Audio Indexing

Gaël Richard
Ecole Nationale Supérieure des Télécommunications (TELECOM ParisTech), France

INTRODUCTION

The enormous amount of unstructured audio data available nowadays and the spread of its use as a data source in many applications are introducing new challenges to researchers in information and signal processing. The continuously growing size of digital audio information increases the difficulty of its access and management, thus hampering its practical usefulness. As a consequence, the need for content-based audio data parsing, indexing and retrieval techniques to make the digital information more readily available to the user is becoming ever more critical.

The lack of proper indexing and retrieval systems is making de facto useless significant portions of existing audio information (and obviously audiovisual information in general). In fact, if generating digital content is easy and cheap, managing and structuring it to produce effective services is clearly not. This applies to the whole range of content providers and broadcasters which can amount to terabytes of audio and audiovisual data. It also applies to the audio content gathered in private collection of digital movies or music files stored in the hard disks of conventional personal computers.

In summary, the goal of an audio indexing system will then be to automatically extract high-level information from the digital raw audio in order to provide new means to navigate and search in large audio databases. Since it is not possible to cover all applications of audio indexing, the basic concepts described in this chapter will be mainly illustrated on the specific problem of musical instrument recognition.

BACKGROUND

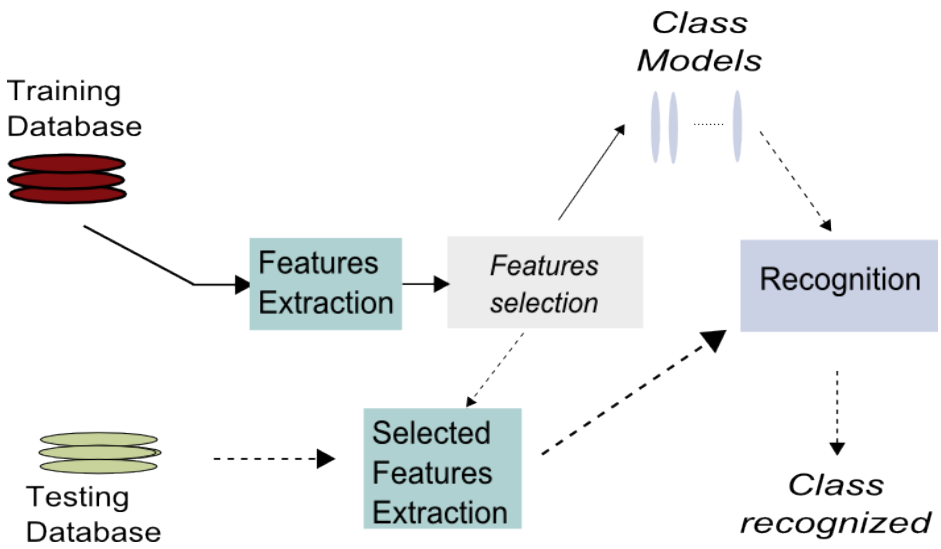
Audio indexing was historically restricted to word spotting in spoken documents. Such an application consists in looking for pre-defined words (such as name of a person, topics of the discussion etc…) in spoken documents by means of Automatic Speech Recognition (ASR) algorithms (see (Rabiner, 1993) for fundamentals of speech recognition). Although this application remains of great importance, the variety of applications of audio indexing now clearly goes beyond this initial scope. In fact, numerous promising applications exist ranging from automatic broadcast audio streams segmentation (Richard & et al., 2007) to automatic music transcription (Klapuri & Davy, 2006). Typical applications can be classified in three major categories depending on the potential users (Content providers, broadcasters or end-user consumers). Such applications include:

- Intelligent browsing of music samples databases for composition (Gillet & Richard, 2005), video scenes retrieval by audio (Gillet & et al., 2007) and automatic playlist production according to user preferences (for **content providers**).
- Automatic podcasting, automatic audio summarization (Peeters & et al., 2002), automatic audio title identification and smart digital DJing (for **broadcasters**).
- Music genre recognition (Tzanetakis & Cook, 2002), music search by similarity (Berenzweig & et al., 2004), personal music database intelligent browsing and query by humming (Dannenberg & et al. 2007) (for **consumers**).

MAIN FOCUS

Depending on the problem tackled different architectures are proposed in the community. For example, for musical tempo estimation and tracking traditional architectures will include a decomposition module which aims at splitting the signal into separate frequency bands (using a filterbank) and a periodicity detection module which aims at estimating the periodicity of a detection function built from the time domain envelope of the signal in each band (Scheirer, 1998)(Alonso & et al., 2007). When tempo or beat tracking is necessary, it will be coupled with onset detection techniques (Bello & et al., 2006) which aim at locating note onsets in

Figure 1. A typical architecture for a statistical audio indexing system based on a traditional bag-of frames approach. In a problem of automatic musical instrument recognition, each class represents an instrument or a family of instruments.



the musical signal. Note that the knowledge of note onset positions allows for other important applications such as Audio-to-Audio alignment or Audio-to-Score alignment.

However a number of different audio indexing tasks will share a similar architecture. In fact, a typical architecture of an audio indexing system includes two or three major components: A feature extraction module sometimes associated with a feature selection module and a classification or decision module. This typical "bag-of-frames" approach is depicted in Figure 1.

These modules are further detailed below.

Feature Extraction

The *feature extraction module* aims at representing the audio signal using a reduced set of features that well characterize the signal properties. The features proposed in the literature can be roughly classified in four categories:

- Temporal features: These features are directly computed on the time domain signal. The advantage of such features is that they are usually straightforward to compute. They include amongst others the crest factor, temporal centroid, zero-crossing rate and envelope amplitude modulation.
- Cepstral features: Such features are widely used in speech recognition or speaker recognition due to a clear consensus on their appropriateness for these applications. This is duly justified by the fact that such features allow to estimate the contribution of the filter (or vocal tract) in a source-filter model of speech production. They are also often used in audio indexing applications since many audio sources also obey a source filter model. The usual features include the Mel-Frequency Cepstral Coefficients (MFCC), and the Linear-Predictive Cepstral Coefficients (LPCC).
- Spectral features: These features are usually computed on the spectrum (magnitude of the Fourier Transform) of the time domain signal. They include the first four spectral statistical moments, namely the spectral centroid, the spectral width, the spectral asymmetry defined from the spectral skewness, and the spectral kurtosis describing the peakedness/flatness of the spectrum. A number of spectral features were also defined in the framework of MPEG-7 such as for example the MPEG-7 Audio Spectrum Flatness and Spectral Crest Factors which are processed over a number of frequency bands (ISO, 2001). Other features