SPEAKER DIARIZATION OF HETEROGENEOUS WEB VIDEO FILES: A PRELIMINARY STUDY

Pierre CLÉMENT[1], *Thierry BAZILLON*[2], *Corinne FREDOUILLE*[1]

[1]Université d'Avignon - Laboratoire Informatique d'Avignon - CERI/LIA - France
{pierre.clement, corinne.fredouille}@univ-avignon.fr
[2]Université d'Aix-Marseille - Laboratoire Informatique Fondamentale - LIF-CNRS - France
thierry.bazillon@lif.univ-mrs.fr

ABSTRACT

In the last ten years, internet as well as its applications changed significantly, mainly thanks to the raising of available personal resources. Concerning multimedia, the most impressive evolution is the continuous growing success of the video sharing websites. But with this success come the difficulties to efficiently search, index and access relevant information about these documents. Speaker diarization is an important task in the overall information retrieval process. This paper describes an audio/video database, especially built for the speaker diarization task, based on different video genres. Through some preliminary experiments, it highlights the difficulties encountered in this context, mainly linked to the database heterogeneity.

Index Terms: speaker diarization, heterogeneous web videos, diarization error rate

1. INTRODUCTION

Many multimedia documents are uploaded daily on different websites, which the most known are YouTube, or Dailymotion. This large amount of multimedia data is still growing each year. It becomes increasingly important to automatically and efficiently search, index, and access the information that is present in the media files, like, for example, the video genre, the spoken language, the linguistic content, the location and context where the video takes place, who is speaking, when, etc. Speaker diarization is an important task in the process of information retrieval, especially through the Rich Transcription process [1]. Indeed, the aim of speaker diarization is to provide automatically the set of the temporal regions in the media stream where a same speaker is speaking, without any prior information on the number of speakers in the document nor their identities. That kind of information is useful for identifying a person, assigning a linguistic message or a specific context to a person, or more simply, for helping other automatic processes like the speech recognition system.

Classically, speaker diarization involves two main steps: the detection of boundaries between speakers, indicating the speaker turns in the document (segmentation step) and the grouping of all same-speaker segments (clustering step). In most of the studies reported recently in the literature, both steps are performed either sequentialy or simultaneously, following two main approaches for the clustering: the bottom-up and top-down strategies. Differences between these approaches rely mainly on the way of dealing with clusters. Initially, the top-down strategy only considers one cluster/speaker from an audio file and tends to detect iteratively new speakers from it, whereas the bottom-up strategy defines a large number of clusters (larger than the one expected) from the audio file and tends to merge them. In both cases, the process aims to reach an optimum number of clusters. Typically based on Hidden Markov Models (HMM), in which each state corresponds to a speaker and is associated with a Gaussian Mixture Model (GMM), the iterative process involves a Viterbi decoding loop associated with a GMM model training phase, in order to determine and refine speaker boundaries.

For about ten years, performance of speaker diarization systems have been measured according to different application domains, like telephone conversations (two speaker audio documents), Broadcast News (BN), and lastly meeting data. Dealing with meeting data remains currently the most difficult task, because of the higher level of speech spontaneity (inducing a large amount of speech overlap, disfluencies, short speaker turns), and a variable quality of recordings (different types of microphones, different levels of noise, background speech). Moreover, each application domain may require some specific system tuning. Confronted to a new kind of application domain like the web videos, it is interesting to study the behavior of such systems (based on audio stream uniquely) in this particular context, where the type of audio documents to process can be very variable. In this paper, the behavior of the LIA[1] speaker diarization system based on a top-down strategy is compared with the one of the LIUM[2], based on a bottom-up strategy. This comparison is carried out through an experimental framework involving five different genres of videos: documentaries, news, movie trailers, commercials, and cartoons.

Thus, section 2 describes the heterogeneous web video database that will be used, exhibiting differences between the different video genres and their intrinsic difficulties. Section 3 details the couple of speaker diarization systems, which experimental results from different datasets are provided in section 4. Finally, some conclusions and perspectives are given in section 5.

2. WEB VIDEO DATABASE

2.1. Global presentation

The LIA multi**GE**nre and multispeake**R** **A**udio/video F**R**ench **D**atabase (LIA GERARD) is built from 856 videos downloaded from different websites. Videos can be classified according to 7 different categories as explained in [2].
In this study, only 5 categories are held: documentaries, movie trailers, cartoons, commercials and news, discarding sport (only audio stream available) and music (only singing voices) categories. This

[1] Laboratoire Informatique d'Avignon

[2] Laboratoire Informatique de l'Université du Maine

Category	Nb of files	Total length	File length (av.)	Part of speech (in %)	Average of speakers nb per file	Speaker turns nb per file (av.)	Speaker turn length (av. in s)
Documentary	29	3:19:06	0:06:51	71.78	8.2	84	3.51
Movie trailer	30	1:07:05	0:02:14	53.73	9.4	35	2.06
Cartoon	30	4:11:40	0:08:23	64.41	10.9	113	2.87
Commercial	10	0:15:40	0:01:34	68.09	5.2	25	2.56
News	30	1:32:40	0:03:05	88.65	4.5	26	6.31

Table 1. *Information issued from the manual annotation relating to the selected videos classified per category: number of files, total duration, % of speech. Per file and category: average duration, average numbers of speakers in the file, of speaker turns and their average length.*

database does not contain "homebrew" videos, to prevent issues related to devices used to record the video (such as mobile phones). The main language of these videos is French (except for some documentaries in English or with voice-overs translation). All the downloaded files are encoded in the same format : a Flash Video format (FLV), in size 320x240, 25 fps, and around 370 kbps for the video streams; MP3, 44 kHz, stereo and 96 kbps for the audio streams.

2.2. Manual annotation

A small part of the database has been manually annotated. The annotation concerns the audio component uniquely (the video annotation will be done later) in order to measure both the speech activity detection and speaker diarization system performance. In this sense, the videos have been segmented, labeling all the speech and speaker turns, marking noise and music and taking the areas of overlap speech into account. In this way, about 130 videos from the 5 selected categories have been annotated and will be used further for experiments.

A numerical description of this corpus per category is given in table 1. It can be noted that the videos selected last from 1 to 10 minutes. The longest category in terms of duration is the cartoons (about 4h12min), followed by the documentaries (about 3h20min). The shortest one is the commercials (about 16min) because only 10 files from the LIA GERARD database respect the minimum duration constraint - 1min - fixed in this study. If statistics are rather variable from one category to another, the average speaker turn length per file is dramatically low. Even the longest one associated with news does not exceed 6.31s.

2.3. Characteristics per video genre

Every video genre treated in this paper has its own level of difficulties. These difficulties, as shown below, can be related to the audio, the video (not treated here) or both.

- **Documentary:** can contain a large amount of speech overlap due to the French video dubbing (because of foreign languages). In the videos selected, the background is very noisy. An off screen voice can also be present, which can be troublesome in the joint use of audio and video streams. Indeed, while the off screen voice is talking, the main speaker is rarely on the picture.
- **Movie trailer:** is very interactive. A larger number of speakers may talk successively (in order to show all the main movie characters) as shown in table 1 where the average number is about 8 for an average file duration of 2min14s. This leads to a large number of very short speaker turns (2s in average). There are many sound effects and music (46% of non-speech), speech over loud music or noise (it is sometimes hard to hear the speaker), and many shot boundaries. Regarding the video, off screen voices or speakers start to speak while the scene does not change. Compared to the others, this category was the most difficult to annotate.
- **Cartoons:** a dubber may dub multiple characters. Consequently, voices of characters are quite similar or the character's voice can vary between two speech shots. In addition, a character may have also two different voices (for example, either for two different kinds of scene, for two different ages, or a scene where the characters are graphically deformed). Like movie trailers, cartoons are very interactive with a lot of music and sound effects (35% of non-speech) as well as speech over music. Compared to other categories, the number of speakers relatively large (11 in average) whilst the speaker turn length is relatively low (about 2.9s), regarding the average duration of files. In addition, characters' voices can be recorded within different background environments. Regarding the video, animals or objects can be likened to some characters with speaking abilities, which can make automatic face detection task more complicated. Similarly, lags in lip synchronization can be present. Finally, the character who is speaking is often not displayed.
- **Commercial:** the style of commercials can be very different from one to another, depending on the kind of product displayed. However, all are very short (1min30s in average), the background is noisy, and the part of non-speech is relatively large (32% of non-speech). Moreover, the number of speakers involved in commercials as well as the number of speaker turns are relatively high, 5.2 and 25 in average per file respectively, regarding the short duration of files. Sometimes, off screen voices can also be used.
- **News:** there are two kind of news, either recorded in studio or in an uncontrolled environment. In the latter, background noise can be significantly present. However, in both cases, speech is mainly present in the files (about 90% of speech). Relatively short (average duration of 3min) compared with cartoons or documentaries, news videos exhibit the largest speaker turn length per file (about 6.3s), and the smallest number of speakers per file (4.5 in average). In most videos, an off screen voice is employed to introduce the main part of the video.

3. SPEAKER DIARIZATION SYSTEM OVERVIEW

3.1. LIA Speaker diarization system

The LIA speaker diarization system, thoroughly described in [3], is based on an Evolutive Hidden Markov Model (E-HMM) using the open-source ALIZE toolkit [4]. It is composed of three main steps:

(1) a Speech Activity Detection (SAD) necessary for detecting non-speech segments, which may affect the speaker diarization process. The SAD algorithm used here is very similar to the one used for the RT'09 evaluation campaign [3]. It is based on a HMM which Viterbi decoding and GMM-based model adaptation are applied on iteratively to refine the segmentation. Depending on the data treated, the number and types of HMM states can be variable: related to speech and silence/noise data for meetings or to music, speech over

	RT'09 eval		**Ester'08 dev**		**Ester'08 eval**		**LIA GERARD subset (auto. speech/non-speech detection)**		**LIA GERARD subset (manual. speech/non-speech detection)**	
Data type	Meeting		BN				Heterogeneous			
Number of files	7		18		26		129			
Total length	3:00:58		9:27:50		7:10:22		9:52:47			
System	LIA	LIUM	LIA	LIUM	LIA	LIUM	LIA	LIUM	LIA	LIUM
DER	18.9%	NA	14.6%	8.8%	15.5%	8.2%	73.2%	55.6%	38.7%	34.3%
E_{missed}	0.5%	NA	1.8%	0.5%	1.3%	0.2%	9.5%	13.9%	0%	0%
E_{fa}	2.9%	NA	0%	2.5%	1.7%	1.6%	27.2%	13%	0%	0%
E_{spkr}	15.5%	NA	12.8%	5.8%	12.5%	6.4%	36.5%	28.7%	38.7%	34.3%

Table 2. *Preliminary results of the LIA and the LIUM speaker diarization systems on different data sets.*

music, narrow speech, etc for BN. The SAD segmentation output is afterwards used as input for the following speaker diarization steps.

(2) a speaker segmentation and clustering process, based on an E-HMM, within which each state characterizes a single speaker and every transition represents a speaker change. In this step, the signal is characterized by 21 coefficients, 20 un-normalized LFCC plus the energy.

The segmentation process begins by initializing the E-HMM with only one state representing the entire audio show (state denoted speaker L0). An iterative process is then started involving the detection of a new speaker and its addition to the E-HMM. This detection is driven by a selection strategy of data segment, labeled as belonging to speaker L0, but further attributed to the new speaker. For the process, successive Viterbi decoding and speaker model training loops attribute speech segments to the different speakers involved in the E-HMM. This iterative process is stopped when no more speaker can be added to the E-HMM. A GMM-based speaker model is assigned to each HMM state, for which the EM/ML (Expectation - Maximization/Maximum Likelihood) algorithm is used for the statistics estimate and a sufficient amount of data has to be available for their robustness. This constraint can be particularly strong for each new speaker detected (only one segment assigned) and quite dependent on the selection strategy mentioned above. Because of that, and inspired from other speaker diarization systems such as the ICSI system [5], a selection based on the largest speech segment available in speaker L0 (with a minimum size fixed to 6s) was proposed in [6] and is used here.

(3) a Resegmentation process, which aims to refine the segmentation outputs and to remove irrelevant speakers (e.g. speakers with too few segments). A HMM is generated from the current segmentation and the Maximum A Posteriori (MAP) adaptation (involving a Universal Background Model) is used instead of EM/ML algorithm to estimate the speaker GMM-based models. In this process, all the boundaries (except speech/non-speech) and segments are reprocessed.

3.2. LIUM Speaker diarization system

The second speaker diarization system involved in this study is developed at the LIUM. Freely distributed, this system performs different steps, thoroughly described in [7]: a HMM-based speech/non-speech segmentation, similar to the LIA SAD process is first applied, coupled afterwards with a gender and bandwidth detection steps. A two-pass segmentation process based on both the GLR and δBIC criteria is then computed, which aims to detect change points corresponding to segment boundaries. A hierarchical agglomerative clustering, still based on the δBIC, followed by a Viterbi decoding are secondly performed to generate a new segmentation. Because the previous segmentation and clustering steps could produce different clusters for a same speaker (because of the background noise which could influence the decisions for instance), a last GMM-based speaker clustering is finally applied to merge clusters of the same speaker. Here, a Universal Background Model (UBM) is used to adapt each cluster and obtain the model of each speaker.

4. RESULTS AND COMPARISON

This section reports results obtained with the LIA and the LIUM speaker diarization systems on different datasets[3]:

- **RT'09 eval**: meeting data issued from the last NIST speaker diarization evaluation campaign [8];
- **ESTER'08 dev and eval**: BN development and evaluation set issued from the French speaker diarization evaluation campaign - ESTER II [9];
- **LIA GERARD subset**: the video corpus subset described in section 2.

The aim of the experimental study is twofold: (1) to observe the behavior of both systems when involved in the multigenre video application domain, and compare it with more classical ones (BN, meetings); (2) to compare the LIA top-down and the LIUM bottom-up strategies in the multigenre video application domain.

4.1. Evaluation

The quality of the speaker diarization segmentation is evaluated thanks to the classical Diarization Error Rate (DER), fully described in [8]. The DER represents the fraction of time that does not correctly match to a speaker or to a non-speech segment and cumulates two types of speech/non speech misclassification errors (E_{fa}, E_{missed}) and one speaker misclassification error (E_{spkr}).

4.2. Results

Table 2 shows the results obtained with both LIA and LIUM speaker diarization systems according to the different data sets mentioned above. It can be pointed out that the LIUM system (bottom-up) outperforms the LIA one (top-down) on the BN domain in terms of speaker error rate, while reaching similar high performance for the SAD process. However, a drastic decrease of performance is observed on both systems when dealing with the multigenre videos (LIA GERARD subset). This decrease is partly due to the SAD process, which seems inefficient in these particular data. For this reason, experiments based on the speech/non-speech segmentation issued from the manual annotation have been carried out (5th column of table 2) which results are reported in the rest of the paper. Here, both systems obtain quite poor performance, between 34% (LIUM) and 38% (LIA) DER (related to speaker error rate only). DER score per file is provided in figure 1 for both systems. This figure presents a large variability in DER score for both cases, underlying different difficulty levels that both systems do not handle necessarily in the

[3] Performance of the LIUM system is not available for the RT'09 data set since it has been configured initially for BN.

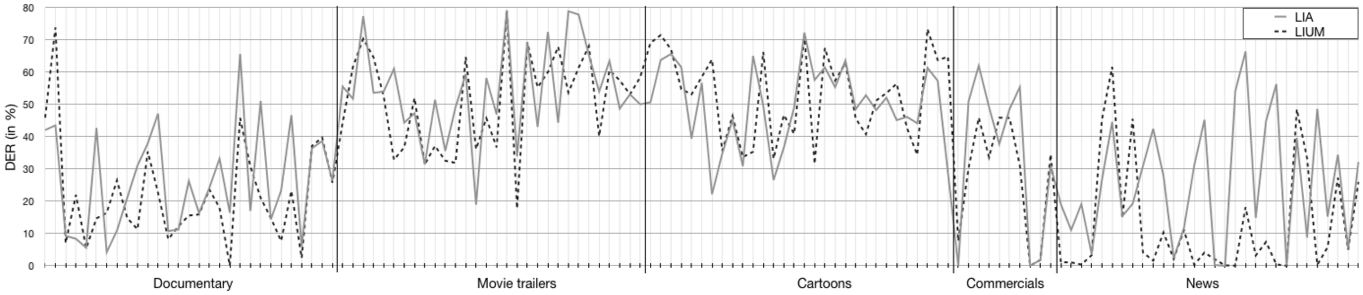

Fig. 1. *DER comparison file by file between LIA and LIUM diarization systems, based on manual speech/non-speech segmentation.*

same manner. In order to measure the influence of the genre – and therefore its characteristics – on the speaker diarization strategies, global performance per genre is shown in table 3, reporting the average number of speakers found, % DER coupled with their minimum and maximum values for the set of video files.

Documentaries and news exhibit the highest performance, with around 26% DER for the LIA, 22% and 12% for the LIUM respectively. These DER can be compared with those of the ESTER'08 eval dataset (LIA: 15.5%, LIUM: 8.2%), as documentaries and news are closer to BN than meetings domain. Based on this comparison, we can observe a difference of performance between news and documentaries, quite stable for the LIA system, but unbalanced for the LIUM. Regarding movie trailers, cartoons, and commercials, DER increase severely (54%, 50%, and 33% for the LIA against 51%, 53% and 28% for the LIUM respectively). This dramatic degradation seems to be due to the joint presence of a very small length of speaker turns and of a relatively large number of speakers considering the short file duration. These characteristics make the detection of speakers in the diarization process more complicated, and the confusion between speakers larger due to the small amount of data available for speaker model estimate. This assumption is augmented in the case of the E-HMM based system (compared with bottom-up strategy) which is well-known to succeed in detecting the main speakers (in terms of speech duration), especially when associated with some long speaker turns. The other reason of the huge performance degradation can be linked to the adverse recording environment (speech over music, sound effects or noises) as well as the intrinsic characteristics related to each category.

Finally, it is interesting to notice that the average number of speakers found by the LIUM bottom-up system (given in table 3) is quite close to the effective number (given in table 1). Nevertheless, the related low performance shows that segments associated with those speakers are not as reliable as expected.

5. CONCLUSION AND PERSPECTIVES

This preliminary study outlines the difficulties encountered by speaker diarization systems, and especially the top-down-based strategy on the video genre application domain. As with the meeting application domain (and the use of Beamforming process, of Time Delay Of Arrival parameters), new techniques will have to be designed for processing web video, notably to deal with the adverse environments. Factor analysis application will be studied in this specific context for instance. Secondly, video contains information about the speakers and could be used to help the speaker diarization process to make its decision, when ambiguity is detected on the audio stream. Nevertheless, the joint use of audio and video information has to focus first the synchronization issues between audio and video streams (speaker may talk, but not be displayed and vice versa).

Category	**Av. nb of spks found**		**DER (in %)**		**Min DER (in %)**		**Max DER (in %)**	
System	LIA	LIUM	LIA	LIUM	LIA	LIUM	LIA	LIUM
News	2.9	4.6	25.7	12.8	0.1	0	66.4	61.6
Documentary	4	7	26.4	22	4.2	0	65.6	73.8
Movie trailer	1.4	2.8	54.3	51.1	18.9	17.7	79.2	79.2
Cartoon	4.3	10.2	49.7	53.1	22.1	31.6	72.2	73.4
Commercial	1.5	3.9	33.6	27.7	0	0	62	45.9

Table 3. *Results per genre given by the two speaker diarization systems : LIA and LIUM.*

6. ACKNOWLEDGEMENTS

This research was supported by the ANR agency (Agence Nationale de la Recherche), RPM2 project (ANR-07-AM-008). We are also grateful to Sylvain MEIGNIER and Teva MERLIN for providing us with their toolkit and their help.

7. REFERENCES

[1] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE TASLP*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] M. Rouvier, G. Linares, and D. Matrouf, "On-the-fly video genre classification by combination of audio features," in *ICASSP 2010*, Dallas, US, 2010.

[3] C. Fredouille, S. Bozonnet, and N. W. D. Evans, "The LIA-EURECOM RT'09 Speaker Diarization System," in *RT'09, NIST Rich Transcription Workshop*, Florida, USA, 2009.

[4] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *Proc. ICASSP'05*, Philadelphia, USA, March 2005, vol. 1, pp. 737–740.

[5] X. Anguera, C. Wooters, and J. Hernando, "Robust speaker diarization for meetings: ICSI RT06s evaluation system," in *Proc. ICSLP*, Pittsburgh, USA, September 2006.

[6] C. Fredouille and N. Evans, "New implementations of the E-HMM-based system for speaker diarisation in meeting rooms," in *Proc. ICASSP'08*, Brisbane, Australia, 2008.

[7] S. Meignier and T. Merlin, "Lium_spkdiarization: An open source toolkit for diarization," *CMU SPUD Workshop*, 2010.

[8] NIST, "The NIST Rich Transcription (RT'09) evaluation," http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/ rt09-meeting-eval-plan-v2.pdf, 2009.

[9] S. Galliano, G. Gravier, and Laura Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Interspeech'09*, Brighton, UK, 2009.