Audio-Indexing : what has been accomplished and the road ahead

Ivan Magrin-Chagnolleau, Nathalie Vallès-Parlangeau

AUDIO INDEXING: WHAT HAS BEEN ACCOMPLISHED AND THE ROAD AHEAD

Ivan Magrin-Chagnolleau[(1)] *and Nathalie Parlangeau-Vallès*[(2)]

[(1)] Laboratoire Dynamique Du Langage (UMR 5596)
Université Lumière Lyon 2 & CNRS – 14, avenue Berthelot – 69363 Lyon Cedex 07 – France

[(2)] LORIA, INRIA Lorraine
Campus Scientifique – BP 239 – 54506 Vandœuvre-les-Nancy – France

ivan@ieee.org - Nathalie.Parlangeau-Valles@loria.fr

ABSTRACT

This paper presents an overview of audio indexing, which has emerged very recently as a research topic with the development of Internet. A lot of data, including audio data, are currently not indexed by web search engines, and audio indexing consists in finding good descriptors of audio documents which can be used as indexes for archiving and search. We discuss speech/music segmentation, language identification, speaker tracking and speaker indexing, and propose some research directions for other audio descriptors which have not been used in the framework of audio indexing, namely key sounds detection, keywords detection, and themes detection. We finally conclude this overview and give a few promising and key perspectives.

1. INTRODUCTION

Internet has become a very important vector of communication during the last few years. There are currently about 400 millions of Internet users, 2 billions of pages directly accessible on 7 millions of public servers[1]. There is an urgent need to classify this huge amount of data. Most of the search engines currently access mainly the HTML pages (or equivalent textual data). But there is an important part of the data which is not accessible, because the data are not indexed, have a dynamic content, or belong to a category which is not easily indexable. All these data belong to what is called the *invisible web*, including audio and video documents.

In this paper, we describe some techniques used to structure, and possibly index, audio documents. Although automatic transcription has reached a good performance level recently, it is not the only kind of information which can be extracted from an audio document. A lot of non-verbal information are also very structuring for an audio document, and can lead to the extraction of pertinent descriptors. For instance, as seen on Fig. 1, we can separate speech segments from music segments, detect key sounds (like jingles), identify the language of a segment, track a known speaker or segment by speakers, detect some keywords, or extract the main themes.

[1] See http://wwwelec.unice.fr/pages/veille/veille.html, in French.

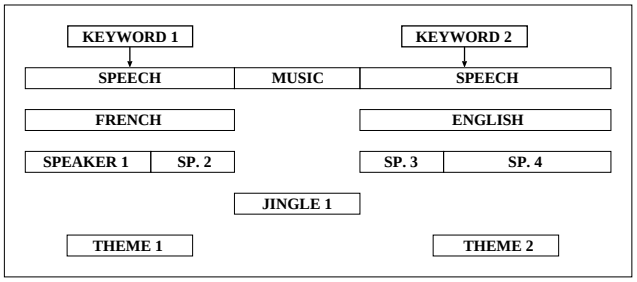

Fig. 1. Some descriptors which can be extracted from an audio document.

2. SPEECH/MUSIC TRACKING AND SEGMENTATION

The first task of interest is the tracking of speech and/or music segments in order to segment an audio document into speech and music portions. Most of broadcast news transcription systems use this kind of separation of speech/music segments in order to confine the application of speech transcription systems on speech segments[1]. This segmentation is often related to a specific structuration of the document (advertisements, jingles, etc.), and it seems important to keep it as a descriptor in an audio indexing system.

Generally, speech/music segmentation is based on the acoustic differences between the two sorts of sounds. [2, 3, 4, 5] have based their discrimination on sets of indices like the mean and the variance of the zero-crossing rate, the spectral "flux" which is defined as the balancing point of the spectral magnitude distribution, etc. All these features usually have interesting discriminating properties as they take very different values for each category of sound.

Various classifiers are commonly used: Gaussian Mixture Models (GMM), k-Nearest-Neighbors (kNN), Neural Networks (NN), etc. In [6, 7], a classical cepstral parame-

terization associated with GMM's gives good results.

What seems important in speech/music separation is the notion of independency between the two tracked cues. In [Fon00], the "differentiated modeling" permits to exploit the structural differences between speech and music. Each class has its own set of parameters (cepstra for speech, filters for music) and models (a 32-component GMM for speech, and a 10-component GMM for music). The decision is totally independent and a module of fusion gives the final segmentation, including mixed segments.

3. LANGUAGE IDENTIFICATION

Language identification consists in identifying the language of an audio document or of a segment in an audio document. This task has emerged recently in the framework of multilingual speech recognition, but has not been used in the framework of audio indexing.

The state-of-the-art approach uses a cepstral-based acoustic analysis and several language-dependent phone recognizers followed by n-gram language models specific of each language to identify [8]. This approach requires that labeled training data be available for several languages, but not necessary for all.

Among the alternative approaches that have been proposed, some researchers work at the level of the acoustic analysis [9, 10], and other propose to explore various cues to identify a language, as rythm [11] for instance.

4. SPEAKER TRACKING

Speaker tracking consists in estimating a beginning and an ending time for each segment in which a target speaker is speaking.

This task is usually tackled by a statistical approach. To detect a target speaker in an audio document, two models are usually built: a target speaker model and a background model which is intended to represent speech from speakers other than the target speaker or other types of sounds. These models are built using feature vectors extracted from labeled segments. In the case of several target speakers, at least one model for each target speaker is needed. Finally, more generally, there could be several background models, representing different types of non-target speaker sounds, and several models for each target speaker, representing different speech qualities. This approach has been adopted for instance in [12, 13].

5. SPEAKER INDEXING

Speaker indexing of an audio database consists in organizing the audio data according to the speakers present in the database. It is composed of three steps (Figure 2). The first step is the segmentation of each audio document by speakers. The segmentation produces a set of speaker-based segmented portions, that is, a set of *speaker utterances*. The second step consists in tying the various speaker utterances among several previously segmented audio documents. During this stage, one label is attributed to all the speaker utterances matched together. The last stage corresponds to the creation of a speaker-based index.

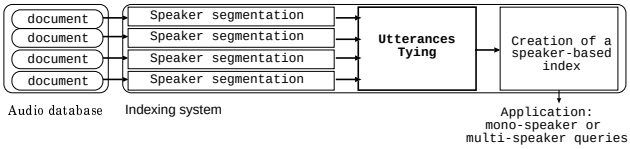

Fig. 2. Block-diagram of a speaker indexing system.

The speaker segmentation problem is usually addressed by one of the two following methods. The first one (described in [14][15][16]) relies on a speaker change detection followed by a clustering step. The second one (see for instance [17][18]) does the segmentation and the clustering simultaneously using a hidden Markov model. In both cases, the system has to determine the set of speakers present within a given audio document as well as the boundaries of each intervention. No *a priori* information on speakers is used in these approaches, *i.e.* the speaker utterance models are built during the segmentation process.

Speaker utterances tying [19] is a classification problem similar to speaker clustering [15][16]. Speaker clustering is usually done inside one audio document, whereas, in speaker utterances tying, utterances are matched among several audio documents. However, similar segments inside one audio document are already matched during the preliminary segmentation process. Moreover, the matched utterances are longer than the segments used in speaker clustering. But the great number of channels in the set of audio documents represents an additional difficulty.

The last step of speaker indexing is the creation of a speaker-based index which remains an open and difficult problem for real applications. There are very few papers on this topic. The aim of a speaker-based index is to organize the matched speakers to make the search in a database more efficient.

6. OTHER TYPES OF INFORMATION

Broadcast news systems are based on a complete transcription of speech segments. As far as indexing systems are concerned, moreover for web search applications, keyword detection is preferred. Keyword spotting consists in detecting a more or less important set of keywords from the speech stream. This process gives the exact time position of a keyword.

Word-spotting systems based on hidden Markov models are considered more efficient at modeling arbitrary speech than template based systems [20, 21]. Two main approaches

are found in the literature. The most obvious is to use a large vocabulary continuous speech recognition system (LVCSR) to produce a word string. Then, search algorithms are applied for keyword detection in that string [22]. This approach is considered as giving the best results [23, 24]. Another common approach is based on the use of keyword and filler models. These latest represent the non-keyword intervals of the utterance [23]. Models can be sub-word keyword models like phonetic models or can be whole-word models. [25] gives a very interesting overview of re-scoring methods applied to these two kinds of models: sub-word models are shown to yield a higher hit rate. An original approach based on a HMM-based acoustic decoder combines a multi-keyword spotter and a RNN prosodic model [26]. Prosodic information intends to give better boundaries of words assuming that speakers emphasize keywords more than non-keywords in a sentence. The last relevant approach is based on a modification of the Viterbi search to compute normalized scores which indicate the matching of the keywords at distinct time positions in the utterance. Scores of the keywords are compared with decision thresholds [27]. This is used efficiently as input in a post-processor stage in [28]. Although LVCSR approaches give better results, it requires task-specific knowledge and large training databases. This have a high computational cost. The filler approach is interesting to avoid this last inconvenient and also to maintain domain independence. But detailed filler models increase the performances and then most of the disadvantages of this approach are lost. Systems based on a real acoustic decoder are interesting in terms of domain independence too. This last method seems to have the advantages of needing quite a small amount of training data and does not require filler models.

The topic detection and tracking (TDT) program proposes a definition of a topic as "a seminal event or activity along with all directly events and activities" [29]. As far as speech is concerned, a topic is commonly defined by a set of relevant words [30]. This implies the use of an automatic speech recognizer as a front-end. Recent approaches consider that phoneme-based methods are more accurate. Major advantages of such an approach have been pointed out in [31]. An interesting article presents a method based on phoneme n-grams [32]. Documents or parts of it will be considered to be "on topic" whenever the story is directly connected to the associated event. We can consider three different tasks: story segmentation, topic detection, and topic tracking. The story segmentation is the task of segmenting the stream of speech into topically cohesive stories. The tracking task consists in associating incoming stories with known topics. Topic detection is the task of detecting and tracking topics not previously known by the system. These three tasks can be integrated into an automatic speech indexing system, even if the most useful is usually the tracking task. The association between a basic unit (word or sequence of phonemes) and a topic is then based on similarity measures, or occurrence frequencies between what is called the conversational vector and the topic-specific vector. A lot of problems are addressed through the topic tracking task and not all problems are yet resolved.

Finally, if speech cues are well known, music segments can also be exploited to find particular cues like rythm [33], musical sentences [34], pitch [35], notes, instrumental parts, singing voices, melody [36, 37], etc.

7. CONCLUSION AND PERSPECTIVES

In this paper, we have presented the current state of research in audio indexing. We have discussed speech/music segmentation, language identification, speaker tracking and speaker indexing, and proposed some research directions for other audio descriptors which have not been used in the framework of audio indexing, namely key sounds detection, keywords detection, themes detection, and other musical descriptors. A lot of work needs to be done in audio indexing. In particular, after the extraction of audio descriptors, there is another important step consisting in organizing all these descriptors to make the search and the navigation in an audio database more efficient. We intend to tackle some of these issues in the ongoing RAIVES project, which is a research project on audio indexing. We also hope that this paper will stimulate some research in that field, which still needs to be largely explored.

8. AKNOWLEDGMENT

We would like to thank the CNRS (French national center for scientific research) for its support to the RAIVES project. We also would like to thank Régine André-Obrecht, Frédéric Bimbot, Jean-François Bonastre, Sylvain Meignier, François Pellegrino, and Christine Sénac for their help and the fructuous discussions.

9. REFERENCES

[1] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda, "Audio partitionning and transcription for broadcast data indexation," in *Proceedings of CBMI 99*, 1999, pp. 67–73, Toulouse, France.

[2] John Saunders, "Real-time discrimination of broadcast speech/music," in *Proceedings of ICASSP 96*, 1996, pp. 993–996, Atlanta, Georgia, United States.

[3] Eric Scheirer and Malcolm Slaney, "Construction and evaluation of a robust multifeatures speech/music discriminator," in *Proceedings of ICASSP 97*, 1997, pp. 1331–1334, Munich, Germany.

[4] Eluned S. Parris, Michael J. Carey, and Harvey Lloyd-Thomas, "Feature fusion for music detection," in *Proceedings of EUROSPEECH 99*, 1999, pp. 2191–2194, Budapest, Hungary.

[5] Michael J. Carey, Eluned S. Parris, and Harvey Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *Proceedings of ICASSP 99*, 1999, pp. 149–152, Phoenix, Arizona, United States.

[6] Mouhamadou Seck, Frédéric Bimbot, Didier Zudaj, and Bernard Delyon, "Two-class segmentation for speech/music detection in audio tracks," in *Proceedings of EUROSPEECH 99*, 1999, pp. 2801–2804, Budapest, Hungary.

[7] Mouhamadou Seck, Ivan Magrin-Chagnolleau, and Frédéric Bimbot, "Experiments on speech tracking in audio documents using gaussian mixture modeling," in *Proceedings of ICASSP 01*, 2001, pp. 601–604, Salt Lake City, Utah, United States.

[8] Marc A. Zissman, "Comparison of four approaches to automatic language identification for telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, January 1996.

[9] Michal Dutat, Ivan Magrin-Chagnolleau, and Frédéric Bimbot, "Language recognition using time-frequency principal component analysis and acoustic modeling," in *Proceedings of ICSLP 00*, Dec. 2000, Beijing, China.

[10] Michel Dutat, Ivan Magrin-Chagnolleau, and Frédéric Bimbot, "Acoustic modeling of spoken languages using time-frequency principal component analysis and hidden markov models: Application to language identification," *Submitted to IEEE Transactions on Speech and Audio Processing*.

[11] Jérôme Farinas and François Pellegrino, "Automatic rythm modeling for language identification," in *Proceedings of EUROSPEECH 01*, 2001, pp. 2539–2542, Aalborg, Denmark.

[12] Aaron E. Rosenberg, Ivan Magrin-Chagnolleau, S. Parthasarathy, and Qian Huang, "Speaker detection in broadcast speech databases," in *Proceedings of ICSLP 98*, 1998, pp. 202–205.

[13] Ivan Magrin-Chagnolleau, Aaron E. Rosenberg, and S. Parthasarathy, "Detection of target speakers in audio databases," in *Proceedings of ICASSP 99*, 1999, pp. 821–824, Phoenix, Arizona, United States.

[14] S. Chen, J.F. Gales, P. Gopalakrishnan, R. Gopinath, H. Printz, D. Kanevsky, P. Olsen, and L. Polymenakos, "IBM's LVCSR system for transcription of broadcast news in the 1997 HUB4 English evaluation," in *Proceedings of the DARPA speech recognition workshop*, 1998, http://www.nist.gov/speech/publications/darpa98/.

[15] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *Proceedings of ICASSP 98*, 1998, Munich, Germany.

[16] D.A. Reynolds, E. Singer, B.A. Carlson, J.J. McLaughlin G.C. O'Leary, and M.A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proceedings of ICSLP 98*, 1998, pp. 610–613, Sydney, Australia.

[17] Sylvain Meignier, Jean-François Bonastre, and Stéphane Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *Proceedings of 2001: A Speaker Odyssey*, June 2001, pp. 175–180, Chania, Crete, Greece.

[18] Lynn D. Wilcox, Don Kimber, and Francine Chen, "Audio indexing using speaker identification," in *Proceedings of the SPIE Conference on Automatic Systems for the Inspection and Identification of Humans*, 1994, pp. 149–157, San Diego, California, United States.

[19] Sylvain Meignier, Jean-François Bonastre, and Ivan Magrin-Chagnolleau, "Speaker utterances tying among speaker segmented audio documents using hierarchical classification: Towards speaker indexing of audio databases," in *Submitted to ICASSP 02*, 2002.

[20] Richard C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," in *Proceedings of ICASSP 90*, 1990, pp. 129–132, Albuquerque, New Mexico, United States.

[21] Lynn D. Wilcox and Marcia A. Bush, "HMM based wordspotting for voice editing and indexing," in *Proceedings of EUROSPEECH 91*, 1991, pp. 25–28, Genova, Italy.

[22] Mitch Weintraub, "Keyword-spotting using SRI's DECIPHER large-vocabulary speech recognition system," in *Proceedings of EUROSPEECH 93*, 1993, pp. 1265–1268, Berlin, Germany.

[23] Alexandros Manos and Victor Zue, "A segment-based worspotter using phonetic filler models," in *Proceedings of ICASSP 97*, 1997, pp. 899–902, Munich, Germany.

[24] Mazin Rahim, "Recognizing connected digits in a natural spoken dialog," in *Proceedings of ICASSP 99*, 1999, pp. 153–156, Phoenix, Arizona, United-States.

[25] K. M. Knill and Steve J. Young, "Speaker dependent keyword spotting for accessing stored speech," Tech. Rep., CUED/F-INFENG/TR-193, Cambridge University Engineering Department, 1994.

[26] Wern-Jun Wang, Chun-Jen Lee, Eng-Fong Huang, and Sin-Horng Chen, "Multi-keyword spotting of telephone speech using orthogonal transform-based sbr and rnn prosodic model," in *Proceedings of EUROSPEECH 01*, 2001, pp. 2773–2776, Aalborg, Denmark.

[27] Jochen Junkawitsch, Günter Ruske, and Harald Höge, "Efficient methods for detecting keywords in continuous speech," in *Proceedings of EUROSPEECH 97*, 1997, pp. 259–262, Patras, Greece.

[28] Luciana Ferrer and Claudio Estienne, "Improving performance of a keyword spotting system by using a new confidence measure," in *Proceedings of EUROSPEECH 01*, 2001, pp. 2561–2564, Aalborg, Denmark.

[29] Yiming Yang, Tom Pierce, and Jaime Carbonell, "A study on retrospective and on-line event detection," in *Proceedings of SIGIR 98, 21st ACM International Conference on Research and Development in Information Retrieval*, 1998, pp. 28–36, Melbourne, Australia.

[30] Hubert Jin, Rich Schwartz, Sreenivasa Sista, and Frederick Walls, "Topic tracking for radio, TV broadcast, and newswire," in *Proceedings of EUROSPEECH 99*, 1999, pp. 2439–2442, Budapest, Hungary.

[31] Jeremy H. Wright, Michael J. Carey, and Eluned S. Parris, "Statistical models for topic identification using phoneme substrings," in *Proceedings of ICASSP 96*, 1996, pp. 307–310, Atlanta, Georgia, United States.

[32] Marius W. Theunissen, Konrad Scheffler, and Johan A. du Preez, "Phoneme-based topic spotting on the switchboard corpus," in *Proceedings of EUROSPEECH 01*, 2001, pp. 283–286, Aalborg, Denmark.

[33] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton, "Classification, search, and retrieval of audio," in *Handbook of Multimedia Computing*. CRC Press, 1999.

[34] Chih-Chin Liu, Jia-Lien Hsu, and Arbee L.P. Chen, "Efficient theme and non-trivial repeating pattern discovering in databases," in *Proceedings of ICDE 99*, 1999, pp. 14–21, Sydney, Australia.

[35] Yoav Medan, Eyal Yair, and Dan Chazan, "Super resolution pitch determination of speech signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 39, no. 1, pp. 40–48, 1991.

[36] Tomonari Sonoda, Masataka Goto, and Yoichi Muraoka, "A www-based melody retrieval system," in *Proceedings of ICMC 98*, 1998, pp. 349–352, Ann Arbor, Michigan, United States.

[37] Rodger J. McNab, Lloyd A. Smith, David Bainbridge, and Ian H. Witten, "The New Zealand digital library: MELody inDEX," *D-Lib Magazine*, May 1997, http://www.dlib.org/.