Current Research in Music Technology at the Audiovisual Institute of the Pompeu Fabra University

Xavier Serra

Music Technology Group, Audiovisual Institute
Pompeu Fabra University, Barcelona (Spain)
http://www.iua.upf.es/mtg, xserra@iua.upf.es

Abstract

The Music Technology Group, MTG, integrated in the Audiovisual Institute of the Universitat Pompeu Fabra of Barcelona, specializes in audio processing technologies and their music and multimedia applications. With more than 40 researchers coming from various disciplines, the MTG carries out research and development projects in areas such as audio processing and synthesis; audio identification; audio content analysis, description and transformation; singing voice processing; interactive systems; and software tools. The MTG was created in 1994 by its current director, Dr. Xavier Serra, as one of the research groups of the Audiovisual Institute, a centre for interdisciplinary research in the different areas of Digital Media.

.

1 Context of the MTG

The Universitat Pompeu Fabra, UPF, (http://www.upf.es) is Catalonia's fourth publicly funded university, it is a young university, yet it already has its own little slice of history: it was founded in 1990 with 316 students taking two degree courses, and today it has 9,000 students, fourteen degree courses and an expanding campus in the heart of Barcelona. It is a young but rapidly expanding public university, having earned a strong reputation in its ten years of existence, some of their degrees being the highest in demand in Catalonia.

The Audiovisual Institute, IUA, (http://www.iua.upf.es) is an interdisciplinary center of the UPF devoted to a range of activities related with digital technologies in the media. The Institute provides a suitable environment for creation, research and dissemination aimed at students, professionals, the industry, and society at large. It is a meeting point for people coming from traditionally separated fields: science, art, engineering, design, computer science, communication, etc.

In terms of education activities, most of the people of the MTG are either involved, as teachers or students, with the Department of Technology of the UPF (http://www.upf.es/dtecn/) or with the Department of Sonology of the Escola Superior de Música de Catalunya, ESMUC, (http://www.esmut.net). The Department of Technology, jointly with the IUA, offers the Doctoral program on Computer Science and Digital Communication. The IUA also offers several graduate programs, for example the Master in Digital Arts (with a Music specialization), the Diploma in Multimedia Programming, or the Diploma in Music Composition.

An important element of the music activity of the IUA is the Phonos Foundation. This Foundation has a long history in Barcelona and since 1994 is very closely related to the IUA. Phonos was established in 1975 and for many years it was the only laboratory of electroacoustic music in Spain. In the courses offered there most of the Spanish musicians that are currently working with electronic technologies were trained. Phonos organizes concerts, conferences and other activities to spread both the works done in the IUA and in similar centers around the world. It also offers grants to artists to produce works in the different IUA labs and in collaboration with the different research groups. Phonos is supported by the main Spanish and Catalan cultural public agencies.

Most of the research carried out in the MTG is funded by either public funds from the EU, Spain or Catalonia or by industrial collaborations. In terms of European projects, the MTG is currently coordinating a large project on semantic audio, SIMAC (http://www.semanticaudio.org) and participating in several other music related projects, such as Semantic-HIFI, OPENDRAMA, AGNULA, ConGAS, HARMOS, $S2S^2$ and AUDIOCLAS.

Next we will briefly summarize the different research themes that are being worked on at the MTG.

2 Audio processing and synthesis

From the initial development of the Sinusoidal plus Residual model (also known as Spectral Modeling Synthesis) as part of the PhD thesis of Xavier Serra at Stanford University (Serra, 1989) and from the improvements carried out since then at the UPF by a number of researchers, the MTG is recognized as a worldwide reference on spectral based audio processing techniques. In fact, many of the research projects are based, or related, to these signal processing techniques.

In the last ten years, and in the context of different research projects, there have been many improvements to the basic spectral model and its implementation. Some of the key developments have been based on the identification and extraction of audio features that are relevant for particular applications or particular sound families. New models have also been proposed and efficient and flexible implementations of these models have been done.

The development of a generic sound synthesis system has been a major goal of the MTG since the beginning. A recent example of this type of work is the SALTO project (Haas, 2001) in which a real-time wind instrument synthesizer was developed. This is an example of applying the basic Sinusoidal plus Residual model to a particular application and working with a particular family of sounds.

A very different spectral based analysis/synthesis approach was required for the development of an automatic near-lossless time stretching system (Bonada, 2000). This technology permits changing the duration of an audio sequence without modifying the timbre and pitch of its content. The basic Sinusoidal plus Residual model is not appropriate for such an application and a model had to be developed for time-scaling any audio signal.

3 Audio content analysis

The efficient management of sound archives (it doesn't matter if they are personal or institutional) requires the usage of indexes and categories that cover different levels of description. Additionally to the traditional manually generated meta-data (format, resolution, channels, author, year, performer, etc.) it is possible to automatically generate some "descriptors" that will capture the sonological or musical features that are embedded in the audio files. The automatic generation of descriptors uses traditional analysis and signal processing techniques, but it also adapts techniques from other fields such as artificial intelligence or data mining. Using those multiple techniques, the signal content is extracted and descriptors with different abstraction levels are generated (more or less understandable for a "non-technical" user). Apart from the descriptors, it is also interesting to study the relations between the different levels of description in order to get a continuous flow from the physical signal to the symbolic labels used to represent its content.

The area of audio content analysis is the one in which the MTG is putting the biggest effort at this time. We are working on many of the key issues of the field, including instrument classification (Herrera et alt., 2003), melodic description (Gómez et alt, 2003) and rhythmic description (Gouyon and Meudic, 2003).

A particular example of a framework in which we carry out all this research is the SIMAC project. SIMAC addresses the study and development of innovative components for a music information retrieval system. The key feature is the usage and exploitation of semantic descriptors of musical content which are automatically extracted from music titles. These descriptors are generated in two ways: as derivations and combinations of lower-level descriptors and as generalizations induced from manually annotated databases by the intensive application of data mining techniques. The project aims also towards the empowering (i.e. adding value, improving effectiveness) of music consumption behaviours, especially of those that are guided by the concept of similarity. The gained knowledge about all this (semantic descriptors, similarity, collection organization, retrieval patterns, etc.), regarding users as individuals but also as members of organized communities, become operational as prototypes for:

- Annotating music items and collections in a way that can be exploited by data mining algorithms,
- Organizing –sonically and visually- music collections,
- Discovering interesting music by exploiting content analysis and user profiling, and
- Interacting with them in ways that will add value to owned music.

These inter-connectable components are devised, developed, and tested in connection with real communities of users and of content distributors.

4 Audio identification

A specific topic in content analysis is audio identification. In this area we have had very successful and practical results (Batlle et alt., 2004). The technology developed in the MTG is based on a low level audio analysis that extracts parameters related to the spectral characteristics of the sound and its temporal evolution. With this analysis an audio signal is coded as a sequence of descriptors, where each one of these descriptors is associated with a specific spectral distribution. This reduces the amount of

information needed to characterize an audio signal to a few thousands of bytes. The main application of this technique is the copyright protection of music on the radio, TV or Internet. It can also be used in applications such as the control and auditing of advertisements, videos or signature tunes of TV and radio programs.

The system developed works in real-time and compares the descriptor sequence ("genetic code") of the audio input signal with the codes stored on a database that has been previously codified. The output of the system is a play-list of songs, or audio fragments, present in the input signal. The system is immune to the noise and distortions caused by the radio transmission.

We are currently working on new analysis and recognition techniques in order to go from identification to the concept of similarity. The system should not only be able to identify songs but to find similarities between songs, based on melodic and/or rhythmic descriptors. From this, the possible applications will be extended and the system will be able to detect plagiarism, to find songs similar to another one, to identify a song in live music, …

5 Singing voice processing

One of our first projects was to develop a specific spectral based model for the processing of the singing voice. Since then we have done quite a bit of research in this area and we have developed a number of systems for the real-time processing and synthesis of the singing voice.

One of these developments has been a real-time system for morphing two voices in the context of a karaoke application (Cano et alt., 2000). As the user sings a pre-established song, his pitch, timbre, vibrato and articulation can be modified to resemble those of a pre-recorded and pre-analyzed recording of the same melody sang by another person. The underlying analysis/synthesis technique is based on the Sinusoidal plus Residual model, to which many changes have been done to better adapt it to the singing voice and the real-time constrains of the system. Also a recognition and alignment module was added for the needed synchronization of the user's voice with the target's voice before the morph is done.

A singing synthesis system has also been developed (Bonada, et alt., 2001). It generates a performance of an artificial singer out of the musical score and the phonetic transcription of a song using a frame-based frequency domain technique. This performance mimics the real voice of a singer that has been previously recorded, analyzed and stored in a database, in which we store his voice characteristics (phonetics) and his low-level expressivity (attacks, releases, note transitions and vibratos). To synthesize such performance the system concatenates a set of elemental synthesis units (phonetic articulations and stationeries). These units are obtained by transposing and time-scaling the database samples. The concatenation of these transformed samples is performed by spreading out the spectral shape and phase discontinuities of the boundaries along a set of transition frames that surround the joint frames. The expression of the singing is applied through a Voice Model built up on top of a Spectral Peak Processing (SPP) technique. SPP considers the spectrum as a set of regions. Each region is made up of one spectral peak and its surroundings and can be shifted both in frequency and amplitude. The Voice Model is based on an improved version of the traditional excitation/filter approach.

6 Software tools

All the MTG projects are developed using CLAM, C++ Library for Audio and Music (Amatriain et alt., 2002). CLAM is a free software framework licensed under GNU-GPL and available from our web site (http://www.iua.upf.es/mtg/clam) that allows to fully develop multiplatform audio applications in C++ using advanced processing algorithms. It does not only support the processing part of the application; it also provides multiplatform solutions for most problems that an audio application should face:

- Accessing audio and MIDI devices,
- Managing threads,
- Serializing objects in formats such XML and SDIF,
- Displaying and controlling the application data,
- Integrating visualization using several multiplatform graphical toolkits,
- Interconecting the application modules in a decoupled way,
- ...

CLAM is able to do complex audio processing involving:

- Management of heterogeneous signal data: not only samples but also spectral data, symbolic, structured data...
- Complex data flows: with asynchronous events (controls), different rates of data feeding...
- Scaling up by composition of smaller processes,
- Dynamic creation and interconnection of processing networks.

And last but not least, it comprises a big repository of audio processing algorithms for topics such as:

- Spectral modeling and transformations

- Feature extraction
- Classification
- ...

7 Interactive systems

An important goal of our research has been to build real systems for making music, that is, system that musicians can interact with, both in real-time and non real-time. For example, a number of interfaces have been developed for the use of spectral based techniques, such as the SMSPerformer (Loscos & Resina, 1998). More recently a special emphasis has been given to real-time collaborative systems.

FMOL (Jordà, 2002) is an interactive music creation platform that allows the meeting of musicians on the net in order to create music on a collective way. Started on 1997, FMOL has been used quite a number of times as a virtual electronic instrument to compose collectively the soundtracks for a Catalan theatre company "la Fura dels Baus", including F@ust 3.0 and fragments of the multimedia opera Don Quijote in Barcelona.

The reacTable* (Jordà, 2003) which is currently under development, is an electronic music instrument, which combines a tangible table-based interface with paradigms such as modular synthesis, visual programming and visual feedback, in order to build a flexible, powerful, intuitive and completely new collaborative music instrument.

References

Serra, X. 1989. A system for sound analysis / transformation / synthesis based on a deterministic plus stochastic decomposition. Ph.D. Thesis. Stanford University.

Serra, X. 1994. "Sound Hybridization Techniques based on a Deterministic plus Stochastic Decomposition Model." ICMC94.

Serra, X. Bonada, J. Herrera, P. Loureiro, R. 1997. "Integrating Complementary Spectral Models in the Design of a Musical Synthesizer." ICMC97.

Arcos, J. et alt. 1997. "Saxex: a Case-Based Reasoning System for Generating Expressive Musical Performances." ICMC97.

Serra, X. Bonada, J. 1998. "Sound Transformations Based on the SMS High Level Attributes." DAFX98.

Loscos, A. Resina, E. 1998. "SMSPerformer: A real-time synthesis interface for SMS." DAFX98.

Herrera, P. Serra, X. Peeters, G. 1999. "Audio Descriptors and Descriptor Schemes in the Context of MPEG-7." ICMC99.

Cano, P., et alt. 2000. "Voice Morphing System for Impersonating in Karaoke Applications." ICMC00.

Bonada, J. 2000. "Automatic Technique in Frequency Domain for Near-Lossless Time-Scale Modification of Audio." ICMC00.

Bonada, J. et alt. 2001. "Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models." ICMC01.

Haas, J. 2001. "SALTO - A Spectral Domain Saxophone Synthesizer." Proceedings of MOSART Workshop 2001.

Amatriain, X. de Boer, M. Robledo, E. Garcia, D. 2002. "CLAM: An OO Framework for Developing Audio and Music Applications." Proceedings of 17th Annual ACM Conference on Object-Oriented Programming, Systems, Languages and Applications Seattle, WA, USA

Jordà, S. 2002. "FMOL: Toward User-Friendly, Sophisticated New Musical Instrument." Computer Music Journal. Vol.26.3 pp 23-39, 2002.

Amatriain, X. Bonada, J. Loscos, A. Arcos, J. Verfaille, V. 2003. "Content-based Transformations." Journal of New Music Research Vol.32 .1

Gómez, E. Klapuri, A. Meudic, B. 2003. "Melody Description and Extraction in the Context of Music Content Processing." Journal of New Music Research Vol.32 .1

Gouyon, F. Meudic, B. 2003. "Towards Rhythmic Content Processing of Musical Signals: Fostering Complementary Approaches." Journal of New Music Research Vol.32 .1

Herrera, P. Peeters, G. Dubnov, S. 2003. "Automatic Classification of Musical Instrument Sounds." Journal of New Music Research Vol.32 .1

Gouyon, F. et alt. 2003. "Rhythmic expressiveness transformations of audio recordings: swing modifications." DAFX03.

Jordà, S. 2003. "Sonigraphical Instruments: From FMOL to the reacTable*." Proceedings of the 3rd Conference on New Instruments for Musical Expression (NIME 03), Montreal, Canada, 2003.

Batlle, E., Jaume Masip, Enric Guaus. 2004. "Amadeus: A Scalable HMM-based Audio Information Retrieval System." First International Symposium on Control, Communications and Signal Processing 21-24 March 2004, Hammamet, Tunisia

Ricard, J., Herrera, P. 2004. "Morphological sound description: computational model and usability evaluation." Proceedings of the 116th AES Convention, May 8–11, Berlin, Germany