pose, facial expressions, lighting, occlusions) makes identification using facial models very unreliable. In addition, maintaining up-to-date large dictionaries of face models is prohibitively expensive. In this paper, we are interested in methods for naming faces in TV content with no face models.

Such person identification methods are often performed in two steps : (1) names are extracted from a range of sources and (2) an association strategy assigns each detected name to a person. In the first step, the identities can be extracted from speech (using Automatic Speech Recognition [2, 3]), image (with Optical Character Recognition [4] on overlaid text) and text content (such as scripts and subtitles [5]). In the second step, the extracted identities are propagated via clustering methods [4, 6]. This step is the focus of our paper. Figure 1 illustrates that process on a video from the *REPERE*[1] corpus [7].

We propose to directly associate OCR and speech detected names with current faces and speakers, and then propagate that information within and cross modalities with face and speaker similarities and talking face detection. This paper is organized as follows : Section 2 describes related work ; Section 5 describes person name acquisition from *OCR* and *ASR* output ; Section 6 similarity measures for speaker and face clustering ; Section 7 presents our identity propagation method based on direct and indirect association. Finally, Section 8 presents the *REPERE* corpus, results of experiments and a discussion.

FIGURE 1 – The *REPERE* corpus. The identity appears in multiple sources.

2. Related work

Several studies have addressed the problem of association-propagation strategies for face identification. Name-it [8] proposed to find face-name associations by maximizing the co-occurrence between similar faces and names extracted from OCR output. [9] proposed to name faces in images using a graphical model for face clustering. Nodes represent detected faces and edges are weighted by SIFT-based similarity. Then, for each name detected in OCR, greedy search is applied to find the sub-graph that maximizes face similarities within the set of faces associated to the name. However, this approach cannot identify faces if no name is detected in the image. In [10], authors proposed to identify faces in *TRECVID* news videos using training data obtained automatically from Google image search. Names were extracted from both OCR and ASR output. In [5], authors proposed to align detected faces with names from the script and used rules based on lip activity and gender detection to resolve ambiguities. In [6], names are extracted from movie scripts and subtitles and associated to faces

1. Reconnaissance de PERsonnes dans des Emissions Audiovisuelles : www.defi-repere.fr

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

according to lip activity. Identities are then propagated using face-level and clothes-level similarities. Although preliminary results are promising, face and clothes variability (pose, expression, color…) hamper the robustness of the similarity measure. In this case, audio information can be used in addition to visual cues to associate names to faces through speaker identity. In fact, in TV content speaker diarization appears to be more robust than face clustering [11]. [2, 3] proposed to extract names using *ASR* output and associated them to speakers using lexical rules on speaker clusters. In [4], names are extracted from OCR output and propagated to speaker clusters in order to maximize co-occurrence.

3. System architecture

The general architecture of the PERCOLI system is displayed in figure 2. The goal of this system is to predict, for each video frame, who is talking (*SPEAKER* hypotheses) and who appears on the screen (*FACE* hypotheses).

There are three main steps in the system :

1. **Person name hypotheses generation** : this step is in charge of producing all the person name identities that can be associated with a voice or a face in a given time window in a video .
2. **Multimodal speaker identification** : this process gives an identity, when possible, to each speaker segment produced during the speaker diarization process thanks to the person name identities given by the previous step.
3. **Multimodal face identification** : this second multimodal fusion process gives an identity, when possible, to each face segment produced by the face tracking process thanks to person name identities, face similarity, and speaker hypothesis provided by the previous step.

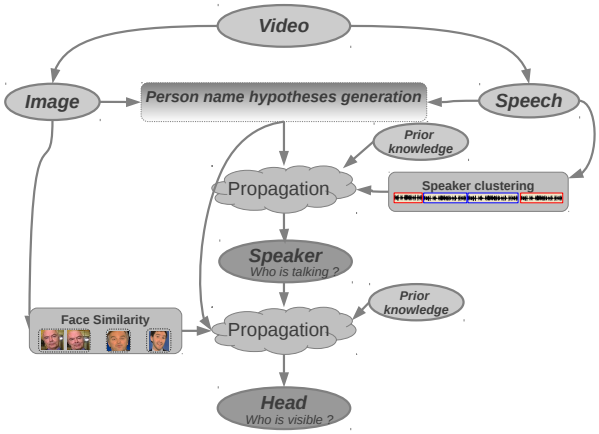


FIGURE 2 – General architecture of the Percoli system

The **Person name hypotheses generation** process is displayed in figure 3. There are three sources of person name identities : person names written in a text box, called *Overlay Person Name* hypotheses, and obtained through an Optical Character Recognition (OCR) process ; person names occurring in the speech channel, called *Utterred Person Name* hypotheses, and extracted from the Automatic Speech Recognition (ASR) of speech segments ; speaker recognition hypotheses obtained thanks to *a priori* speaker models corresponding to the main presenters, journalists and politicians likely to occur in the news. For the first two sources of identities, an *entity linking* process is needed in order to obtain a full identity from the occurrences of names detected in either text or speech.

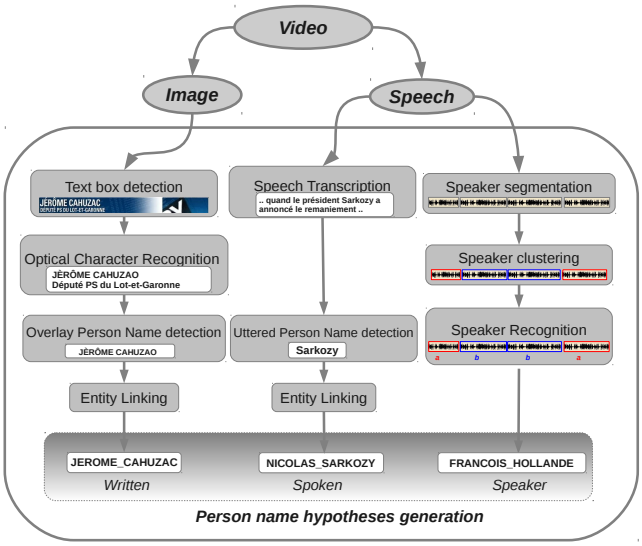


FIGURE 3 – Person name hypotheses generation according to three modalities : text in overlay text box ; speech ; speaker

4. Prior knowledge

Our system relies on three knowledge sources for performing identification : person name databases, speaker models and show-specific knowledge. From these sources, speaker models are the only biometric source of identity which is not allowed in unsupervised REPERE tasks.

4.1. Person name linking

Person name linking consists in linking name hypotheses manipulated in the system with a large database, in order to (1) discard unlikely names, and (2) account for meta-data such as the gender or spoken language of a person. In particular, we are interested at determining names which are likely to appear in the show being processed. This likelihood is estimated from entity linking and the aired date of the show. Entity linking is performed in two steps : clustering of person name variants and matching with database entries. The first step is achieve by collecting newswire articles around the dates of the broadcast shows, detecting names with a Named Entity Tagger [12] and clustering mentions. We applied this process to a large corpus of newswire spreading from 2004 to 2012. About 9.2M NEs were automatically detected. After the clustering and filtering process we obtained 117K clusters containing 162K mentions of person names. The second step consists in linking each cluster with a unique identifier. We used for this purpose the *ALEDA* entity database [13], which is a structured version of Wikipedia (about 225K person entities). From this database, heuristics are used to extract biographical elements such as gender, spoken language, topic, and whether the person is alive. When processing name hypotheses, they are linked to the mention which corresponds to the most frequent person cluster.

4.2. Speaker models

The speaker identification system is a standard GMM/UBM system (512 gaussians). We have collected audio for 345 speakers, mainly on journalists and politicians, from REPERE training data and various BN sources. Speakers with less than 30 seconds of speech are discarded. The generated models cover 30% of the training data speakers, 50% of the development data speakers and 54% of the test data speakers. Post-campaign evaluation has shown that the system has robustness issues because speakers with moderate quantity of training data are three times more likely to be incorrect on the test set than they are on the development set.

4.3. Show-specific constraints

The idea is to build models of who is likely to appear in recurring TV shows, and take advantage of show structure in order to capture names that would otherwise be difficult to detect. In particular, our system relies on two sources of information : lists of per-show presenters, journalists, columnists, commentators, and the setting of a show as its type (talk-show on a fixed stage, news with field reports), the number of invited speakers. In addition, for specific stage shows, an online component uses the order in which guests are presented to determine their location on stage and deduce probable coocurrence on a give camera angle.

5. Name detection

In addition to prior knowledge sources, person names are extracted from overlaid texts by using optical character recognition and from spoken content thanks to automatic speech recognition.

5.1. Optical Character Recognition

The Overlay Person Name (OPN) recognition process is made of 3 steps in our approach : text box detection ; Optical Character Recognition producing a confusion network of characters ; person name recognition in the character hypotheses.

Text box detection is achieved with a convolutional neural net approach described in [14], then OCR is performed with Tesseract [2], a standard open-source OCR system. Frame-to-frame tracked text boxes lead to different OCR hypotheses because of background changes and animations. The consecutive transcripts are merged in a Confusion Network (CN) in order to compute the most maximum posterior probability character sequence on the whole track. A few heuristics are used to locate actual person names in text boxes that contain other information (occupation, etc) and hypothesized names are filtered according to their Levenstein distance, computed efficiently with finite state transducers, to the large list of person names described in Section 4. If linking the name to the database fails, we back off to a web search and filter names returning less than 400 hits.

5.2. Automatic Speech Recognition

Automatic transcription of all speech content is not adequate for finding person names because word error rate can be relatively high (near 30% in our system) and names tend to be out-of-vocabulary and therefore never hypothesised by the system. Our name spotting component searches for names in phoneme confusion networks generated by a first pass of ASR. Given a list of potential person names likely to appear in the processed show, the system ranks them according to the average phonetic posterior after alignment of the phonetic transcription of the name (with a cutoff on the Levenstein distance). Name spotting and ASR 1-best are hybridized in order to retrieve first names which are easier for ASR.

2. https ://code.google.com/p/tesseract-ocr/

6. Speaker and Face Diarization

The task of diarization aims at determining for each pair of (visual or acoustic) frames whether it contains the same person. This task is often referred to as clustering.

6.1. Speaker diarization

The diarization system used in this work is the one presented in [15]. It is a sequential processing using firstly Bayesian Information Criterion and then Cross-likelihood Criterion, with special attention paid for overlapped speech. Overlapped speech segments are first detected and discarded from the clustering process, and then reassigned to the 2 nearest speakers, in terms of temporal distance between speech segments. Processing overlapped speech is particularly interesting for shows including debates.

6.2. Face diarization

Faces are detected using OpenCV's cascade classifier [16] for frontal and profile faces. The resulting detections are tracked until shot boundaries using bounding box overlap. Then, the upper body is detected using a background subtraction algorithm based on Grabcut [17], initialized with detected face. The background subtraction algorithm yields a very accurate silhouette of the person, even in presence of a dynamic background. Each extracted person is then modelled using a space-time color histogram [18]. This model stores color along with geometric and time information. It allows to retain the aspect of the person as it moves throughout the shot. A similarity matrix is built between person tracks using a combination of Bhattacharyya coefficient and Mahalanobis distance [18]. In the PERCOLI system, the similarity matrix is directly used in the face identification process as described in section 7.2 without requiring a specific clustering process.

7. Multimodal Fusion

As mentioned in section 3, our system identifies speaker identities in a first step and identifies face identities in a second step. Both identification steps are achieved thanks to a multimodal fusion system composed of two modules :

- Local identities propagation
- Show-specific post-processing

The following subsections describe the nature of local identities that are propagated for both steps, along with the generic propagation approach and the specific post-processing stage. Note that the different strategies have been designed in order to minimize the EGER metric (defined in section 8) for which all types of errors are equally weighted. In particular, substitutions and omissions having the same cost, we have chosen to try to give an identity to every detected speaker or face.

7.1. Multimodal speaker identification

Local identities for speaker identification are OPN hypothesis and scored speaker recognition (SR) hypothesis. The score

of an SR hypothesis is obtained thanks to a re-ranking process applied on the speaker recognition n-best list provided by speaker models. Re-ranking is based on the acoustical score and the presence of the speaker name in overlaid text as described in details in [19]. UPNs extracted from the spoken content are used in the post-processing and validation steps, along with show-specific *a priori* knowledge.

The core **identity propagation** method consists in attributing local identities to speakers in the following way :

1. direct identification : SR hypothesis are attributed to their corresponding speaker turns if their score is above a given threshold ; then for each local OPN hypothesis, the speaker turn which has the maximum duration overlap with the OPN span is given the identity carried by the OPN hypothesis ;
2. indirect identification : each unidentified speaker turn is given the identity of its speaker cluster, i.e. the OPN hypothesis which has the maximum duration overlap with the whole cluster or the SR which has the maximal score over the whole cluster if there is no such OPN hypothesis along the cluster.

The first **post-processing** step applies on speaker turns that have not been identified in the previous propagation step. It consists in using specific knowledge about the shows to identify speakers that cannot be identified by the core propagation step. It is applied for the unsupervised system where some speaker turns can remain unnamed after the propagation step. It is particularly designed for the identification of voice-overs for shows that contain reports commented by journalists (LCP_CaVousRegarde, LCP_LCPInfos, BFMTV_BFMStory, BFMTV_PlaneteShowbiz). The identity of such journalists is usually not displayed in the overlaid text and can only be retrieved from the spoken content. To this purpose, we use a predefined list of potential journalists for each type of show, and perform a specific name spotting in the audio content leading to a set of specific UPN hypotheses. A show can potentially contain several voice-over reports and we make the assumption the voice-over journalists are introduced before their report. Hence, after each specific UPN hypothesis, we attribute the corresponding identity to every unidentified speaker turn until the next specific UPN hypothesis.

Finaly a *global validation* step is performed for two particular shows for which the number of speakers is known in advance. It is the case of debate shows that only contain on-stage debates without any additional reports (LCP_PileEtFace with only three speaker and LCP_EntreLesLignes with five speakers). If the overall number of speaker identity hypotheses is above the a priori number of speakers N, the N most frequent hypotheses are kepts and the others are simply replaced by the most frequent hypothesis.

7.2. Multimodal face identification

Local identities for face identification are OPN hypothesis and speaker identities output by the previous multimodal speaker identification stage. UPNs extracted from the spoken content are used in the post-processing step, along with show-specific *a priori* knowledge.

The core **identity propagation** method is also the succession of a direct and an indirect identification steps. The *direct face identification* step follows the assumption that most OPNs occur while the corresponding face appears on the screen. Statistics on the *REPERE* corpus corroborate this idea, showing that 98.5% of the annotated frames containing an OPN also contain the corresponding face. Consequently in unambiguous shots where only one face is detected, we locally propagate the OPN to the face track. Then, for ambiguous shots where multiple faces could be identified by an OPN, we make a global decision using bipartite matching. For a given OPN, potential face sets are formed by gathering all face tracks that do not co-occur in the same shot. Then, that name is associated to the purest cluster containing all shots it occurs in.

The *indirect face identification* approach implemented in the PERCOLI system includes $Face \rightarrow Face$ propagation and cross-modal $Speaker \rightarrow Face$ propagation. First, the $Face \rightarrow Face$ propagation corresponds to the "*similarity based*" approach described in [20]. It is based on the principle that directly-named faces are very reliable, and can be considered as *models* in an open-set face identification paradigm. Let $\hat{g}_n$ be the set of faces directly associated to name n, for each face f with no direct naming, a distance D is defined between this face and $\hat{g}_n$. The name $\hat{n}(f)$ given to face f is the name for which the distance is minimal, if the distance $< \theta_1$.

$$\hat{n}(f) = \begin{cases} \text{argmin}_{n \in N} D(\hat{g}_n, f) \text{ if } D(\hat{g}_n, f) < \theta_1 \\ \emptyset \text{ otherwise} \end{cases}$$

with $D(g, f) = \frac{1}{|g|} \sum_{f_i \in g} d(f, f_i)$ being computed on the basis of the similarity matrix described in section 6.2. Note that for the moment, it has been applied only for shows from LCP channel (except LCP_TopQuestions) because similarity matrix were not reliable enough for the other shows that have a much more complex video editing policy.

The second aspect of our *indirect face identification* approach consists in attributing an identity to the remaining face tracks from the speaker modality. The identity of the speaker which has the maximal temporal overlap with the face track is given to the track. Note that if the system had to be designed in order to optimize identity precision, this $Speaker \rightarrow Face$ propagation should be refined and conditioned for instance to the response of a talking face detector.

The **post-processing** module relies on *a priori* knowledge of the shows, regarding their structure and their staging. If the structure of reports is difficult to model, studio set parts of a show usually follow a regular staging process. In order to exploit this specific knowledge, a set of rules has been manually designed for 4 different shows.

BFMTV_PLANETESHOWBIZ is dedicated to show business news : the show starts with an introduction by two anchor journalists in studio followed by several voice-over reports

→if one of the anchors is speaking and two faces are detected, then the second face is given the identity of the second anchor.

LCP_LCPINFO is a classical Brodcast News show with an alternation of reports introduced by an anchor speaker (identified as the first OPN hypothesis), and studio interviews between the anchor and the principal guest (identified as the most frequent OPN hypothesis). The staging of this show implies that during the interviews, the anchor and the guest can appear simultaneously on screen with smaller face sizes.

→if the principal guest is identified by the identity propagation step and has a small size, a second face track corresponding to the anchor is systematically added.

LCP_PILEETFACE is a debate between two politicians, their names are detected as being the two most frequent OPN hypotheses along the show. They appear on screen whether alone or together with a smaller head size.

→if one politician face track is identified by the identity propagation step and has a small size, a second face track corresponding to the second guest is systematically added.

LCP_ENTRELESLIGNES is a debate between four journalists which are sitting two by two on both sides of a square table. Their names are detected as being the four most frequent OPN hypotheses along the show. A specific spotting of these four names in the audio content of the very beginning of the show (when they are presented by the debate animator), allows to infer their position around the table. Actually they are always presented in the same order, and it is possible to infer who is sitting next to who and who is facing who.

→if one guest face track is identified by the identity propagation step and has a small size, a second face track corresponding to his neighbour is systematically added.

These rules are very specific but can cover a large proportion of shots in a studio show which follows a regular and structured staging.

8. Evaluation

Metric	**EGER**			**Pre.**	**Rec.**	**F-m**
Modality	**speak.**	**head**	**s+h**	**speaker+head**		
Sup. local	24.4	53.5	40.2	75.4	62.7	68.5
+ post-proc	24.5	50.4	**38.6**	75.7	64.6	69.7
Unsup. local	36.3	58.8	48.5	81.6	51.9	63.4
+ post-proc	34.1	55.4	**45.7**	67.5	57.6	62.2

FIGURE 4 – Global results in terms of EGER and F-measure for the two fusion strategies (local fusion and local fusion + post-processing) for the supervised and unsupervised tasks. The post-processed systems correspond to REPERE submissions.

EGER	**local**	**+ post-proc**
BFMTV_BFMStory	46.0	43.9
BFMTV_CultureEtVous	93.5	81.5
LCP_CaVousRegarde	65.8	67.5
LCP_EntreLesLignes	61.6	57.1
LCP_LCPInfo	51.5	48.0
LCP_PileEtFace	51.5	38.0
LCP_TopQuestions	64.8	69.3
All	58.8	55.4

FIGURE 5 – Impact of the fusion strategy on each show, for the head modality (unsupervised mode).

The results presented in this section have been obtained on the 2013 REPERE test corpus during the challenge. The output of the automatic systems participating to the challenge is a list, for each video file, of temporal segments representing the identities of detected persons in the video with the corresponding modality, such as :

```
s1 227.6  240.1  speaker Valerie_PECRESSE
s1 237.9  256    head    Nicolas_SARKOZY
s1 249.2  252.7  speaker Nicolas_SARKOZY
s1 282.2  284.1  head    Valerie_PECRESSE
```

The first field is the show id, then the time window, the modality (*head, speaker*) and the name in a normalized form (first name/last name). The 2013 REPERE test corpus contains 2 hours of video from 2 TV channels and 7 different shows. The evaluation was performed on 1187 manually annotated key-frames containing 1165 speaker identities and 1386 face identities. The official scoring metric of the REPERE challenge is an error metric called *Estimated Global Error Rate* (EGER). This metric compares the list of person names produced by the automatic systems on the key-frames with the reference list. One or several modalities can be considered in the scoring. EGER computes the error rate by adding three kinds of errors : *confusion*, *false alarm* and *missed detection*. The cost of each error is set to 1 and the following score is computed :

$$EGER(m) = \frac{Conf(m) + FA(m) + Miss(m)}{\#\ of\ person\ name\ in\ modality\ m}$$

In the official results, the main results are given for the *head+speaker* modality. The performance of the PERCOLI system presented in this paper are displayed in table 4 in term of EGER and F-score. We present two variants : one only with only local propagation, and the submitted output with the post-processing presented in section 7. Compared to other participants, a stratified shuffling test [21] shows that our supervised system is significantly worse than the best participant ($\Delta = 5.1$, $p = 0.026$), and that our unsupervised submission is not significantly different from the best submission ($\Delta = 3.4$, $p = 0.433$). In addition, Table 5 shows the impact of post-processing on each show. Clearly, the strategy only pays off for a few shows which have a stable structure. Figure 6 shows the origin of face naming decisions in the system for the subset of shows we were able to process with face similarity matrix. OCR direct naming and Face similarity-based naming are very accurate but only cover a third of the faces that we were able to identify. Naming from the co-occurent speaker is not very precise, but enables to name a large set of faces, for which no other information is available. Finally, the show specific post-processing is fairly accurate for these shows.

PRIMARY	**Supervised**		**Unsupervised**	
Origin	**%Test**	**%Corr.**	**%Test**	**%Corr.**
Direct OCR	12.7	98.5	12.7	98.5
Face similarity	20.8	84.3	20.8	84.3
Speaker →Face	49.7	67.8	49.7	59.1
Post-processing	16.6	86.0	16.6	84.0
Total	100.0	77.3	100.0	72.7

FIGURE 6 – Origin of face identities in the primary system output for LCP (except LCP_TopQuestions).

9. Conclusions

This paper presents the PERCOL system for the first phase of the REPERE challenge. The system is focused on the unsupervised task which precludes the use of prior biometric models. Person identification is achieved by (1) detecting names in overlaid text and speech, (2) linking those name hypotheses to large databases of known people, (3) propagating them to detected speakers and faces through clustering and show-specific heuristics. In addition the variant of the system for supervised identification uses speaker models as another source of identity. In particular, propagation is achieved first on speakers, and then on faces, because of the confidence we have in those two modalities. In the official evaluation, this approach performed on par with the best system on the main unsupervised task and not significantly worse than the second best system on the main supervised task. Future work includes taking advantage of training data to learn how to merge identity hypotheses from the various components of the system, as well as inserting face identification in the pipeline.

10. Acknowledgements

This work is supported by the ANR project PERCOL, under ANR contract number 2010-CORD-102-01.

11. References

[1] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus : a multimodal corpus for person recognition," in *LREC*, 2012.

[2] S. E. Tranter, "Who really spoke when ? finding speaker turns and identities in broadcast news audio," *ICASSP*, 2006.

[3] F. Liu and Y. Liu, "Identification of soundbite and its speaker name using transcripts of broadcast news speech," *ACM*, 2010.

[4] J. Poignant, H. Bredin, V.-B. Le, L. Besacier, C. Barras, and G. Quénot, "Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast," *Interspeech*, 2012.

[5] T. Cour, C. Jordan, and B. Taskar, "Learning from ambiguously labeled images," *CVPR*, 2009.

[6] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in tv video," *Image Vision Comput.*, 2009.

[7] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P. Joly, "A presentation of the repere challenge," in *CBMI*, 2012.

[8] S. Satoh and T. Kanade, "Name-it : Association of face and name in video," *CVPR*, 1997.

[9] D. Ozkan and P. Duygulu, "A graph based approach for naming faces in news photos," *CVPR*, 2006.

[10] C. Liu, S. Jiang, and Q. Huang, "Naming faces in broadcast news video by image google," *ACM*, 2008.

[11] H. Bredin, J. Poignant, M. Tapaswi, G. Fortier, V. B. Le, T. Napoléon, G. Hua, C. Barras, S. Rosset, L. Besacier *et al.*, "Fusion of speech, faces and text for person identification in tv broadcast," in *ECCV Workshop on Information fusion in Computer Vision for Concept Recognition*, 2012.

[12] F. Bechet and E. Charton, "Unsupervised knowledge acquisition for extracting named entities from speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, Dallas, USA, 2010.

[13] B. Sagot and R. Stern, "Aleda, a free large-scale entity database for French," in *Proceedings of LREC 2012*, Istanbul, Turquie, 2012, p. 4 pages. [Online]. Available : http ://hal.archives-ouvertes.fr/hal-00699300

[14] M. Delakis and C. Garcia, "Text detection with convolutional neural networks," *VISAPP*, 2008.

[15] D. Charlet, C. Barras, and J. Lienard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," *ICASSP*, 2013.

[16] P. Viola and M. Jones, "Robust real-time object detection," *IJCV*, 2002.

[17] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut : Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, 2004.

[18] R. Auguste, A. Aissaoui, J. Martinet, and C. Djeraba, "Les histogrammes spatio-temporels pour la ré-identification de personnes dans les journaux télévisés," *CORESA*, 2012.

[19] D. Charlet, C. Fredouille, G. Damnati, and G. Senay, "Improving speaker identification in tv-shows using person name detection in overlaid text and speech," *submitted to Interspeech*, 2013.

[20] M. Bendris, B. Favre, D. Charlet, G. Damnati, G. Senay, R. Auguste, and J. Martinet, "Unsupervised face identification in tv content using audio-visual sources," *CBMI*, 2013.

[21] E. Yücesan, "Evaluating alternative system configurations using simulation : A nonparametric approach," *Annals of Operations Research*, vol. 53, no. 1, pp. 471–484, 1994.

Named Entity Recognition in Speech Transcripts following an Extended Taxonomy

Mohamed Hatmi[1], *Christine Jacquin*[1], *Emmanuel Morin*[1], *Sylvain Meignier*[2]

[1] LINA, University of Nantes, France,
[2]LIUM, University of Le Mans, France
{mohamed.hatmi, christine.jacquin, emmanuel.morin}@univ-nantes.fr
sylvain.meignier@lium.univ-lemans.fr

Abstract

In this paper, we present a French named entity recognition (NER) system that was first developed as part of our participation in the ETAPE 2012 evaluation campaign and then extended to cover more entity types. The ETAPE 2012 evaluation campaign considers an hierarchical and compositional taxonomy that makes the NER task more complex. We present a multi-level methodology based on conditional random fields (CRFs). With respect to existing systems, our methodology allows a fine-grained annotation. Experiments were conducted using the manually annotated training and evaluation corpora provided by the organizers of the campaign. The obtained results are presented and discussed.

Index Terms: Named Entity Recognition, Structured Named Entities, CRF model.

1. Introduction

Named Entities (NEs) are defined as autonomous monoreferential linguistic expressions. They cover traditionally the names of all the person, organization and location. There are two most widespread approaches for the Named Entity Recognition (NER): symbolic approaches which rely on hand-coded grammar and gazetteers, and learning-based approaches which require large quantities of manually-annotated corpus [1].

NER from speech is mainly performed by transcribing speech and then applying NER approaches to transcripts. NER systems are adapted to fit in with the characteristics of automatic speech transcripts such as speech disfluencies, automatic speech recognition errors and out-of-vocabulary (OOV) problems. To that is added the problem of lack of some important NER features such as capitalization and punctuation. In order to improve speech NER, previous work has included restoring punctuation and capitalization in transcripts [2], using the Part-of-speech (POS) tags as features [3], incorporating indicative OOV words and ASR confidence features [4, 5, 6]. The ESTER 2 evaluation campaign [7, 8] has shown that the symbolic systems produce best results on manual transcripts whereas the learning-based systems show best results on automatic transcripts [9, 3].

NER systems require manually transcribed and annotated data, whether for performance evaluation or learning an annotation model. The adopted annotation schema has a direct impact on NER performance. For example, flat and relatively small entity types and granularity can achieve good results. The problem becomes more complex by using a fine-grained hierarchical taxonomy. As in [3], we propose a CRF-based approach that integrates the POS tags as features. However, the fundamental difference in our approach is that the adopted taxonomy is hierarchical and compositional.

In this paper, we present a French NER system that was first developed as part of our participation in the ETAPE 2012 evaluation campaign and then extended to cover more entity types. We propose a multi-level methodology which allows NER annotation following a fine-grained taxonomy. Three levels of annotation are defined : the first level consists of annotating the main categories, the second level has to do with the annotation of components and the last level deals with the problem of nested named entities.

This paper is organized as follows: Section 2 briefly presents the ETAPE evaluation campaign. Section 3 describes the Quaero extended taxonomy adopted in this campaign. Section 4 presents the corpora and the metrics used for evaluation. Section 5 presents the method used. Section 6 reports experimental results, while Section 7 concludes and presents future work.

2. The ETAPE evaluation campaign

The ETAPE evaluation campaign aimed to measure the performance of speech technologies for the French language [10]. Three main tasks were considered in this campaign: segmentation, transcription and information extraction. The evaluation concerned a variety of TV materials with various level of spontaneous speech and overlapping speech from multiple speakers. We are interested in the information extraction task that consists of detecting and categorizing all direct mentions of named entities following the Quaero named entity taxonomy.

3. Quaero named entity taxonomy

The Quaero annotation schema [11, 12] adopts a fine-grained hierarchical taxonomy. Named entity tagset is composed of 7 main categories and 32 sub-categories:

- **person**: invidual person (pers.ind), collectivity of persons (pers.coll),
- **location**: administrative location (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup), physical location (loc.phys.geo, loc.phys.hydro, loc.phys.astro),
- **organization**: services (org.ent), administration (org.adm),
- **function**: individual function (func.ind), collectivity of functions (func.coll),
- **human production**: manufactury object (prod.object), art products (prod.art), media products (prod.media), financial products (prod.fin), software (prod.soft), award

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

(prod.award), transportation route (prod.serv), doctrine (prod.doctr), law (prod.rule),

- **time**: absolute date (time.date.abs), date relative to the discourse (time.date.rel), absolute hour (time.hour.abs), hour relative to the discourse (time.hour.rel),
- **amount**

Entity tags are organized in a structured way so that a named entity can include another one. For example, in the named entity "<func.ind>Minister of <org.adm>Education</org.adm></func.ind>", the *func.ind* type includes the *org.adm* type.

In addition, the elements inside the named entities are categorized using components. A named entity includes at least one component. For example, a street name can be composed of a kind and a name : <loc.oro> <kind> rue </kind> de <name> Vaugirard </name> </loc.oro> (<loc.oro> <kind> *street* </kind> *of* <name> *Vaugirard* </name> </loc.oro>). There are two kinds of components:

- Transverse components that can fit each type of entity: name, kind, qualifier, demonym, val, unit, object, range-mark,
- Specific components which are only used for a reduced set of components: pers.ind (name.last, name.first, name.middle, title), loc.add.phys (address.number, po-box, zip-code, other-address-component), and time.date (week, day, month, year, century, millenium, reference-era, time-modifier)

In cases of metonymy, the named entity is double annotated with the type to which the entity intrinsically belongs and with the type to which the entity belongs according to the context. For example, the named entity "Roland Garros" is annotated as *loc.fac* and *pers.ind* in the sentence "We are in Roland Garros"

4. Corpora and metrics description

We first present the corpora used in this work, then we present the different metrics used for evaluation.

4.1. Corpora

To carry out the experiments, we used the ETAPE and ESTER 2 data which have been made available to the participants in the ETAPE evaluation campaign. The ETAPE corpus consists of 42.5 hours of data recorded from different French speaking radio and TV stations which are BFM TV, La Chaîne Parlementaire and TV8. The ESTER 2 corpus comprises about 100 hours of radio broadcast from various French speaking radios which are France Inter, Radio France International, France Culture, Radio Classique, Africa number one, Radio Congo and Radio Television du Maroc. These corpora have been manually transcribed and annotated following the Quaero named entity taxonomy.

The ETAPE and ESTER data jointly are divided into three parts. The first part (1,761,677 words) is used to train various CRF models. The second part (108,340 words) is used as development corpus to experiment with and adjust some parameters. The remaining (106,803 words) is used in the final evaluation.

4.2. Evaluation metrics

The evaluation of the NER performance is performed using the SER and the F-measure.

The SER [13] (cf. equation 1) combines different types of error: insertions (I), deletions (D) and substitutions (errors both in span and in type (S_{ST}), errors in span (S_S), errors in type (S_T)). The corresponding equation is given by:

$$SER = \frac{D + I + S_{ST} + 0.5 \times (S_S + S_T)}{\#\ of\ entities\ in\ the\ reference} \quad (1)$$

The F-measure (cf. equation 4) combines precision and recall. Precision (cf. equation 3) represents the percentage of annotated entities that are correct. Recall (cf. equation 2) represents the percentage of correct entities that are annotated. The corresponding equations are given by:

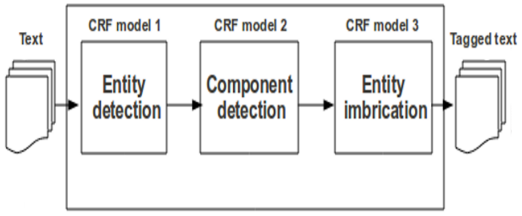
$$Recall = \frac{\#\ of\ correct\ annotated\ entities}{\#\ of\ annotated\ entities\ in\ the\ reference} \quad (2)$$

$$Precision = \frac{\#\ of\ correct\ annotated\ entities}{\#\ of\ annotated\ entities\ in\ the\ hypothesis} \quad (3)$$

$$F{-}measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4)$$

5. Method used

Figure 1: *Architecture of the NER system*



Several machine learning methods have been used for annotating named entities in text. Annotation is considered as sequence labeling task. Each word in the sequence is labeled with its appropriate tag. Tags include the category of the named entity and the location of the word within the named entities (BIO annotation). The first word in a named entity is tagged with "entity-tag-B", and further named entity words are tagged with "entity-tag-I". Words outside named entities are tagged with "O" (Other). Several studies [14] have shown that discriminant methods like Maximum Entropy Markov Model (MEMM) [15] or CRF [16] overcome the difficulties encountered in generative methods like Hidden Markov Model (HMM) [17]. Discriminative models allow to relax the independence assumptions needed by generative models and to include much more features in the model. The machine learning method employed in this work is CRF which is a discriminative undirected graphical model.

Named entities in the training data are organized in a structured way as shown in section 3. Named entities contain nested tagging of other named entities and components. Therefore, words constituting the named entities can belong at

the same time to one or more categories. This is a problem for the preparation of the training data for classification because each word must be assigned to just one category. Here is an example of a sentence in training corpus:

Vous êtes <func.ind> <kind> directeur </kind> de l' <org.ent> école nationale d' assurance </org.ent> </func.ind>
(*You are the* <func.ind> <kind> *director* </kind> *of the* <org.ent> *national insurance school* </org.ent> </func.ind>)

In order to handle structured tagging, we defined three levels of annotation. The first level consists of annotating the 32 categories in a flat way. The second level has to do with the annotation of components. The last level allows overlapping annotation when a category includes another category. We trained a CRF model for each level of annotation. Figure 1 shows the architecture of the NER system.

We used the open source implementation of CRF CRF++ toolkit[1] to implement the different models.

5.1. Entity detection

The first CRF model aims to annotate a text with the 32 categories. To achieve this, we presented the training data in a flat way by separating nested annotations and eliminating component tags. Here is how the sample sentence given in section 5 is presented in training corpus:

Vous êtes <func.ind> directeur de l' </func.ind> <org.ent> école nationale d' assurance </org.ent>
(*You are the* <func.ind> *director of the* </func.ind> <org.ent> *national insurance school* </org.ent>)

We then encoded the obtained corpus in BIO notation (Begin, Inside, Outside) and train the CRF model. Two types of features are used to predict if a word is part of a named entity or not:

- Contextual information: we took a context window of $[-2, +2]$ and consider unigram, bigram and trigram combinations.
- Semantic and syntactic information: we used the French POS tagger LIA_Tagg[2] to assign a POS tag to each word. LIA_Tagg is a free tool based on HMM. The POS tags are enriched with four semantic labels for proper names: person, organization, location and product. We augmented the lexicon of LIA_Tagg with 111,600 new named entities extracted form the Web (30,300 persons, 18,700 organizations et 62,600 locations). This allows a first lexicon-based level of annotation.

Figure 2 shows an example of the CRF training corpus.

5.2. Component detection

The second CRF model is applied on the output of the first CRF model. Here, the goal is to predict a component label to each word of a text. Two kinds of features are used to train the CRF model:

Figure 2: Example of the CRF training corpus for entity annotation

Vous	PPERP	O
êtes	VEP	O
directeur	NMS	func.ind-b
de	DETFS	func.ind-i
l'	DETFS	func.ind-i
école	NFS	org.ent-b
nationale	AFS	org.ent-i
d'	PREPADE	org.ent-i
assurance	NFS	org.ent-i

Figure 3: Example of the CRF training corpus for component annotation

Vous	O	O
êtes	O	O
directeur	func.ind-b	kind-b
de	func.ind-i	O
l'	func.ind-i	O
école	org.ent-b	O
nationale	org.ent-i	O
d'	org.ent-i	O
assurance	org.ent-i	O

- Contextual information: we took a context window of $[-2, +2]$.
- Semantic information: we used the first CRF model to assign a named entity tag to each word of a text.

Figure 3 shows an example of the CRF training corpus.

5.3. Entity imbrication

The third CRF model is applied on the output of the second CRF model in order to deal with the problem of nested named entities. In fact, named entities are annotated in a flat way in the second level. Therefore, we need to change the boundaries of certain named entities in order to overlap other ones. For example, in the sentence given in section 5.1, we need to move the closing tag of the function entity to overlap the organization entity. We used two types of features to train the CRF model in order to learn the imbrication rules:

- Contextual information: we took a context window of $[-2, +2]$.
- Semantic information: we used the first and the second CRF models to assign a named entity and a component tag to each word of a text.

Figure 4 shows an example of the CRF training corpus.

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

[2] http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/bechet/download_fred.html

Figure 4: Example of the CRF training corpus for entity imbrication

Vous	O	O	O
êtes	O	O	O
directeur	func.ind-b	kind-b	func.ind-b
de	func.ind-i	O	func.ind-i
l'	func.ind-i	O	func.ind-i
école	org.ent-b	O	func.ind-i
nationale	org.ent-i	O	func.ind-i
d'	org.ent-i	O	func.ind-i
assurance	org.ent-i	O	func.ind-i

Table 1: *Global NER results for the 32 categories computed on the manually-transcribed test corpus and ASR output. (F: F-measure, P: precision, R: recall)*

	Manual transcriptions	**ASR output (WER=23)**	**ASR output (WER=30)**
	F (R/P) (%)	**F (R/P) (%)**	**F (R/P) (%)**
S1	71.1 (62.5/18.7)	52.3 (43.7/65.1)	55.4 (46/69.6)
LL	**68** (64.5/71.7)	**49.9** (44.7/56.3)	**53.1** (46.59/61.91)
S2	66.4 (65.7/67.3)	40 (33.8/49.1)	41.4 (34.3/52.1)
S3	66.2 (65.6/77)	43.4 (41/46.1)	46.2 (44/48.6)
S4	66.2 (64.4/68.1)	43.7 (37.6/52.2)	48.8 (41.2/59.8)
S5	46.1 (67/61.4)	41.5 (40.7/42.4)	45 (43.3/46.8)
S6	55.5 (61/50.8)	23 (21.6/24.5)	27.7 (26.6/28.8)
S7	34.8 (28.3/45.1)	6.1 (19.1/9.2)	7.8 (23/11.6)

6. Results

We used the ETAPE test corpus to evaluate the performance of our system. This corpus contains 5,705 named entity occurrences and 7,174 component occurrences.

The NER system we used to participate in the ETAPE 2012 evaluation campaign annotates only the categories without components. It uses the first and the third CRF models. Table 1 shows the results obtained by our system, named *LL*, and the results of other participating NER systems for the 32 categories without components. The evaluation is performed on the manual transcriptions and ASR output with different Word Error Rate (WER). The ASR output is obtained from different ASR systems. The proposed approach achieves the second best F-measure on manual and automatic transcriptions for the 32 categories. Obviously, the F-measure decreases when dealing with automatic transcriptions. The NER features used for the well-written text appear insufficient to deal with noisy text and new specific ASR features are needed to be added.

After the ETAPE evaluation campaign, we extended our NER system to annotate also the components using the second CRF model. The system shows 37.5 % of SER on the manually-transcribed test corpus and 62.2 % of SER on the ASR output (WER=23). Table 2 shows the NER results by category. The results show good performance for some standard categories such as *pers.ind* and *loc.nat*, and poorer performances for others such as *loc.fac* and *prod.object*. These are characterized by a poor recall. This is mainly due to a low frequency in the training corpus. In addition, we observe some categorization errors particularly for the entities with metonymic sense (Paris as a town or as an organization) and between certain sub-categories of location and product. There is also some annotation ambiguity problems

Table 2: *NER results by category and component computed on the manually-transcribed test corpus and ASR output. (F: F-measure, P: precision, R: recall).*

	Manual transcriptions	**ASR output (WER=23)**	**Entities in reference**
	F (%)	**F (%)**	
amount	65.5	49.6	705
pers.ind	84.8	54.7	1,398
pers.coll	49.7	40.3	177
pers.other	0	0	1
time.date.abs	42.8	33	192
time.date.rel	74.7	65.5	348
time.hour.abs	58.1	32.8	46
time.hour.rel	74.2	62	84
loc.oro	0	50	2
loc.fac	13.7	11.3	81
loc.add.phys	0	0	4
loc.add.elec	83.3	46.15	18
loc.adm.town	71.4	44.5	279
loc.adm.reg	56.5	51.8	47
loc.adm.nat	86.9	77.9	276
loc.adm.sup	76.4	53.3	33
loc.phys.geo	23.5	0	30
loc.phys.hydro	0	0	5
loc.phys.astro	0	0	0
prod.object	6.2	0	58
prod.art	16.4	0	87
prod.media	68.1	56.8	164
prod.fin	19.6	11.6	84
prod.soft	0	0	0
prod.award	44.4	37.5	13
prod.serv	0	0	0
prod.doctr	0	33.3	3
prod.rule	54.5	0	5
prod.other	0	0	7
prod.unk	0	0	2
func.ind	59.8	51.6	383
func.coll	47.6	30.8	243
org.ent	45.7	38.3	307
org.adm	54.9	45.9	286
kind	50.4	42.3	1,163
extractor	40	33.3	4
qualifier	31.7	28.2	250
title	38.1	31.5	75
val	85	70.1	808
unit	87.6	75.4	463
object	61.7	49	225
range-mark	77	59.3	71
day	85.7	71.8	36
week	84.9	77.7	39
month	69	59.8	74
year	88.8	72.4	95
century	100	80	3
reference-era	25	33.3	2
time-modifier	64.3	57.2	337
award-cat	0	0	2
demonym	57.9	49	206
name	68.1	57.3	1,593
name.last	82.9	41.2	928
name.first	86.4	53.5	1,032
name.nickname	30.6	28.2	71
all	69.7	52.5	12,879

which concerns particularly some named entities composed of common nouns such as for *pers.coll* (e.g. classes populaires (working classes)), *func.coll* (e.g. sentinelles citoyennes (sentinel citizens)) and *prod.art* (e.g. devons-nous payer 100 % des tudes des futurs traders ? (should we pay 100 % of the studies of future traders)).

7. Conclusions

In this paper, we have presented a French NER system using CRF. We have proposed a multi-level method that annotates named entities following a fine-grained hierarchical taxonomy. The evaluation has shown good results on manual and automatic transcriptions. Future work will concentrate on improving the annotation of some categories and components that shows a weak performance. This is due to their limited appearance in the training corpus. We also intend to explore new features gathered from the ASR process to improve NER in automatic transcriptions.

8. References

[1] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticae Investigationes*, vol. 30, pp. 3–26, January 2007.

[2] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *Proceedings of ICASSP '09*, Taipei, Taiwan, 2009, pp. 4741–4744.

[3] F. Béchet and E. Charton, "Unsupervised knowledge acquisition for Extracting Named Entities from speech," in *Proceedings of ICASSP '10*, Dallas, Texas, USA, 2010, pp. 5338–5341.

[4] D. D. Palmer and M. Ostendorf, "Improving information extraction by modeling errors in speech recognizer output," in *Proceedings of HLT '01*, San Diego, California, USA, 2001, pp. 1–5.

[5] K. Sudoh, H. Tsukada, and H. Isozaki, "Incorporating speech recognition confidence into discriminative named entity recognition of speech data," in *Proceedings of ACL '06*, Sydney, Australia, 2006, pp. 617–624.

[6] C. Parada, M. Dredze, and F. Jelinek, "OOV Sensitive Named-Entity Recognition in Speech." in *Proceedings of INTERSPEECH '11*, Florence, Italy, 2011, pp. 2085–2088.

[7] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Proceedings of INTERSPEECH '09*, Brighton, UK, 2009, pp. 2583–2586.

[8] A. Zidouni, S. Rosset, and H. Glotin, "Efficient combined approach for named entity recognition in spoken language," in *Proceedings of INTERSPEECH '10*, Makuhari, Japan, 2010, pp. 1293–1296.

[9] J.-h. Kim and P. Woodland, "A Rule-Based Named Entity Recognition System for Speech Input," in *Proceedings of ICSLP '00*, Beijing, China, 2000, pp. 521–524.

[10] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *Proceedings of LREC '12*, Istanbul, Turkey, 2012, pp. 114–118.

[11] S. Rosset, C. Grouin, and P. Zweigenbaum, "Entités nommées structurées : guide d'annotation quaero," in *Technical Report*, LIMSI-CNRS, France, 2011.

[12] O. Galibert, S. Rosset, C. Grouin, P. Zweigenbaum, and L. Quintard, "Structured and extended named entity evaluation in automatic speech transcriptions," in *Proceedings of IJCNLP '11*, Chiang Mai, Thailand, 2011, pp. 518–526.

[13] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proceedings of DARPA Broadcast News Workshop*, Herndon, Virginia, USA, 1999, pp. 249–252.

[14] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding." in *Proceedings of INTERSPEECH '07*, Antwerp, Belgium, 2007, pp. 1605–1608.

[15] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, pp. 39–71, 1996.

[16] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of ICML'01*, Williamstown, MA, USA, 2001, pp. 282–289.

[17] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE '89*, 1989, pp. 257–286.

Speaker Role Recognition on TV Broadcast Documents

Benjamin Bigot, Corinne Fredouille

University of Avignon, LIA, France

firstname.lastname@univ-avignon.fr

Delphine Charlet

Orange Labs, Lannion, France

delphine.charlet@orange.com

Abstract

In this paper, we present the results obtained by a state-of-the-art system for Speaker Role Recognition (SRR) on the TV broadcast documents issued from the REPERE Multimedia Challenge. This SRR system is based on the assumption that cues about speaker roles may be extracted from a set of 36 low level features issued from the outputs of a Speaker Diarization process. Starting from manually annotated speaker segments, we first evaluate the performance of the SRR system, formerly evaluated on Broadcast radio recordings, on this heterogeneous set of TV shows. Consequently, we propose a new classification strategy, by observing how building show-dependent models improves SRR. The system is then applied on some speaker segmentation outputs issued from an automatic system, enabling us to investigate the influence of the errors introduced by this front-end process on Role Recognition. In these different contexts, the system is able to correctly classify 86.9% of speaker roles while being applied on manual speaker segmentations and 74.5% on automatic Speaker Diarization outputs.

Index Terms: speaker role recognition, speech processing, content-based indexing of audiovisual documents.

1. Introduction

Maintaining efficient means of access to the information held in the huge mass of audiovisual documents broadcast everyday by TV and radio channels is very challenging. The increasing number of projects and evaluation campaigns puts to the fore the important work currently achieved in order to propose automatic methods dedicated to information extraction, content-based indexing and structuring in audiovisual documents.

The REPERE Multimedia Challenge [1] (2010-2014) is dedicated to the specific task of person identification in TV programs. It provides a framework (corpora and evaluation protocols) to support research on this topic in multimodal conditions. The work presented in this paper is part of the PERCOL project which is one of the three consortia chosen to participate at this challenge. In the context of the first official phase of the REPERE campaign, the scientific partners involved in PERCOL have proposed several systems dedicated to speaker identities through the recognition of pronounced names and speaker identification in speech, person name detection in overlaid texts and face recognition in video.

We are interested in bringing information relative to speaker roles in a speaker identification perspective. In well structured documents, as in broadcast news programs, several studies have already taken advantage of the links between speaker roles (like anchor, journalist, guest or interview participant) and content structure. Role information has previously been used in [2] to summarize broadcast news documents, for topic indexing [3] and for story segmentation, relying either on the detection of the anchorman [4] or journalists [5].

This paper is dedicated to the application of a state-of-the-art speaker role recognition system on the document set of the REPERE challenge since there are several potential benefits in using information brought by speaker roles in a speaker identification perspective. For example, a TV show presenter is expected to introduce his guests and chroniclers by citing their names. Speaker roles may as well bring confidence in the outputs of a speaker recognition process.

The paper is structured as follows. In section 2 we present a brief description of related works on speaker role recognition. The document set as well as both the Speaker Diarization and Speaker Role Recognition systems used in this study are presented in sections 3 and 4 respectively. Section 5 is dedicated to the experiments carried out. Finally, we conclude this paper with some perspectives.

2. Related Works

First contributions to SRR [6, 7] are methods based on the outputs of Automatic Speech Recognition (ASR). A second category of approaches concern works based on Social Network Analysis and Social Affiliation Network applied on Speaker Diarization [8, 9]. In this second case, prior knowledge about the structure of the show is taken into account to determine relevant roles. These methods are mainly based on three classic roles: *anchorman*, *journalist*, *other* (or *guest*). The number of roles could be greater, but on very specific corpora (bulletins from a same news program). A more detailed survey of the state-of-the-art concerning these methods can be found in [10]. More recent contributions tend to benefit from speech transcriptions as well as from the temporal organization of speaker turns. In [11], authors integrate information relative to speaking style and *a priori* information about turn-taking patterns of conversations in a Dynamic Bayesian Network (DBN). The method in [12] assumes that cues about speaker roles are available in the way speakers formulate their questions. In [13] both structural and lexical features are used together. The approach proposed in [14] classifies speaker segments among the three classical roles. A first step based on the temporal distribution of speech segments and on the average Bayesian Information Criterion realizes the detection of *anchorman*. A second step, based on textual information achieves the classification of *journalist* and *other*. In [15] the authors investigate the links between speech spontaneity markers and speaker roles. The work of [16] is based on prosodic and temporal features calculated for every speaker segment obtained from a Speaker Diarization. The decision step is achieved using Conditional Random Fields (CRF). Because of the important diversity among these contributions (methods, features, language, corpus and metrics), these results are difficult to compare together.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

3. Corpus

At each step of the REPERE challenge, a set of TV broadcast documents and its annotations is delivered to the participants. The training set of the first phase has been annotated with speaker roles. We present in details this corpus as well as its features from a speaker role recognition point of view.

3.1. REPERE Broadcast TV document set

This data set is composed of 135 documents corresponding to several recordings of 7 different TV programs taken from two different TV channels. It corresponds to an overall speech duration of 24 hours distributed among these different programs as reported in Table 1. This set can be divided into three categories of programs :

- **Broadcast News** (13.8h) with *BFMStory*, *Showbiz* and *LCPInfo*. Documents belonging to this category count for more than the half of the entire data set duration. The program *Showbiz* is slightly different from the other ones and is a People News and gossip program.
- **Debates** (6.6h) : among this set, *EntreLesLignes* is dedicated to journalistic questions, while *ÇaVousRegarde* focuses on society questions. *PileEtFace* is a head-to-head political debate.
- The last category (3.6h) is for *TopQuestions*. These documents are **recordings of the parliamentary sessions of the French National Assembly**.

Name	nb. of shows	doc. type	speech duration
BFMStory	14	News	7.9h
Showbiz	66	News	1.9h
LCPInfo	15	News	4h
ÇaVousRegarde	6	Debate	2.2h
EntreLesLignes	7	Debate	2.2h
PileEtFace	9	Debate	2.2h
TopQuestion	18	Nat. Ass.	3.6h
	135		24h

Table 1: The REPERE data set and its speech distribution among various programs

3.2. Role definitions

Manual speaker segmentation has been enriched with annotations relative to speaker roles. Five types of roles have been consensually defined by the REPERE participants:

- **type R1** is for the anchorman persons, presenters and TV newscasters. Only one R1-type person is typically expected in each program, except for *Showbiz* where the shows are presented by 2 anchormen. As presented in table 2, R1 is the only role present in all the programs.
- **type R2** is for journalists and chroniclers. According to the chosen definition, these speakers must appear physically on the television studio set. We can see in tables 2 and 3 that these speakers are present in only three types of programs.

	R1	R2	R3	R4	R5	spk #
BFMStory	6.2	4.8	16.5	19.4	53.1	273
Showbiz	21.7	0	12.6	0	65.7	563
LCPInfo	5.5	2.2	19	5.8	67.5	274
ÇaVousRegarde	11.5	0	7.7	50	30.8	52
EntreLesLignes	20	80	0	0	0	35
PileEtFace	34.6	0	0	65.4	0	26
TopQuestions	12.9	0	0	0	87.1	140
overall (%)	14.2	3.5	12.6	8.2	61.5	
overall (# spk)	194	47	172	112	838	1363

Table 2: speaker role proportion for every type of program and total number of speakers

	R1	R2	R3	R4	R5
BFMStory	25.1	14.2	11	38.7	11
Showbiz	14	0	52.1	0	33.9
LCPInfo	27.2	3.8	20	27.8	21.2
ÇaVousRegarde	29	0	3.3	63	4.7
EntreLesLignes	25.8	74.2	0	0	0
PileEtFace	21.8	0	0	78.2	0
TopQuestions	3.9	0	0	0	96.1
overall (%)	21.6	12.1	11.3	30.4	24.6
oveall (h)	5.17	2.91	2.72	7.28	5.90

Table 3: speech proportion in percent of the speech duration per program depending on speaker roles

- **type R3** is for journalists who are not present on the television studio set and for voice-over journalists during reports. This role occurs in the News programs and in the reports of *ÇaVousRegarde* (cf. Table 2).
- **type R4** stands in a large manner for the guests and the experts present in the shows, and more precisely to any person that is not working for the TV show. These speakers may be on the television studio set, or on live by telephone. As shown in Table 3, this is the most important role in terms of speech duration.
- **type R5** contains all remaining speakers. It gathers anonymous persons, sound-bites and archives taken from press conference, spectators asking questions during a show, person interviewed during a report and politicians talking during a session at the National Assembly. This large category gathers 838 speakers but does not stand for the larger proportion of speech duration.

Speaker role distribution in terms of speaker numbers and speech duration is significantly different over the programs of the data set as depicted in Tables 2 and 3. For instance *BFMStory* and *LCPInfo* are the only programs that contain five roles. On the contrary *EntreLesLignes*, *PileEtFace* and *TopQuestions* contain only two roles with an important disproportion in terms of speaker numbers and speech duration.

4. Speaker Role Recognition Architecture

The Speaker Role Recognition system used in this study has been previously presented in [10]. This systems has reached the good score of 92% of roles correctly attributed to the speakers of the broadcast radio programs composing the EPAC project

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

corpus. It is initially dedicated to the recognition of 5 roles (*anchorman*, *punctual and recurrent journalists*, and *punctual and recurrent others*). The terms *punctual* and *recurrent* characterize speakers activity in one given document. It has been adapted for the need of this study to the role categories presented above. We first briefly present the speaker diarization system used in this work and then depict the classification procedure applied to the SRR system.

4.1. Speaker Diarization

The diarization system used in this work is the one presented in [17]. It is a sequential processing using firstly Bayesian Information Criterion and then Cross-likelihood Criterion, with special attention paid for overlapped speech for TV-debates, where the amount of overlapped speech is significant. For these shows, overlapped speech segments are first detected and discarded from the clustering process, and then reassigned to the 2 nearest speakers, in terms of temporal distance between speech segments. For news shows, overlapped speech is considered negligible, and this process is not applied.

4.2. Speaker Role Recognition system

This system relies on the assumption that cues about speaker roles can be extracted from a set of 36 low-level features (14 temporal features, 10 acoustic and 12 prosodic ones) computed from speech signal and from the temporal organization of speech turns available from a speaker diarization process. These features are used in a second time to model speaker roles using a supervised classification approach.

In [10] we put to the fore the efficiency of a hierarchical classification process where each classification step is reduced to a two-class problem. At each step, the redundancy or correlation of features for a given problem is reduced using a Principal Component Analysis and a discriminant model is learnt using a Linear SVM classifier. In this current study, the successive steps of the classification are adapted to the current problem. As presented in figure 1 a first step concerns the classification of the role R1. Then the classified speakers found as "not R1" are directed to a second classification step that considers R2 and R3 roles versus R4 and R5. Finally the last classification steps are done in parallel and separate R2 from R3 and R4 from R5.

5. Experiments

This classification is achieved at the scale of a speaker cluster and we assume that one speaker in one document has exactly one role. In order to deal with the quite limited number of samples in the corpus, the classification process is done in a leave-one-out fashion. Therefore, one document is used for test while the other documents are used to learn models.

		predicted	
		C1	¬ C1
ref.	C1	TP	FN
	¬ C1	FP	TN

Table 4: confusion matrix

Performance is reported in terms of Correct Classification Rate CCR defined according to the confusion matrix in Table 4:

$$CCR = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

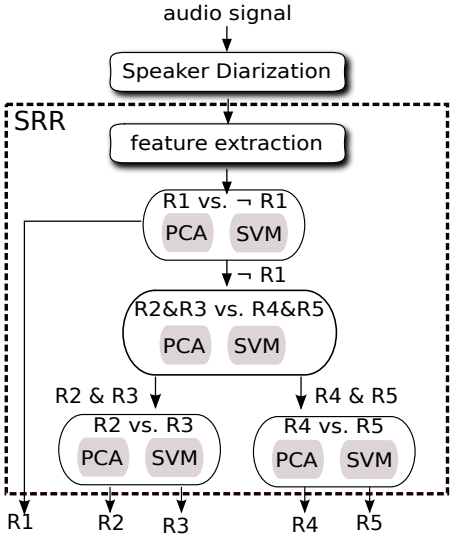


Figure 1: the speaker role recognition architecture

Recall and Precision measures defined respectively as follows are also reported :

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

in which TP, FP, and FN stand for true positive, false positive and false negative respectively.

5.1. Baseline-SRR using show-independent role models

This first experiment is done with manual speaker and role annotations, by applying a leave-one-out over the 135 documents of the data set independently of the type of program. Thus, while one document is processed for recognition, the 134 other documents are used for training.

The overall Correct Classification Rate (CCR) (cf. Table 5) is equal to 73.14% of speaker roles and 66.1% of the processed speech duration. A maximum 81.4% CCR is reached on the documents belonging to the program *TopQuestions*. The worst CCR value has been obtained on the programs *EntreLesLignes*. Globally, best performance has been reached on broadcast News programs. Role recognition seems less efficient on the debate programs where precision and recall values are particularly low for the role R1. In the one hand, we can assume that debate presenters and News broadcasters do not share similar temporal, prosodic and acoustic characteristics. On the other hand, R1 speakers in news programs are more numerous than R1 in debates. This may have led to R1 models more adapted for news programs. We have also observed that all R1 speakers in *Showbiz* have been attributed to the role R5. This confusion may be caused by overlapping music during speech interventions of R1 speakers in *Showbiz* which makes them more similar to R5 speakers. Finally this baseline system presents several major issues. First, this system allows confusion between role types that do not exist in test document. For instance, while processing recognition on a *TopQuestions* document, recognition may conduct to attribute speakers to the roles R2, R3 or

		R1		R2		R3		R4		R5	
	CCR	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
BFMStory	78.4	71.4	58.8	54.5	46.1	65.6	46.6	81.2	73.6	82.1	95.1
Showbiz	73.7	0	0	0	0	92.3	67.6	0	0	71.8	99.2
LCPInfo	77	80	33.3	0	0	71.4	19.2	60.9	87.5	79.5	98.9
ÇaVousRegarde	48.1	30	50	0	0	40	50	75	23	50	87.5
EntreLesLignes	25.7	28.5	57.1	100	17.8	0	0	0	0	0	0
PileEtFace	34.6	60	33.3	0	0	0	0	85.7	41.1	0	0
TopQuestions	81.4	100	66.7	0	0	0	0	0	0	94.4	83.6
overall (%)	73.14	60	18.6	52.4	23.4	62.8	47	69.9	58	75.8	95.9

Table 5: SRR performance in terms of Correct Classification Rate, Precision and Recall for every role and program type using the baseline architecture on manual speaker and role references

		R1		R2		R3		R4		R5	
	CCR	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
BFMStory	83.5	100	58.8	80	61.5	76.5	57.8	80.8	79.2	85	97.9
Showbiz	90.4	81.6	76.2	X	X	97.1	94.4	X	X	91.8	94.3
LCPInfo	87.2	92.3	80	0	0	76	73.1	91.7	68.8	89.9	96.2
ÇaVousRegarde	48.1	0	0	X	X	0	0	58.1	69.2	33.3	43.8
EntreLesLignes	85.7	66.7	57.1	89.7	92.9	X	X	X	X	X	X
PileEtFace	65.4	50	44.4	X	X	X	X	72.2	76.64	X	X
TopQuestions	97.8	94.1	88.9	X	X	X	X	X	X	98.4	99.2
overall (%)	86.9	82.7	71.6	85	72.3	85.6	76.2	74.3	75	89.7	95.1

Table 6: SRR performance in terms of Correct Classification Rate, Precision and Recall for every role and program type using a program dependent architecture on manual speaker and role references

R4 even if *TopQuestions* does not contain these roles. A second issue lays in the difference observed among speakers belonging to a given role type. For instance, we have observed low performance in R1 role recognition. There also exists an important confusion between R2 and R4 because these speakers seem to share similar characteristics. Finally, the size and the variety of speakers held in the R5 role may tend to unbalance the classification process. To overcome these limitations, we have first used program names as complementary features in the classification methods. Better performances have been reached in a second experiment, presented above, where role models are learnt depending on the type of programs.

5.2. SRR using show-dependent role models

In this experiment, the leave-one-out process is limited to the documents corresponding to one given program. For instance, to recognize speaker roles in *TopQuestions* we only use the 18 documents available. This configuration reduces the amount of data used to model speaker roles. Compared to the previous experiment, one benefit provided by this program-dependent approach is to make impossible several role confusions since the classifier will only learn SVM models for roles that really occur in these programs.

Overall CCR reaches 86.9% of speakers as reported in Table 6. This corresponds to 83% of the duration correctly labelled. Performance has been globally improved for every show except for *ÇaVousRegarde*. In the latter, CCR value remains unchanged compared to the one reached in the previous experiment and all R1 and R3 speakers have been attributed to the roles R4 and R5. Globally we observe that in debate programs, the confusion between R1 speakers and the other roles remains important. This puts to the fore a possible lack of efficiency of the actual low level features used for these documents. Another explanation may stand in the fact that the speaker and role annotations provided for these programs do not correspond to entire shows. In the next experiment, we apply this program-dependent strategy to automatic speaker segmentations provided by the speaker diarization system described in section 4.

5.3. SRR using show-dependent role models on Speaker Diarization outputs

Performance of the automatic Speaker Diarization system is reported in terms of Diarization Error Rate (DER). The overall performance on the corpus is equal to 12.1%, including overlapped speech in the evaluation. In Table 8 we observe that DER values depend on the type of programs. *Showbiz* presents the most important error rate value. This is mainly due to the high level of background music and noise, which gives a high miss detection rate, and also makes the clustering process more difficult for the detected speech. *EntreLesLignes* also presents an important DER value. This program contains several sequences of overlapping speech between chroniclers of the show.

To produce a ground truth for the evaluation of speaker role on the automatic speaker clusters, we first produce an alignment between the manual speaker segments and the outputs of the automatic speaker diarization system. Using the toolbox provided by NIST during a previous evaluation campaign we apply the Hungarian algorithm in order to associate automatic clusters with reference speakers. Then using these associations, we project the manual annotations for speaker roles over the automatic speaker clusters.

		R1		R2		R3		R4		R5	
	CCR	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
BFMStory	69.3	54.5	70.6	62.5	38.5	50.9	67.4	77.4	46.1	80.4	86
Showbiz	80.1	61.1	81.5	X	X	92.1	87.9	X	X	89.4	72.8
LCPInfo	76.5	52.2	80	0	0	55.4	76.6	80	25	92.1	84.8
ÇaVousRegarde	44.2	0	0	X	X	0	0	51.6	60.9	31.3	50
EntreLesLignes	79.4	50	42.9	85.7	88.9	X	X	X	X	X	X
PileEtFace	52	28.6	22.2	X	X	X	X	61.1	68.8	X	X
TopQuestions	83.8	83.8	41.7	55.6	X	X	X	X	X	92.9	88.1
overall (%)	74.5	53.9	65.9	78.4	64.4	66.5	76.9	65.4	49.5	86.7	83

Table 7: SRR performance in terms of Correct Classification Rate, Precision and Recall for every role and program type using a program dependent architecture on automatic speaker diarization outputs

sources	DER
BFMStory	12.6
Showbiz	34.4
LCPInfo	8.9
ÇaVousRegarde	12.2
EntreLesLignes	17.0
PileEtFace	10.5
TopQuestions	4.0
total	12.1

Table 8: Diarization Error Rate for each program type

One consequence of the automatic process is that several speakers of the reference do not match an "automatic" speaker. These differences are reported in Table 9 and 10. We can see that among the 1363 speakers, only 885 are associated with a corresponding cluster. This loss is directly related to the DER since most of these lost speakers are from the *Showbiz* program. Considering the speech duration lost during this process, we have evaluated (cf Table 10) that it represents only 6.25% ($1.5h$) of the initial document set. The impact of the process is as well important on the speakers belonging to the class R1. Their overall speech duration in the outputs of the automatic clustering is equal to $4.29h$ instead of $5.17h$ in the reference data.

	R1	R2	R3	R4	R5	# spk
BFMStory	7.6	5.8	19.1	23.1	44.4	225
Showbiz	26.9	0	32.8	0	40.3	201
LCPInfo	6.8	2.3	21.3	7.2	62.4	221
ÇaVousRegarde	14	0	9.3	53.5	23.2	43
EntreLesLignes	20.6	79.4	0	0	0	34
PileEtFace	36	0	0	64	0	25
TopQuestions	13.2	0	0	0	86.8	136
overall (%)	14.2	5.1	18.1	12.1	50.5	
overall (number)	126	45	160	107	447	885

Table 9: Speaker role in the diarization outputs in terms of speaker population

Speaker Role Recognition is then performed and an overall CCR equal to 74.5% is reached (cf Table 7). In term of speech duratio, it corresponds to 89% of the processed speech (22.5h) and 84.5% of the entire data set (24h). Program ranking in terms of CCR is the same as the one observed on the manual segmentations. The best classification rate has been obtained on *TopQuestions* with 83.8% of speaker roles correctly attributed.

	R1	R2	R3	R4	R5	dur.
BFMStory	20.1	15	12.2	41	11.7	7.27
Showbiz	14.6	0	66.2	0	19.2	1.51
LCPInfo	24.9	3.7	19.9	29.7	21.8	3.77
ÇaVousRegarde	27.9	0	3.4	64.9	3.8	2.08
EntreLesLignes	26.6	73.4	0	0	0	2.14
PileEtFace	19	0	0	81	0	2.1
TopQuestions	3.3	0	0	0	96.7	3.62
overall (%)	19.1	12.4	12	31.8	24.6	
overall (h)	4.29	2.8	2.71	7.15	5.54	22.5

Table 10: Speaker role in the diarization outputs in terms of speech duration

6. Conclusion

In this paper we have presented an experimental study dedicated to the application of a state-of-the-art Speaker Role Recognition system on the audiovisual document set provided in the context of the Multimedia Challenge REPERE. We first have carried out a baseline experiment by modelling speaker roles independently from the TV program processed. This system has been able to correctly attribute 73.14% of speaker roles, by using manual speaker segmentation. We have then proposed to build program dependent models of speaker roles. This second system has reported a correct classification rate of 86.9% on the same conditions. A third experiment has consisted in applying program-dependent models on the outputs of a speaker diarization system with a DER equal to 12.1%. On this difficult document set, our system has been able to correctly recognize 74.5% of speaker roles. In the context of the PERCOL project involved in the REPERE Challenge, our future work will be directed on the use of speaker role information in combination with speaker recognition system. We will as well investigate relations existing between speaker roles and the presence of person names pronounced in speech data in combination with systems like those presented in [18].

7. Acknowledgement

The authors thank the financial supports ANR 2010-CORD-102-02 from the French National Research Agency (ANR).

8. References

[1] Kahn, J., Galibert, O., Quintard, L, Carre, M., Giraudel, A. and Joly, P. "A presentation of the repere challenge," in CBMI, 2012, pp. 1–6.

[2] Maskey, S. and Hirschberg, J. "Automatic speech summarization of broadcast news using structural features". in Proc. of EUROSPEECH 2003.

[3] Amaral, R. and Trancoso, I. "Topic Indexing of TV Broadcast News Programs" in Proc. of PROPOR 2003.

[4] Kolluru, B. and Gotoh, Y. "Speaker role based structural classification of broadcast news stories" in Proc. of INTERSPEECH 2007

[5] Ma, C., Byun, B., Kim, I. and Lee, C.-H. " A detection-based approach to broadcast news video story segmentation" in Proc. of ICASSP 2009.

[6] Barzilay, R., Collins, M., Hirschberg, J. and Wittaker, S. "The rules behind the roles: identifying speaker role in radio broadcast" in Proc. of 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence. AAAI Press/The MIT Press, 2000, pp. 679–684.

[7] Liu, Y. "Initial study on automatic identification of speaker role in broadcast news speech," in Proc. of Human Language Technology Conference of the NAACL, 2006, pp. 81–84.

[8] Vinciarelli, A. "Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling," in IEEE Transactions on Multimedia, vol. 9, no. 6, pp. 1215–1226, 2007.

[9] Salamin, H., Favre, S. and Vinciarelli, A. "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," IEEE Transactions on Multimedia, vol. 11, no. 7, pp.1373–1380, 2009.

[10] Bigot, B., Ferrané, I., Pinquier, J. and R. André-Obrecht, "Speaker role recognition to help spontaneous conversational speech detection," in Proc. of Int. Workshop on Searching Spontaneous Conversational Speech , pp. 5–10, 2010.

[11] Yaman, S., Hakkani-Tür, D., and Tür, G. "Social role discovery from spoken language using dynamic bayesian networks," in Proc. of INTERSPEECH, 2010, pp. 2870–2873.

[12] Bazillon, T., Maza, B., Rouvier, M., Bechet, F. and Nasr, A. "Speaker role recognition using question detection and characterization," in Proc. of INTERSPEECH, 2011, pp. 1333–1336.

[13] Hutchinson, B., Zhang, B. and Ostendorf, M. "Unsupervised broadcast conversation speaker role labeling," in Proc. of IEEE ICASSP, 2010, pp. 5322 – 5325.

[14] Damnati, G. and Charlet, D. "Robust speaker turn role labeling of TV broadcast news shows," in Proc. of IEEE ICASSP, 2011, pp. 5684 – 5687.

[15] Dufour, R., Estève, Y. and Deléglise, P. "Investigation of spontaneous speech characterization applied to speaker role recognition," in Proc. of INTERSPEECH, 2011, pp. 917–920.

[16] Salamin, H., Truong, K., Mohammadi, G. and Vinciarelli, A. "Automatic Role Recognition Based on Conversational and Prosodic Behaviour," in Proc. of ACM Multimedia, 2010, pp. 847–850.

[17] D. Charlet, C. Barras and J.S. Lienard, "Impact of Overlapping Speech Detection on Speaker Diarization for Broadcast News and Debates", in Proc. of ICASSP, 2013.

[18] Bigot, B., Senay, G. , Linar'es, G., Fredouille, C. and Dufour,R. "Person Name Recognition in ASR outputs using Continuous Context Models" in Proc. of ICASSP, 2013.

Speaker Attribution of Australian Broadcast News Data

Houman Ghaemmaghami, David Dean, Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia
{houman.ghaemmaghami,d.dean,s.sridharan}@qut.edu.au

Abstract

Speaker attribution is the task of annotating a spoken audio archive based on speaker identities. This can be achieved using speaker diarization and speaker linking. In our previous work, we proposed an efficient attribution system, using complete-linkage clustering, for conducting attribution of large sets of two-speaker telephone data. In this paper, we build on our proposed approach to achieve a robust system, applicable to multiple recording domains. To do this, we first extend the diarization module of our system to accommodate multi-speaker (>2) recordings. We achieve this through using a robust cross-likelihood ratio (CLR) threshold stopping criterion for clustering, as opposed to the original stopping criterion of two speakers used for telephone data. We evaluate this baseline diarization module across a dataset of Australian broadcast news recordings, showing a significant lack of diarization accuracy without previous knowledge of the true number of speakers within a recording. We thus propose applying an additional pass of complete-linkage clustering to the diarization module, demonstrating an absolute improvement of 20% in diarization error rate (DER). We then evaluate our proposed multi-domain attribution system across the broadcast news data, demonstrating achievable attribution error rates (AER) as low as 17%.

Index Terms: speaker attribution, diarization, linking, complete linkage, broadcast news.

1. Introduction

The recent developments in speaker modeling and recognition techniques, such as joint factor analysis (JFA) modeling [1] and i-vector speaker modeling [2], have brought about great improvements to the field of speaker diarization [3, 4, 5]. This has motivated the proposal of *speaker attribution* as a recent field of research [4, 5, 6, 7, 8, 9]. Speaker attribution is the process of automatically annotating a typically large archive of spoken recordings based on the unique speaker identities that are present within the analysed archive of recordings, without any prior knowledge of the present speaker identities. This annotation can then be employed to search and index the recording archive based on speaker identity. A typical speaker attribution system can be divided into the two independent modules of speaker diarization and speaker linking [4, 5, 9]. In such a system, the set of recordings are first processed using speaker diarization to ideally extract a set of speaker-homogeneous segments from within each recording [10, 11]. These segments are then passed to the speaker linking module of the attribution system, where they are linked to identify segments belonging to the same speaker identities across multiple recordings [6, 8].

One of the main challenges with speaker attribution is the problem of session variation between the analysed set of recordings. Session variability can degrade the performance of speaker linking when attempting to cluster inter-session segments belonging to the same identity. In our previous work, we demonstrated the erroneous effects of inter-session variability on the tasks of speaker linking and attribution, and proposed the use of JFA modeling to overcome this issue [7]. JFA and i-vector modeling have since been the only speaker modeling techniques employed for conducting attribution [4, 5, 6, 9].

As speaker attribution is often employed to process large sets of data [4, 5, 6], it is of great importance to carry out this process in an efficient manner. The most obvious area for gaining efficiency is the clustering module of attribution. In diarization, clustering is typically based on a computationally expensive, agglomerative merging and retraining scheme [10, 11, 12, 13]. This may not pose a problem to diarization efficiency when processing short recordings, however this is highly inefficient for conducting speaker linking in large datasets. For this reason, van Leeuwen proposed an agglomerative clustering approach, without retraining, for speaker linking [6]. We then proposed a complete-linkage approach to clustering, for both diarization and speaker linking using JFA modeling and cross-likelihood ratio (CLR) scoring, and demonstrated that our complete-linkage clustering approach is more efficient and more accurate than traditional agglomerative clustering with retraining and the method proposed by van Leeuwen [7, 5, 8].

State-of-the-art attribution technology has largely dealt with two-speaker telephone recordings [4, 7, 5, 8], with recent work conducted by Ferras and Bourlard on attribution of meeting room data with poor results [9]. In this paper we extend our previously proposed telephone data attribution system [5], to a robust attribution method applicable to multiple recording domains. To do this, we collected a set of real, and publically available, Australian broadcast news recordings, with the topic of the recordings centered around related events to ensure multiple occurrences of identities across recordings. We then carried out a manual annotation of this dataset to obtain the ground-truth diarization labels for evaluation purposes.

As a common assumption in speaker diarization of telephone recordings [4, 5, 3], our previously proposed diarization module employed a stopping criterion of two speakers for the clustering process. We thus need to modify our diarization module to accommodate recordings with an arbitrary number of unique speaker identities. To do this we propose a CLR threshold stopping criterion for speaker clustering in our baseline diarization module. We justify our choice of this threshold value based on the computation of the CLR metric. We then evaluate this baseline diarization module across the broadcast news data and propose an additional pass of the clustering stage to improve the baseline system. We demonstrate an absolute improvement of 20% in DER over the baseline performance through the application of this additional pass of the clustering stage. We then evaluate our proposed speaker attribution system across the broadcast data to reveal an achievable AER of 17%, given an ideal speaker diarization module.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

2. Speaker modeling and clustering

To carry out robust and efficient speaker attribution of inter-session spoken recordings, we draw from our previous work and employ a JFA speaker modeling approach with session compensation [14, 15]. We compare the modeled speaker segments using the pairwise CLR metric [10]. The pairwise CLR scores are then used to conduct a single stage complete-linkage clustering of the speaker segments without retraining. [5, 8]. This section provides the theory behind JFA speaker modeling, pairwise CLR scoring and complete-linkage clustering.

2.1. JFA speaker modeling

We perform JFA modeling with session compensation using a combined gender universal background model (UBM) [14, 15]. To do this, we introduce a constrained offset of the speaker-dependent, session-independent, Gaussian mixture model (GMM) mean supervector, $\boldsymbol{m}$,

$$\boldsymbol{m}_i(s) = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}(s) + \boldsymbol{D}\boldsymbol{z}(s) + \boldsymbol{U}\boldsymbol{x}_i(s), \tag{1}$$

where $\boldsymbol{m}$ is the speaker- and session-independent GMM-UBM mean supervector of dimension $CL \times 1$, with C being the number of mixture components used in the GMM-UBM and L the dimension of the features. $\boldsymbol{x}_i(s)$ is a low-dimensional representation of variability in session i, and $\boldsymbol{U}$ is a low-rank transformation matrix from the session subspace to the GMM-UBM mean supervector space. $\boldsymbol{y}(s)$ is the speaker factors, representing the speaker in a specified subspace with a standard normal distribution [15]. $\boldsymbol{V}$ is a low-rank transformation matrix from the speaker subspace to the GMM-UBM mean supervector space. $\boldsymbol{D}\boldsymbol{z}(s)$ is the residual variability not captured by the speaker subspace, where $\boldsymbol{z}(s)$ is a vector of hidden variables with a standard Gaussian distribution, $N(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$. $\boldsymbol{D}$ is the diagonal relevance maximum *a posteriori* (MAP) loading matrix [16].

To conduct JFA modeling it is necessary to estimate the speaker independent hyperparameters $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{D}$, $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$. In our work, we employ the coupled expectation-maximization (EM) algorithm hyperparameter training proposed by Vogt et al. [15].

2.2. CLR model comparison

After JFA modeling of the initial speaker segments, a robust metric is required to perform a pairwise comparison of the speaker models prior to clustering. We use the CLR metric as it has been shown to be a robust measure of pairwise similarity between models adapted using a UBM [10]. To do this, given two speaker segments i and j, and their corresponding feature vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively, the CLR score a_{ij} is computed as,

$$a_{ij} = \frac{1}{K_i} log \frac{p(\boldsymbol{x}_i|M_j)}{p(\boldsymbol{x}_i|M_B)} + \frac{1}{K_j} log \frac{p(\boldsymbol{x}_j|M_i)}{p(\boldsymbol{x}_j|M_B)}, \tag{2}$$

where, K_i and K_j represent the number of observations in $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively. M_i and M_j are the adapted models, and $p(\boldsymbol{x}|M)$ is the likelihood of $\boldsymbol{x}$, given model M, with M_B representing the GMM-UBM.

We then use the work by Glembek et al. [17], to accommodate CLR scoring into the JFA framework, calculating the likelihood function of model M, given data $\boldsymbol{x}$, using,

$$logp(\boldsymbol{x}|M) = \boldsymbol{Z}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{F} + \frac{1}{2} \boldsymbol{Z}^* \boldsymbol{N} \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}\,, \tag{3}$$

where, $\boldsymbol{\Sigma}$ is a $CP \times CP$ diagonal covariance matrix containing c, GMM components' diagonal covariance matrices, $\boldsymbol{\Sigma}_c$ of dimension $P \times P$. $\boldsymbol{N}$ is a $CP \times CP$ dimensional diagonal matrix consisting of each component's zeroth order Baum-Welch statistics, and $\boldsymbol{F}$ is a $CP \times 1$ dimensional vector achieved by concatenating the first order Baum-Welch statistics of each component. In our work, $\boldsymbol{F}$ was centralised on the GMM-UBM (M_B) mean mixture components.

2.3. Complete-linkage clustering

In our previous work we have demonstrated the efficiency and robustness of complete-linkage clustering [5], and have shown that this clustering method outperforms the traditional agglomerative cluster merging and retraining approach that is extensively used in speaker diarization [11, 18, 12, 13], as well as the alternative technique proposed by van Leeuwen [6], for carrying out agglomerative speaker clustering without retraining.

Complete-linkage clustering is a form of hierarchical clustering, in which the pairwise distance between clusters is employed to construct a clustering tree that represents the relationship between all speakers/clusters. The obtained tree can then be employed to merge clusters based on the complete-linkage criterion, and the final clustering outcome is then acquired using a distance threshold or the desired number of clusters [19].

In complete-linkage clustering models are initially merged based on a highest similarity, or lowest distance, score. As this clustering technique does not conduct retraining after each cluster merge, the pairwise scores between clusters are updated after a merge to indicate the distance between their most dissimilar elements. This approach thus takes into account the *best worst-case scenario* scores and assesses the relationship between all elements within two compared clusters, allowing for a more robust clustering decision.

To carry out complete-linkage clustering we first obtain the upper-triangular matrix $\boldsymbol{A}$, known as the attribution matrix [5], containing the pairwise CLR scores a_{ij} between all compared speaker models. As complete-linkage clustering is designed to compare distance values, as with our previous work [7, 5], from $\boldsymbol{A}$ we first compute an upper-triangular matrix $\boldsymbol{L}$, containing the corresponding pairwise distance scores l_{ij}, computed from the a_{ij} CLR scores using,

$$l_{ij} = \begin{cases} e^{(-a_{ij})}, & (i \neq j), \\ 0, & (i = j). \end{cases} \tag{4}$$

We then perform complete-linkage clustering using the distance attribution matrix $\boldsymbol{L}$, in the following manner:

1. Initialize $C{=}N$ clusters, assigning segment i to C_i.
2. Find the minimum distance score, l_{ij} and its corresponding clusters C_i and C_j.
3. Merge segments i and j by merging C_i and C_j into $C_{i'} = \{C_i, C_j\}$, and removing rows and columns i and j from $\boldsymbol{L}$.
4. Obtain the new $(N-1){\times}(N-1)$ matrix $\boldsymbol{L}$, by computing the distance between newly formed cluster and remaining clusters using the complete-linkage rule:

$$l_{i'r} = max(l_{ir}, l_{jr}) \tag{5}$$

5. If the stopping criterion is satisfied stop clustering, else repeat from step 2.

3. The SAIVT-BNEWS dataset

As speaker attribution is a recent area of research, there is a lack of availability of *suitable* datasets for evaluating proposed speaker attribution technology. A suitable evaluation corpus is one that provides reference diarization labels for each recording in the dataset, with multiple occurrences of speaker identities across recordings. In addition, a speaker identity key is required to ensure that each speaker, within each recording, can be mapped to a unique identity across the entire set of recordings. For this reason, in our previous work [7, 5, 8], we employed the National Institute of Standards and Technology (NIST) SRE 2008 summed channel telephone conversation test corpus [20]. This telephone corpus provides a range of inter-session data and allows for the convenience of employing a two-speaker stopping threshold for the diarization of each recording [3, 4, 5].

In this work, we collected a set of publically available Australian broadcast news recordings from a media website providing up to 100 broadcast news videos per day. We used this data to create a suitable attribution evaluation dataset, referred to as the SAIVT-BNEWS corpus. We did this to allow for free access to the data by other researchers active in the field of speaker attribution. We first collected a subset of the broadcast news data. This subset contained 55 broadcast news videos, centered on the same news topic and its related events. We selected the videos in this manner to ensure that the dataset contains multiple occurrences of unique speaker identities across recordings. We then extracted the audio, from the broadcast news videos, and manually produced reference diarization labels for each recording. To then identify the unique speaker identities across the set of recordings, we utilised the information in the video to label speakers across the recordings, allowing for the evaluation of speaker attribution across this subset of 55 recordings.

The 55 recordings collected range from 47 seconds to 5 minutes and 47 seconds in length. Each recording contains a different number of unique speaker identities, ranging from 1 speaker to a maximum of 9 speakers per analysed recording. As the recordings are from the broadcast news domain, a wide range of channel variations are observed both within and between recordings. Using reference diarization labels, a total of 175 initial speaker homogeneous segments are obtained, which can be linked to a total of 92 unique speaker identities across the entire dataset, consisting of 64 male and 28 female speakers.

A large variety of speakers are present in this dataset, such as reporters, politicians, children, elderly people and more. The presence of music in some videos and overlapping speech from different speakers provides an excellent corpus for evaluating the performance of attribution technology, as well as the possibility of addressing other new challenges. To obtain the SAIVT-BNEWS dataset, and its corresponding reference labels, the last author of this paper may be contacted by email.

4. Evaluation and results

In our previous work, we proposed a full speaker attribution system for conducting robust and efficient attribution of large datasets containing two-speaker telephone conversation recordings [7, 5, 8]. In this section we propose and evaluate a robust and efficient attribution approach that is applicable to multiple recording domains, with an arbitrary number of speakers within each recording. We begin by employing our telephone-data attribution system [5], and modify the diarization module of this system to accommodate recordings with any number of speakers, rather than only two speakers assumed for telephone conversations. We evaluate this baseline diarization approach on the SAIVT-BNEWS dataset (detailed in Section 3) to measure the performance of our previously proposed telephone-data diarization scheme, and reveal its robustness on a significantly different audio domain. We then analyse the shortcomings of our baseline diarization system and propose a simple modification to significantly improve the performance of this module.

After speaker diarization of the data, speaker linking is required to complete the task of speaker attribution. In this section, we propose employing our telephone-data speaker linking module [5, 8], to complete our multi-domain attribution system. We then evaluate our proposed attribution approach across the broadcast news dataset to demonstrate our system's performance across this corpus.

We evaluate the speaker diarization systems using the standard diarization error rate (DER) metric, as defined by NIST [20]. To evaluate our proposed speaker attribution system, we employ our previously proposed attribution error rate (AER) metric [5, 8]. In the studies conducted by van Leeuwen [6], and Vaquero et al. [4], cluster purity and coverage are used for evaluating speaker linking and attribution. We previously employed these measures to evaluate our system [7], however it is necessary to employ an error metric that reflects diarization errors, as well as the speaker linking errors. We believe the AER is a more appropriate metric for evaluating the task of attribution. The AER can be described as an extension to the standard DER measure, from a single recording, to a collection of recordings. The AER thus represents the percentage of time that a speaker identity is misattributed within recordings, as well as across recordings. To compute the AER it is necessary to first concatenate the reference diarization labels into a single label file and to then ensure that each unique speaker identity is labeled using a unique label across the entire concatenated reference label file. This can be referred to as the attribution reference label. The same process is then required to generate the attribution system label file, but this time based on the system's decision of the diarization output and the linked speaker identities. The two label files can then be compared using the NIST DER metric [20], however as this measured error is now representative of the DER per recording, as well as the speaker errors across recordings, we refer to it as the AER.

For JFA modeling the speaker and session subspaces were obtained using a coupled EM algorithm, with a 50-dimensional session and 200-dimensional speaker subspace [15]. The features we employed for speaker modeling were 13 MFCCs with 0^{th} order coefficient, deltas and feature warping [21], extracted using a 20 bin Mel-filterbank, 32 ms Hamming window and a 10 ms window shift. For the segmentation stages of our diarization module, as will be detailed in this section, we use 20 MFCCs with 0^{th} order coefficient, no deltas or feature warping, extracted in a similar manner. It is important to note that for JFA modeling of speaker segments, in both the diarization and speaker linking modules, we employ a previously trained combined gender GMM-UBM, consisting of 512 mixture components, trained using telephone speech data, as detailed in our previous work [7]. This means that our modeling approach is expected to perform better when dealing with telephone domain data. This work thus reveals the robustness of our attribution approach with respect to processing of multi-domain data.

4.1. Speaker diarization

As our baseline diarization system, we employ our previously proposed telephone-data speaker diarization module [5]. This

system was designed to perform robust and efficient diarization of two-speaker telephone conversation recordings. In this system, we followed the common practice of telephone-data diarization [4, 3], and employed our prior knowledge of the number of speakers within each recording as the stopping criterion to the clustering stage of our diarization module. We now require a method of dealing with an arbitrary number of speakers. Recall from Section 2.3, complete-linkage clustering can be carried out using the desired number of output clusters, or a distance threshold, as the stopping criterion to the clustering process. As we have no prior knowledge of the number of speakers within each recording, we propose using a suitable CLR threshold as the stopping criterion to the clustering phase of diarization. We thus go back to the CLR computation in (2),

$$a_{ij} = \overbrace{\frac{1}{K_i} log \frac{p(\boldsymbol{x}_i|M_j)}{p(\boldsymbol{x}_i|M_B)}}^{\delta_i} + \overbrace{\frac{1}{K_j} log \frac{p(\boldsymbol{x}_j|M_i)}{p(\boldsymbol{x}_j|M_B)}}^{\delta_j}, \qquad (6)$$

where (6) displays two splits of the CLR measure, δ_i and δ_j. δ_i represents likelihood that the data for speaker i is produced by the competing speaker model M_j, compared to the likelihood of this data being produced by the general speaker population (GMM-UBM). δ_j is the same measure, but for speaker j. From (6), a_{ij} will be negative if the general speaker population better models a speaker than its competing model, and a positive a_{ij} signifies that the speaker data in i and j are more similar to each other compared to the general speaker population. If ideal models are used, we would not expect δ_i and δ_j to have opposite signs and high absolute values, as it does not make sense for speaker i to be very similar to j but for j to be very different to speaker i. For these reasons, $a_{ij} = 0$ would serve as a suitable theoretical CLR threshold. We thus employ $a_{ij} \leq 0$ as the stopping criterion to the clustering stage of our diarization module to deal with an arbitrary number of speakers.

4.1.1. Baseline diarization system

We previously proposed a speaker diarization method using complete-linkage clustering for conducting efficient diarization within our proposed speaker attribution system [5] In this diarization system, we employ the hybrid voice activity detection (VAD) and the ergodic hidden Markov model (HMM) Viterbi resegmentation approach presented in [11]. We first use Viterbi segmentation to achieve an initial segmentation of the recordings, and then carry out modeling and clustering of these segments to complete the diarization process. We then apply a final Viterbi segmentation of the output speaker/clusters to refine the segment boundaries. In this work, we employ this system as our baseline diarization module and apply the CLR threshold stopping criterion, discussed in Section 4.1.

Our baseline system consists of the following stages:

1. Linear segmentation of the audio into 4 second segments and 3 iterations of Viterbi using 32 component GMMs to model each segment.
2. VAD to remove non-speech regions, followed by JFA modeling with session compensation.
3. Clustering of the speaker segment models using complete-linkage clustering until the CLR stopping threshold of $a_{ij} \leq 0$.
4. 3 iterations of Viterbi using 32 component GMMs to model final speaker/cluster, and a single Gaussian to model non-speech regions.

Table 1: *DER of baseline and proposed diarization systems.*

Diarization system	DER
Baseline	33.1%
Baseline + (1 iteration CLC)	**13.3%**
Baseline + (2 iterations CLC)	16.7%

4.1.2. Proposed diarization system and results

We evaluated our baseline diarization approach on the Australian broadcast news data, detailed in Section 3. The result of this evaluation can be seen in Table 1. It can be seen that our baseline diarization module is highly erroneous. We thus investigated the output of the baseline system to understand the underlying cause of the high DER obtained across the broadcast data. Through this investigation we found that our baseline system was under-clustering the speaker segments provided by the initial Viterbi segmentation and VAD stages. This may be addressed by knowing the desired number of output speakers, or by applying a different CLR stopping threshold (than 0) to the clustering process for each recording. However, this would mean having to abandon the convenience of employing a robust and theoretically ideal CLR threshold for any given recording. As our previous work on attribution [5], and particularly linking [8], had suggested that a CLR threshold value of 0 would serve as a robust stopping criterion, we concluded that the system was failing to robustly cluster speaker models as the initial segmentation did not provide sufficient data for modeled segments.

To overcome this, we propose using an additional pass of the complete-linkage clustering stage followed by Viterbi refinement. For convenience, we call the combination of these stages (steps 3 and 4 from Section 4.1) CLC, for complete-linkage clustering. We thus utilise the full baseline system to conduct a reliable initial segmentation of the recording, producing larger speaker homogeneous segments of data. We then apply a single iteration of CLC to the output of the baseline system. From Table 1 it can be seen that an absolute improvement of almost 20% is observed with respect to the DER measure.

This motivated our evaluation of another diarization system using the baseline system plus two additional passes of CLC. This system displayed a higher error rate than our proposed system using only one additional iteration of CLC. After observing the results, we found that a second additional iteration of CLC did not over-cluster the results, but it was rather the extra Viterbi refinement iterations that led to a higher DER measure, which reinforces our choice of the CLR stopping criterion of $a_{ij} \leq 0$. We thus propose employing our (baseline + CLC) diarization module for conducting robust speaker attribution.

4.2. Speaker attribution

In this section we employ our diarization system proposed in Section 4.1. As our previously proposed speaker linking system using complete-linkage clustering [5, 8], can be applied to this task without further modifications, we employ this linking module together with our proposed diarization method to carry out speaker attribution of the broadcast news data.

To conduct attribution, our proposed linking system obtains an initial set of (ideally) speaker homogeneous segments from the output of the diarization module across the collection of recordings. Each segment represents a unique speaker identity within its associated recording. These segments are then mod-

eled using JFA with session compensation, compared using the CLR metric and clustered using complete-linkage clustering.

We carried out the speaker attribution of the SAIVT-BNEWS data using our proposed multi-domain attribution system, which we will refer to as the **D-L** system, for diarization and linking. For evaluation purposes, we also carried out speaker attribution using reference diarization labels (DER = 0%) to initialise the speaker segment models in the linking phase of attribution. We did this for evaluation purposes and to reveal the potential of our attribution approach, should an ideal diarization module is used. To distinguish this system from our attribution approach, we will refer to this system as the **REF-L** system, for reference diarization and linking.

Figure 1 displays the AER of each system at all possible CLR threshold values. The horizontal axis has been reversed to display, from left to right, the clustering of the initial speakers/clusters into a single cluster. The oracle AER point of each system, obtained at its corresponding CLR threshold, has been marked on both the **D-L** and **REF-L** plots. It can be seen that as more speakers are correctly clustered a low AER region appears in the performance plot of each system. A lower valley, with respect to the vertical axis, indicates a higher accuracy associated with the analysed attribution system. In addition, the robustness of the systems is directly proportional to the width of the low AER region, and inversely proportional to the absolute value of the slope to the right of the oracle AER point, as marked on each plot. This slope is formed as each attribution system achieves its oracle AER point and then begins to attribute incorrect speaker identities to the already obtained clusters, creating a rise in the AER measure until all speakers are merged into a single cluster and maximum AER of the system achieved.

Table 2 displays the details associated with the oracle AER point of the two attribution systems. For reference, 92 unique speakers are present in the dataset, as detailed in Section 3. It can be seen that, as expected, the **REF-L** performs better than the **D-L** attribution system. This is also the case in Figure 1, which demonstrates that the **REF-L** system consistently performs better than the **D-L** attribution system. In addition, the CLR thresholds at which the oracle AER points of the two systems are achieved are both close to 0, thus further reinforcing the robustness of this CLR threshold as a stopping criterion to the task of clustering.

From Figure 1 and Table 2, it can be seen that the difference in the oracle AER of the two systems is almost equal to the DER displayed by our diarization module (Section 4.1). As the AER metric measures both the DER and the linking errors, and the fact that this difference in the oracle AER points of the two systems is almost equal to our achieved DER across the data, and as both systems achieve the same number of unique speaker identities across the dataset, it can be concluded that our linking module has been robust enough to deal with the erroneous diarization output. This suggests that any improvements to the DER achieved by our proposed diarization approach will directly apply to the AER obtained by our **D-L** system, potentially achieving a minimum AER of 17%, as obtained by our **REF-L** attribution system.

5. Discussion

Compared to our previous work on attribution of two speaker telephone-data [7, 5, 8], our multi-domain speaker attribution system proposed in this paper demonstrates similar results across the Australian broadcast news dataset. This is while our system remains largely unchanged, with the exception of the modification applied to the diarization module (Section 4.1) to accommodate an arbitrary number of speakers. Most importantly, as discussed in Section 4 and detailed in our previous work [7], our proposed multi-domain system employs a 512 component combined gender GMM-UBM, trained on telephone-data, for JFA modeling. This is indicative of the robustness of our attribution approach and suggests that our system may be improved even further through utilising a GMM-UBM trained on data from a broadcast news domain.

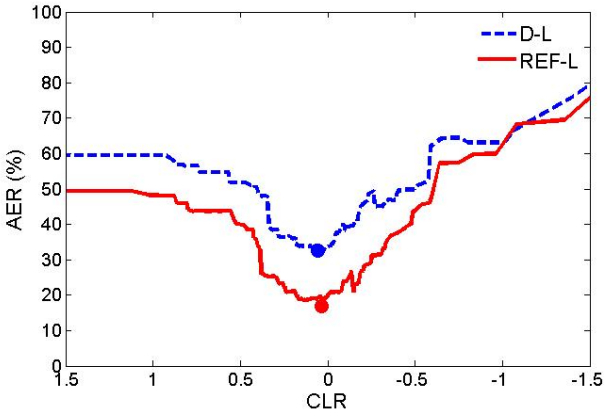


Figure 1: AER versus CLR for **REF-L** and **D-L** attribution.

Table 2: *Oracle attribution using* **REF-L** *and* **D-L** *systems.*

Attribution system	AER	Obtained speakers	CLR
REF-L	**17.0%**	77	0.03
D-L	32.6%	77	0.05

6. Conclusion

In this paper we proposed a robust and efficient speaker attribution approach, applicable to multiple audio domains, with the ability to conduct automatic diarization and attribution of multiple recordings, each containing speech from an arbitrary number of speakers. We did this by extending our previously proposed telephone-data speaker attribution approach. In this work, we proposed using a theoretically suitable CLR stopping threshold for complete-linkage clustering in diarization and linking. We demonstrated that, even in diarization where small segments are required to be clustered, this stopping threshold can be employed as a robust stopping criterion. Our work in this paper, and previous studies, suggests that this stopping threshold is robust across different audio domains when employed in the same manner as our multi-domain attribution approach. Finally, we demonstrated achievable AERs as low as 17%, across the broadcast news data, using our attribution system.

7. Acknowledgments

This paper was based on research conducted through the Australian Research Council (ARC) Linkage Grant No: LP0991238 and the follow-up applied research based on Australian broadcast data conducted through the Cooperative Research Centre for Smart Services.

8. References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435 –1447, may 2007.

[2] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *INTERSPEECH*, 2009, pp. 1559–1562.

[3] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 6, pp. 1059 –1070, 2010.

[4] C. Vaquero, A. Ortega, and E. Lleida, "Partitioning of two-speaker conversation datasets," in *Interspeech 2011*, August 28-31 2011, pp. 385–388.

[5] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, march 2012, pp. 4185 –4188.

[6] D. A. V. Leeuwen, "Speaker linking in large data sets," in *Odyssey2010, the Speaker Language and Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 202–208.

[7] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Extending the task of diarization to speaker attribution," in *Interspeech2011*, Florence, Italy, August 2011. [Online]. Available: http://eprints.qut.edu.au/43351/

[8] H. Ghaemmaghami, D. Dean, and S. Sridharan, "Speaker linking using complete-linkage clustering," in *to be presented in Australian International Conference on Speech Science and Technology (SST2012)*, 2012.

[9] M. Ferras and H. Bourlard, "Speaker diarization and linking of large corpora," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, Dec., pp. 280–285.

[10] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage speaker diarization of broadcast news," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1505 –1512, 2006.

[11] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*. Springer Berlin / Heidelberg, 2008.

[12] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, nov.-3 dec. 2003, pp. 411 – 416.

[13] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557 –1565, 2006.

[14] P. Kenny. "Joint factor analysis of speaker and session variability: Theory and algorithms". [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[15] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech 2008*, 2008, pp. 853–856.

[16] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, 2000, p. 2000.

[17] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 0, pp. 4057–4060, 2009.

[18] X. Anguera, S. Bozonnet, N. W. D. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech & Language Processing*, pp. 356–370, 2012.

[19] A. Jain, A. Topchy, M. Law, and J. Buhmann, "Landscape of clustering algorithms," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 1, 2004, pp. 260 – 263 Vol.1.

[20] (2007) The NIST rich transcription website. http://www.nist.gov/speech/tests/rt/.

[21] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, June 18-22 2001, pp. 213–218.

Semi-Supervised and Unsupervised Data Extraction Targeting Speakers: From Speaker Roles to Fame?

Carole Lailler, Grégor Dupuy, Mickael Rouvier, Sylvain Meignier

LUNAM Université, LIUM, Le Mans, France

first.lastname@lium.univ-lemans.fr

Abstract

Speaker identification is based on classification methods and acoustic models. Acoustic models are learned from audio data related to the speakers to be modeled. However, recording and annotating such data is time-consuming and labor-intensive. In this paper we propose to use data available on video-sharing websites like *YouTube* and *Dailymotion* to learn speaker-specific acoustic models. This process raises two questions: on the one hand, which are the speakers that can be identified through this kind of knowledge and, in the other hand, how to extract these data from such a noisy corpus that is the Web. Two approaches are considered in order to extract and to annotate the data: the first is semi-supervised and requires a human annotator to control the process, the second is totally unsupervised. Speakers models created from the proposed approaches were experimented on the REPERE 2012 TV shows test corpus. The identification results have been analyzed in terms of speaker roles and *fame*, which is a subjective concept introduced to estimate the ease to model speakers.

Index Terms: speaker identification, JFA, semi- and unsupervised speaker modeling, speaker roles, *fame*

1. Introduction

REPERE is a French evaluation campaign in the field of multimedia people in television documents. The main purpose of this challenge is to answer the questions "who is speaking ?" and "who is seen ?" at any time of the videos. The targets are both television professionals and guests, which can refer either to experts in a specific field, or to politicians, or celebrities. This paper is only concerned in the "who is speaking?" question. In this context, the identification task aims to determine the identity of the speakers, at any time.

The system presented in this paper is a two-levels architecture that uses both speaker diarization and speaker identification to process the shows. The speaker diarization level aims to partition the input audio stream into homogeneous segments, and group these segments according to the identity of the speakers. The purpose of the speaker identification level is to annotate the segments with the true identity of the speakers. However, available data in the training corpus are insufficient to learn specific and robust speaker models for each of the persons appearing in the videos: the coverage in the training corpus, in terms of number of speakers, is too low.

A solution to address the problem of insufficient coverage is to enhance the training corpus with data matching persons who are not already present. Nevertheless, the creation of such annotated corpora is time-consuming and labor-intensive. With the advent of video-sharing websites on the Internet, like *YouTube* and *Dailymotion*, it is now possible to collect innumerable data on speakers. The downside is that such data are noisy and poorly annotated, only the title and the description of the videos help to determine the topic: find satisfactory video content is not easy because either no information have been provided, or information are untrue, inaccurate or incomplete. In addition, the available videos are often of different qualities: some of them are professional videos while the others look like homemade movies shared by amateurs. Also, different recording situations (indoor or outdoor, with one person or with a group, ...) make data mining challenging: it is easier to exploit data from a politician show where a single man appears on the screen than data from a show where many speakers are interacting.

The use of Internet to build up corpora has lately been the subject of many research, especially in the field of speaker identification. In the video field, various works attempted to associate names to faces for a special type of web pictures. [1, 2, 3, 4, 5] focused on the face-name association in news photographs. [1] and [6] applied a face detector on the pictures and a named entity detector on the captions, then tried to find associations between detected names and faces. In the audio field, the main method focuses on learn a consistent association of speech and face from videos [7]. All the proposed approaches were focused on unsupervised methods applied to the identification of celebrities.

In this paper, two methods are proposed (semi-supervised and unsupervised) to build up specific speaker models from data from video-sharing websites and thus, to get round teh lack of data. The videos are retrieved using a list of speakers that may appear in the TV news. The semi-supervised approach needs a human annotator in order to control the automatic extraction of the data that are supposed match the targeted speaker. Thus, human interventions are greatly minimized. The unsupervised approach allows to automatically extract the data corresponding to the targeted speaker without any control from the human annotator. These methods are evaluated with the TV shows that compose the test corpus of the French evaluation campaign REPERE 2012. The evaluation focuses on the quality of the speaker models extracted from the data obtained through the semi- and unsupervised approaches. In addition, an analysis based on the subjective concept of people *fame* was conducted to understand relationship between speaker roles and identification results.

In the next section, we briefly describe the initial training and test corpora. Then, we present the semi- and unsupervised approaches used to model speakers using non-annotated data in section 3. The implementation of the two-levels architecture is described in section 4, and evaluation metrics as well as experiments results are given in section 5. In the section 6, the aim is to answer to the question: what is the nature of modeled people? by an analysis of the speaker role and *fame*. This section

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

is followed by some conclusions.

2. Corpus

This work in speaker identification was conducted as part of the REPERE 2012 evaluation campaign [8]. As such, experiments were performed on the test corpus of this evaluation campaign, composed of 3 hours of data. This data are drawn from 28 TV shows, recorded from French TV channels: BFM and LCP. The corpus is balanced between prepared speech, with 7 broadcast news from French radio stations, and spontaneous speech, with 21 political discussions or street interviews. Only a part of the recordings are annotated, giving a total duration of 3 hours.

The purpose of this work is to identify people who frequently appear in the news, so a list of 580 people was manually built with either people appearing in the media, or people likely to be present in the news, people who might appear. This list contains anchors, journalists, celebrities such as ministers, actors, singers, *etc.* 152 people from this list can be modeled using the training corpus. The training corpus is composed of every annotated data distributed during the French ESTER-2, ETAPE and REPERE evaluation campaigns. Among the 152 extracted models, 30.1% match people present in the test corpus.

Despite the amount of annotated data used as training corpus, 428 people from the list can not be modeled. External data is needed. Thus we propose to use data available on video-sharing websites like *YouTube* and *Dailymotion* to learn speaker-specific acoustic models.

3. Data extraction and speaker modeling

Video-sharing websites provides access to a considerable amount of data. However, data mining is challenging because of various factors: the quality of the media, the recording situation (indoor/outdoor, single speaker/group of speakers, *etc.*), the quality of annotations (inaccurate, incomplete or nonexistent). Building up a corpus is performed in two steps: data extraction then data annotation. The extraction is performed by retrieving videos on the video-sharing websites according to a request. The process is as follows:

1. *Request:* the request is composed of the name of the speaker to be modeled
2. *Filter:* all the videos in which the title do not include the name of the speaker to be modeled are put aside
3. *Download:* the first twenty videos are downloaded

Two different approaches are presented to annotate these data in terms of speaker identity, in order to learn speaker-specific acoustic models. The unsupervised method automatically takes the decisions. The semi-supervised method involves a human annotator to help the choices made by the system.

3.1. Unsupervised

The unsupervised approach aims for automatically select the segments that match the targeted speaker without any control from the human annotator. The main difficulty is to automatically detect if the person talking is the one sought. We made the assumption that the targeted speaker participate in each of the video extracted from the video-sharing websites. Indeed, this assumption has been validated on a portion of the training corpus listening to selected segments (the segments of less than 300 frames were not taken into account in this corpus).

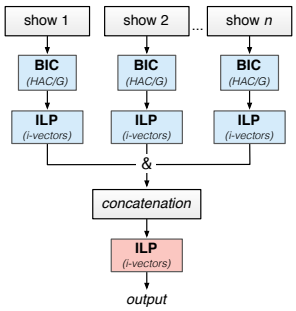


Figure 1: The audio cross-show speaker diarization architecture used to identify the cross-show speakers among the collection of videos.

Based on this assumption, an audio cross-show speaker diarization system is used to detect the speakers appearing across the multiple videos of the collection [9, 10, 11]. As illustrated in Figure 1, the cross-show diarization system first processes each video individually, by using a single-show speaker diarization system based on a Bayesian Information Criterion (BIC) segmentation followed by a clustering expressed as an Integer Linear Programming (ILP) problem. Then, the system attempts to identify speakers reappearing in several videos within the collection, by performing an overall ILP clustering [11].

After this cross-show diarization process, only the main cluster is considered. A filtering step is then performed to stop the process if not enough data are available to create the speaker model: if the speaker associated to the main cluster is appearing at least in three videos, and if the length of it interventions is longer than 2 minutes, then the acoustic model is created.

3.2. Semi-supervised

The aim of the semi-supervised method is to annotate the data extracted from the video-sharing websites while minimizing human efforts. We assume that the speaker to be modeled is present in each of the video collected, and that this speaker is the one who talk the most. An audio single-show speaker diarization system, as described in section 4.1, is run on each of the video. In order to correct the resulting clustering, that may not be perfect, a human annotator has to verify and invalidate the erroneous clusters. The purpose is to obtain the maximum number of segments that represent the targeted speaker, while maximizing the purity of the data by putting aside erroneous clusters. To minimize the human annotator effort, a validation of the audio segmentation according to the corresponding image is proposed: we have considered that the person appearing on the image is the one who is speaking because in the REPERE training corpus, the targeted speaker appears in about 80% of cases. The full process of the semi-supervised is as follow:

1. An audio speaker diarization system is run on each video,
2. Only the main cluster is considered (making the assumption that the main cluster match to the targeted speaker),
3. The images in the middle of each of the segment from the main cluster are extracted,
4. Human annotator verifies the speaker clustering by invalidating segments (so the picture) that do not contain the targeted speaker.

Figure 2: The application that help the annotator to invalidate audio segments according to the person appearing on the image.

An application has been developed to help the human annotator, by clicking on the images provided, to invalidate the segments that do not show the targeted speaker (Figure 2). More than 1900 hours of videos have been downloaded, and it took 224 hours to process the annotation. Finally, 480 hours of data have been annotated with the speaker identities. The ratio between the duration of the data to be annotated and the duration of the annotation itself is about 0.11. In [12], the manual annotation of a 2h08 corpus lasted 1h17, the ratio was about 0.60. Although it is less accurate, the duration of the annotation process, with the *semi-supervised* method is 6 times faster than a fully manual annotation.

4. Architecture of the identification system

In this paper we present a two-levels architecture that combines a speaker diarization system with a speaker identification system. The speaker diarization task aims to answer the question "who spoke, when ?", by partitioning an input audio stream into segments, and by clustering those segments according to the identity of the speakers. Experiments were carried out using the *LIUM_SpkDiarization* toolkit[1]. The speaker identification system consist in identifying each of the clusters with the real name of the speaker. This system is based on Joint Factor Analysis (JFA).

4.1. Speaker Diarization

The speaker diarization system is composed of an acoustic Bayesian Information Criterion (BIC) segmentation followed by a BIC hierarchical clustering using BIC both as similarity measure between speakers and as stop criterion for the merging process. Each speaker is modeled by a Gaussian distribution with a full covariance matrix. A Viterbi decoding is used to adjust the segment boundaries using Gaussian Mixture Models (GMMs) with 8 diagonal components, trained by Expectation-Maximization (EM) algorithm on the data of each speaker. Segmentation, clustering and decoding are performed using 12 MFCC+E, computed with a 10ms frame rate. Music and jingle regions are removed using Viterbi decoding with 8 one-state HMMs: 1 music model, 1 jingles model , 2 silence models (wide/narrow band), 1 narrow band speech model, and 3 wide band speech models (clean/over noise/over music). Each state is represented by a 64 diagonal GMM.

[1] http://www-lium.univ-lemans.fr/en/content/liumspkdiarization

In the previous steps, features were used unnormalized (to preserve information on the background environment). At this point, each speaker is not necessarily represented by a single cluster. The contribution of the background environment is removed through a feature normalization and the system then performs an ILP clustering dealing with i-vectors speaker models [13].

In order to identify the cross-show speakers in the *unsupervised* method, a final ILP clustering is performed from the concatenation of the single-show diarization outputs [11].

4.2. Speaker Identification

The speaker identification system aims to identify the real name of speaker for each cluster given by the speaker diarization system. The speaker identification system is based on the Joint Factor Analysis (JFA) framework [14, 15]. The purpose of JFA is to decompose the speaker-specific model into three different components: a speaker-session-independent component, a speaker-dependent component and a session-dependent component (each recording corresponding to one of these session). A supervector is defined as the concatenation of the GMM means components. Let D be the dimension of the feature space, the dimension of a supervector mean is $M.D$, where M is the number of components in the GMM. For a speaker s belonging in session h, the factor analysis model can be formulated as:

$$\mathbf{m}_{(h,s)} = \mathbf{m} + \mathbf{D}\mathbf{y}_s + \mathbf{U}\mathbf{x}_{(h,s)}, \quad (1)$$

where $\mathbf{m}_{(h,s)}$ is the session-speaker dependent supervector mean, $\mathbf{D}$ is $M.D \times M.D$ diagonal matrix, $\mathbf{y}_s$ the speaker vector (a $M.D$ vector), $\mathbf{U}$ is the session variability matrix of low rank R (a $M.D \times R$ matrix), and $\mathbf{x}_{(h,s)}$ are the channel factors, a R vector. All parameters of the JFA model are estimated by using the Maximum Likelihood criterion and the EM algorithm. Several sessions corresponding to each speaker have to be used for an accurate estimation of JFA parameters. 60-dimensional acoustic features were computed, with a 10ms frame rate. The features are composed of 19 MFCCs + log energy, and augmented by their first and second-order derivatives. The GMM-UBM is a gender- and channel-independent GMM composed of 1024 Gaussians. The dimension of R is 40.

5. Experiments

Experiments were performed on the test corpus of the REPERE 2012 evaluation campaign. This corpus is composed of 3 hours of data, drawn from 28 TV shows, recorded from French TV channels: BFM and LCP. The corpus is balanced between prepared speech, with 7 broadcast news from French radio stations, and spontaneous speech, with 21 political discussions or street interviews. Only a part of the recordings are annotated, giving a total duration of 3 hours.

5.1. Evaluation metrics

The Diarization Error Rate (DER) is the metric used to measure performance in the speaker diarization task. DER was introduced by the NIST as the fraction of speaking time which is not attributed to the correct speaker using the best match between references and hypothesis speaker labels.

The evaluation metric chosen to measure identification performance is the official REPERE Estimated Global Error Rate (EGER). This metric is defined as follow:

	Supervised	Supervised + Semi-supervised	Supervised + Unsupervised	Semi-supervised	Unsupervised
BFMStory	58.2%	**55.8%**	56.5%	90.3%	89.1%
CultureEtVous	56.1%	**53.7%**	**53.7%**	100.0%	100.0%
CaVousRegarde	62.4%	**56.4%**	**56.4%**	90.1%	90.1%
EntreLesLignes	**13.5%**	**13.5%**	**13.5%**	40.8%	63.5%
LCPInfo	52.7%	**50.5%**	51.1%	65.6%	89.1%
PileEtFace	51.2%	**22.3%**	42.1%	45.8%	66.1%
TopQuestions	35.3%	35.3%	**35.2%**	41.3%	53.2%
# of speaker models	152	410	397	377	343
% useful in the test corpus	30.1%	40.4%	39.7%	28.7%	23.9%
REPERE	46.5%	**41.9%**	44.2%	67.7%	77.2%

Table 1: EGER on the REPERE 2012 test corpus with the *semi-* and *unsupervised* methods, combined or not with the speaker models from the training corpus (*supervised* method). The number of speaker models extracted, as well as the coverage (% of speaker models really matching a speaker in the test corpus), are also presented.

$$EGER = \frac{\#fa + \#miss + \#conf}{\#total} \quad (2)$$

where $\#total$ is the number of person utterances to be detected, $\#conf$ the number of utterances wrongly identified, $\#miss$ the number of missed utterances and $\#fa$ the number of false alarms. Both DER and EGER are computed using the scoring tool developed by the LNE[2] as part of the ETAPE and the REPERE campaigns.

5.2. Speaker diarization results

Single-show Diarization Error Rates obtained on the REPERE 2012 test corpus are reported in Table 2. DER of each show was computed from the output of the first level of the architecture presented in Paragraph 4. The variability of the results directly depends on the type of video processed. The DER is approximately 7% on broadcast news videos (BFM story, LCP Info), 11% to 16% on political discussions videos, and 28% on people/entertainment videos (Culture Et Vous). This system obtained the best results during the ETAPE 2012 and REPERE 2013 evaluation campaigns [16].

	%Miss	%F.A.	%Sub.	%DER
BFMStory	0.48	1.45	5.91	7.86
CultureEtVous	4.21	2.99	21.74	28.95
CaVousRegarde	2.02	0.10	12.78	14.91
EntreLesLignes	0.00	0.46	11.23	11.70
LCPInfo	0.42	0.95	5.97	7.35
PileEtFace	0.04	0.39	16.27	16.71
TopQuestions	1.34	3.04	10.60	14.99
REPERE	0.95	1.41	9.92	12.30

Table 2: Single-show DER on the REPERE 2012 test corpus.

5.3. Speaker identification results

Estimated Global Error Rates obtained on the REPERE 2012 test corpus are presented in Table 1. The "supervised" column shows the results obtained with the 152 speaker models extracted from the training corpus. Other columns present results obtained with the *semi-* and *unsupervised* methods, combined or not with the speaker models from the training corpus. EGER of the *semi-* and *unsupervised* methods, when combined with speaker models from the *supervised* method, are 41.9% and 44.2%, respectively.

EGER obtained with both method is improved because of the increase of speaker models. The *supervised+semi-supervised* method gives the best results. The resulting speaker models are more robust because of the verification made by the human annotator. The *unsupervised* method (without the *supervised* data) gives a coverage of 23.9% (343 speaker models were automatically extracted), and a EGER of 77.2 %.

[2]The French National Laboratory of Metrology and Testing

6. Speaker roles influence

Two methods were proposed (*semi-supervised* and *unsupervised*) to increase the number of speaker models, or improve the existing models. This section present an analysis which focuses on the relationship between the speakers models and the role of the speakers.

6.1. Roles description

Five roles are described in the REPERE evaluation campaign which are commonly used in the literature [17, 18]. In this analysis, R4 and R5 have been merge because of their similarity.

- **R1**: The anchors. These speakers are characterized by their presence throughout the show, without discontinuity.
- **R2**: The journalists. They are TV professionals appearing one time or more during the show.
- **R3**: The reporters. Similar to the role R2, they are correspondents covering events outside the set of the show.
- **R4+R5**: The guests (R4). They are invited to interact with the actualities. They were asked for their knowledge or their *fame* to discuss under the guidance of the anchor. They are neither part of the organization committee, nor the leaders of debates. They can be present in different TV shows, especially during a highly publicized event. R5 role refers to everyone else that could appear, like interviewed people in a report.

6.2. Results and comments

Table 3 shows the EGER and the coverage (% of speaker models really matching a speaker in the test corpus) of each of the roles (R1, R2, R3 and R4+R5), for each of the systems that have been presented in paragraph 5.3. The column "Reference" only shows the role distribution of the manually built list of 580 speakers used to collect the videos from the video-sharing websites. For example, this list contains 90.9% of anchors (i.e. R1 role) who are present in the test corpus.

Regarding the *supervised* system, a EGER of 8.6% and 12.2% were obtained for the R1 and the R2 roles, respectively. These low error rates are essentially due to the presence of the R1 and R2 speakers both in the training and test corpora of

	Reference	Supervised	Supervised + Semi-supervised	Supervised + Unsupervised	Semi-supervised	Unsupervised
R1	*(90.9%)*	**8.6%** *(81.8%)*	9.5% *(81.8%)*	9.0% (81.8%)	83.8% *(18.1%)*	100.0% *(0.0%)*
R2	*(85.7%)*	**12.2%** *(85.7%)*	**12.2%** *(85.7%)*	**12.2%** *(85.7%)*	43.2% *(42.8%)*	65.6% *(28.5%)*
R3	*(50.0%)*	**43.4%** *(50.0%)*	**43.4%** *(50.0%)*	**43.4%** *(50.0%)*	100.0% *(0.0%)*	100.0% *(0.0%)*
R4+R5	*(35.9%)*	64.7% *(20.1%)*	**57.5%** *(32.4%)*	61.1% *(30.2%)*	63.0% *(31.5%)*	70.7% *(28.0%)*
# of speaker models	512	152	410	397	377	343

Table 3: EGER comparison between the roles (R1, R2, R3 and R4+R5) and the speaker models of each system in the REPERE 2012 test corpus. Values in parentheses indicate the number of speakers with the corresponding role divided by the number of speaker models in each system.

the REPERE campaign. R3 and R4+R5 EGER are 43.4% and 64.7%, respectively. These rates are consistent with the frequency of appearance of the corresponding speakers. Less the speaker takes part in the training corpus, more it is difficult to detect him in the test corpus. 81.8% of R1 speaker of the test corpus have a model (85.7% for R2 speaker). It is particularly true for the R4+R5 speakers, corresponding to the guests in a broad sense, 20.1% of those speakers have a model.

Supervised+semi-supervised and *supervised+unsupervised* methods allow to better detect the R4+R5 role. Compared to the *supervised* method, the EGER of the *supervised+semi-supervised* decrease from 64.7% to 57.5% (-7.2% absolute), and the EGER of the *supervised+unsupervised* decrease from 64.7% to 61.1% (-3.6% absolute). The difference between the two methods is explained by the fact that the *Supervised+semi-supervised* method have more data to learn models. Indeed, the models coming from the *Supervised* method is learned with more data. Moreover, 14 new speakers models are added.

We introduce the subjective notion of *fame* of a speaker. A speaker has a significant *fame* if his presence on TV is going to beyond the scope of a channel. The celebrities, politicians or artists are easily recognizable by their large representation in various shows: their interviews are widely diffused. Thus, they have a wide *fame*. On the other hand, people like TV professionals, only appearing in the TV channel they work for, have a limited *fame*. It is easy to find data on video-sharing websites for famous people; it is difficult for not famous ones.

Thus, in the list of 580 speakers we have build, 9.1% [3] of R1 role misses in order to obtain all the anchors, 14.3% of R2 role for the journalists, 50% of R3 role for reporters and 64.1% of R4+R5 roles for the guests. Experiments show that the *unsupervised* method does not help to identify a single anchor, and the *semi-supervised* method has only found 18.1% of them. The half of the speakers list of 580 speakers is labelled as R3 role, and none of the two methods helps to identify this category of people (0%). Conversely, R4+54 roles (guests) only represents 35.9% of the 580 speakers list. The *semi-supervised* method allows to successfully identify 31.5% of the guests, and the *unsupervised* method, in which speakers models are automatically created without any human supervision, is able to identify 28% of them.

Aside from the headliners, the broadcast news programs are uniquely composed of TV professionals who only officiate on that channel. The influence of these individuals is lower and it becomes more difficult to trace, because their name is often associated with a single channel or to a single show. They always appear in the same situation, and fulfill the same role each time. This set includes the R1, R2 and R3 roles.

However, differences can be identified within these three following roles:

- **R1**: They always appear with the same appearance and the same clothes, in the same context. They finally have a relative important *fame*: again, it is difficult to obtain relevant and reliable data to produce robust speaker models with the proposed method.
- **R2**: This role is ultimately less difficult to identify because journalists often have a presence on several channels, in different situations. This variability leads to a better identification. The fact that they appear on several channels increase their *fame*. For example, the Semi-supervised method has a recovery rate of 42.8%.
- **R3**: This role corresponds to individuals usually appearing outside, and very occasionally. That imply real difficulties to obtain relevant data. In addition, little information flows because the contexts in which they appear are usually very different. Moreover, these individuals often work in noisy environments which increase the difficulties to obtain reliable acoustic information.

[3] This percentage comes from table 3, it corresponds to 100%-90.9%, etc.

7. Conclusions

In this paper, the various approaches proposed help to quickly obtain data in order to produce models for speaker identification. Regardless the method used, people with an important *fame* like celebrities are easy to model because of the ease to find related data. However, the proposed methods do not provide sufficient data to model anchors or journalists (unless they have an activity outside the channel). The *semi-supervised* approach has obtained better results than the *unsupervised* approach. Speaker models produced are more robust because of the controls made by the human annotator.

Aside from celebrities which are true "headliners", the shows are composed of a group of TV professionals who officiate on the channel in question (sometimes exclusively). The *fame* of these persons is limited; it becomes difficult to find related data on video-sharing websites because their presence is unfrequent, and restricted to a particular situation.

8. References

[1] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth, "Whos in the picture," in *NIPS*, 2004.

[2] D. Ozkan and P. Duygulu, "A graph based approach for naming faces in news photos," in *CVPR (2)*, 2006, pp. 1477–1482.

[3] M. Guillaumin, T. Mensink, J. J. Verbeek, and C. Schmid, "Automatic face naming with caption-based supervision," in *CVPR*, 2008.

[4] P. T. Pham, M.-F. Moens, and T. Tuytelaars, "Cross-media alignment of names and faces," *IEEE Transactions on Multimedia*, vol. 12, no. 1, 2010.

[5] J. Luo, B. Caputo, and V. Ferrari, "Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation," in *NIPS*, 2009, pp. 1168–1176.

[6] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. G. Learned-Miller, and D. A. Forsyth, "Names and faces in the news," in *CVPR (2)*, 2004, pp. 848–854.

[7] M. E. Sargin, H. Aradhye, P. J. Moreno, and M. Zhao, "Audiovisual celebrity recognition in unconstrained web videos," in *ICASSP*, 2009, pp. 1977–1980.

[8] J. Kahn, O. Galibert, M. Carré, A. Giraudel, P. Joly, and L. Quintard, "The repere challenge: Finding people in a multimodal context," in *Odyssey 2012 - The Speaker and Language Recognition Workshop*, 2012.

[9] V.-A. Tran, V. B. Le, C. Barras, and L. Lamel, "Comparing multistage approaches for cross-show speaker diarization," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[10] Q. Yang, Q. Jin, and T. Schultz, "Investigation of cross-show speaker diarization," in *Proceedings of Interspeech*, Florence, Italy, 2011.

[11] G. Dupuy, M. Rouvier, S. Meignier, and Y. Estève, "I-vectors and ILP clustering adapted to cross-show speaker diarization," in *Proceedings of Interspeech*, Portland, Oregon (USA), 2012.

[12] Y. E. Thierry Bazillon and D. Luzzati, "Manual vs assisted transcription of prepared and spontaneous speech," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), may 2008.

[13] M. Rouvier and S. Meignier, "A global optimization framework for speaker diarization," in *Odyssey Workshop*, Singapore, 2012.

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

[15] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Proc. Interspeech*, 2007.

[16] O. Galibert and J. Kahn, "The first official repere evaluation," in *SLAM 2013*, France, Marseille, 2013.

[17] R. Barzilay, M. Collins, J. Hirschberg, and S. Wittaker, "The rules behind roles: Identifying speaker role in radio broadcasts," in *Proceedings of the National Conference on Artificial Intelligence*, 2000, pp. 679–684.

[18] T. Bazillon, B. Maza, M. Rouvier, F. Bechet, and A. Nasr, "Speaker role recognition using question detection and characterization," in *Interspeech*, 2011.

Towards a better integration of written names for unsupervised speakers identification in videos

Johann Poignant[1], *Hervé Bredin*[2]
Laurent Besacier[1], *Georges Quénot*[1], *Claude Barras*[2]

[1]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
[2]Univ Paris-Sud, LIMSI-CNRS, Spoken Language Processing Group, BP 133, 91403, Orsay, France
[1]first.lastname@imag.fr, [2]first.lastname@limsi.fr

Abstract

Existing methods for unsupervised identification of speakers in TV broadcast usually rely on the output of a speaker diarization module and try to name each cluster using names provided by another source of information: we call it "late naming". Hence, written names extracted from title blocks tend to lead to high precision identification, although they cannot correct errors made during the clustering step.

In this paper, we extend our previous "late naming" approach in two ways: "integrated naming" and "early naming". While "late naming" relies on a speaker diarization module optimized for speaker diarization, "integrated naming" jointly optimize speaker diarization and name propagation in terms of identification errors. "Early naming" modifies the speaker diarization module by adding constraints preventing two clusters with different written names to be merged together.

While "integrated naming" yields similar identification performance as "late naming" (with better precision), "early naming" improves over this baseline both in terms of identification error rate and stability of the clustering stopping criterion.

Index Terms: speaker identification, speaker diarization, written names, multimodal fusion, TV broadcast.

1. Introduction

Knowing "who said what" in broadcast TV programs is very useful to provide efficient information access to large video collections. Therefore, the identification of speakers is important for the search and browsing in this type of data. Conventional approaches are supervised with the use of voice biometric models. However, the use of biometric models faces two main problems: 1) manual annotations: generating biometric models is very costly because of the great number of recognizable persons in video collections; 2) lack of prior knowledge on persons appearing in videos (except for journalists and anchors): a very large amount of a priori trained speaker models (several hundreds or more) is needed for covering only a decent percentage of speakers in a show.

A solution to these problems is to use other information sources for naming speakers in a video. This is called unsupervised naming of speakers and most approaches for that can be decomposed into the three steps:

1. Speaker clustering (or diarization),
2. Extraction of hypothesis names from the video (or from the collection of videos),
3. Mapping (or association) between hypothesis names and speaker clusters.

Speaker diarization is the process of partitioning the audio stream into homogeneous clusters without prior knowledge on the speakers' voice. Each cluster must correspond to only one speaker and *vice versa*. Most systems use a bottom-up approach which tries to merge speech turns into clusters that are the purest as possible using a distance metric (with a distance-based criterion to stop the clustering).

Two modalities, intrinsic to the video, can provide the name of speakers in broadcast TV: pronounced names and names written on the screen (see figure 1). Most state-of-the-art approaches rely on pronounced names due to the poor quality of written names transcription observed in the past. Naming speakers with pronounced names has been proposed by *Canseco et al.* [1, 2] and *Charhad et al.* [3]. Manually-designed linguistic patterns indicate whether a name refers to the speaker of the current speech turn, the following or the previous one. *Tranter et al.* [4] learn these patterns as sequences of *n-grams*. *Mauclair et al.* [5] use semantic classification trees (SCT) to match names and speaker turns. *Estève et al.* [6] compare these two techniques. They conclude that SCTs are less sensitive to automatic speech transcriptions errors than sequences of *n-grams*. *Jousse et al.* [7] improved over the SCT baseline: first, each name is attached locally to a nearby speech turn; names are then propagated globally to speaker clusters. They also show a performance drop from 19.5% to 85% in speaker identification error rate when using automatic speech transcription instead of (perfect) manual transcriptions and named entities detection. More recently, we proposed three propagation methods to propagate written names to speaker clusters [8]. These unsupervised multi-modal methods yield much better performance than mono-modal ones. We also show that these methods lead to 98.9% accuracy with perfect speaker diarization.

(a) (b)

Figure 1: Example of written names on the screen

The use of automatically extracted pronounced names faces several challenges: (i) transcription errors; (ii) named entity detection errors (missing first/last name, false alarms, etc.); (iii)

This work was partly realized as part of the Quaero Program and the QCompere project, respectively funded by OSEO (French State agency for innovation) and ANR (French national research agency).

mapping errors (current, previous or next speech turn).

The use of automatically extracted written names faces similar difficulties: (a) transcription errors – though better video quality reduces these errors; (b) detection errors – fewer because each TV show uses specific spatial position for title blocks[1]; (c) mapping errors – though a name is usually written on the screen while the person is talking, yielding easier affiliation.

This paper addresses other errors that can impact results: the errors made during the clustering process (during speaker diarization). For instance, the incorrect merging of two clusters containing different speakers can severely impact the speaker naming performance. Tuning the stopping criterion for hierarchical clustering is important to avoid such a problem. In this paper, we rely on the hypothesis that the high precision of written names to identify the current speaker can help us improve the diarization process in order to avoid the problems mentioned earlier. We limit our study to the use of written names for unsupervised speaker identification in videos and propose an extension of [8]. In this previous work, we proposed three methods for "late naming" of speakers which are highly dependent on the quality of speaker diarization. In this article, we present two novel approaches to overcome this issue: "integrated naming" aims at better choosing the value of the stopping criterion in order to minimize the speaker identification error while "early naming" adds written names-driven constraints to speaker diarization.

The outline of the paper is as follows. Section 2 presents the experimental setup as well as the speaker diarization module and the written names extraction module used in our experiments. Then, we describe our speaker naming methods in Section 3. Section 4 presents our experiments. Finally, we conclude this work and give some perspectives.

2. Experimental setup

The REPERE [9] evaluation campaign phase 1 took place in January 2013. The main objective of this challenge is to answer the two following questions at any instant of the video: *"who is speaking?" "who is seen?"*. In this paper, we try to answer the first question in an unsupervised way.

2.1. REPERE Corpus

The dataset used in our experiments is extracted from a corpus created for the REPERE challenge [10], which addresses multimodal person identification in videos. Videos are recorded from seven different shows (including news and talk shows) broadcasted on two French TV channels. An overview of the data is presented in Table 1.

	Train	Test
Raw video	58h	15h
Annotated part	24h	3h
Number of annotated frames	8766	1229

Table 1: Train and test sets statistics

Though raw videos were provided to the participants (including the whole show, adverts and part of surrounding shows), only excerpts of the target shows were manually annotated for the evaluation.

[1]Title block: spatial position used in the TV show to write a name and introduce the corresponding person.

Our evaluation is performed on test set. It is important to note that, although the whole test set is processed, the performance is measured only on the annotated frames. Figure 2 shows some statistics of the test set (duration and number of videos) for each TV show available in the REPERE corpus.

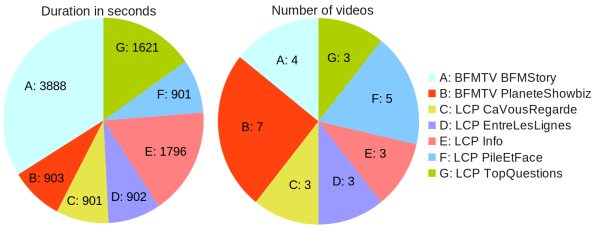


Figure 2: Duration and number of videos for the various TV shows available in the REPERE collection

2.2. Evaluation Metrics

Alongside the usual precision P and recall R, the official REPERE metric is also used for evaluation. It is called the Estimated Global Error Rate (EGER). This metric is defined as:

$$\text{EGER} = \frac{\#\text{fa} + \#\text{miss} + \#\text{conf}}{\#\text{total}}$$

where #total is the number of person utterances to be detected, #conf the number of utterances wrongly identified, #miss the number of missed utterances and #fa the number of false alarms.

To evaluate speaker diarization performance, we also used the diarization error rate (DER) defined by:

$$\text{DER} = \frac{d_{fa} + d_{miss} + d_{conf}}{d_{total}}$$

where d_{total} is the total speech time, d_{fa} the duration of false alarm, d_{miss} the duration of missed speech and d_{conf} the duration of the speech time where hypothesis and reference disagree. As identities of speakers are not considered, hypothesis and reference are aligned 1-to-1 to minimized d_{conf}.

2.3. Audio and Video Processing Modules

2.3.1. Speaker Diarization

Speaker diarization consists in segmenting the audio stream into speaker turns and tagging each turn with a label specific of the speaker. Given that no a priori knowledge of the speaker's voice is available in the unsupervised condition, only anonymous speaker labels can be provided at this stage.

After splitting the signal into acoustically homogeneous segments, we calculate a similarity score matrix between each pair of segments using the BIC criterion [11] with single full-covariance Gaussians. This similarity matrix is then given as input of a complete-link agglomerative clustering. Depending on the similarity threshold used as stopping criterion, several clustering results can be obtained.

It is worth mentioning that the matrix is not updated after each merging of clusters, as this is usually the case for regular BIC clustering.

We are aware that hierarchical clustering based on BIC distance is less efficient than hierarchical clustering with CLR distance [12] but our goal, here, is to do a fair comparison of several speaker naming methods, independently of the similarity measure (BIC or CLR).

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

2.3.2. Written names extraction

To detect the names written on the screen used to introduce a person, a detection and transcription system is needed. For this task we used LOOV [13] (LIG Overlaid OCR in Video). This system has been previously evaluated on another broadcast news corpus with low-resolution videos. We obtained a character error rate (CER) of 4.6% for any type of text and of 2.6% for names written on the screen to introduce a person.
From the transcriptions, we use a simple technique in order to detect the spatial positions of title blocks. This technique compares each transcript with a list of famous names (list extracted from Wikipedia, 175k names). Whenever a transcription corresponds to a famous name, we add its spatial position to a list. With the repeating positions in this list we find the spatial positions of title blocks used to introduce a person. However, these text boxes detected do not always contain a name. A simple filtering based on some linguistic rules allows us to filter false positives.

3. Unsupervised Naming of Speakers

We propose three methods for unsupervised (i.e. with no prior biometric models) naming of speakers with written names.

3.1. Late naming (`LN`)

Late naming is based on our previous work [8] (method M3). Speaker diarization and overlaid names recognition are run independently from each other. Speaker diarization is tuned to achieve the best diarization performance (i.e. minimize the diarization error rate, DER) as shown in Figure 3.

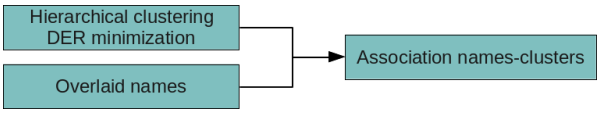


Figure 3: Late naming

The mapping between written names and speaker clusters is based on the following observations:

- when only one name is written on screen, any co-occurring speech turn is very likely (95% precision according to the train set) to be uttered by this person;
- the speaker diarization system can produce over-segmented speaker clusters, i.e. split speech turns from one speaker into two or more clusters.

Therefore, this method proceeds in two steps. First, speech turns with exactly one co-occurring name are tagged. Then, each remaining unnamed speech turn is tagged cluster-wise using the following criteria:

$$f(s) = \underset{n \in \mathcal{N}}{\operatorname{argmax}} \, \mathrm{TF}\,(s, n) \cdot \mathrm{IDF}\,(n)$$

where the *Term-Frequency Inverse Document Frequency* (TF-IDF)[14, 15] coefficient – made popular by the information retrieval research community – is adapted to our problem as follows:

$$\mathrm{TF}\,(s, n) = \frac{\text{duration of name } n \text{ in cluster } s}{\text{total duration of all names in cluster } s}$$

$$\mathrm{IDF}\,(n) = \frac{\text{\# speaker clusters}}{\text{\# speaker clusters co-occurring with } n}$$

In other words, speaker clusters are analogous to textual documents, whose words are detected written names.

Late naming is based on this method but there is a slight update that needs to be mentioned: we reduce the temporal scope of each written name to the more co-occurring speech turn, this can correct the time offset between audio and written names segmentation. It is important to note that the diarization can be different before and after the name-clusters association: some clusters may be merged (same name) or split (speech turn with a different name). Therefore, the scoring of the diarization can marginally change.

3.2. Integrated naming (`IN`)

One limitation of the late naming method is that the threshold used to stop hierarchical clustering is optimized in terms of diarization error rate (DER), while the ultimate objective is speaker identification, not diarization. Obviously, optimizing DER does not necessarily lead to the lower identification error rate (EGER). Therefore, "integrated naming" is a simple extension of "late naming" where the stopping criterion threshold is tuned in order to minimize the EGER. We will show later in the experiments that the resulting threshold is generally higher than the one selected to minimize DER (i.e. agglomerative clustering is stopped earlier)

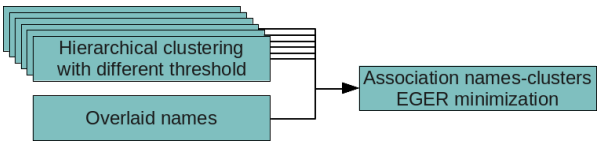


Figure 4: Integrated naming

In practice, as shown in Figure 4, we keep multiple clustering outputs, on which we apply the same method as in the "late naming" strategy described before. The threshold optimizing EGER on the training set is chosen.

3.3. Early naming (`EN`)

As already stated, when one or more names are written on the screen, there is a very high probability that the name of the current speaker corresponds to the written name on screen. Therefore, in "early naming", we use the information provided by written names during the clustering process.

Before clustering, we associate each written name n to the more co-occurring speech turns. At this stage, a speech turn can have several names if several names are written on the screen at the same time. Then, regular agglomerative clustering (based on speech turn similarity) is performed with the constraint that merging two clusters s without at least one name n in common is forbidden. For example, two clusters s_1 and s_2 **can** be merged into a new one s_{new} in the following case (the list of associated names is shown between brackets):

- $s_1(\emptyset) \cup s_2(\emptyset) \Rightarrow s_{new}(\emptyset)$
- $s_1(n_1) \cup s_2(\emptyset) \Rightarrow s_{new}(n_1)$
- $s_1(n_1, n_2) \cup s_2(\emptyset) \Rightarrow s_{new}(n_1, n_2)$
- $s_1(n_1, n_2) \cup s_2(n_1) \Rightarrow s_{new}(n_1)$

Below are examples where the two clusters **cannot** be merged:

- $s_1(n_1) \cup s_2(n_2) \Rightarrow$ Forbidden
- $s_1(n_1, n_3) \cup s_2(n_2) \Rightarrow$ Forbidden

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

The clustering is stopped according to the optimal (minimizing EGER) threshold learned on the training set.

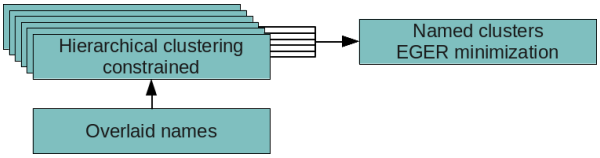


Figure 5: Early naming

4. Results

In this section we compare the ability of our naming methods to correctly identify speakers in TV broadcast and more particularly their sensitivity to the value of the stopping criterion threshold.

4.1. Learning the threshold as stop criterion

We used the training set to learn the stopping criterion threshold. However, in order to be less dependent on manual annotations, we did not use the whole 24 hours training set and selected 100 subsets randomly from it. These subsets were chosen to match the test set (duration, balance between shows, and number of videos for each show).

Naming strategy	median	min	max	standard deviation
LN: lower DER	1540	1440	1680	54
IN: lower EGER	1620	1520	1740	44
EN: lower EGER	1260	300	1640	277

Table 2: Threshold learned on 100 subsets of the train set, to minimize the DER or the EGER, LN: Late naming, IN: Integrate naming, EN: Early naming

As expected, Table 2 shows that the optimal threshold for IN is higher than those for LN. It means that IN stops earlier in the agglomerative clustering though split clusters may end up with the same name.

The constrained clustering of EN stops at a lower threshold. The standard deviation for EN is very high compared to the two others methods, it is possible to interpreted that EN is less sensitive to the threshold value. For the rest of the paper, we chose to use the median as global threshold.

4.2. Speaker Identification

For all the following experiences, it is important to note that the stopping criterion thresholds are learned on the training set while the results are displayed for the test set. Figure 6 shows the evolution of EGER with respect to the selected threshold and should be read from right to left as a smaller threshold value means that the agglomerative clustering stops later. LN and IN curves overlap but differ in the optimal stopping criterion threshold: threshold ⓐ aims at minimizing the DER (late naming) while ⓑ focuses on minimizing EGER (integrated naming). EN behaves very differently. ① shows the impact of the written name constraints and ⓒ the threshold learned to minimize the EGER.

Table 3 summarizes the performance of the three methods. The integrated naming has a lower EGER but the difference is very small, yet this method has better precision due to its higher threshold. As far as EN is concerned, the clustering constraint helps keeping the same precision (80.4%) though the threshold is lower. It allows to correctly merge some additional clusters and therefore increases the recall to 68.3%. For IN and EN, minimizing the EGER still allows to maximize other metrics like precision, providing at least enough speech duration to build speakers models.

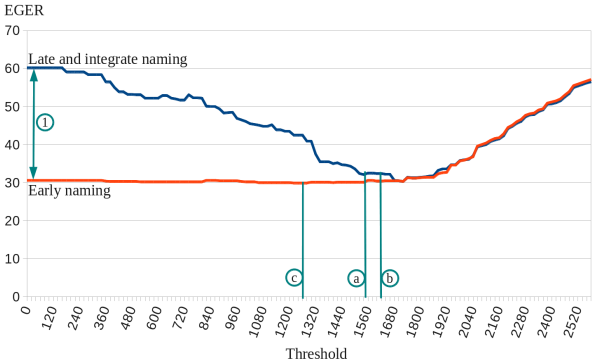


Figure 6: Influence of the stopping criterion threshold (ⓐ, ⓑ, ⓒ learned on train set) on identification error rate on test set, for the three naming strategies.

Naming strategy	Thr.	EGER (%)	P (%)	R (%)
Late (LN)	ⓐ 1540	32.1	80.4	66.0
Integrated (IN)	ⓑ 1620	32.4	81.5	65.3
Early (EN)	ⓒ 1260	29.9	80.4	68.3

Table 3: Trained stopping criterion threshold learned on the train set and the corresponding identification error (EGER), precision (P) and recall (R) obtained on test set.

4.3. Speaker Diarization

Figure 7 shows the evolution of DER as a function of the threshold. The baseline "before naming" corresponds to an audio-only diarization. As explained in section 3.2 the diarization is different before and after the late naming.

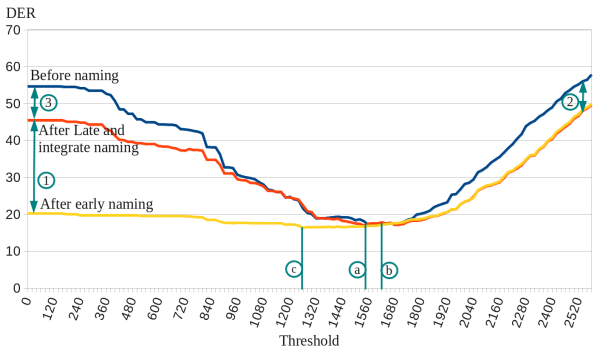


Figure 7: Influence of the stopping criterion threshold on diarization error rate on test set, before and after naming.

② and ③ show the influence of the direct speech turn tagging step. At the start of the clustering ②, this step merges

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

speech turns with the same name. At the end of the clustering ③, this step removes from clusters some speech turns with a different name. ① shows the effect of the constraints preventing clusters with different names from being merged.

ⓐ corresponds to the threshold tuned to minimize the DER. We obtain an 18.11% DER on the test set without written names (see Table 4). "Integrated naming" has a higher threshold but some clusters end up merged (thanks to their identical associated names), leading to a lower DER of 17.5%. The constrained clustering shows only a small variation of DER (from 18.7% to 20.2%, with a minimum of 16.37%) over the [0-1800] threshold range: it appears to be much less sensitive to the threshold choice (see figure 7).

	Thr	DER
Before naming	ⓐ 1540	18.11
After late and integrated naming	ⓑ 1620	17.51
After early naming	ⓒ 1260	16.37

Table 4: DER depending on the threshold

4.4. Sensitivity to the training set

Threshold tuning is achieved by randomly selecting 100 subsets from the training set and choosing the best threshold value for each of them.

The x-axis of Figure 8 summarizes the range of variation of this optimal threshold over the 100 training subsets (e.g. 1440 to 1680 for late naming strategy), as already introduced in Table 2. The y-axis reports the corresponding average identification error rate (EGER) and its standard deviation on the test set.

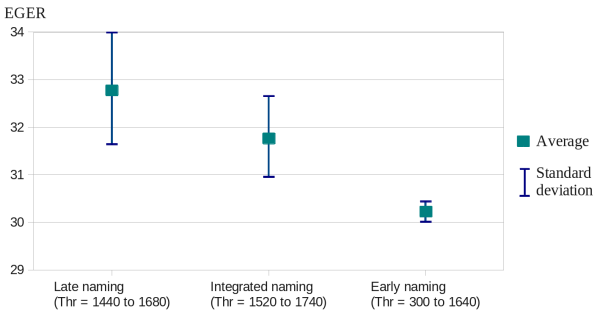


Figure 8: Average and standard deviation of the EGER on test set depending on the subsets used to learn the threshold

This figure points out that late and integrated naming strategies are more dependent on the training set and may therefore suffer from over fitting. Their respective identification error rates (EGER) has a standard deviation of 1.2% and 0.8%, while standard deviation of early naming EGER is only 0.2% (though the range of optimal thresholds over the 100 training subsets is much bigger).

4.5. Show-dependent threshold

The test corpus is composed of seven different types of shows (as illustrated in Figure 2). While a global show-independent threshold (*Thr. corpus*) can be trained, we also investigate the use of a show-dependent threshold (*Thr. per show*) and report the outcome of this experiment in Figure 9. *Thr. oracle* corresponds to the best possible performance in case an oracle is able to predict the best threshold. The robustness of a particular naming strategy can be inferred by the difference between the thresholds tuned on the whole training set (*Thr. corpus* and *Thr. per show*) and the optimal threshold (*Thr oracle*).

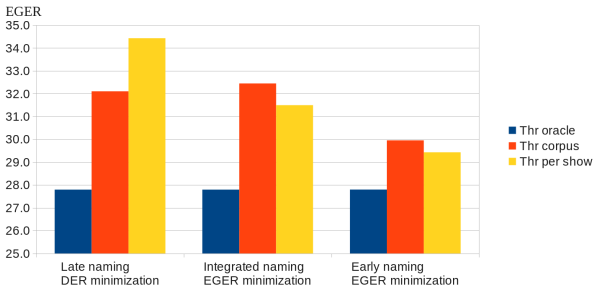


Figure 9: Identification error rate (EGER) for a show-dependent or show-independent stopping criterion.

Figure 9 shows that there is a difference of behavior between DER minimization (*late naming*) or EGER minimization (*integrated* or *early naming*). On one hand, DER minimization aims at associating one specific cluster to each speaker, whether they can be named or not. On the other hand, EGER minimization tries to associate its name to every speaker. Anonymous speakers can remain in the same cluster or split into several clusters as it has no influence on the final value of the identification error rate (EGER).

The REPERE corpus is composed of various types of shows. Some contains numerous speakers (up to 18 for news show *BFM Story*) whose names are usually displayed only once. Others, like the debate *Pile Et Face*, only have three speakers (two guests and the anchor) whose names are displayed 24 times on average over the duration of each show. For this particular type of show, the optimal DER threshold is 1300 while the EGER one is 1560. As a matter of fact, since speaker names are written multiple times, it is not worth trying to get exactly one cluster per speaker. A speaker cluster can be split into multiple smaller clusters as long as those clusters are named correctly.

Finally, we highlight that oracle results show almost identical performance for the three strategies. However, since early naming is less sensitive to the chosen threshold, it leads to much better identification performance (very close to the oracle one).

5. Conclusions

In this paper, we introduced and analyzed two naming strategies for unsupervised speaker identification in TV broadcast. *Integrated naming* is a simple extension of our previous work [8] that improves precision (+1.1%) while keeping the same identification error rate (32.4%). *Early naming* relies on the knowledge of overlaid names during the clustering process. This information is used to constrain clustering by preventing two clusters named by different written names from being merged. This method leads to better identification error rate (29.9%) and is less sensitive to the choice of the stopping criterion threshold. These two methods allow maximizing the metric associated to the target task. Future works will focus on the integration of additional sources of information like pronounced names or face clustering.

6. References

[1] Canseco-Rodriguez L., Lamel L., Gauvain J.-L., Speaker diarization from speech transcripts, *the 5th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2004, p. 1272-1275, Jeju Island, Korea.

[2] Canseco L., Lamel L., Gauvain J.-L., A Comparative Study Using Manual and Automatic Transcriptions for Diarization, *Workshop on Automatic Speech Recognition and Understanding (ASRU*, 2005, p. 415-419, Cancun, Mexico.

[3] Charhad M., Moraru D., Ayache S., Quénot G., Speaker Identity Indexing In Audio-Visual Documents, *Content-Based Multimedia Indexing (CBMI)*, 2005, Riga, Latvia, 2005

[4] Tranter S. E., Who Really Spoke When? Finding Speaker Turns and Identities in Broadcast News Audio, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, p. 1013-1016, Toulouse, France.

[5] Mauclair J., Meignier S., Estève Y., Speaker diarization : about whom the speaker is talking?, *IEEE Odyssey 2006 - The Speaker and Language Recognition Workshop*, 2006, p. 1-6, San Juan, Porto Rico.

[6] Estève Y., Meignier S., Deléglise P., Mauclair J., Extracting true speaker identities from transcriptions, *the 8th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2007, p. 2601-2604, Antwerp, Belgium.

[7] Jousse V., Petit-Renaud S., Meignier S., Estève Y., Jacquin C., Automatic named identification of speakers using diarization and ASR systems, *the 34th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2009, p. 4557-4560, Taipei, Taiwan.

[8] Poignant J., Bredin H., Le V.B., Besacier L., Barras C., Quénot G., Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast, *the 13rd Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2012, Portland, USA.

[9] Kahn J., Galibert O., Quintard L., Carr M., Giraudel A., Joly P., A presentation of the REPERE challenge, *Workshop on Content-Based Multimedia Indexing (CBMI)*, 2012, p 1-6, Annecy, France.

[10] Giraudel A., Carré M., Mapelli V., Kahn J., Galibert O., Quintard L., The REPERE Corpus : a Multimodal Corpus for Person Recognition, *the 8th International Conference on Language Resources and Evaluation (LREC)*, 2012, p. 1102-1107, Istanbul, Turkey.

[11] Chen S. S. and Gopalakrishnan P., Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion, in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, Virginia, USA.

[12] Barras C., Zhu X., Meignier S., Gauvain J.-L., Multi-Stage Speaker Diarization of Broadcast News, *IEEE Transactions on Audio, Speech and Language Processing*, 2006, vol. 14, no. 5, pp. 1505–1512.

[13] Poignant J., Besacier L., Quénot G., Thollard F., From Text Detection in Videos to Person Identification, IEEE *International Conference on Multimedia & Expo (ICME)*, 2012, p. 854-859, Melbourne, Australia.

[14] Robertson S. E., Jones K.S., Relevance weighting of search terms, Journal of the American Society for Information Science, 1976, p. 129-146.

[15] Fang H., Tao T., Zhai C., A formal study of information retrieval heuristics, the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, p. 49-56, Sheffield, UK

Narrative-driven Multimedia Tagging and Retrieval: Investigating Design and Practice for Speech-based Mobile Applications

Abhigyan Singh[1]*, Martha Larson*[2]

[1] Industrial Design, Delft University of Technology, Netherlands
[2] Electrical Engineering, Math and Computer Science, Delft University of Technology, Netherlands
a.singh@tudelft.nl, m.a.larson@tudelft.nl

Abstract

This paper presents a design concept for speech-based mobile applications that is based on the use of a narrative storyline. Its main contribution is to introduce the idea of conceptualizing speech-based mobile multimedia tagging and retrieval applications as a story that develops via interaction of the user with characters representing elements of the system. The aim of this paper is to encourage and support the research community to further explore and develop this concept into mature systems that allow for the accumulation and access of large quantities of speech-annotated images. We provide two resources intended to facilitate such work: First, we describe two applications, together referred as the 'Verbals Mobile System', that we have developed on the basis of this design concept, and implemented on Android platform 2.2 (API level 8) using Google's Speech Recognition service, Text-to-Speech Engine and Flickr API. The code for these applications has been made publically available to encourage further extension. Second, we distill our practical findings into a discussion of technology limitations and guidelines for the design of speech-based mobile applications, in an effort to support researchers to build on our work, while avoiding known pitfalls.

Index Terms: Mobile Speech Application, Multimedia Service, Narrative-driven design, Image Retrieval, Image Tagging, Android, Flickr

1. Introduction

Speech-based mobile applications, and specifically those that make use of a combination of speech and multimedia, are evolving rapidly. The data that such applications generate represent an interesting challenge for the research area of speech, language and audio in multimedia. However, speech has not yet established itself as a mainstream annotation modality for users who capture, save and share multimedia with their mobile phones. Instead, if users tag photos, they generally rely on text-based input forms. Before speech processing technology becomes truly relevant in the area of mobile multimedia, applications that allow users to use speech to tag and retrieve photos must establish their foothold. As long as use of such applications is not yet widespread, researchers will lack the critical mass of speech-annotated multimedia data that is necessary in order to address the question of how speech processing technology may support mobile speech-based image tagging and retrieval. This paper is motivated by the need for applications that break with existing conventions and practices in order to allow speech processing technology to realize its full potential to support users in capturing, sharing, viewing, and retrieving multimedia using mobile devices.

Upon first consideration, it appears nearly trivial to bring a photo-sharing website such as Flickr to a mobile phone by exploiting speech technology and using spoken tags and spoken queries to replace text-based user interaction. The fact that the telephone was originally developed as a device to support spoken communication, suggests, a priori, that voice input would be widely accepted by users as a mode for interacting with an image retrieval system. However, the idea that replacing text-tags with speech-tags is a simple switch soon reveals itself to be an overhasty assumption. Despite the research work that has been devoted to the topic of using speech to tag images, for example [1], [2], [3], there does not exist a single mobile application enabling speech-based tagging and retrieval of multimedia that has managed to enter mainstream use. In this paper, we take the standpoint that the challenges faced by such applications are not exclusively technical in nature, but rather have a critical dependency on the expectations and needs of users. The goal of this paper is to introduce a new design concept for mobile speech-based multimedia tagging and retrieval applications that tackles the challenges of handling the interaction between technological limitations and how users expect the system to work and what they want to achieve.

The key insight of our design concept is that the interaction of the user with the tagging and retrieval application can be conceptualized as a narrative. Within the storyline of this narrative different characters interact. These characters correspond to different elements of the system, and have personalities that reflect the function, speed and reliability of these elements. If the storyline is designed so that it is both engaging and easy to understand, the user will be able to formulate a useful model to understand the inner mechanics of the system that will promote acceptance and patience.

The rest of the paper is organized as follows. In Section 2, we discuss the preliminary investigation that led us to be concerned with the interaction of user expectations and technological limitations and to arrive at the concept of narrative-driven multimedia tagging and retrieval. In Section 3, we present background information on narrative structure and also our narrative-based design concept. Then, in Section 4, we describe our experiences in applying this design concept in the development of a particular mobile application for speech-based tagging and retrieval of images, called the 'Verbals Mobile System' (where 'verbal' refers to a voice tag). Finally, in Section 5, we distill our experiences into a series of lessons learned that will inform the development of other speech-based mobile applications for multimedia that apply the same design concept. We finish with a conclusion and outlook onto future work.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

2. Preliminary Investigation

The troublesomeness of speech recognition errors for systems that allow users to speech-tag their photos has been pointed out in the literature [2], [4]. The focus of our investigations was set specifically on speech-tagging in mobile environments, and we began with an informal study of existing mobile applications and a field study. During the field study, we tested mobile speech recognition technology in a series of real-world environments. Our investigation identified a set of key design challenges for the design and implementation of mobile 'voice' interfaces that connect with a remote image sharing server. The first set of challenges is technical in nature: Speech recognition in mobile environments (i.e., settings characterized by every-day noise conditions) falls far short of the recognition levels of human listeners in the same environments. Mobile environments are more challenging than other environments in which speech recognition technology is used because they are uncontrolled and highly variable (including background babble, birds, traffic, music, construction, wind). The strength of the signal received by a smartphone can vary with environment (e.g., at high-altitude, in forests, or in basements), meaning that connection with a remote image-sharing server is not stable. Mobile speech recognition technology that requires an Internet connection also suffers from delay and disconnection. The second set of challenges is related to users: The use of speech input is restricted by social and cultural norms that apply in both public and private spaces. Speech is for this reason not always acceptable in certain civic settings (e.g., in certain parts of hospitals or during a lecture/meeting) and may also be more disruptive than text input when a user is simultaneously engaged in social interactions with other people. Misrecognitions and delays caused by the technical challenges mentioned above can be a source of enormous frustration for users, causing them to experience the application as tedious to use or not worth the effort.

A critical characteristic of these design challenges is that the technical challenges are not separated from the user challenges, but rather the two sets of challenges are interdependent. Based on this observation, we concluded speech-based mobile applications must arise from a convergence of design and technology. We then set for ourselves the objective of developing a design concept that would accommodate both the needs of users and the technical restrictions of mobile speech, focusing on two main questions: (1) How to improve the tolerance of users for mobile multimedia applications which are considered technically to be 'not-perfect' or 'still-evolving'? (2) How to deal with possible errors, introduced by automatic speech recognition into tags and queries, in a manner acceptable to users?

As the basis for our design concept, we chose to conceptualize user interaction with the mobile application as a narrative. The aim of the narrative is to engage users and to provide useful explanations for the system's unexpected behavior. In the following section, we describe the 'Verbals Mobile System', which we developed as an application of this design to the task of image tagging and retrieval.

3. Narrative-Driven Mobile Speech Design Concept

Narrative structure could be understood as consisting of two parts: Content and Form. While content consists of story, i.e., characters, events, and conflicts, the form deals with plot, i.e., how the story is told or narrated. A narrative adds to the meaning generation process and engages an audience by facilitating interpretation of the story [5], [5], [6], [7]. Narrative structures are widely utilized in variety of media like newspapers, television advertisements, documentaries, films, and games to communicate with the audience [6], [7].

Upon first consideration, there are a seemingly endless number of stories, making it difficult to decide which storyline would be most appropriate for use in a mobile speech-based application. Ideally, the story chosen should be appropriate for any possible user of the system. Closer examination of literary and anthropological research reveals that narratives generally fall into broad patterns that share wide appeal. The folklorists and anthropologists that have analyzed and discussed narrative patterns in folklores and myths include Tvzentan Todorov, Vladimir Propp, Claude Levi-Strass, Roland Barthes and Joseph Campbell. Vladimir Propp's analyzed Russian folktales and discussed 31 narrative functions in his seminal book, 'Morphology of Folk tales', first published in 1928. He defined narrative functions as smallest possible unit of a narrative and described a certain order of appearance narrative functions in Russian folktales [8]. Propp's work has influenced work of many theoreticians and practitioners including Joseph Campbell, who analyzed myths in various cultures in his book 'The Hero with a Thousand Faces' [9], [10]. Campbell describes a pattern called Hero's Journey, which has been used in many Hollywood movies including Star Wars. It is the Hero's Journey narrative pattern that we chose to build upon in our design concept because it involves a single protagonist (who is represented in our concept by the user), overcoming unexpected obstacles (which are represented by unpredictable behavior of system elements such as speech recognition) in order to reach a goal (tagging or retrieval of images).

Various mediums offer varying benefits and limitations for incorporating narrative structures. The narratives could be adapted to these strengths and limitations [6], [7], [11]. For instance, a three-hours adventure film may have multiple subplots and numerous characters, whereas a thirty-second television commercial may have a single plot and a character. Narrative structure for mobile phones has not been explored much as yet. However, with their increasing ubiquity and technological advancement, mobile phones as a medium present an evolving and challenging platform for use of narrative structures [6], [7], [12]. Mobile phones facilitate various modes of interactions not possible in mediums like films or television. Mobile phones, like the World Wide Web itself, is an interactive medium rich with possibilities for multi-modal interactions (e.g., text, speech, touch) and the simultaneous use of multimodal features (e.g., audio, video, images). Further, the mobile phone is usually seen as a personalized object primarily used by a single user. This provides possibilities for a much more intimate interaction with the user. Due to these characteristics, mobile phones offer a rich platform for exploring narrative structures.

Our design concept enhances the tolerance of users for imperfect mobile phone technology by using a narrative in order to provide an explanation for unexpected behavior, such as disconnections, delays, and speech recognition errors. Because they are engaged with the narrative, users will also experience the time needed for processing and transmission to be less frustrating. In the next section, we apply our design concept to develop the 'Verbals Mobile System'.

4. Verbals Mobile System

This section first describes the narrative structure that was developed for the 'Verbals Mobile System', which we developed as an application of this design to the task of image tagging and retrieval.

4.1. Narrative Structure

As the basis for our design concept, we chose to conceptualize user interaction with the mobile application as a narrative, which we create by overlaying a dialogue with characters and a storyline. As previously mentioned, we build on the narrative pattern of 'Hero's Journey'. We brainstormed on various possibilities for a story and a plot and narrowed them down to the theme of travelling back in time in combination with a common human interest to view images of events in the past. As a result, the narrative of 'Verbals Mobile System' is based on the theme of 'Communication with the Past' and the genre of 'adventure'.

The narrative emerges via the interaction of a set of characters who pursue individual goals and each have their own personalities. Different realizations of the interaction are conceivable. For example, it would be possible for all characters to be explicitly instantiated by animation or voices seen and heard by the user and to interact with the user directly. We choose for a realization in which the user interacts with only a single, central character and that the other characters are not directly represented. Rather their actions are reported to the user by the central character. This choice simplifies the user interaction with the system and allows us to map the narrative onto a dialogue between two persons, the user, who is the protagonist of the narrative and the central character, who is directed by the user.

In the remainder of this section, we discuss the characters in our narrative (summarized in Table 1) and the mechanics of the plot. Note that the narrative-based designed concept that we proposed is not restricted to use of these characters or mechanisms. Rather, the specifics that we present here serve as an example of how the cast of characters can be established, and what the correspondence should be between the characters and the elements of the system.

Central character: 'Pica' is the central character and is a bird that can be directed by the users of our applications. We selected a bird since the human users of the applications could naturally associate birds with flying and with the activity of sending messages or communication with distant lands. The type of bird is a European Magpie, which was chosen because it is a common bird in the Netherlands and is also well known to for its habit of carrying shiny objects off and for its mischievous intelligence. We characterized 'Pica' as having a special ability to fly to 'The Past' and some ability to understand human voice. A user interacts with Pica and can direct her to travel back in time to access the human memories in the 'Past'. When users start the 'Verbals Push' application, they hear the voice of Pica saying, 'I am Pica, the magpie, I can carry your tags and images to the invisible Land of the Past. Select an images that you wish to send to the Garden of Human memories in the Past'.

Mentor: 'Google', is characterized as the omnipresent mentor of Pica, who helps her in understanding and interpreting human voice. A user can speak to Pica and send her on a journey to the 'Past'. The user can direct Pica to bring images from 'Past' and can also send images and some tags to 'Past'. The 'Past' appears in the narrative as a distant vast land holding shared human memories. Users all live in the 'Present' and any moment before the 'Present' is in the 'Past'. Please see Table-1 for the list of characters and their roles in the Verbals Mobile System.

Table 1. *List of characters and their roles.*

Character	Role in narrative	Role in the application
Pica [Protagonist]	Pica, a female magpie has the ability to fly back in time to reach 'The Past'. Pica could carry images and tags	Interaction with the user
Google [Mentor]	An omnipresent mentor, who helps Pica in understanding human voice	Accessing Google speech recognition service and obtaining results
Flickr [Guardian]	The guardian of an invisible valley in the distant clouds having an entrance to 'The Past'. To enter 'The Past' for the first time Pica has to provide a secret code	Fetch/Post images and tags from Flickr, Flickr account authentication
Zaat [Antagonist]	A wind demon in 'The Past' who shoots fast vertical winds to slow-down, disorient or snatch images from Pica in her journey	To deal with delay or error in image fetch/post
Bazooka [Antagonist]	A sonic demon that fires rapid noise beams to disrupt Pica's communication with Google. Bazooka does not like silence and human voice	To deal with incorrect speech recognition of a user's voice input

Quest: To enter the 'Past', Pica has to provide secret code to 'Flickr'. 'Flickr' is the guardian of an invisible valley in the distant clouds having an entrance to the 'Past'.

Conflict and challenges: Bazooka, a sonic demon does not like silence and human voice. Bazooka tries to attack with rapid noise beams whenever it gets to know of a communication between a human (user) and Pica. The noise beams could disrupt Pica's communication with 'Google' and hence lead to misinterpretations. So, the first challenge for a user is to successfully communicate with Pica so that Bazooka does not get to break the communication. Zaat, a wind demon in the 'Past' shoots fast vertical winds to slow-down, disorient or snatch images from Pica in her journey. So, the second challenge for a user is to succeed in either sending images and some tags to the 'Past' or retrieving images from the 'Past'.

Reward: Users who succeed in dealing with the conflicts and the challenges get to see ten images from the 'Past' as identified by their spoken tags (Verbals Pull) or their selected image and spoken tags successfully stored in 'The Past' (Verbals Push).

4.2. Architecture of the Verbals Mobile System

In this section, we discuss the Verbals Mobile System's architecture. The system consists of two separate applications: one for speech tagging (referred to as 'Verbals Push') and for speech search (referred to as 'Verbals Pull'). We have implemented the architecture as an Android phone application using Google Speech API, and Flickr API. The open source source-code of the Verbals Mobile System, including both the Verbal Push Application[1] and the Verbal Pull Application[2], is available on line. We have also made a demo video[3] available.

Verbals Push Application: The architecture of the Verbals Push application for speech tagging is depicted in Figure 1. The dialogue manager is the heart of the architecture. The user interacts (through the mobile phone's user interface) with the dialogue manager and the dialogue manager based on the current state of the application decides on subsequent actions and provides user feedbacks. For instance, when a user starts the application the dialogue manager introduces the application (including the character of 'Pica' as part of the narrative) and facilitates the user to select an image from phone's memory. Further, the dialogue manager uses Google Speech API to identify user's spoken-tags (or 'verbals'). Google Speech API sends user spoken-input to Google Speech servers and receives results. The speech recognition results are shown on phone's interface and dialogue manager requests for an implicit confirmation from the user. On receiving a 'go-ahead' from the user, dialogue manager connects with Flickr API, sends user authentication information and then pushes the selected image and the spoken-tag to the users Flickr account. Now, the user should see the image and the tags on his/her Flickr photostream. The dialogue manager uses the phone's text-to-speech engine to provide spoken feedback to the user and facilitates dialogue delivery for the narrative. The dialogues are implemented in an XML file format.

Verbals Pull Application: The architecture of the Verbals Pull application for speech-based image retrieval is depicted in Figure 2. The dialogue manager is, again, the heart of the architecture. The user interacts (through the mobile phone's user interface) with the dialogue manager and the dialogue manager based on the current state of the application decides on subsequent actions and provides user feedbacks. For instance, when a user starts the application the dialogue manager introduces the application (including the character of 'Pica' as part of the narrative) and encourages the user to speak a location. Further, the dialogue manager uses Google Speech API to identify user's spoken-tag (in this case a location name). Google Speech API sends user spoken-input to Google Speech servers and receives results. The speech recognition results are shown on phone's interface and dialogue manager requests for an implicit confirmation from the user. On receiving a 'go-ahead' from the user, dialogue manager connects with Flickr API, and then requests for ten random public images on Flickr that are tagged with the user's spoken-tag (a location name). The Flickr API returns the images and the images are shown on the phone's user interface. Again here the dialogue manger uses phone's text-to-speech engine to provide spoken feedback to the user and facilitates dialogue delivery as part of the narrative. The dialogues are implemented in an XML file format.

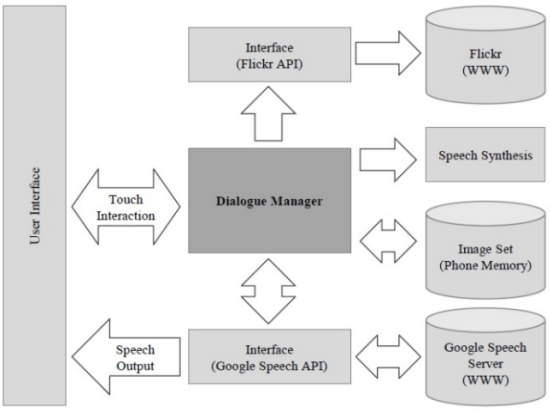


Figure 1: *Verbals Push Application.*

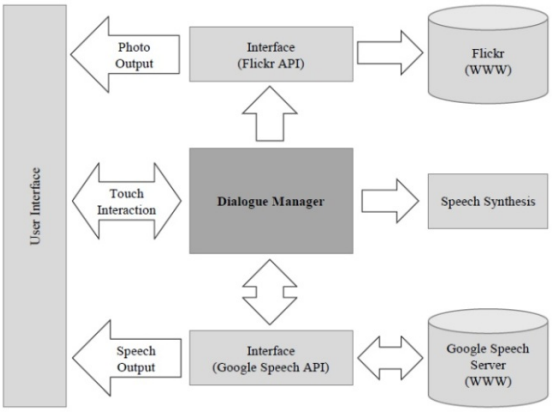


Figure 2: *Verbals Pull Application.*

5. Mobile speech in practice

In this section, we briefly present our general observations, first covering technical and then design aspects.

5.1. Technical Challenges and Limitations

Android platform: Before starting development it is important to realize that the Android platform based speech application development will require a real mobile phone in the development process as the Android emulator does not cater to voice input.

Dialogue manager: Implementing a dialogue manager is the key aspect for a narrative driven mobile speech applications. Android (2.2) framework requires a low-level implementation of dialogue manager. Many conventional server-side dialogue-driven applications like IVR systems use VoiceXML, which structures and simplifies the implementation of dialogues manager by facilitating a high-level implementation. However, VoiceXML (version 2.0 and 2.1) is primarily designed for server-side implementations and not suitable for client-side implementations as in case of Verbal's Mobile System. The upcoming version of VoiceXML may address this issue.

Mobile speech recognition: The Google Speech API, as of the time of writing, offers limited functionalities, a factor that

[1] https://github.com/abhigyan/Verbals-Push
[2] https://github.com/abhigyan/Verbals-Pull
[3] http://youtu.be/rnXtBsVgEII

constrains the features and interaction models that can be implemented in mobile speech application by application developers. For instance, it is not possible to record the user's voice input while using the Speech Recognition service or to retrieve the recorded audio-file from speech servers. A user click is needed to start the Speech Recognition as the mobile speech recognition cannot be in 'always on' mode. For this reason, design for speech-enabled application on mobile phone needs to balance 'speech' and 'touch' input. Google Speech Recognition's language options are evolving and at the time of writing there are limitations in vocabulary. For instance, during the time of implementation of Verbals Mobile System, Dutch Language and English (Dutch) Locale were not available. Implementing and designing mobile speech using Android (API 2.2) requires an Internet connection as the user's speech input is sent and results are received from Google's speech servers. This dependence on communication with speech servers brings possibilities of delays, errors and disconnections.

Mobile speech synthesis: The Android's default Text-to-Speech Engine (PICO) provides limited options. For instance, for speech synthesis only a 'female' voice option is available. Similarly, various speech synthesis audio-effects like echo, tempo, and accents are not yet available. This limits the use of narrative based on multiple characters as such a scenario would require variation in dialogue delivery in terms of pace, voice, pitch to depict contrast, variation in moods and personalities of the characters.

Multimedia sharing platform: Certain functionalities (like Image Post) using Flickr API requires user authentication. At the time of implementation of the Verbals Mobile System, the Flickr API was migrating to OAuth authentication scheme. Although this new mechanism is enhanced and more secure compared to the earlier version of authentication, it requires investing additional implementation time. Time and care is needed in selecting an Android-Java Flickr library. There are varieties of options available but many libraries offer incomplete functionalities and some are not well documented.

5.2. Guidelines for Mobile Speech Applications

Our guidelines for mobile speech application design emphasize user aspects, since we anticipate that specific aspects of the needs of users will endure, even after technical limitations have been addressed. First, we note that users needs in a certain sense run ahead of technology. Even as mobile phone coverage continues to improve globally, users will continue to find new places to capture and share multimedia (e.g., underground, under water). Multimedia content contains increasingly more information, for example, we can anticipate a move from images to video to 3D video. It is important not to assume that bandwidth is cheap, and its price might be the limitation on use for some users.

Second, we note that narrative-driven dialogue design has an enormous potential to engage users with speech based systems and improve their experience. However, a variety of factors shape the aesthetics of dialogues and narrative structure for that are appropriate for mobile platforms. Long and complex dialogues that maybe fine for interactive applications on the Web or for films, but these could create significant problems and reduce user experience when used in mobile speech applications. The aesthetics of vocal delivery is important: 'machine-like', 'young adult', 'feminine' and having 'native English accent' can all contribute to building the narrative. However, these can also have an impact on users' individual emotional response and must be taken into consideration. In order to target broad appeal, various popular genres of narrative like humor, horror or mystery could be leveraged.

6. Conclusions and Outlook

We have presented a narrative-based design that we have developed in order to reconcile technological limitations of mobile speech-based applications with user expectations. We start from the observation that the key to wider uptake of mobile speech-based applications for multimedia tagging and retrieval is not exclusively technical in nature. Rather, such a solution requires careful consideration of what users expect applications to do, and what they actually can do, given technological limitations.

We have presented an example narrative that builds on the 'Hero's Journey' narrative pattern. The individual characters involved in the narrative correspond to individual elements of the application, and are given personalities that reflect the unexpected behavior of these elements, so that they can serve to 'explain' to users why speech-based tagging and retrieval does not always precede along a smooth and predictable path, but rather encounters technical limitations.

We would like to note that we anticipate our approach will remain relevant, even as speech-based mobile technology continues to develop. A notion of "error free" speech recognition is difficult to formulate, since a speech recognition system that never makes an error must recognize speech better than a human being. However, even should "error free" speech recognition measured by any reasonable notion be achieved, tagging and retrieval systems still stand to benefit from collecting more and richer data from the users, in terms tags and queries that are more specific. Here, the engagement of a narrative-based system could help to extend users' patience and direct their formulation of tags and queries.

We would like to point out that smartphones offer possibilities for multi-modal interactions for narrative structure, going far beyond what we have covered here. Moving forward, we see opportunities for narrative-driven speech applications to be enhanced by integrating rich user context information (e.g., geo-location) and sensor information (e.g., from accelerometers) as narrative elements. We also are interested in the question of speech-based annotation and retrieval of video. The fact that video is a temporally continuous medium opens new challenges e.g., will users associate speech annotation with particular parts of the video, or with the whole video? The fact that videos may already contain speech might make it undesirable to add a second layer of spoken annotation. For example, users might be able to easily "speak over" existing speech to add or listen to annotations.

The number and variety of challenges left open by our work provides a rich field for future investigation, which must include design and implementation of mature mobile speech-based tagging and retrieval applications as well as testing and refinement. We hope that with this work, we have opened the door for such a future and have provided tools and resources that might prove useful to support it.

7. Acknowledgments

We thank the STITPRO Foundation http://www.stitpro.nl/ for their generous funding that supported this work.

8. References

[1] Chen, Jiayi, Tan, Tele, Mulhem, P., and Kankanhalli M. "An Improved Method for Image Retrieval Using Speech Annotation." In Proceedings of the International Conference on Multi-Media Modelling, 12-30, 2003.

[2] Hazen, T.J. Sherry, B. and Adler, M. "Speech-Based Annotation and Retrieval of Digital Photographs," In Proceedings of Interspeech 2007, 2165-2168, 2007.

[3] Kalashnikov, D.V., Mehrotra, S., Jie Xu, and Venkatasubramanian, N., "A Semantics-Based Approach for Speech Annotation of Images," IEEE Transactions on Knowledge and Data Engineering, 23(9):1373-1387, 2011.

[4] Rodden, K. and Wood, K.R. "How do people manage their digital photographs?" In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems *(CHI '03).* ACM, New York, NY, USA, 409-416, 2003.

[5] Wilson, K., "Narrative", Online: http://mediaknowall.com/as_alevel/alevkeyconcepts/alevelkeycon.php?pageID=narrative, accessed on 1 July 2013.Block, B., "The Visual Story: Creating the Visual Structure of Film, TV and Digital Media", Second edition, Focal Press, Oxford, UK, 2007.

[6] Miller, C. H., "Digital Storytelling: A creator's guide to interactive entertainment", Second edition, Focal Press, 2008.

[7] Alexander, B., "The New Digital Storytelling: Creating Narratives with New Media", Praeger, California, 2011.

[8] Propp, V., "Morphology of the Folktale", L. Scott [Translator], Second edition, University of Texas Press, Austin, Texas, 1968.

[9] Wilson, K., "Propp's Analysis of Folk Tales", Online: http://www.mediaknowall.com/as_alevel/alevkeyconcepts/alevelkeycon.php?pageID=propp, accessed on 1 July 2013.

[10] Campbell, J., "The Hero with a Thousand Faces", Third edition New World Library, California, 2008.

[11] Lindley, C. A., "Story and Narrative Structures in Computer Games", in B. Bushoff [Ed], Developing Interactive Narrative Content, München: High Text Verlag, 2005.

[12] Epstein, M. and Vergani, S., "History Unwired: Mobile Narrative in Historic Cities", AVI '06, Venezia, Italy, 302-305, 2006.

Multi-Modal Conversational Search and Browse

Larry Heck[1]*, Dilek Hakkani-Tür*[1]*, Madhu Chinthakunta*[1]*,Gokhan Tur*[1]
Rukmini Iyer[2]*, Partha Parthasarathy*[2]*, Lisa Stifelman*[2]*, Elizabeth Shriberg*[1]*, Ashley Fidler*[2]

[1]Microsoft Research, Mountain View, CA
[2]Microsoft, Sunnyvale, CA
larry.heck@ieee.org

Abstract

In this paper, we create an open-domain conversational system by combining the power of internet browser interfaces with multi-modal inputs and data mined from web search and browser logs. The work focuses on two novel components: (1) dynamic contextual adaptation of speech recognition and understanding models using visual context, and (2) fusion of users' speech and gesture inputs to understand their intents and associated arguments. The system was evaluated in a living room setup with live test subjects on a real-time implementation of the multimodal dialog system. Users interacted with a television browser using gestures and speech. Gestures were captured by Microsoft Kinect skeleton tracking and speech was recorded by a Kinect microphone array. Results show a 16% error rate reduction (ERR) for contextual ASR adaptation to clickable web page content, and 7-10% ERR when using gestures with speech. Analysis of the results suggest a strategy for selection of multimodal intent when users clearly and persistently indicate pointing intent (e.g., eye gaze), giving a 54.7% ERR over lexical features.

Index Terms: spoken dialog systems, spoken language understanding, multi-modal fusion, conversational search, conversational browsing.

1. Introduction

Spoken dialog (conversational) systems have seen considerable advancements over the past two decades [1]. A variety of practical goal-oriented conversational systems have been built and deployed. The goal of these systems is to automatically identify the intent of the user as expressed in natural language, extract associated arguments or slots, and take actions accordingly to satisfy the user's requests.

A major limitation of conversational systems is their narrow scope; conversational systems are constrained to operate over a small number of narrowly defined, known domains, with hand-crafted domain-dependent schemas (ontologies). As a result, there has been an increased level of interest by the research community to create open-domain conversational systems. These systems utilize very broad vocabularies, grammars, and intent models. However, the breadth of domain coverage comes at the cost of lower accuracy; without the constraints of limited tasks, speech-enabled systems are often unable to cope with the complexity open-domain speech recognition and understanding.

Advances in hand-held devices, touch displays and vision processing technology provide an opportunity for the speech community to increase the domain coverage of conversational systems. Rather than relying on spoken input only, systems can exploit the *visual constraints* introduced by touch, gesture, and eye gaze. For example, pointing gestures can be used to narrow the focus of attention to sub-region of the visual presentation, giving the conversational system useful priors on what to expect the user to say (e.g., selecting an item by pointing at it and saying "that one"). Since Bolt's seminal work on voice and gesture at the graphics interface [2], several studies investigated use of multi-modality for conversational interactions with a machine. Previous studies investigated the use of pointing gestures [3], touch gestures (including selection of items or an area on the screen for example with a remote control [4, 5], with finger [6] or with a pen [7]), and gaze and head-pose [8].

Another promising source of constraints for open-domain conversational systems is data from web search and internet browsers [9]. Web search engines and browsers are perhaps the most pervasive, ubiquitous open-domain tools available to people today to find information and complete transactions. In many ways, search and browse have elements of automated conversational interactions, or the "interactive, spontaneous communication between two or more [agents] who are following rules of etiquette" [10]. Search and browse conversations are interactive because the system responds to what has previously been communicated. The conversations are spontaneous because the user is not constrained by domain. Developers of search engines and browsers place considerable emphasis on the design of interactions. These interaction models in many ways are patterned after rules of etiquette of human-human conversations, with designs considering how to maximize information flow while minimizing unpleasant interruptions (e.g., relevance versus monetization).

Early work on leveraging search engines and browsers focused on exploiting offline information in the user logs: queries and corresponding clicks on links (documents) from search engines and browsers capturing interactions over many hundreds of millions of users and sessions. Work on exploiting the query-click graphs include [11–16]. More recent work has focused on human-computer addressee detection for conversational browsing [17], as well as methods to exploit the combination of search logs and semantic graphs [18–21].

In this paper, we create an open-domain conversational system by combining the power of internet browser interfaces with multi-modal inputs and data mined from web search and browser logs. We focus on two input modes, speech and gesture, and combine them to interact with browser and web page interfaces and page elements (e.g., links, drop-down menus, forms). By utilizing the pre-existing interaction mechanisms of web pages, we are able to by-pass the requirement to craft interactive user experiences for each domain of interest. In this way, the system inherits the open-domain designs and protocols of internet searching and browsing.

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

2. Conversational Scenario

In the conversational search and browse scenario, a user is free to navigate and interact with any page on the web through natural conversations with the machine. The user can speak with no constraints on their vocabulary, grammar, or choice of intent. As the user is browsing, they may choose to refer to content on the current page or not. Users may select links of the page contents in at least 3 ways:

1. **Explicit clicks:** User utterance refers to a link on the page, such as "show me Il Fornaio" or "Il Fornaio" in Figure 1. The utterance may be accompanied with hand gestures and eye focus.
2. **Location referrals:** User's utterance may include the relative position of the hyper-link on the page, such as "click on the top one". These may again be accompanied with gestures and eye focus.
3. **Gesture and speech:** Users may click on a link by gesturing in combination with speech, for example, pointing to the link and saying "that one", where the spoken utterance does not overlap with the anchor text of the hyper-link.

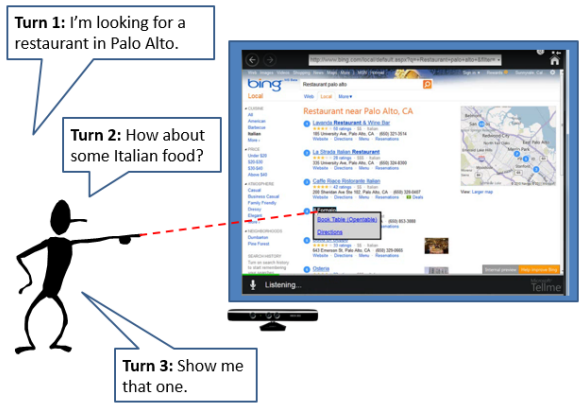


Figure 1: *Example multimodal (speech + gesture) scenario.*

Developing a system to enable the above scenario presents several technical challenges. First, the system must decide whether the user is referring to content on the current page or another page. If the user refers to the current page, the system must capture the intent: click, fill a form, scroll up/down, etc. In Figure 1, the user's intent in Turn 1 (from another web page) was to navigate to this current page by saying "I'm looking for a restaurant in Palo Alto". Turn 2 refines the content on the page to only show Italian restaurants. And finally, in Turn 3, the intent was to select the restaurant link they gestured towards. If the user had said "now show me what's playing at the closest theater", the system would need to recognize the shift in user intent/task as well as understand that the user is not referring to any content on the page, but rather wants to navigate to a movie theater listings web page.

3. Context Adaptation

A particularly effective method to reduce complexity of conversational systems is adapting to context. The context is in multiple forms. Some of the more common examples include:

- **Visual Context**: used to increase the prior likelihood the user will refer to entities/relations on the page
- **Dialog Context**: used for grounding, co-reference resolution, as well as potentially more complex inference and reasoning
- **Personal Context**: used to increase the prior of choices based on personal preferences from histories, geographically, etc.

In this paper, we leverage visual context. We use maximum-a-posteriori (MAP) unsupervised adaptation to adapt the statistical language model (SLM) of the speech recognizer to the content on the page [22]. The adaptation text can either be extracted from the page links (anchor text/titles) and/or landing page content. The extraction can be completed at either run-time or during an offline web page crawl procedure. For the example in Figure 1, the listed restaurant names, street names, food genre are all extracted from the scrape of the link/anchor text. The SLM probabilities and lexicons for names such as "Il Fornaio" can be increased to reflect the given visual context. The details of the restaurant found on the landing page of the link can be included in the adaptation data as well.

In addition to adapting the speech recognizer, the visual context can be used to adapt the semantic components of the system. For the example of Figure 1, priors of intents related to restaurants would be increased (reservations, reviews, etc.). Each of the links represents a new intent that can be dynamically added to the system.

An advantage of adapting to the page content at run-time versus crawl-time is the scalability of the solution: the system is always "fresh" and able to support conversational interaction on a page even if its content has recently changed. This is particularly important for dynamic pages (restaurant/movie reviews, breaking news, sporting results).

By following the above procedure to dynamically adapt to the visual context, the system in effect scales to the breadth of the web. By adjusting priors based on the visual content of the page, as well as related/connected pages (landing pages), the system can achieve this scale *robustly*, as will be demonstrated in the experiments of Section 5.

4. Multimodal Click Intent Detection

In addition to expressing intent verbally, a user may find it more natural in certain situations to express their intent visually. The simultaneous combination of two modes of intent expression is referred to as multimodal intent. This paper focuses on the combination of speech and hand gesture. Specifically, we study the effect of speaking while pointing at an object, such as a link on a web page. The scenario in Figure1 shows a user pointing at a restaurant link and saying "Show me that one." Multimodal interactions such as these are powerful, saving time by reducing dialog turns as well as intent/speech recognition errors.

In the following, we discuss each mode of intent capture separately, and then how they are combined. Then we show experimental results that illustrate the power of the resulting multimodal user interface.

4.1. Lexical Intent

Given the dynamic nature of web pages, we seek an effective lexical intent similarity measure that can be implemented without the requirement for supervised training. For this purpose, we utilize the well known *term frequency-inverse document*

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

frequency (TF-IDF) similarity measure from web search relevance [23].

For our purposes, we treat the k-th actionable element on the web page, p_k, (e.g., link, drop-down menu, form) as a document. We will refer to the user's utterance as a query, q. The TF-IDF similarity between the query, q, and the page element, p_k, is given as

$$\text{TF-IDF}(q, p_k) = \sum_{t \in q} \text{tf-idf}_{t,p_k} \quad (1)$$

where t denotes each term (word) in the query, TF is the number of occurrences of the term in p_k, and IDF is the log inverse of the number of page elements that contain the term t. The IDF factor is especially important for our task, since many terms on a given page will have little or no semantic discriminating power. For example, anchor text from links on a restaurant web page are likely to have the term *restaurant* in almost every link.

4.2. Gesture Intent

As with lexical intent, we seek a measure to capture the simultaneous voice and visual gesture intent of the user. For both speech recognition and hand movement detection, we use the popular and low-cost sensor Kinect. Kinect is a microphone array and a skeletal tracking motion sensing input device by Microsoft for the Xbox video game console and Windows PCs. The sensor has adequate resolution and software to accurately track hand movements. However, additional processing is required to discriminate intentional hand gestures such as pointing from unintentional hand movements.

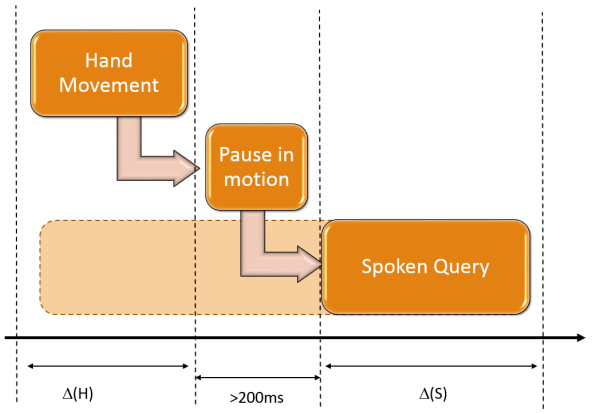


Figure 2: *Pointing gesture intent model*

For this work, we employ a simple model of pointing intent; a sequence starting with the hand motion, a brief pause with the hand still, followed by a spoken query. Figure 2 represents this model. Typically (as with Kinect) the hand gesture controls a cursor. The simplest method to determine the intended object selection is to compute the shortest straight line distance from the cursor to the (bounding box around) the page element. To decrease the chance of false positives, a *gesture focus region* may also be used. Gesture focus regions can be implemented with a weighting function, typically based on the inverse distance of the cursor to the object. Figure 3 shows the family of exponential inverse distance weighting (IDW) functions used in the experiments

$$\text{Gesture Score} = \exp\left(\frac{-|d|^a}{10^b}\right). \quad (2)$$

The IDW is used to specify the region of focus around the gesture's cursor. The goal of the IDW is to help balance the precision-recall of the gesture detection: a narrower region around the cursor (e.g., $a = 2, b = 0$) decreases the false alarms by reducing the affect of nearby objects, while a wider region ($a = 1, b = 1$) decreases the chance of incorrectly missing the intended object. The distance is measured in pixels from the gesture cursor to the object on the screen (e.g., web link, drop-down menu, form region). The IDW functions map the distances to $[0, 1]$, with a distance of 0 (the user is pointing directly at the object of interest) mapped to the maximum weight of 1.

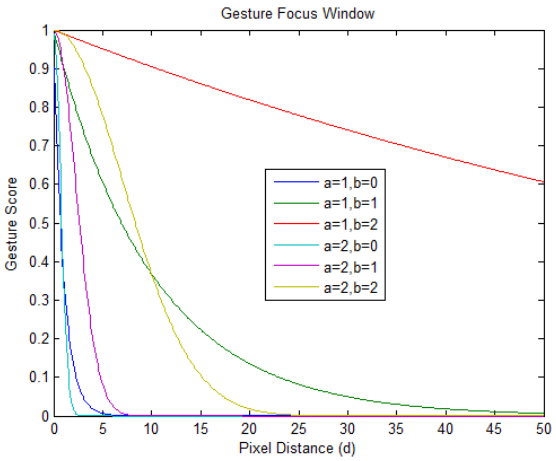


Figure 3: *Gesture focus windows around the cursor.*

4.3. Combining Intents

To form a single multimodal score for the kth page element, S_{M_k}, we use linear interpolation to combine the lexical score and the gesture score

$$\begin{aligned} S_{M_k} &= (1-\alpha)\ \text{TF-IDF}(q, p_k) + \alpha \cdot \text{Gesture Score} \\ &= (1-\alpha) \sum_{t \in q} \text{tf-idf}_{t,p_k} + \alpha \cdot \exp\left(\frac{-|d|^a}{10^b}\right) \quad (3) \end{aligned}$$

The values for α, a, and b are determined experimentally. Once the multimodal intent score is computed, it can be used for detection of intent by thresholding the score

$$\Lambda(S_{M_k}) \underset{reject}{\overset{accept}{\gtrless}} \theta. \quad (4)$$

The threshold θ can be optimized to minimize the cost of the error types: *false accept*, where the system incorrectly detected the presence of a user intent, and *miss*, where the system failed to detect the intent of the user.

5. Experiments and Results

5.1. Data Sets

We collected two data sets from 8 speakers during 25 sessions (set 1) and 7 speakers during 14 sessions (set 2). Both collections were performed in a living room set up, where users were seated on a couch approximately 5-6 feet away from a television screen. At the beginning of their first session, users were shown a short tutorial video demonstrating how the system can

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

be used, and were asked to improvise open-domain usage scenarios. In the first collection, the tutorial video included example usage scenarios with explicit (voice) clicks (as defined in Section 2) whereas in the second collection, the tutorial video included examples of using gesture with speech to click on a link. In both collections, users searched and browsed the web over open-domain tasks (e.g., shopping).

The total number of user turns in the first collection is 2,868, and 917 (31.9%) of these have a click intent. The second set includes 1,101 user turns, and 284 (25.8%) of these have a click intent. For the second set, we also computed the number of different types of clicks: 87.3% of the clicks are explicit clicks, 1.1% are location referrals and 11.6% include combined gesture and speech.

While hand pointing gestures were used for "click" intents, the collected data also includes cases of false gestures and false alarms by the system, such as a user lifting their arm to reach something on the coffee table. Hence, we further analyzed the usage of multi-modal input on a subset of the second collection. First, we separated out all utterances that control the display, such as "scroll down", as these can be captured with a high precision using the user's spoken utterances. Table 1 summarizes this analysis. 558 user turns are split into two: ones that are accompanied with a hand gesture and no gesture. The intent of these turns are categorized into click intents and non-click intents. In this analysis, we merged the gesture and speech clicks with location clicks as the second group is very infrequent, and named them "Click other". In this data subset, 22.8% of user utterances did not have a click intent, and yet a gesture was captured falsely. Similarly, 18.3% of the click utterances (excluding the explicit clicks) did not include a pointing gesture.

Table 1: *Statistics of user turns with/without hand gestures.*

	Gesture Found	No Gesture Found	TOTAL
Click "that one"	15 (2.7%)	1 (0.1%)	16 (2.8%)
Click other	25 (4.5%)	102 (18.3%)	127 (22.8%)
Non-Click	127 (22.8%)	288 (51.6%)	415 (74.4%)
TOTAL	167 (30.0%)	391 (70.0%)	558

5.2. Results

To examine the effectiveness of contextual adaptation for ASR, we used 2,868 utterances (9,346 words) from the first collection and completed tests on statistical language model (SLM) adaptation. While the average utterance length in this set looks short (3.3 words), this is mainly because this set contains all user turns in a session, including commands to change the display, which are usually 1-2 words (such as "scroll down" and "back"). About 40% of the utterances are such commands, 32% are click utterances, and 28% are the rest.

Table 2 shows the word error rate (WER) results from these experiments, where we compare a generic large vocabulary 400K word conversational speech recognition language model (LVCSR-LM) with its dynamically adapted version. The table also includes an analysis of impact on performance of click (32%) and non-click (68%) utterance subsets. Overall, with an out-of-vocabulary (OOV) rate of 0.25% and adapting the language models to the visual context improved the WER of the LVCSR-LM from 20.6% to 19.2%. WER for the context-related click utterance subset improved from 28.2% to 23.7% (a relative improvement of 16%), without a degradation on the performance of the rest of the turns. The small improvement (from 17.4% to 17.1%) on the non-click turns can be partially due to domain adaptation as a side effect of adapting to the visual content.

Table 2: *ASR WER with contextual adaptation.*

LM	WER overall	WER Click subset	WER Non-Click subset
LVCSR-LM	20.6	28.2	17.4
LVCSR-LM + adaptation	19.2	23.7	17.1

To study the effects of the gesture intent signal *independent* of how often it is used and the quality of the gesture detector, we complete simulations where all components/measures are real *except* the gesture. Table 3 shows results on a held-out random sample of 75% of the turns in data sets 1 and 2. The table shows the probability of the error types (false accept and miss) using the multimodal score and the intent detector of Equations 3-4. Results are computed for both manual transcription of the speech and automatic speech recognition using the contextual adaptation. The parameters of the detector are varied to show the affects of the size and shape of the gesture focus window (a and b) and the interpolation weight (α) between lexical and gesture-based intent. Since we normalized the scores of the lexical and gesture intent detectors to be [0-1], α can be interpreted as the relative importance of the gesture score in the combination.

For these experiments, we also simulated the human user's gesture intent to control for gesture precision. The simulation places the gesture cursor on an equidistant curve from the intended page element (link). The gesture precision distance, R, is the number of pixels that the cursor is away from the desired object (e.g., web link) and the page. We simulated gestures for two different gesture precisions: $R = 0$ and $R = 20$ pixels. The probability of missing the multimodal intent, P_{miss}, is computed in the operating region where the probability of falsely detecting an intent is low ($P_{fa} = 1\%$). We focus on this operating region due to the sensitivity of users to false positives and the objectionable user experience of the system incorrectly taking actions (clicking).

The best performing multimodal intent detector uses a balanced blend of lexical and gesture ($\alpha = 0.5$) and a broad gesture focus window ($a = 1$,$b = 2$). At these settings, with perfect speech recognition, perfect gesture precision ($R = 0$), and the user gesturing towards the intended page element (link) for 100% of the trials, the P_{miss}(@P_{fa}=1%) $= 8.1\%$. This represents an upper bound on the performance and is a 68.2% error rate reduction (ERR) compared to the single mode lexical intent detector ('No Gesture") with P_{miss}(@P_{fa}=1%) $= 25.5\%$. With the same settings for the gesture focus window, with automatic speech recognition (ASR), and with a gesture precision of $R = 20$, the P_{miss}(@P_{fa}=1%) $= 16.9\%$. This is a 50.3% error rate reduction (ERR) over lexical intent alone.

Using the same development data set used to compute the results in Table 3, we conducted experiments with real human gestures and an automated gesture detection model. For this test, we extracted gesture positions from the data logs that were generated using the model shown in Figure 2. We examined two cases: (1) all logged gestures and (2) only gestures where the user also said "that one". The first case includes all cases where a gesture was detected, which is approximately 30.0% of the test cases (see Table 1). The second case was used to isolate human gesture precision from the errors introduced by the gesture detection model.

Table 3: *Summary of multi-modal intent detection with simulated gestures.*

				$R = 0$		$R = 20$	
	IDW a	IDW b	Gesture α	$P_{miss}@P_{fa}$=1% Manual	$P_{miss}@P_{fa}$=1% ASR	$P_{miss}@P_{fa}$=1% Manual	$P_{miss}@P_{fa}$=1% ASR
No Gesture	-	-	-	-	-	25.5%	34.0%
Gesture	1	0	0.25	10.2%	21.1%	25.5%	34.0%
	1	0	0.50	9.5%	19.4%	25.5%	34.0%
	1	0	0.75	10.0%	19.9%	25.5%	34.0%
	1	1	0.25	10.2%	21.1%	21.5%	31.5%
	1	1	0.50	8.3%	18.3%	16.9%	27.1%
	1	1	0.75	8.6%	18.5%	8.3%	20.1%
	1	2	0.25	10.9%	21.3%	12.3%	23.1%
	1	**2**	**0.50**	**8.1%**	**18.3%**	**7.2%**	**16.9%**
	1	2	0.75	8.3%	18.3%	7.6%	17.1%
	2	0	0.25	10.4%	21.3%	25.5%	34.0%
	2	0	0.50	9.5%	19.4%	25.5%	34.0%
	2	0	0.75	10.0%	19.9%	25.5%	34.0%
	2	1	0.25	10.2%	21.1%	25.5%	34.0%
	2	1	0.50	8.3%	18.3%	25.5%	34.0%
	2	1	0.75	10.0%	19.9%	25.5%	34.0%
	2	2	0.25	10.2%	21.1%	25.0%	33.3%
	2	2	0.50	8.3%	18.3%	23.4%	32.2%
	2	2	0.75	8.3%	18.3%	20.6%	30.8%

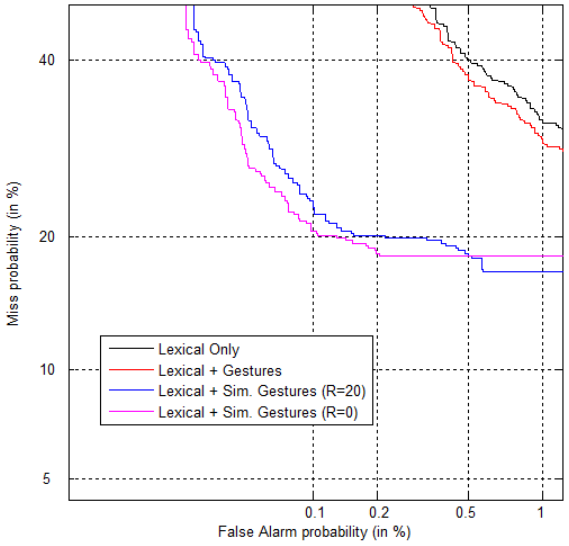


Figure 4: *DET results for multi-modal intent detection*

The results are shown in Table 4. With the introduction of errors due to both human gesture precision and the gesture detection model, the performance over all the trials was $P_{miss}(@P_{fa}\text{=}1\%) = 22.9\%$ (ERR=10.2%) and 31.7% (ERR=6.8%) for manual transcriptions and ASR, respectively. For the case where the user clearly indicated the pointing intention with the phrase "that one" while gesturing, the $P_{miss}(@P_{fa}\text{=}1\%) = 15.4\%$ for both manual and ASR (perfect recognition of the phrase), which is ERR=39.6% and ERR=54.7%, respectively. For this second case, we computed the average gesture precision for humans. Referring to Table 3, the gesture precision (R) for humans was is in the range of 16.4 to 28.6 pixels, depending on the density of the visual content on the screen. In other words, humans are able to precisely gesture towards the intended element. The drop in performance, therefore, was a result of (1) humans only gesturing toward the intended page element 30.0% of the time (see Table 1) and (2) errors in the gesture detection model (see Figure 2).

Figure 4 summarizes the performance of the same experiment for ASR and compares human performance to the upper bound with perfect gesture detection and 100% user participation in gesturing towards the intended page element (link) when speaking. The top two curves show the performance of the real multimodal lexical+gesture detector compared to the baseline (lexical only).

Table 4: *Multi-modal intent detection with real gestures.*

	IDF a	IDF b	Gesture α	$P_{miss}@P_{fa}$=1% Manual	$P_{miss}@P_{fa}$=1% ASR
No Gesture	-	-	-	25.5%	34.0%
All	1	2	0.50	22.9%	31.7%
Gestures + "that one"	1	2	0.50	15.4%	15.4%

6. Conclusions

This paper described the development of a multi-modal dialog system for conversational web search and internet browsing. The work focused on two novel components: dynamic contextual adaptation of speech recognition and spoken language understanding models using multi-modal conversational context, and fusion of users' multi-modal speech and gesture inputs for understanding their intents and associated arguments. The system was evaluated in a living room setup with live test subjects on a real-time implementation of the multimodal dialog system. Results showed a 16% error rate reduction (ERR) for contextual ASR adaptation to clickable web page content, and 7-10% ERR when using gestures with speech. Analysis of the results showed that when users clearly and always indicate pointing intent while simultaneously using voice, the combination of modalities yields a 54.7% ERR over lexical features. While we observed users only point with hand gesture 30% of the time, the result suggests that other, more persistent modalities (e.g., eye gaze) could be used to yield substantial gains over speech alone.

7. Acknowledgements

The authors would like to thank Malcolm Slaney for helpful discussions related to this work.

8. References

[1] G. Tur and R. De Mori, Eds., *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, John Wiley and Sons, New York, NY, 2011.

[2] R. A. Bolt, "Put-that-there: Voice and gesture at the graphics interface," *Computer Graphics*, vol. 14, no. 3, pp. 262, 1980.

[3] G. Taylor, R. Frederiksen, J. Crossman, J. Voigt, and K. Aron, "A smart interaction device for multi-modal human-robot dialogue," *Ann Arbor*, pp. 190–191, 2012.

[4] R. Balchandran, M.E. Epstein, G. Potamianos, and L. Seredi, "A multi-modal spoken dialog system for interactive tv," in *Proceedings of the 10th international conference on Multimodal interfaces*, 2008, pp. 191–192.

[5] A. Ibrahim and P. Johansson, "Multimodal dialogue systems for interactive tv applications," in *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces*, 2002, pp. 117–122.

[6] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, "MATCH: an architecture for multimodal dialogue systems," in *Proceedings of the ACL*, Philadelphia, PA, July 2002.

[7] S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro, "Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions," *Human-computer interaction*, vol. 15, no. 4, pp. 263–322, 2000.

[8] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004, pp. 2422–2427.

[9] L. Heck, "The conversational web," in *IEEE Spoken Language Technology Workshop(SLT), Keynote*, 2012.

[10] Wikipedia The Free Encyclopedia, "Conversation," http://en.wikipedia.org/wiki/Conversation.

[11] D. Hakkani-Tür, G. Tur, L. Heck, A. Celikyilmaz, A. Fidler, D. Hillard, R. Iyer, and S. Parthasarathy, "Employing web search query click logs for multi-domain spoken language understanding," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011, pp. 419–424.

[12] D. Hakkani-Tür, G. Tur, L. Heck, and E. Shriberg, "Bootstrapping domain detection using query click logs for new domains," in *Interspeech*, 2011.

[13] G. Tur, D. Hakkani-Tür, D. Hillard, and A. Celikyilmaz, "Towards unsupervised spoken language understanding: Exploiting query click logs for slot filling," in *Interspeech*, 2011.

[14] D. Hakkani-Tür, L. Heck, and G. Tur, "Exploiting query click logs for utterance domain detection in spoken language understanding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5636–5639.

[15] D. Hakkani-Tür, G. Tur, R. Iyer, and L. Heck, "Translating natural language utterances to search queries for slu domain detection using query click logs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4953–4956.

[16] G. Tur, D. Hakkani-Tür, L. Heck, and S. Parthasarathy, "Sentence simplification for spoken language understanding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[17] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and L. Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog," in *Interspeech*, 2012.

[18] G. Tur, Minwoo Jeong, Ye-Yi Wang, D. Hakkani-Tür, and L. Heck, "Exploiting semantic web for unsupervised statistical natural language semantic parsing," in *Proceedings of Interspeech*, 2012.

[19] L. Heck and D. Hakkani-Tür, "Exploiting the semantic web for unsupervised spoken language understanding," in *Proceedings of the IEEE SLT Workshop*, Miami, FL, 2012.

[20] D. Hakkani-Tür, L. Heck, and G. Tur, "Using a knowledge graph and query click logs for unsupervised learning of relation detection," in *Proceedings of the ICASSP*, 2013.

[21] L. Heck, D. Hakkani-Tür, and G. Tur, "Leveraging knowledge graphs for web-scale unsupervised semantic parsing," in *Interspeech*, 2013.

[22] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[23] C.D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

LMELECTURES: A MULTIMEDIA CORPUS OF ACADEMIC SPOKEN ENGLISH

K. Riedhammer, M. Gropp, T. Bocklet, F. Hönig, E. Nöth, S. Steidl

Pattern Recognition Lab, University of Erlangen-Nuremberg, GERMANY

noeth@cs.fau.de

Abstract

This paper describes the acquisition, transcription and annotation of a multi-media corpus of academic spoken English, the *LMELectures*. It consists of two lecture series that were read in the summer term 2009 at the computer science department of the University of Erlangen-Nuremberg, covering topics in pattern analysis, machine learning and interventional medical image processing. In total, about 40 hours of high-definition audio and video of a single speaker was acquired in a constant recording environment. In addition to the recordings, the presentation slides are available in machine readable (PDF) format. The manual annotations include a suggested segmentation into speech turns and a complete manual transcription that was done using BLITZSCRIBE2, a new tool for the rapid transcription. For one lecture series, the lecturer assigned key words to each recordings; one recording of that series was further annotated with a list of ranked key phrases by five human annotators each. The corpus is available for non-commercial purpose upon request.

Index Terms: corpus description, academic spoken English, e-learning

1. Introduction

The *LMELectures* corpus of academic spoken English consists of high-definition audio and video recordings of two graduate level lecture series read in the summer term 2009 at the computer science department of the University of Erlangen-Nuremberg. The *pattern analysis (PA)* series consists of 18 recordings covering topics in pattern analysis, pattern recognition and machine learning. The *interventional medical image processing (IMIP)* series consists of 18 recordings covering topics in medical image reconstruction, registration and analysis. The lectures are read by a single, non-native but proficient speaker, and acquired in the *E-Studio*[1] which ensures a constant recording environment in the same room using a clip-on cordless close-talking microphone. The recordings were professionally edited to achieve a constant high audio and video quality. Note that not all lectures are consecutive; some recordings had to be dropped from the corpus because of a different speaker, sole use of German language, or technical issues such as a misplaced or defect close-talking microphone.

This paper documents the acquisition of the audio and video data (Sec. 2), the semi-automatic segmentation (Sec. 3), the subsequent manual transcription (Sec. 4), and the additional annotations (Sec. 5). Sec. 6 lists possible uses of the *LMELectures* and places the corpus in context with other corpora of academic spoken English. Sec. 7 suggests a partitioning of the data that is recommended for research on automatic speech recognition and key phrase extraction.

[1] RRZE MultiMediaZentrum, http://www.rrze.uni-erlangen.de/dienste/arbeiten-rechnen/multimedia/

2. Audio and Video Data

The audio data was acquired at a sampling rate of 48 kHz and 16 bit quantization, and stored in the *Audio Interchange File Format (AIFF)*. A 16 kHz version for the use with speech recognition systems was produced using down-sampling. The cordless close-talking microphone was able to reduce most of the room acoustics and background noises.

The video was acquired using an HD camera with manually controlled viewpoint and zoom setting to track the lecturer. Furthermore, the currently displayed presentation slide and, if applicable, on-screen writings is captured seperately. The video data is available in two formats:

- Presenter only, 640 x 360 pixel resolution, H.264 encoded (see Fig. 1, inset on the top left).
- Presenter, currently displayed slide and on-screen writings and lecture title, 1280 x 600 pixel resolution, H.264 encoded (see Fig. 1).

In total, 39.5 hours of audio and video data was acquired from 36 lecture recordings. The video recordings feature an AAC encoded audio stream based on the original 48 kHz data.

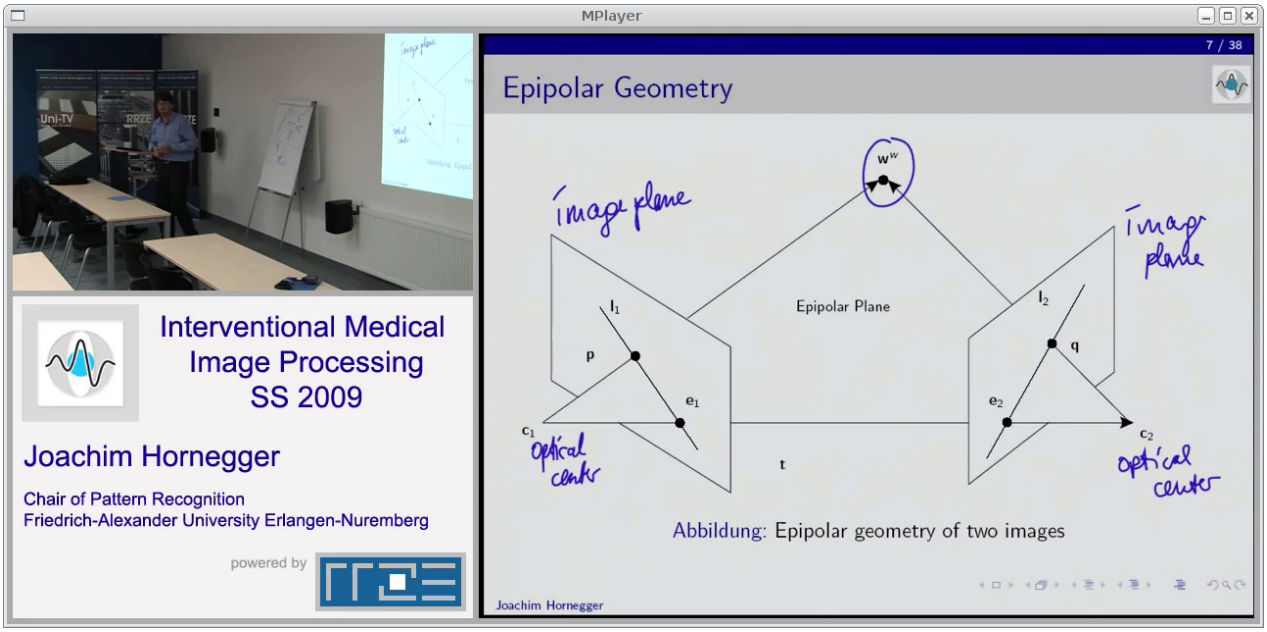


Figure 1: Example image from the video of lecture *IMIP01*. The left side shows the lecturer (top) and the lecture title (bottom), the right side shows the current slide and on-screen writings.

3. Semi-Automatic Segmentation

For the manual transcription, as well as for most speech recognition and understanding tasks, long recordings are typically split into short segments of speech. Another benefit is that longer periods of silence are removed from the data. The segmentation of the *LMELectures* is based on the time alignments of a Hungarian phoneme recognizer [1] that has been successfully used for speech/non-speech detection in various speaker and language identification tasks. The rich phonetic alphabet of the Hungarian language was found to be advantageous in the presence of various languages (here German and English) or wrong pronunciations. The set of phoneme strings was reduced by mapping the 61 original symbols to two groups: the pause (*pau*), noise (*int*, *e.g.*, a door slam) and speaker noise (*spk*, only if following *pau*, *e.g.*, cough) symbols were mapped to *silence* and the remaining symbols to *speech*. Merging adjacent segments of *silence* and *speech* results in an initial speech/non-speech segmentation (*cf.* Fig. 2).

Due to the design of the phoneme recognizer, the resulting segmentation has very sharp cut-offs and does not necessarily reflect the actual utterance or sentence structure, as even a very short pause may terminate a speech segment. With the aim of producing speech segments of an average length of four to five seconds[2], consecutive speech segments are merged based on certain criteria regarding segment lengths and intermediate silence (*cf.* Tab. 1).

Algorithm 1: Merge of consecutive segments based on their duration and interleaving silence.

```
for all segments i do
    if Pau(i, i+1) < min. pau or Dur(i) <
    min. dur then
        required ← true
        while required or Dur(i) < max. dur do
            if ! required then
                if Dur(i) > med. dur or
                Dur(Merge(i, i+1)) >
                max. dur or Pau(i, i+1) >
                max. pau then break
            end
            i ← Merge(i, i+1)
            required ← (Pau(i, i+1) <
            min. pau)
        end
    end
end
```

Algorithm 1 outlines the greedy merging procedure. 150 ms were added to the end of each segment to ease the sharp cut-offs. Given the desired target length, the major control variables are the pauses. Allowing too long pauses within a segment (`max.pau`) may lead to segments that contain the end and beginning of two separate

[2] as suggested by previous experiences of the group with manual transcription and speech recognition system training and evaluation

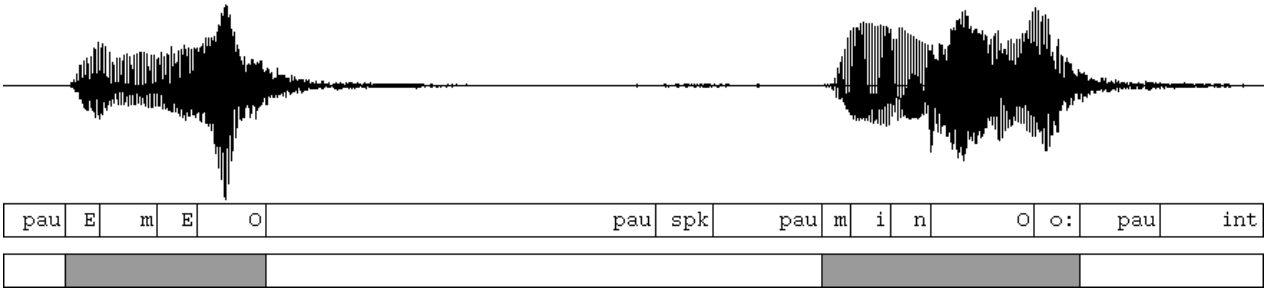


Figure 2: ≫*And then (breath) we know*≪. Adjacent segments of *silence* or *speech* phonemes are merged to an initial speech (gray) and non-speech (white) segmentation.

quantity	*description*	*value*
`min.dur`	if segment is shorter than *min. dur*, merge with following	2 s
`med.dur`	stop if merged segment is longer than *med. dur*	4 s
`max.dur`	only merge if resulting segment is shorter than *max. dur*	6 s
`max.pau`	maximum duration of pause within a segment	1 s
`min.pau`	minimum duration of pause between two segments	0.5 s

Table 1: Final merging criteria for consecutive speech segments.

utterances. Requiring long silences between segments (`min.pau`) leads to unnaturally long segments.

The segmentation closest to the desired characteristics comprises 23 857 speech turns with an average duration of 4.4 seconds, and a total of about 29 hours of speech. Note that these segments are for the purpose of recognition, and do not necessarily resemble dialog acts or "actual" speech turns. The right column of Tab. 1 shows the respective merging criteria. The typically 0.5 s to 3 s of silence between speech segments accumulate to about 10 hours.

4. Manual Transcription

The manual transcription of speech typically requires about ten to 50 times the duration of speech using professional tools like TRANSCRIBER [2, 3]. TRANSCRIBER, similar to other tools, allows to work on long recordings by identifying segments of speech, noise and other acoustic events. Furthermore, higher level information like speaker, speech or language attributes can be annotated. However, this higher level information regarding the data at hand is usually known in advance, and lectures are typically very dense in terms of speech, thus reducing the main task to the (desirably) fast transcription of the speech segments.

The segments were manually transcribed using BLITZSCRIBE2,[3] a platform independent graphical user interface specifically designed for the rapid transcription of large amounts of speech data. It is inspired by research of Roy *et al.* [3] and is publicly available as part of the Java Speech Toolkit (JSTK) [4].[4] Fig. 3 shows the interface that displays the waveform of the currently selected speech segment, a progress bar indicating the current playback position, an input text field to type the transcription, and a list of turns, optionally with prior transcription.

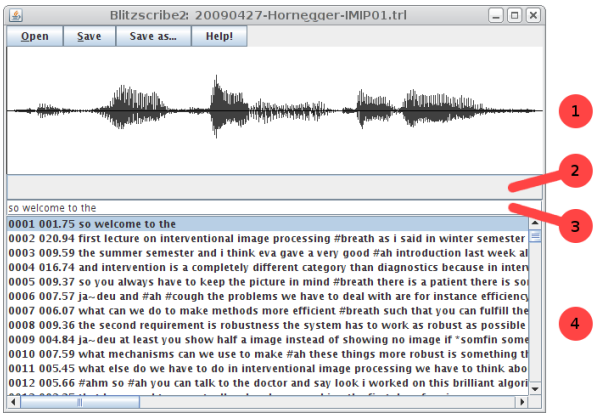


Figure 3: Screenshot of the BLITZSCRIBE2 transcription tool; (1) waveform of the currently selected speech segment, (2) progress bar indicating the current playback position, (3) text field for the transcription, (4) list of segments with transcription (if available).

The key idea to speed up the transcription is to simplify the way the user interacts with the program: although the mouse may be used to select certain turns for transcription or replay the audio at a desired time, the most frequent commands are accessed via keyboard shortcuts listed in Tab. 2.

For a typical segment, the transcriber types the transcription as he listens to the audio, pauses the playback if necessary (CTRL+SPACE), and hits ENTER to save the transcription, which loads the next segment and starts the playback. This process is very ergonomic as the hands

[3] `http://www5.informatik.uni-erlangen.de/en/research/software/blitzscribe2/`

[4] `http://code.google.com/p/jstk`

key combination		*command*
ENTER		save transcript, load **and play** next segment
SHIFT	+BACKSPACE	save transcript, load previous segment
SHIFT	+ENTER	save transcript, load next segment
CTRL	+SPACE	start/pause/resume/restart playback
CTRL	+BACKSPACE	rewind audio and restart playback
ALT	+S	save transcription file

Table 2: Keyboard shortcuts for fast user interactions in BLITZSCRIBE2.

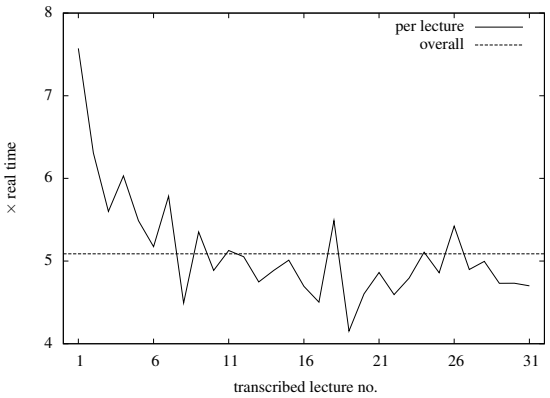


Figure 4: Change of the median transcription real time factor required by transcriber 1 throughout the transcription process.

remain on the keyboard during all times.

The lectures were transcribed by two transcribers. The work was shared among the transcribers and no lecture was transcribed twice. As the language is very technical, a list of common abbreviations and technical terms was provided along with the annotation guidelines. The overall median time required to transcribe a segment was about five times real time, which is a significant improvement over traditional transcription tools. Fig. 4 shows the decreasing transcription real time factor of one transcriber while adapting to the BLITZSCRIBE2 tool.

In total, about 300 500 words were transcribed with an average of 14 words per speech segment. Intermittent German words were transcribed and marked; those typically include greetings or short back-channel. Other foreign, mispronounced or fragmented words were transcribed as closely as possible, and marked for later special treatment. The resulting vocabulary size is 5 383 including multiple forms of words (*e.g.*, plural, composita), but excluding words in foreign languages and mispronounced or word fragments.

5. Further Manual Annotations

The presentation slides are available in machine readable (PDF) format, however, only the video provides accurate information about the display times. The lecturer added key words to each of the lecture recordings in series *PA*.

Lecturer's Phrases	*Annotator 1*	*Annotator 2*	*Annotator 3*	*Annotator 4*	*Annotator 5*
linear regression	●	●	●	●	●
norms	●		●	●	○
dep. linear regression			○		○
ridge regression	●		●		●
discriminant analysis	○		○	○	
motivation					
AP(5)	0.90				
NDCG(5)	0.73				

Table 3: Master key phrases of lecture *PA06* assigned by the lecturer, coverage indicators (●) for the human annotators, and phrase rank of the automatic rankings, if applicable. The empty bullets (○) indicate a partial match, *e.g.*, "linear discriminant analysis" satisfies "discriminant analysis."

The individual lecture *PA06* was further annotated with a ranked list of key phrases by five human subjects that have either attended the lecture or a similar lecture in a different term. The annotators furthermore graded the phrases present in their ranking in terms of quality from 1 – "sehr relevant" *(very relevant)* to 6 – "nutzlos" *(useless)*. This additional annotation can be used to assess the quality of automatic rankings using measures such as average precision (AP) [5] or normalized distributed cumulative gain (NDCG) [6, 7], two measures popular in the search engine and information retrieval community.

Tab. 3 shows, for *PA06*, the lecturer's phrases, whether the raters also extracted them, and the average AP and NDCG when comparing each rater to the remaining ones when considering the top five ranked terms.

6. Intended Use and Distinction from Other Corpora of Academic Spoken English

The corpus, with its annotations, is an excellent resource for various mono- and multi-modal research. The roughly 30 hours of speech of a single speaker provide a great base to work on acoustic and language modeling, speaker adaptation, prosodic analysis and key phrase extraction. The spoken language is somewhere in between read text and spontaneous speech, with passages of well-structured and articulated speech followed by a mumbled utterance with disfluencies and hesitations. At a higher level, the video can be used to determine slide timings, on-screen writing and other interactions of the lecturer. The two series of consecutive lectures provide a good scenario to work on automatic vocabulary extension and language model adaptation as required for a production system.

name	*duration*	*# turns*	*# words*	*% OOV*
train	24h 31m 55s	20 214	250 536	—
dev	2h 07m 28s	1 802	21 909	0.87 %
test	2h 12m 30s	1 750	23 497	0.99 %

Table 4: Data partitioning for the *LMELectures* corpus; the number of words excludes word fragments and foreign words. The percentage of OOV words is given with respect to the words present in the *train* partition.

The two main corpora of academic spoken English are the BASE corpus,[5] and the Michigan Corpus of Academic Spoken English (MICASE) [8]. Although both corpora cover more than 150 hours of speech, their setting is different from the *LMELectures*. The BASE corpus covers 160 lectures and 40 seminars from four broad disciplinary groups (Arts and Humanities, Life and Medical Sciences, Physical Sciences, Social Sciences). Audio, video and transcription material are available for licensing. The MICASE corpus features a wide variety of recordings of academic events including lectures, colloquia, meetings, dissertation defenses, *etc.*. Again, audio and transcripts are subject to licensing, but video data is unavailable.

The main distinction of the *LMELectures* is however the technical homogeneity in terms of recording environment, speaker, and topic of the two lecture series.

7. Suggested Data Partitioning

For experiments on speech recognition and key phrase extraction, the authors suggest to partition the data in three parts. The development set, *devel*, consists of the four lecture sessions IMIP13, IMIP17, PA15 and PA17, and has a total duration of about two hours. The test set, *test*, consists of the four lecture sessions IMIP05, IMIP09, PA06 and PA08, and has also a total duration of about two hours. The remaining 28 lecture sessions form the training set, *train*, with a total of about 24 hours. Tab. 4 summarizes the partitioning and lists details on the duration, number of segments and words, and out-of-vocabulary (OOV) rate with respect to a lexicon based on the training set. A baseline speech recognition experiments using the KALDI toolkit resulted in a word error rate of about 11 % on the test set [9]. For any other partitioning, the authors suggest to include PA06 in the test set as it was annotated with key phrases.

8. Summary

This paper describes the collection and annotation of a new corpus of academic spoken English that consists of audio/video recordings of two series of computer science lectures at the graduate level. The data was acquired in high definition, and was edited to achieve a constant quality; there are two versions of the video available: one that shows only the presenter (including accidental parts of the blackboard and projector canvas), and a combined view that shows both the presenter and the currently displayed slide including on-screen writing. The PDF slides are available, although there exists no exact lecture to slide set alignment: some slide sets overlap multiple sessions, some sessions focus on classic blackboard oriented teaching.

In addition to the plain data, several manual annotations are available:

- The newly developed BLITZSCRIBE2 was used to transcribe the roughly 30 hours of speech in about five times real time instead of ten to 50 times real time as reported for other transcription tools. BLITZSCRIBE2 is freely available as part of the JSTK.
- The lecturer assigned a rough set of key phrases to each lecture, which can be considered a ground truth from a teaching perspective.
- For an individual lecture *PA06*, five human annotators that either observed that very lecture or a similar one in previous years extracted and ranked a set of key phrases.

The collected corpus forms a good base for future research on ASR for lecture-style, non-native speech (a significant percentage throughout the world), supervised and unsupervised key phrase extraction, topic segmentation, slide to speech alignment, and other e-learning related issues. The corpus is available for non-commercial use upon request, please contact the authors for details. Further details of the transcription and annotation process can be found in [10].

9. Acknowledgments

The authors would like to thank Prof. Dr.-Ing. Joachim Hornegger for authorizing the release of the lecture recordings and related PDF slide material. The recording, editing, media encoding and data export was done by the Regionales Rechenzentrum Erlangen (RRZE). The authors would furthermore like to thank Dr. Anton Batliner for his very valuable advice on how to structure, organize and execute a large scale data set acquisition.

[5] The British Academic Spoken English (BASE) corpus project. Developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson.

10. References

[1] P. Matejka, P. Schwarz, J. Cernocky, and P. Chytil, "Phonotactic Language Identification using High Quality Phoneme Recognition," in *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, 2005, pp. 2237–2240.

[2] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: Development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1-2, pp. 5–22, 2001.

[3] B. Roy and D. Roy, "Fast transcription of unstructured audio recordings," in *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, 2009, pp. 1647–1650.

[4] S. Steidl, K. Riedhammer, T. Bocklet, F. Hönig, and E. Nöth, "Java Visual Speech Components for Rapid Application Development of GUI based Speech Processing Applications," in *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, 2011, pp. 3257–3260.

[5] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[6] K. Järvelin and J. Kekäläinen, "IR Evaluation Methods for Retrieving Highly Relevant Documents," 2000, pp. 41–48.

[7] K. Järvelin and J. Kekäläinen, "Cumulated Gain-Based Evaluation of IR Techniques," vol. 20, no. 4, pp. 422–446, 2002.

[8] R. C. Simpson, S. L. Briggs, J. Ovens, and J. M. Swales, "The michigan corpus of academic spoken english," Tech. Rep., University of Ann Arbor, MI, USA, 2002.

[9] K. Riedhammer, M. Gropp, and E. Nöth, "The FAU Video Lecture Browser system," in *Proc. IEEE Workshop on Spoken Language Technologies (SLT)*, 2012, pp. 392–397.

[10] K. Riedhammer, *Interactive Approaches to Video Lecture Assessment*, Logos Verlag Berlin, 2012.