Speaker Identity Indexing In Audio-Visual Documents

Mbarek Charhad, Daniel Moraru, Stéphane Ayache, Georges Quénot

SPEAKER IDENTITY INDEXING IN AUDIO-VISUAL DOCUMENTS

Mbarek Charhad, Daniel Moraru, Stéphane Ayache and Georges Quénot

CLIPS-IMAG
BP 53, 38041 Grenoble cedex 9, France
Georges.Quenot@imag.fr

ABSTRACT

The identity of persons in audiovisual documents represents very important semantic information for content-based indexing and retrieval. The task of speaker's identity detection can be carried out by exploiting data elements resulting from different modalities (text, image and audio). In this article, we propose an approach for speaker identity indexing in broadcast news using audio content. After a speaker segmentation phase, an identity is given to speech segments by applying linguistic patterns to their transcription from speech recognition. Three types of patterns are used to predict the speaker in the previous, current and next speech segments. Predictions are then propagated to other segments by similarity at the acoustic level. Evaluations have been conducted on part of the TREC 2003 corpus: a speaker identity could be assigned to 53% of the annotated corpus with an 82% precision.

1. INTRODUCTION

Thanks to technological progress, digital video databases continuously become more and more common as well as larger and larger. Efficient content-based access to them becomes more and more a critical issue and new and efficient tools are always needed for this.

The great variety of video document contents yields a large number of possible keys for indexing and retrieval. An indexing key can be for instance a simple word appearing in the discourse as well as a topic or category (sports, economy, politics, …) [1]. In the second case, all words possibly related to a topic will be searched for. Another indexing key that we find very useful is the identity of the speaker. We considered speaker identity indexing in the context of TV news. This type of video document usually exhibits a well defined structure that eases the process of speaker identity detection and tracking.

Approaches based on the use of speech transcription obtained from automatic speech recognition and acoustic-based speaker segmentation are especially interesting since they allows for speaker identity indexing without needing the building of acoustic models of speakers (which is costly and restrict the indexing to learned speakers). These approaches can also be applied to purely audio documents like radio news unlike visual ones based on face recognition and lips motion detection. The visual methods are also limited to learned people and can even miss them if they do not appear from the right angle or at the right scale.

Alternatively again, OCR-based text recognition in the images could provide useful information about speaker identity but the current performance of OCR tools in video images (often degraded due to compression) does not appear sufficient for reliable extraction. The problem of attaching a face identity and/or a speaker identity to a name extracted by OCR also remains difficult.

The speaker identity indexing approach proposed here is base on the use of linguistic patterns applied to a transcription of the audio tracked obtained by automatic speech recognition (ASR). It is similar to the one proposed in [2]. The search for parameterized regular expression in the transcription permits to propose speaker identity in speech segments in which it appears, in speech segments following the one in which it appears or in speech segments preceding the one in which it appears. This approach was completed by the propagation of the identity attributed to some segments to other segments attributed to the same speaker by acoustic similarity. Quantitative evaluations were performed using a part of the TRECVID 2003 corpus.

This rest of this article is organized as follows: in section 2, we present some previous work on the use of audio information for content based indexing and retrieval. In section 3, we detail our approach for speaker identification using linguistic patterns. We present an evaluation of it in section 4 and we discuss the results and the approach in section 5. Finally, we draw some conclusions and consider some perspectives in section 6.

ISBN: 952-15-1364-0

2. CONTEXT

The audio track in audio-visual document is a very rich source of information and contains a multitude of indexing keys at both the signal and semantic levels. Audio information constitutes the basis of a large number of approaches in the literature. These approaches distinguish themselves by the specificity of their applicative framework, usually restricted to specific domains.

Audio information is often used alone with great success for audio-visual indexing and retrieval [3][4]. From a semantic point of view, speech is more used than other audio information like music or noise [5]. In [6] the authors propose a method for detecting and tracking different speakers in conversational speech using hidden Markov models. They exploit detection results for document segmentation.

An important use of audio information is the automatic transcription of speech [7][8]. This information brought in a text form gives a lot of information about the content at a level very close to semantics. It permits the use of text based retrieval techniques within audio-visual document bases. This technique has been used in many research works like for instance the ANTS[1] system [9]. The exploitation of the speech transcription and speaker segmentation in order to identify the speaker remains to be done.

2.1. Document Structure

TV news usually has a well defined structure. This structure can help to generate indexing keys. The speaker is often one of the main actors in these documents. We aim at exploiting this information for indexing segments of TV journals. From a general point of view, there exist two ways to perform a segmentation of TV news. The first one in by considering the thematic content: the journal will be seen as a sequence of topics (politics, sports, weather) often separated by jingles, commercials, etc. The second one exploits the studio / reporting structure to produce the segmentation. This is often during this type of segmentation that there are transitions and speaker turns.

Starting from this structure, we aim at identifying the name of the speaker for each segment. We propose to use for this purpose the results of speech transcription and speaker segmentation

1. ANTS [1] (Automatic News Transcription System) developped in the context of the ESTER project

2.2. Audio Segmentation

The objective of the audio segmentation is to obtain coherent audio segments. Each segment has a specific content: silence, noise, music, speech, etc. Many approaches for automatic segmentation have been proposed in the literature. For instance, Kwon and Narayanan have proposed in [11] a technique for document segmentation based on the computation of weighted distances of speaker change points in the audio stream. This approach is useful, especially for the detection of several distinct speakers. Other segmentation techniques, more general, exploit audio characteristics like silences or jingles to infer transitions.

The most useful segmentation for our approach is the segmentation into speakers. Such segmentation is considered as a task belonging to the domain of speech processing. It is generally conducted as a two-step process: the first step is the detection of changes between speakers, often using a Bayesian Information Criteria, and the second step is the merging of similar segments, adjacent or not. Not only the transitions between speakers are detected but also all the segments from a same speaker are identified as such (generally with some errors).

2.3. Linguistic Patterns

In addition to its physical structure (studio / reporting) a TV journal has a linguistic structure. This linguistic structure appears in many occasions when there is a speaker change. Journalists often used marked expressions in their discourse to pass the speech turn to each other or to intervening people. We propose to exploit this structure to set up a technique for speaker identity recognition. The expressions often used for speaker transitions will be treated as linguistic patterns for speaker identification.

The advantage of this type of patterns is that they readily permit to distinguish between the identities of speakers and the identities of all other persons appearing in the speech transcriptions.

3. SPEAKER IDENTITY DETECTION USING LINGUISTIC PATTERNS

Speaker identity detection is obtained from an automatic transcription of what is said [10] and from an automatic speaker segmentation [10][12]. The speaker segmentation detects speech turns between speakers but it also determines whether different segments are from a same speaker. It must be noticed that the speaker segmentation output does not contain any information about speaker identity. No acoustic model of speaker is used here.

Besides the transcription and speaker segmentation, it is also necessary to have access to general knowledge in order to be able to identify named entities. Lists of names, locations and organizations can be used. We got a lot of them from public sources (like Wikipedia).

Our approach is based on a linguistic analysis of the speech transcription in order to identify passages (patterns) from which we can infer a speaker identity.

3.1. Direct Detection

Linguistic patterns are applied to pieces of text that correspond to the transcription of the audio-visual document restricted to a speech segment obtained from speaker segmentation. They correspond to regular expression (usually parameterized) expected to indicate the identity of the person who speaks, who just spoke or who is about to speak. They are applied to each speech segment and, when they are triggered, they permit to predict the speaker identity in the current, previous or next segment. A pattern category is built for each case:

- The first category is for the detection of the identity of the person who is speaking. For instance, a speaker introduces himself: "… This is CNN news I'm [name] …" or during reporting, generally at the end, the person who speaks mentions its identity (often associated to the location or the organization from where he speaks).
- The second category is for the detection of the identity of the person that just spoke, for instance: "Thank you (very much) [name] …".
- The third category is for the detection of the identity of the person which is about to speak. This transition often corresponds to a transition from studio environment to a reporting. For instance, at the end of the discourse of the anchor person: "… [name] has the latest" or "… [name] reports".

A list of patterns has been built for each category. The use of appropriate linguistic patterns permits to discriminate between the identity of a person simply mentioned in the discourse and the identity of an actual speaker. Table 1 shows some example of linguistic patterns from the three classes. Some of these patterns are specific to the corpus that we use (TRECVID 2003 and 2004: TV news from ABC and CNN).

The "Good morning" linguistic pattern is often used to indicate a speaker either in the previous or in the next speech segment. In this case, this is the location where the pattern has been found (close to the beginning or close to the end) that makes the decision.

Patterns for current segment
[name] for ABC news I'm [name] [name] CNN [name] ABC
Patterns for previous segment
thank you … [name] thanks … [name] [name] reporting good morning [name]
Patterns for next segment
tonight with [name] ABC's [name] [name] reports good morning [name]

Table 1: examples of linguistic patterns

Our linguistic patterns are all parameterized. They have at least a parameter for a name, a surname or a couple (name, surname). They can also include parameters for geographic locations, for organizations (CNN or ABC for instance) and/or for expressions like "good morning", "good evening" or "thanks".

For each segment, we analyze the content in order to detect patterns providing the identity of the speaker. Patterns are generally found at the end or at the beginning of speech segments. For instance, the following pattern infers a speech turn between the anchor person and a reporter "… ABC's Sheila Macvicar has the latest."

For the recognition of people's first name, we use a list of 12,400 first names (each tagged as male or female). Family names are negatively filtered from a list of (non-name) English words. Family names almost always appear after a first name. We used a total of 20 linguistic patterns.

3.2. Propagation

In the case of TV news, the speaker usually introduce himself only once during its first speech turn. Therefore, the linguistic pattern approach alone permits to index only small parts of video documents. We have therefore completed it by propagating detected identities to additional speech segments using similarity at the acoustic level. In this part, we propose to propagate a speaker identity to all the segments attributed to a same speaker at the acoustic level (during the speaker segmentation phase) each time an identity has been assigned to any of them by the application of linguistic patterns.

4. EVALUATION

Evaluations have been conducted on a part of the TREC video 2003 corpus [13]. We used the speech transcription and the speaker segmentation distributed with the TREC video corpus and produced by LIMSI [10]. The word error rate has not been evaluated on this corpus but from similar sources, it is estimated to be between 10 and 15%. The speaker segmentation quality has not been evaluated. It includes errors both at the speaker transition level (some transitions between speakers are missed and some are wrongly inserted) and at the speaker clustering level (some segments from different speaker are attributed to a single speaker and some segments from a same speaker are labeled as coming from different speakers). The global quality of the speaker segmentation and clustering was found quite good however and it bas been used as it was provided.

Speaker identity recognition is not an official task of the TRECVID 2003 evaluation campaign. However, this corpus along with its various associated data sets contains everything that is needed for the evaluation of this task, except ground truth about speaker identity. We manually annotated it each time it was available in four TV journals (two from ABC and two from CNN) of about half an hour each. Table 2 summarize the characteristics of the used (sub)corpus.

Total video duration	7009.0 s
Total speech duration	5249.1 s
Total news duration	4250.3 s
Speech within news	3767.1 s
Annotated speech within news	3677.5 s

Table 2: characteristics of the corpus used for evaluation

Table 3 shows the results obtained for identity detection by linguistic patterns. Patterns specialized in the detection of the identity of the speaker in the previous, current and next speech segments permits a prediction in 1.0 %, 6.8 % and 7.0 % of cases (in speech duration) respectively. A direct prediction is therefore made in 14.8 % of cases. Propagation of these predictions using acoustic similarities permits a prediction in 52.7 % of cases. The rates of correctness (accuracy) of the indexed parts are of 100 %, 90.2 %, 74.1 %, 83.3 % and 82.4 % respectively. The source of the documents (ABC or CNN) was not used during the identification process.

Prediction	Predicted duration		Correct duration	
Previous	37.2	1.0 %	37.2	100 %
Current	250.3	6.8 %	225.9	90.2 %
Next	258.2	7.0 %	191.4	74.1 %
Direct	545.8	**14.8** %	454.6	**83.3** %
Propagated	1936.8	**52.7** %	1595.9	**82.4** %

Table 3: Results for speaker identity indexing. The percentage of the predicted duration is relative to the total duration manually annotated. The percentage of the correct duration is relative to the predicted duration.

It can be noticed that the propagation of predictions using acoustic similarities introduce very little errors while it significantly increases the proportion of indexed signal. Errors are due to bad location of speech segment boundaries, to bad associations of segments to a same speaker or to errors in people's name in the speech transcription.

5. DISCUSSION AND FUTURE WORK

The transcription of speech is a very rich source of semantic information. In the case of audio-visual documents, there is generally a relation between audio information (what is said) and visual information (what appears on the screen). In the approach presented in this paper we have exploited the results of speech transcription. The detection of the speaker identity is part of the larger problematic of the modeling of the semantic audio-visual contents for indexing and retrieval.

We plan to extend the linguistic pattern based approach in order to extract additional information about the speaker including the location (city name, country name) from where he speaks, the topics he is talking about, his function, etc. We also consider using this approach for detecting topic changes in the document.

In order to exploit this approach in the context of a system for content-based search in audio-visual documents, it is interesting to integrate the audio and visual information in a way similar to what has been proposed in [14]. However, there is not always a correspondence between the name of a speaker and the image content. Thanks to the structure of TV news, it is possible (for this type of documents) to estimate the conditions under which one can see and listen to a person simultaneously. This information is useful in order to estimate the relevance of systems' answers for users as well as the results' accuracy when speakers are not well known.

The temporal alignment of the audio transcription eases the detection process and permits to determine when speaker turns occur. This is similar to what is done on the visual side during shot segmentation. The duration of a video shot is generally short compared to speech segments. In our approach, we propose to exploit the temporal information in order to associate to each audio segment the corresponding key frame(s) in the visual stream. This eases the evaluation of the results by the user. The user can for instance make a query for "passages in which James Walker speaks" and can quickly extract the relevant answers by looking at the displayed key frames.

6. CONCLUSION

We have proposed, implemented and evaluated an approach for indexing speaker identity into audio-visual documents for content-based retrieval. We defined a set of linguistic patterns for speaker identity detection. These patterns are of three types in order to detect the identity of the speaker in the previous, current or next speech segment (relative to the segment in which they appear).

The main part of the work has been to analyze the text of the transcription coming from ASR. Several semantic information are exploitable in the transcription such as locations, organizations and people.

Evaluations conducted on a part of the TRECVID 2003 corpus have shown that the proposed method permit a speaker identity indexing for about 53 % of speech time in TV news with an accuracy of about 82 % (47 % of speech signal is not indexed, 43 % is correctly indexed and 10 % is wrongly indexed).

Speaker indexing permits both a conceptual indexing and a relational indexing: the knowledge of speaker identity allows to infer the relations "speaks" or "speaks about".

In the case of TV news, it is sometimes very difficult to determine from the audio track only who the speaker is. For instance, the identity of intervening people is often displayed in the image track and rarely cited in the discourse. Therefore, we plan to complete our detection approach using additional information sources like automatic text recognition in images. Also, detecting the identity of intervening people can still be done using linguistic patterns but less frequent ones.

7. ACKNOWLEDGEMENTS

This work has been sponsored by an INPG BQR project and by the PENG – IST-004597 European project.

8. REFERENCES

[1] Fiscus, J., Doddington, G., Garofolo, J., and Martin, A.: Topic Detection and Tracking, Evaluation (TDT), Fifth European Conf. On Speech Comm. and Tech., Vol. 4, pp. 247-250, 1998.

[2] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain: "Speaker Diarization from Speech Transcripts". In International Conference on Speech and Language Processing, pages 1272-1275, Jeju Island, October 2004.

[3] T. Kemp and A. Waibel: Reducing the Oov Rate In Broadcast News Speech Recognition, In Proceedings of the ICSLP, Sydney, Australia, 1998.

[4] A. Albiol, L. Torres, E. J. Delp: Video Preprocessing for Audiovisual Indexing. In 5th IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI) April 2002.

[5] J. Pinquier, C. Sénac and R. André-Obrecht: "Speech and music classification in audio documents", in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2002) Orlando, Floride, Mai 2002.

[6] M. K. Sönmez, L. Heck, M. Weintraub: ``Speaker Tracking and Detection with Multiple Speakers," in Proc. EUROSPEECH 99, Volume 5, Page 2219-2222 Budapest, Hungary, September 1999.

[7] Garofolo John S., Ellen M. Voorhees, Cedric G. P. Anzanne, and Vincent M. Stanford: "Spoken document retrieval: 1998 evaluation and investigation of new metrics". In Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio, pages 1-7, Cambridge, UK, April, 1999.

[8] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman: "Transcriber: development and use of a tool for assisting speech corpora production", Speech Communication special issue on Speech Annotation and Corpus Tools, Vol 33, No 1-2, January 2000.

[9] A. Brun, C. Cerisara, D. Fohr, I. Illina, D. Langlois, O. Mella, K. Smaïli : "ANTS : le système de transcription automatiquedu LORIA" Journées d'Etude sur la Parole (JEP'04) se tiendra, du 19 au 22 avril 2004 Fes, Maroc.

[10] J.L. Gauvain, L. Lamel, and G. Adda: The LIMSI Broadcast News transcription system, in Speech Communication 37, 2002, pp. 89-108.

[11] S. Kwon and S. Narayanan: "Speaker Change Detection Using a New Weighted Distance Measure", In Proceedings of International Conference Spoken Language Processing. Denver, Colorado, U.S.A., September 16-20, 2002.

[12] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J-F Bonastre, Segmentation selon le locuteur : les activités du Consortium ELISA dans le cadre de Nist RT03", JEP 2004, Fès, Maroc, 19-22 avril 2004.

[13] Smeaton A., Kraaij W., and Over P.: "TRECVID 2003 - An Introduction", Text REtrieval Conference TRECVID Workshop, Gaithersburg, Maryland, 17-18 November 2003.

[14] J. Yang, A. Hauptmann, M-Y. Chen: "Finding Person X: Correlating Names with Visual Appearances", International Conference on Image and Video Retrieval (CIVR'04), Dublin City University, Ireland, July 21-23, 2004.