CHAPTER THREE

Audio Retrieval and Navigation Interfaces

Philippe Aigrain
European Commission[1] INFSO/E2

SELECTING and presenting the essential readings in multimedia cannot be done without some working definition of multimedia research. This requirement is even more obvious when introducing a chapter on audio retrieval and navigation interfaces. For 30 years, some very fine researchers have developed and matured the technologies of speech and music processing; have studied and modeled auditory perception; and have invented ways of analyzing, representing, and synthesizing sound signals. Without these advances, and without more general work on information retrieval and human-computer interaction, none of the papers presented in this chapter could have been written, and their authors would not have even imagined attempting what they have set as a goal. The recent developments of audio retrieval and navigation interfaces, of which this chapter presents highlights, is the history of the meeting between multimedia research and music and speech technology research. In multimedia, we have witnessed the application-driven integration and redefinition of processing technologies with interaction and retrieval techniques. In parallel, in each of the individual fields, researchers have tried to find ways to eliminate the barriers that prevent their techniques from delivering full benefits. For each of the papers reproduced in this chapter, there are important counterparts in specialized speech, music, human-computer interaction, or information retrieval literature. One of the purposes of this introduction is do pay justice to the preexisting specialized work that has made mutimedia content processing technology possible. The introduction also refers to several papers, in particular in emerging domains, that would have been included in this collection if size limitations had not prevented it. A recent survey of audio retrieval technology, more specifically focused on speech-based processing, can be found in the work of J. Foote [1]. A specific survey of technology and applications of content-based retrieval of music can be found in my paper of 1999 [2].

A SHORT HISTORY

Multimedia research was initially slow to tackle audio retrieval and representation. Part of the reason lies with the privilege of video in the general social imagination of our societies. Whatever measure you use, audio and, in particular, music are no less important than video in terms of business revenues, entertainment and culture, employment, or communication. But attention was focused on video and television. Even within multimedia applied to video and television, the importance of audio soundtrack processing and representation was only very slowly recognized. Nonetheless, if we look back at the initial developments of multimedia, in particular in the context of hypertext, some of the challenges were clearly identified

[1] *Views presented in this paper are only the author's and do not necessarily represent the official position of the European Commission.*

early on. Christodoulakis and Graham [3] tried to tackle in a pioneering paper the difficult issue of how to represent sequences of time-based media contents in hypertext navigation interfaces. This line of research has been very fruitful in the moving image domain, in which it led to the mosaic image and video icon techniques for representing a video shot statically. But in the audio world, this remains largely an unsolved problem, and as we will see, recent work seems to point at playback in sound space as a possible alternative to static image representation of a sound sequence.

The next stage of identification of problems was the progressive recognition that the general audio retrieval and navigation category actually covered a number of quite different areas:

- retrieval of and navigation between segments of a sound document
- indexing and retrieval of short individual sounds in sound databases
- retrieval of sound documents in a relatively consistent set of documents (for instance, songs)
- retrieval of and navigation between segments of a video document using audio, speech, or music information
- retrieval by word spotting in speech databases or video databases with speech indexing

The more general problem of unrestricted audio retrieval (for example, "here is a sound extract, find me all similar ones") is still largely unexplored. It is also ill defined, because retrieval can occur only when it is possible to specify the nature and scale of components that are searched for. One of the reasons for multimedia research's success in delivering new technology and finding overlooked potential of existing technology is that it has worked within precisely defined paradigms of usage. This chapter provides the reader with essential references on a variety of these domains.

In recent months (following the selection for papers in this chapter), the main trend has been toward strengthening technique and evaluating them in a more systematic manner. The MPEG-7 standardization effort is playing an important role in that matter, in particular because it has reached the "core experiments" stage. Researchers in this rapidly growing research community are experimenting with techniques for spoken content retrieval, timbre classification, sound effects classification, and segmentation.

GENERAL SEGMENTATION AND LABELING TECHNIQUES

Hawley [4] was the first to clearly set out a research program for automatically producing structured representations of audio documents that are fit for summarization, indexing, and retrieval. He also suggested some techniques for making first steps in the direction of speech/music discrimination and developed several music indexing features. In line with Hawley's work, the fourth paper in this chapter, "Toward content-based audio indexing and retrieval and a new speaker descrimination technique," by L. Wyse and S. W. Smoliar, originally presented in 1995, describes a number of techniques for segmenting a sound stream or document and associates content-based labels (such as speech/music discrimination and locutor discrimination) with the segments. The notion of a global processing chain for segmenting and labeling audio content has been pursued more recently in work conducted at IRCAM [5].

BROWSING AND NAVIGATING WITHIN A MUSIC DOCUMENT

Video browsing is a well-established domain of multimedia research and is covered in Chapter 5 of this book. What makes video browsers useful is that they provide a structured access and navigation interface for dealing with the opacity of video documents and for coping with their temporal nature. This opacity is even more problematic in the case of audio documents for reasons that can be illustrated by considering the impossibility of stopping on sound[2] while keeping any meaningful hindsight on the contents. The general privilege of visual human-computer interaction also contributes to the problem. As early as in the 19th century, researchers have tried to develop static visual representations of sound productions. This line of work has been particularly active within ethnomusicology [6], and spectrographical imaging [7] and has brought about new developments, including its association with user annotation [8]. When audio content can be associated with a pre-

[2] *You can loop on short stationary sound extracts, and this feature can be useful for analyzing some very low-level features, but certainly is not for representing a more general context.*

scriptive representation, whether a musical score or a speech transcript, they can be used as a representation upon which a browsing interface can be based. But this approach faces the extreme difficulty of automatic transcription and also suffers from the fact that the representation hides some important features of the content (for instance, voices for speech or performance for music). Static representations of audio content take all their value when they can be presented at different scales and levels of abstraction and directly associated with sound production.

The first paper in this chapter, "Representation-based user interfaces for the Audiovisual Library of Year 2000,"by myself, P. Joly, P. Lepain, and V. Longueville, proposed a framework for defining music browsers, using a detailed comparison with a similar approach on video. At the time of its writing (1995) some aspects of the paper were still "works in progress," but they were further evolved into a complete piece of software [9], that was experimentally used by music specialists [10]. One key challenge for establishing a realistic parallel between video and music browsing and indexing is the ability to define significant segments in music, which can then be used as units for representation, interaction (for instance, selection and gravity), navigation, and matching.

MELODY RETRIEVAL AND THE KARAOKE PROBLEM

Querying for a given melodic *motif* is a problem that has attracted many researchers, for reasons that arise from the privilege that Western music gives to pitch and melody. The Karaoke situation has provided a simplified context in which this problem could be explored. In this context a singer hums a melody (or starts singing it) and the system is supposed to answer by recognizing which song is being sung and by then providing the corresponding musical accompaniment. Japanese research laboratories started work on this problem in the early 1990s, and an example of such work was reported in by Kageyama and colleagues [11]. A query-by-humming system consists of a pitch tracker (including segmentation in notes) and a matching algorithm in a database of melodies. Approximate matching has to be used, to deal with deletions or insertions of notes resulting from ornamentation or simply from errors of either the singer or the pitch-tracking algorithm. The representation of the melody to be matched is also an essential aspect. The melody is always represented as a sequence of relative intervals to make it independent of key changes. Kageyama et al. used intervals expressed in semi-tones and computed a matching distance using the number of matches with one semi-tone difference. The second paper in this chapter "Query by humming: Music information retrieval in an audio database," by A. Ghias and colleagues, describes a technique using melodic contour (that is, a succession of up and down or zero intervals that ignores the size of the up and down intervals). Researchers in auditory perception have pointed out that melodic contour is indeed an essential representation for the human matching of melodies [12]. Ghias and colleagues also proposed a simple evaluation framework. In these simplified contexts, the melody database is monophonic. Of course the next challenge is to be able to match against polyphonic content. Assuming that polyphonic pitch detection is available[3], querying for melodies—from an example or humming—calls for extracting some salient melody either by voice segregation, such as that proposed by Uitdenbogerd and Zobel [13] or by harmonic analysis. Matching in the presence of ornamentation and other modifications to the searched motif, or matching using rules that fit whose structure and tonality varies greatly from that of Western music are difficult challenges. Melodic matching in large databases of musical recordings is still very much unexplored, and we presently lack a clear idea of the scalability of the existing methods. Most recent approaches aim at matching multiple features rather than melodic information only.

RETRIEVAL IN SOUND DATABASES

Content-based retrieval in large audio databases is an easier problem for databases of short sounds, such as the Foley sounds that are used for soundtracks in video or film. The third paper in this chapter, "Content-based classification, search, and retrieval of audio," by E. Wold and colleagues, reports work conducted within the company Muscle Fish. Wold and colleagues use

[3] *See the* Journal of New Music Research *special issue,* Content Processing of Music for Multimedia Application *27, no. 4 (1999), for a review and a number of state-of-the-art papers on the subject.*

multiple features to index short sounds. The choice of features is based on psychoacoustical knowledge, and classification/retrieval is based on classical data analysis techniques. The paper discusses integration with browsers and databases and contains a short discussion of the extension of the proposed methods to retrieval of longer sounds (phrases, video soundtracks, raw audio). A preliminary version of the paper appeared as a communication at the IJCAI '95 Workshop on Intelligent Multimedia Information Retrieval and was also published in 1997 [14]. Multifeature indexing is being extended in recent work to enable the identification of pre-indexed musical recordings from a short extract. For an example, see the RAA project Web site [15]. This process has applications in intellectual property rights management, online music sales, and musical archives.

INDEXING AND NAVIGATING VIDEO SOUNDTRACKS

Researchers long recognized that the audio soundtrack of video contains essential information for retrieval of relevant sequences. In the first paper in the chapter: Aigrain and colleagues proposed the representation of the audio track in a video browser based on classification of music, speech, and noise segments. Shahraray and Gibbon [16] and Maybury and colleagues [17] demonstrated the potential of speech transcripts such as closed captions for browsing, retrieval ,and summarization. But it was through work initiated in 1995 within the Informedia project at Carnegie-Mellon [18], through research on video browsers at FX-PAL [19], and through the Virage Audio logger™ [20] (using technologies from IBM and Muscle Fish) that audio soundtrack information truly became integrated in video indexing and browsing.

RETRIEVAL IN SPEECH DATABASES

Retrieval in speech databases was first explored on real-scale databases for single-speaker databases, using speaker-dependent training of Hidden Markov Models. The fifth paper in this chapter, "Open-vocabulary speech indexing for voice and video mail retrieval," by M.G. Brown and colleagues, generalized these techniques to obtain speaker-independent word spotting capabilities on open-vocabulary databases. In the case of this paper, these techniques are used for indexing and retrieval in voice and video mail systems. Similar techniques were developed by Meteer and colleagues [21] at BBN/GTE and integrated in applications such as broadcast news soundtrack retrieval.

BROWSING IN SOUND SPACE AND SONIC BROWSING

Sound spatialization techniques open new dimensions to listening. The first obvious application is to allow the listener to explore sound space itself, for instance, by moving closer to some sound sources. Pachet and Delerue [22] have developed an intriguing interactive listening system based on constrained control of a sound spatialization system. The constraints express, for instance, the need to maintain some balance between sound sources. Browsing musical spaces can also be used for sound information retrieval. Fernstrom and Bannon [23] have proposed an approach to sonic browsing with a clear generic potential, though its range of applicability remains to be explored. Sound documents (for instance musical recordings) are analyzed and the features extracted are projected in 3-D space. A document becomes a point in this space, which is assumed to be absorbent for sound. The user can then browse through this space, using a sound spatializer, hearing each document with intensity according to its distance. The quality of such browsing interfaces rests on the relevance of the extracted features and on the ability of the user to navigate meaningfully in the projected feature space.

CONCLUSION

Despite recent developments, content-based retrieval of audio still faces major challenges. Efficient indexing and retrieval for large databases of music and for speech databases with many different locutors has still to be demonstrated. An approach combinating navigation techniques and query/retrieval techniques is still in its infancy. Nevertheless, a core set of techniques has been developed, from which more systematic work can proceed.

REFERENCES

[1] J. Foote, "An overview of audio retrieval," *Multimedia Systems* 7, no. 1 (1999): 2–10.

[2] P. Aigrain, "New applications of content processing of music," *Journal of New Music*

Research 28, no. 4 (1999): 271–80.

[3] S. Christodoulakis and S. Graham, "Browsing within time-driven multimedia documents," *ACM SIG-OIS* 9, nos. 2–3 (1988): 219–27.

[4] M. Hawley, "Structure out of sound," Ph. D. dissertation Cambridge: MIT Media Lab, 1993.

[5] S. Rossignol, X. Rodet, J. Soumagne, J-L. Colette, and P. Depalle, "Automatic characterization of musical signals: Feature extraction and temporal segmentation," *Journal of New Music Research* 28, no. 4 (1999): 281-295.

[6] T. Ellington, "Transcription," in H. Myers, ed., *Ethno-Musicology: An Introduction* (London: W.W. Norton, 1992), 110–52.

[7] R. Cogan, *New Images of Musical Sounds* (Cambridge: Harvard University Press, 1984).

[8] D. Besson, "La transcription des musiques électro-acoustiques: Que noter, comment et pourquoi," *Analyse Musicale* 24 (1991): 37–41.

[9] P. Lepain, "SATIE: An interactive software for listening to musical recordings," in *Proceedings of the 4th ACM International Multimedia Conference* (ACM, 1996): 413–14.

[10] P. Aigrain, and P. Lepain, "Le groupe Ecoute Interactive de la Musique de la Bibliothèque Nationale de France," *Actes des Journées d'Informatique Musicale*, (May 1996): 128–38).

[11] T. Kageyama, K. Mochisuki, and Y. Takashima, "Melody retrieval by humming," in *Proceedings of the International Computer Music Conference 1993* (San Francisco: International Computer Music Association, 1993), 349–51.

[12] S. Handel, *Listening: An introduction to the perception of auditory events* (Cambridge: MIT Press, 1989).

[13] A. Uitdenbogerd and J. Zobel "Manipulation of music for melody matching," in *Proceedings of the 6th ACM International Multimedia Conference* (ACM, 1998), 235–40.

[14] T. Blum, D. Keislar, J. Wheaton, and E. Wold, "Audio databases with content-based retrieval," in Mark Maybury, ed., *Intelligent Multimedia Information Retrieval* (AAAI Press/MIT Press, 1997), 113–35.

[15] RAA, http://raa.joanneum.ac.at, 2000.

[16] B. Shahraray and D. C. Gibbon "Automatic generation of pictorial transcripts of video programs," in P*roceedings of the IS&T/SPIE 95—Digital Video Compression: Algorithms and Technologies*, SPIE proceedings 2419, 512–19.

[17] M. Maybury, A. Merlino, and D. Morey, "Broadcast news navigation using story segments," in *Proceedings of the 5th ACM International Multimedia Conference* (ACM, 1997), 381–91.

[18] A. G. Hauptmann and M. J. Witbrock, *Informedia News-On-Demand: Using Speech Recognition to Create a Digital Video Library*, http://www.informedia.cs.cmu.edu/pubs/aaai-info-haupt.pdf, 1997.

[19] J. Foote, J. Borecsky, A.Girgensohn, and L. Wilcox, "An intelligent media browser using multimodal analysis," in *Proceedings of the 6th ACM International Multimedia Conference* (1998), 375–80.

[20] Virage, http://www.virage.com/products/audiologger.html, 1999.[23] M. Fernstrom, and L. Bannon, "Explorations in sonic browsing," *Proceedings of HCI 97* (1997). http://www.ul.ie/~idc/library/papersreports/MikaelFernstrom/hciuk97/sonicbrowse.html.

[21] M. Meteer, H. Gish, F. Kubala, R. Schwartz, and R. Weischedel, "Gisting gets the most out of your speech in the least amount of time," *Speech Technology Magazine* 13 (June–July 1998): http://www.speechtechmag.com/st13/justfact.htm

[22] F. Pachet and O. Delerue, "MidiSpace: A constraint-based music spatializer," in *Proceedings of the 6th ACM International Multimedia Conference* (ACM, 1998), 351–60.