AUDIO

AUDIO INDEXING

Gaël RICHARD

Ecole Nationale Supérieure des Télécommunications (ENST)

Speech and Image Processing Department

37-39, rue Dareau, 75014 Paris, France

Audio Indexing

Gaël RICHARD

Ecole Nationale Supérieure des Télécommunications, Paris, 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 & 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 consumer). Such applications include:

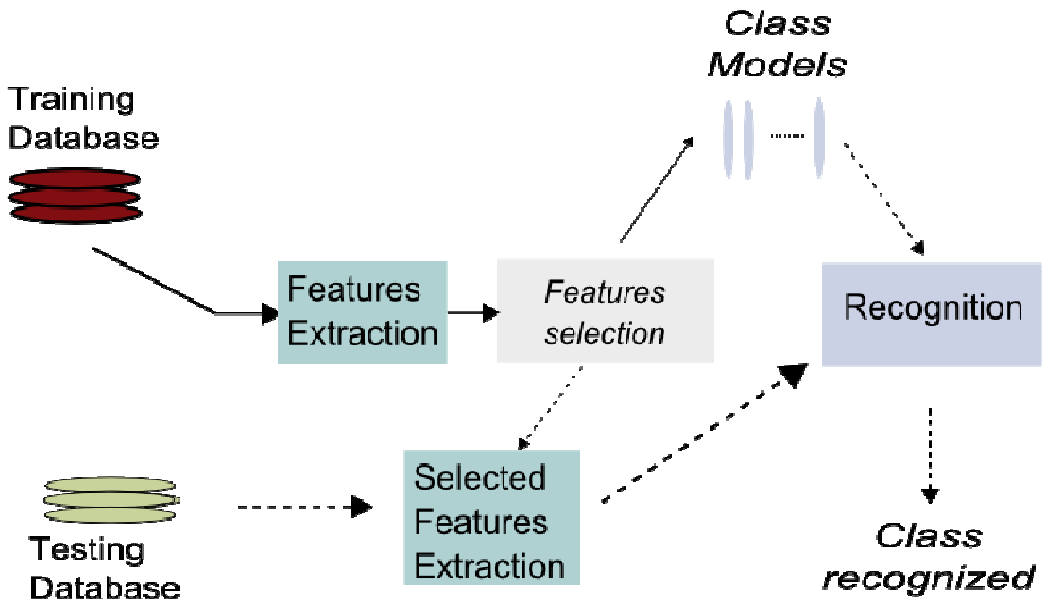
- intelligent browsing of music samples databases for composition (Gillet & Richard, 2005), video scenes retrieval by audio (Gillet & al., 2007) and automatic playlist production according to user preferences (for **content providers**)

- Automatic podcasting, automatic audio summarization (Peeters & al., 2002), automatic audio title identification and smart digital DJing (for **broadcasters**)
- Music genre recognition (Tzanetakis & Cook, 2002), music search by similarity (Berenzweig & al., 2004), personal music database intelligent browsing and query by humming (Dannenberg & 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 & al. 2007). When tempo or beat tracking is necessary, it will be coupled with onset detection techniques (Bello & al.2006) which aim at locating note onsets in 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 the figure below:



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.

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 proposed include the Spectral slope, the the spectral variation and the frequency cutoff. Some specific parameters were also introduced by (Essid & al. 2006a) for music instrument recognition to capture in a rough manner the power distribution of the different harmonics of a musical sound without recurring to pitch-detection techniques: the Octave Band Signal Intensities and Octave Band Signal Intensities Ratios.
- Perceptual features : Typical features of this class include the relative specific loudness representing a sort of equalization curve of the sound, the sharpness - as a perceptual alternative to the spectral centroid based on specific loudness measures-

and the spread, being the distance between the largest specific loudness and the total loudness.

For all these features, it is also rather common to consider their variation over time through their first and second derivatives.

It is also worth to mention that due to their different dynamic it is often necessary to normalize each feature. A commonly used transformation scheme consists in applying a linear transformation to each computed feature to obtain centered and unit variance features. This normalization scheme is known to be more robust to outliers than a mapping of the feature dynamic range to a predefined interval such as [-1 : 1]. More details on most of these common features can be found in (Peeters, 2004) and in (Essid, 2005).

Features selection

As mentioned above, when a large number of features is chosen, it becomes necessary to use *feature selection techniques* to reduce the size of the feature set (Guyon & Eliseeff, 2003). Feature selection techniques will consist in selecting the features that are the most discriminative for separating the different classes. A popular scheme is based on the Fisher Information Criterion which is expressed as the ratio of the inter-class spread to the intra-class spread. As such, a high value of the criterion for a given feature corresponds to a high separability of the class. The appropriate features can therefore be chosen by selecting those with the highest ratios.

Classification

The *classification module* aims at classifying or labelling a given audio segment. This module usually needs a training step where the characteristics of each class are learned. Popular supervised classification approaches for this task include K-nearest neighbours, Gaussian Mixture Models, Support Vector Machines (SVM) and Hidden Markov models (Burges, 1998), (Duda & al., 2000).

For example, in a problem of automatic musical instrument recognition (Essid & al., 2006a), a state of the art system will compute a large number of features (over 500), use feature selection and combine multiple binary SVM classifiers. When a large number of instruments is considered (or when polyphonic music involving more than one instrument playing at a time, as in (Eggink and Brown, 2004)), hierarchical approaches aiming first at recognising an instrument family (or group of instruments) are becoming very efficient (Essid & al. 2006b).

Future trends

Future trends in audio indexing are targeting robust and automatic extraction of high level semantic information in polyphonic music signals. Such information for a given piece of music could include the main melody line; the musical emotions carried by the musical piece, its genre or tonality; the number and type of musical instruments that are active. All these tasks which have already interesting solutions for solo music (e.g. for mono-instrumental music) become particularly difficult to solve in the context of real recordings of polyphonic and multi-instrumental music. Amongst the interesting directions, a promising path is provided by methods that try to go beyond the traditional "bag-of-frames" approach described above. In particular, sparse representation approaches that rely on a signal model (Leveau &

al. 2008) or techniques based on mathematical decomposition such as Non-Negative Matrix factorization (Bertin & al. 2007) have already obtained very promising results in Audio-to-Score transcription tasks.

Conclusion

Nowadays, there is a continuously growing interest of the community for audio indexing and Music Information Retrieval (MIR). If a large number of applications already exist, this field is still in its infancy and a lot of effort is still needed to bridge the “semantic gap” between a low-level representation that a machine can obtain and the high level interpretation that a human can achieve.

KEY TERMS AND THEIR DEFINITIONS

Support Vector Machines: Support Vector Machines (SVM) are powerful classifiers arising from Structural Risk Minimization Theory that have proven to be efficient for various classification tasks, including speaker identification, text categorization and musical instrument recognition.

Features: features aimed at capturing one or several characteristics of the incoming signal. Typical features include the energy, the Mel-frequency cepstral coefficients, …

Spectral Centroid: Spectral centroid is the first statistical moment of the magnitude spectrum components (obtained from the magnitude of the Fourier transform of a signal segment).

Spectral Slope: is obtained as the slope of a line segment fit to the magnitude spectrum.

Spectral variation: represents the variation of the magnitude spectrum over time.

Frequency cutoff (or Roll-off): is computed as the frequency below which 99% of the total spectrum energy is concentrated.

Mel-Frequency Cepstral Coefficients (MFCC): are very common features in audio indexing and speech recognition applications. It is very common to keep only the first few coefficients (typically 13) so that they mostly represent the spectral envelope of the signal.

Semantic gap: refers to the gap between the low-level information that can be easily extracted from a raw signal and the high level semantic information carried by the signal that a human can easily interpret.

Octave Band Signal Intensities: These features are computed as the log-energy of the signal in overlapping octave bands.

Octave Band Signal Intensities ratios: These features are computed as the logarithm of the energy ratio of each subband to the previous (e.g. lower) subband.

Musical instrument recognition: is the task to automatically identify from a music signal which instruments are playing. We often distinguish the situation where a single instrument is

playing with the more complex but more realistic problem of recognizing all instruments of real recordings of polyphonic music.

Non-Negative Matrix factorization: This technique permits to represent the data (e.g. the magnitude spectrogram) as a linear combination of elementary spectra, or atoms and to find from the data both the decomposition and the atoms of this decomposition (see [Lee & al., 2001] for more details).

Sparse representation based on a signal model: Such methods aim at representing the signal as an explicit linear combination of sound sources, which can be adapted to better fit the analyzed signal. This decomposition of the signal can be done using elementary sound templates of musical instruments.

References

M. Alonso, G. Richard and B. David (2007) "Accurate tempo estimation based on harmonic+noise decomposition", *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 82795, 14 pages. 2007.

J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, (2005) "A tutorial on onset detection in musical signals," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047. 2005

A. Berenzweig, B. Logan, D. Ellis, B. Whitman (2004). A large-scale evaluation of acoustic and subjective music-similarity measures, *Computer Music Journal*, 28(2), pp. 63-76, June 2004.

N. Bertin, R. Badeau and G. Richard, (2007) "Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark", *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'07*, Honolulu, Hawaii, USA, 15-20 april 2007.

C. J. Burges, (1998) "A tutorial on support vector machines for pattern recognition," *Journal of Data Mining and knowledge Discovery*, vol. 2, no. 2, pp. 1–43, 1998.

R. Dannenberg, W. Birmingham, B. Pardo, N. Hu, C. Meek and G. Tzanetakis, (2007) "A comparative evaluation of search techniques for query by humming using the MUSART testbed." *Journal of the American Society for Information Science and Technology* 58, 3, Feb. 2007.

R. Duda, P. Hart and D. Stork, (2000) *Pattern Classification*,. Wiley-Interscience. John Wiley and Sons, (2nd Edition) 2000.

J. Eggink and G. J. Brown, (2004) "Instrument recognition in accompanied sonatas and concertos",. in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*), Montreal, Canada, May 2004, pp. 217.220.

S. Essid, G. Richard and B. David, (2006) “Musical Instrument Recognition by pairwise classification strategies”, *IEEE Transactions on Speech, Audio and Language Processing*, Volume 14, Issue 4, July 2006 Page(s):1401 - 1412.

S. Essid, G. Richard and B. David, (2006), "Instrument recognition in polyphonic music based on automatic taxonomies", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, N. 1, pp. 68-80

S. Essid, (2005) *Automatic Classification of Audio Signals: Machine Recognition of Musical Instruments*. PhD thesis ,Université Pierre et Marie Curie. December 2005 (In French)

O. Gillet, S. Essid and G. Richard, (2007) “On the Correlation of Automatic Audio and Visual Segmentations of Music Videos”, *IEEE Transaction On Circuit and Systems for Video Technology*, Vol. 17, N. 3, March 2007.

O. Gillet and G. Richard, (2005) “Drum loops retrieval from spoken queries”, *Journal of Intelligent Information Systems - Special issue on Intelligent Multimedia Applications*, vol. 24, n° 2/3, pp. 159-177, March 2005.

I. Guyon and A. Elisseeff, (2003) An introduction to feature and variable selection,. *Journal of Machine Learning Research*, vol. 3, pp. 1157.1182, 2003.

ISO, (2001). Information technology - multimedia content description interface - part 4: Audio,. ISO/IEC, International Standard ISO/IEC FDIS 15938-4:2001(E), jun 2001.

A. Klapuri and M. Davy, editors. (2006) *Signal Processing methods for the automatic transcription of music*. Springer, New-York, 2006.

D.D. Lee and H.S. Seung, (2001) Algorithms for non-negative matrix factorization, *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

P. Leveau, E. Vincent, G. Richard, L. Daudet. (2008) Instrument-specific harmonic atoms for midlevel music representation. *To appear in IEEE Trans. on Audio, Speech and Language Processing*, 2008.

G. Peeters, A. La Burthe, X. Rodet, (2002) Toward Automatic Music Audio Summary Generation from Signal Analysis, in *Proceedings of the International Conference of Music Information Retrieval (ISMIR)*, 2002.

G. Peeters, (2004) "A large set of audio features for sound description (similarity and classification) in the cuidado project," *IRCAM, Technical Report*, 2004.

L. R. Rabiner, (1993) *Fundamentals of Speech Processing*, ser. Prentice Hall Signal Processing Series. PTR Prentice-Hall, Inc., 1993.

G. Richard, M. Ramona and S. Essid, (2007) "Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams, *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, 2007.

E. D. Scheirer. (1998) Tempo and Beat Analysis of Acoustic Music Signals. *Journal of Acoustical Society of America,* 103 :588-601, janvier 1998.

G. Tzanetakis and P. Cook, (2002) Musical genre classification of audio signals,. *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, July 2002.