AUDIO TRANSFORMATION TECHNOLOGIES APPLIED TO VIDEO GAMES

OSCAR MAYOR[1], JORDI BONADA[1], AND JORDI JANER[1]

[1] *Music Technology Group (MTG), Universitat Pompeu Fabra, SPAIN*
oscar.mayor@upf.edu
jordi.bonada@upf.edu
jordi.janer@upf.edu

In this paper we present a set of spectral domain audio transformation technologies and some existing and proposed applications of these technologies in video games. The technologies include real-time voice transformation and tempo/pitch transformation for music. Voice Transformation is mainly used to create new character's voices modifying existing recorded material but also for transforming the player's voice in real-time in a singing game. Tempo/Pitch transformations are more focused in transforming background music to create different sound atmosphere for different scenes in a video-game. Audio Transformation techniques are very useful to reduce sound creation/recording costs but also for reducing the size of the game without compromising the quality of the sound. Moreover, these real-time techniques can offer new attractive sound features to games.

INTRODUCTION

New high quality audio transformation technologies can provide a lot of applications to the video game industry. These technologies can be very valuable for the video game sound designers, but also for the players if the technologies are integrated as real-time transformations to enrich the sound features of the game. Last generation consoles have enough processing power capabilities to allow integrating high quality and complex digital signal processing algorithms into their games, something unbelievable during the last decades.

In this paper we present audio transformation technologies developed by the authors based on spectral domain algorithms that manipulate voice and music material and allow to intelligently transforming it. These technologies include, in one hand, voice transformation allowing to control pitch and timbre of the voice and, in other hand, general audio transformations allowing to change speed, rhythm and pitch of music preserving its original timbre and quality.

The technologies presented in this paper have been successfully integrated by some video game companies in their games showing a great potential and offering worth and attractive functionalities to the games.

1 STATE OF THE ART

In video games it is common to use audio middlewares to manage all the audio assets of the game, the most well-known and widely used are FMOD from Firelight Technologies [1], Wwise from Audiokinetic [2] and XACT from Microsoft [3], some of them include transformations that could be compared in some way to the ones presented in this paper, nevertheless the transformations explained here, if integrated within an audio middleware would offer new and worth features to the audio game designer, resulting in more attractive games.

Wwise, for instance allows creating interactive music structures combining segments of music and even allows changing parameters like volume, reverb, 3D position & pitch of every sound in real time (RTPC) but does not allow for intelligent transformations like gender change for voices or time stretching or pitch shifting preserving the original timbre quality of the sound. FMOD and XACT offer pitch shifting capabilities for sounds but the quality is not comparable to what the technologies explained in this paper can offer.

Apart from these middlewares, several research centres and companies have developed similar technologies. For instance IRCAM are very active in developing their high quality voice transformation & conversion technologies and successfully applying them to the film and video-game industry [4], they have also included the voice transformation technology in the T.R.A.X Transformer plug-in recently released by the French company Flux. The company TC-Helicon also offers a wide range of professional hardware based technologies for high quality voice manipulation like the Voice Pro

Professional Studio [5] which offers a wide range of effects and transformations including gender and age change with amazing quality.

Regarding general audio transformation technologies, Jean Laroche from Creative Labs & Emu Systems published some good scientific publications in the late 90's related to using the phase-vocoder technique for pitch shifting and time-scaling for both voice and music material [6], which is a similar method to the one described in this paper. At the industry level, the German company z-Plane licenses the technology Élastique PRO [7], which offers high quality time scaling and pitch shifting. This technology is included in many commercial software sequencers because of its high quality and efficiency. The main reference for time-stretching and pitch-shifting software for film post-production is Serato's Pitch'n'Time Pro [8] a protools plug-in widely used by audio professionals in the film and music industry over 10 years.

2 TECHNOLOGIES

The following technologies have been developed partly within the SALERO EU project [9]. During the project, some video game companies that were also partners of the project, including Blitz Games Studios and Pepper's Ghost Productions evaluated and integrated the technologies in their game prototypes as it is explained with more details in section 3.

2.1 VOICE TRANSFORMATION TECHNOLOGY (KALEIVOICECOPE)

KaleiVoiceCope [10] is the name of a real-time voice transformation technology developed by the authors. An input voice signal is first analyzed in the spectral domain extracting several spectral descriptors. Based on a set of parameters, a new voice is generated changing its timbre, amplitude, pitch, and other spectral and physical characteristics. This transformation allows a wide range of possibilities, for instance, changing the gender of a voice from male to female or transforming a teenager to an old woman. Also more exotic transformations are possible such as robotizing the voice, converting the voice in order to be used in a cartoon character or giving the voice an alien character as it was taken from a science fiction film.

2.1.1 Processing Technique

The voice transformation technology implements the Wide-Band Harmonic Sinusoidal Modeling (WBHSM) technique [11] [12]. This method works in wide-band conditions, i.e. it uses analysis windows that cover only one period of signal. WBHSM is able to model voice pulses in frequency domain with a set of sinusoids, which represent both harmonic and noisy components. It provides an independent control of each single pulse, thus allowing pulse sequence transformations with easy. This ability is typical of time-domain methods, but complex to achieve in frequency domain, since it implies dealing with complex subharmonics patterns [13]. At the same time, WBHSM's sinusoidal representation of the signal allows an independent control of each single harmonic component, this way overcoming typical limitations of time-domain techniques. In this sense, WBHSM combines some of the main pros of both time and frequency-domain methods while avoids some of their main drawbacks.

The proposed method can be divided in three main phases, namely analysis, transformation and synthesis, as shown in figure 1. The analysis involves first estimating the voice pulse sequence. Then the wide-band spectrum of each detected voice pulse is computed by means of periodization. Finally, sinusoidal components are estimated from the spectral peaks.

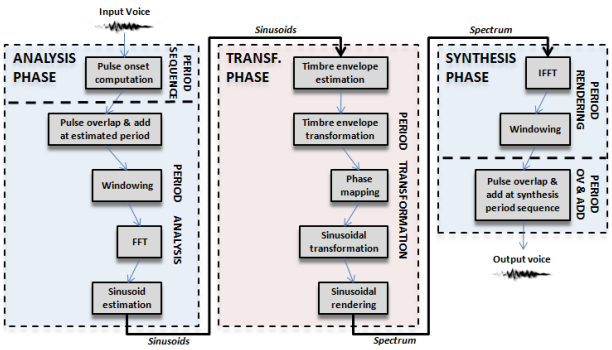

Figure 1: Voice Transformation Processing Framework.

There are two main types of transformations, the ones related to the period onset sequence and the ones related to each individual period.

Thinking of the traditional source-filter voice model, we could say that the former group of transformations are related to the voice source whereas the latter to the vocal tract. Traditional transformations such as time-scaling and pitch transposition involve scaling the period onset sequence, and repeating, removing or interpolating periods, in the same way as done in typical time-domain Pitch Synchronous Overlap-and-Add (PSOLA) techniques. However, pitch transposition also requires modifying the harmonic components of each period in order to match the target fundamental frequency; although phase continuation is not needed since consecutive period onsets are distant by one period. Conversely, timbre transformations work as in

typical frequency-domain techniques, by modifying the individual frequency components. Initially, both amplitude and phase spectral envelopes are computed by interpolating the estimated sinusoid parameters properly modified. Then, synthesis sinusoidal components are computed out of the target fundamental frequency and both amplitude and phase envelopes.

The synthesis phase involves firstly rendering the spectrum with the window transform convolved by each of the sinusoids, and secondly computing its corresponding time-domain signal. The resulting signal contains the modelled voice pulse repeated several times. Therefore, a single voice pulse is extracted and overlapped to the actual output signal.

Many high level parameters based on analysis descriptors can be controlled in order to transform the character of a voice.

2.1.2 Controllable parameters

The Voice Transformation technology (KaleiVoiceCope) allows control over a wide range of parameters, in the next subsections these parameters are explained grouped by type of transformation.

2.1.2.1 Tuning transformations

Pitch transposition: controls the amount of transposition applied to the input pitch.
Pitch Quantization: allows quantizing or not the input pitch to the closest semitone in a desired tonality, it's mainly used for singing voice.
Tremolo/Vibrato transformations: allow to add vibrato and tremolo to the output voice, controlling vibrato depth and frequency and tremolo frequency.

2.1.2.2 Sinusoidal Transformations

Frequency stretch: controls the amount of stretch applied to the frequency spectrum sinusoids.
Frequency Shift: controls the amount of shift applied to the frequency spectrum sinusoids.
Odd/Even Harmonics: balances between the amplitude of Odd and Even harmonics.

2.1.2.3 Excitation Transformations

Roughness: controls the amount of rough added to the output voice, it is commonly known as the Tom Waits effect, as the output transformation remains to the singer's voice.
Breathiness: controls the amount of breath added to the output voice.
Robotizer: adds a robot effect to the voice with metallic sound and constant pitch.
Alienator: changes the voice to an outer space sounding voice.
Whisper: adds a whisper effect to the output voice.

2.1.2.4 Timbre Transformations

The most relevant information that characterizes the human voice is the timbre, represented by the spectral harmonic envelope of the voice. Modifying the timbre of the voice combined with pitch transposition allows to easily applying gender transformations to the voice. The KaleiVoiceCope technology allows doing timbre modifications of the voice by means of a timbre mapping function. Adding (x,y) points in a two dimensional space create the timbre mapping function applied to the spectrum of the original timbre to create the timbre of the transformed voice. The curvature of the timbre mapping function determines if the lower frequency part of the spectral envelope (first formants) or the higher frequency part (higher formants) are compressed or expanded, resulting in a masculine or feminine/childish voice.

2.1.3 Implementation

The technology is implemented as a C++ software library and has been already implemented in some commercial products and prototypes. Next figures show some examples where the technology has been applied.

Figure 2 shows an screenshot of a VST audio plug-in that uses the KaleiVoiceCope technology and allows to control any parameter explained in section 2.1.2.

Figure 2. Voice Transformation VST plug-in.

Figure 3. KaleiVoiceCope online application.

Other examples that implement the technology are, for instance, online applications like the one shown in figure 3, where an applet running in the internet browser records the voice of the user, transforms it using the preset selected by the user (male, female, baby, old man, monster, alien, robot or clown) and plays back the transformed voice. Also several real-time installations for museums like the one in figure 4 were built on top of the KaleiVoiceCope technology. In these interactive installations the user speaks to a microphone, selects the desired voice transformation to be applied from a set of presets and is able to listen in real-time his voice transformed. This installations are gathering positive feedback, demonstrating that the technology is reliable and robust in 24/7 situations. More details about these and other applications including the online application can be found in [14] or in the webpage http://mtg.upf.edu/project/kaleivoicecope.

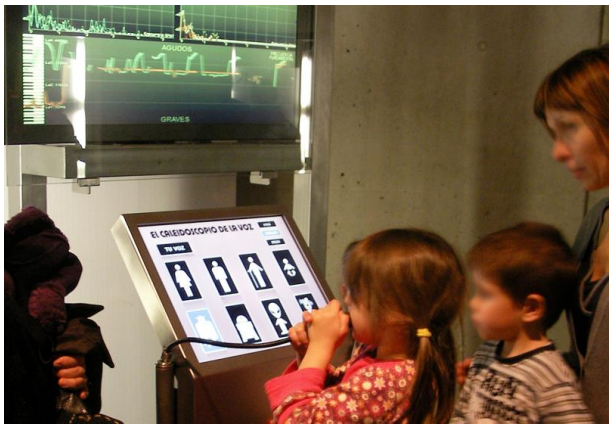

Figure 4. Museum installation for kids using the KaleiVoiceCope technology.

2.2 AUDIO TRANSFORMATION TECHNOLOGY (FLEXIBLE AUDIO)

Flexible Audio is a technology, also developed by the authors, which allows transforming polyphonic audio/music by changing the tempo, the pitch or even modifying the swing of the music when combined with a BPM (beats per minute) detector algorithm also developed by the authors. Transforming the tempo means to play the music faster or slower but without altering the timbre of the music and minimizing degradation in the sound quality. Transforming the pitch of the music means to transpose the music to a higher or lower musical key, but also without affecting the quality or the tempo of the music. These kinds of transformations are commonly used in platform video games to remind the player that is running out of time to finish the stage by playing the music faster or higher in pitch. The technology offers also an innovative swing transformation that intelligently advances or delays the beat positions of each musical bar to add a swing effect to any piece of music.

2.2.1 Processing Technique

The Flexible Audio technology allows to time-scale and pitch-shift a mono or stereo audio signal with high quality [15]. It features transient detection and preservation, channel phase coherence preservation, and formant preservation when pitch-shifting.

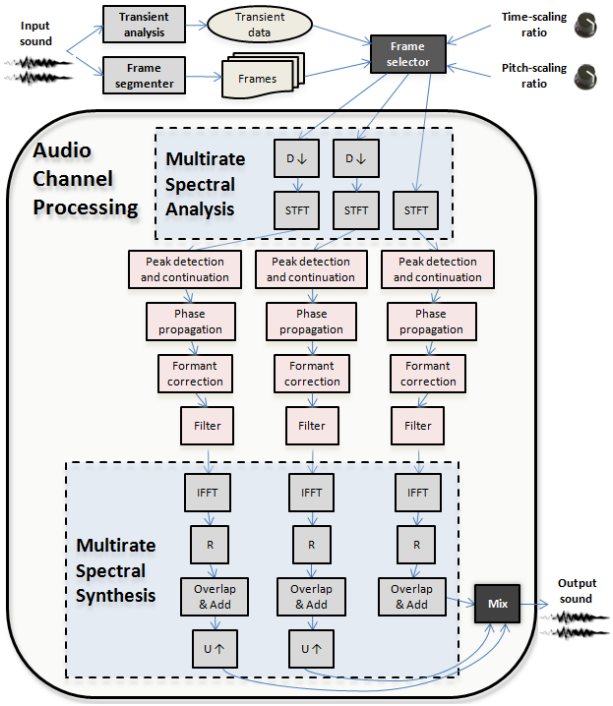

Figure 5. Block diagram of the processes involved in time-scaling and pitch-shifting operations.

In figure 5 the processing blocks involved when transforming an input audio signal are depicted. First,

transients of the input signal are detected and characterized. When processing stereo files, transients of left and right channels are synchronized. The input signal is segmented into overlapping consecutive frames with a constant hop size. The "Frame selector" module selects the appropriate input frame at each synthesis step according to the transient data and both time-scaling and pitch-shifting controls. Note that time-scaling is achieved by repeating or dropping frames and partial phase propagation, whereas pitch-shifting is achieved by time-domain resampling and partial phase propagation.

Then, each output frame to synthesize is processed by one or two "Audio channel processing" modules depending on whether the input audio is mono or stereo. The audio frame to process is analyzed by the "Multi-rate spectral analysis" module that splits the audio into several frequency bands and performs an optimized multi-rate analysis combining down sampling blocks with polyphase filters.

For each frequency band, spectral peaks are detected and continued. Next, phase corrections are estimated for avoiding partial discontinuities. Then the "Formant correction module" computes and applies a filter to minimize the spectral envelope scaling due to the pitch-shifting operation. Next, a filter is applied to limit the frequency content of each band and ensure appropriate signal reconstruction.

Then each frequency band is rendered to the original time-domain representation by the "Multirate spectral synthesis" module. This involves several steps: IFFT, a Resampling module to perform the pitch-shifting, overlap-add and upsampling if necessary. Finally, all time-domain signals are mixed together to obtain the output transformed signal corresponding to each audio channel.

2.2.2 Controllable Parameters

The Audio Transformation technology (Flexible Audio) is controlled mainly by two parameters: pitch scaling ratio and time-scaling ratio. Note that both parameters can be used together to allow more degree of transformations.

2.2.2.1 Pitch-Scaling transformations

Pitch scaling transformations refer to changing the pitch of the audio without changing its tempo and without degrading the quality of it. The easiest way to change the pitch of music is to resample the audio but then the audio will just play faster or slower like when changing the rpm speed to the wrong value in an old vinyl, resulting in a high degradation of the quality of the sound, for instance transients will be shorted or enlarged and they won't sound natural.

The pitch scaling ratio is often expressed with the number of semitones we are transposing the sound. For instance setting the parameter to 3 means that the output transformed sound will be pitched 3 semitones higher than the original, while a value of -12 means that the output sound is transposed down 1 octave (12 semitones). It is not recommended to use values below -6 or above +6, over these limits some artefacts may appear in the transformed sound.

2.2.2.2 Time-Scaling transformations

Time scaling transformations refer to change the tempo/speed of the audio without changing the pitch. We can also change the speed of a sound by resampling it but as explained in the previous section if we want to preserve the timbre and the transients of the original sound we have to do it in an intelligent way as explained in section 2.2.

The time scaling ratio is often expressed with a real number where 1.0 means no transformation, 2.0 means to play the sound at double speed and 0.5 means to play the sound at half speed, 4.0 means quad speed and 0.25, to play it at 1/4 of the original speed. It is not recommended to use values below 0.5 or above 2.0, over these limits some artefacts may appear in the transformed sound.

Swing ratio is a parameter that allows adding swing to any music like explained in [16]. To achieve this, first of all, the audio needs to be analyzed to detect all beat positions and then for each of them the procedure is to time-stretch the first half of each beat while time-shrinking the second half to create an off-beat syncopated music structure. The Swing ratio, expressed between 0 and 1 defines the amount of time-scaling that is applied to each half of the beat so the higher the value is the more swing is applied to the output sound.

2.2.3 Implementation

The technology is implemented as a C++ software library and has been already implemented in some prototypes developed by the authors. Some examples include audio effects plug-ins (see figure 6), online services (see figure 7) and it is planned to be included in the Music technology online platform called CANORIS (http://www.canoris.com) as a web service where users can upload a file to the server, specify the pitch-scale and time-scale ratio and the server return a link to the file transformed.

In figure 6, a screenshot of one prototype developed by the authors using the Flexible Audio library is shown. It consists of a VST instrument, where the user can load a music file, the plug-in analyses it computing the musical beats and the BPM (beats per minute) of the audio and then the user can change the playback tempo of the song or even add swing to the music by adjusting the speed and the swing sliders. All this process of time-scaling uses the Flexible Audio technology, preserving the original quality and timbre [17].

Figure 6. Tempo/Swing Transformation VST plug-in.

In figure 7 an online applet developed by the authors is shown. This applet developed in java runs in any internet browser and allows the user to upload a sound file, the waveform of the file is represented and the user is able to define points of an envelope function for pitch scaling and time scaling parameters over the waveform, this way, each envelope contains pair values of (transformation parameter, time position) that will be applied to the original sound by the server running the Flexible Audio technology and then the user will be able to listen to the transformed sound.

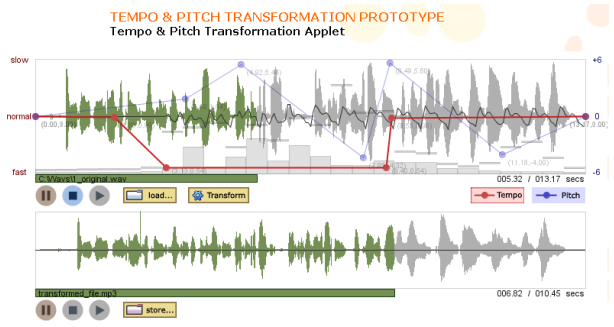

Figure 7. Tempo & Pitch transformation online prototype using Flexible Audio technology.

3 APPLICATIONS IN VIDEO GAMES

The above explained technologies have been analyzed and successfully integrated in some video games in the context of the SALERO EU project.

The first example is a serious training game developed by TruSim (http://www.seriousgames.com), a division of the UK video game company Blitz Games Studios, where doctors are trained to cure patients with breathing problems where the diagnostic is based in the regularity and speed of the breath, as well as the breathing sound the patient produces. The tempo/pitch transformation technology was used to generate sequences of breath sounds by transforming the tempo and pitch of a pre-recorded template sound to simulate several breathing malfunctions. The main advantage that the technology provided is that with only one real recording of a breathing sound sequence, using the technology, we had the possibility to generate infinite breath sequences to simulate several cases of diseases with enough quality.

There was the need to not only transform the speed of the breath sequence but also the regularity, the pitch of each breath as well as the inhale/exhale ratio for certain diseases. The videos of casualties' breath sequences were generated and the audio had to be synchronized with the image (figure 8). Using the Flexible Audio technology we could:

- Use the minimum number of breath sound files.
- Reuse content by transformation.
- Get full control of inhale/exhale ratio and breaths per minute (bpm), allowing progressive variation in real-time.

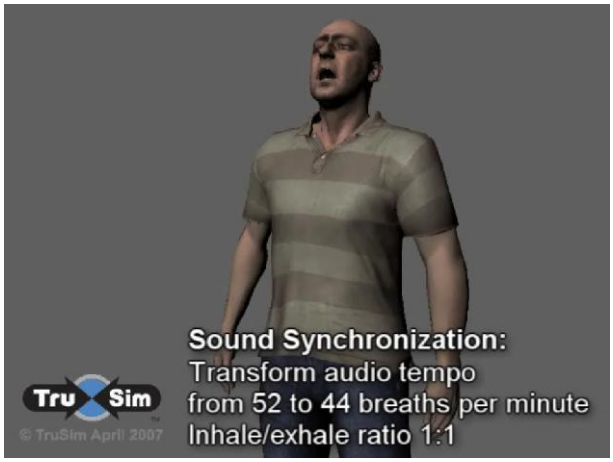

Figure 8. Video example of breath synchronization, courtesy of Blitz Games Studios.

Another example was the integration of the audio transformation technologies in the online game "My

Tiny Planets" (http://mytinyplanets.com) developed by the UK company Pepper's Ghost Productions. This game used the voice transformation technology to create character voices for robots and aliens appearing in the game by transforming pre-recorded human voices and also synthetic voices generated with a text-to-speech technology. The video game company recorded some speech from a real actor for the voice of some alien and robot characters of the game, then these recordings were transformed using the fiction character preset transformations of the KaleiVoiceCope technology to sound like aliens and robots but still being plausible and intelligible to be included in the game. Also some voices from some characters appearing in the game were synthesized using a text-to-speech (TTS) system that only provided a female voice. This process allowed the video game designer to have a wide range of synthetic speech to add new dialogs just entering sentences in the form of text. The output of the TTS was transformed using the Voice Transformation technology to create unlimited voices with different timbres for each character appearing in the game.

Also the tempo and pitch transformation technology was used in the same game to create new background music for all the scenes of the video-game transforming existing music they already had. This allowed them to reuse existing licensed music material without having to create new music, and also to adapt the music to better fit in each scene. For instance in an scene that things have to be done fast because for instance the main character is in a hurry or, to play the music faster helps to transmit this feeling to the user.

Figure 9. My Tiny Planets screenshot courtesy of PGP.

Moreover, the voice transformation technology can be easily applied in entertainment software transforming the user's voice in an online multiplayer game with voice chat like the existing ones in the Nintendo Wi-Fi Connection or the Xbox Live service. Here the user can choose a transformation and the other online players will listen to him with the voice transformed; it can be funny to sound like a robot in an online voice chat. Although this is funny within the context of a game, it would have ethical implications when used with a purpose that go far beyond entertainment.

Other possible applications include transforming the voice of the user to improve the timbre or the tuning in a karaoke-like game, or change the gender of the singer. Here you can perform a song and then go to the virtual post-production studio to improve your recorded voice to sound like a professional singer, or apply funny transformations to it.

The Flexible Audio technology can be also used, in a platform video game, to inform the player that is running out of time to complete the stage by playing the music faster, or in a puzzle game we can increase the pitch of the music to inform the player that he has achieved a new record or that the difficulty level of the game has increased.

Apart from the above applications related to video-games, we can also enumerate other applications where the voice transformation has been applied from a wide range of fields from professional musicians to film makers. A professional musician used the Voice Transformation VST plug.-in to create an electro acoustic music piece using multiple simultaneous voices transformed with different timbres creating an interesting atmosphere, also a major film company used the voice transformation offline authoring tool, to rejuvenate the voice of an adult speaker that gave the voice to an animated child character in a Hollywood film to be finished by 2011.

4 EVALUATION

The quality, plausibility and naturalness of the voice transformation has been evaluated by a perceptual hearing test where 50 users have listened to 10 triplets of excerpts of audio including a real voice recording and two transformed versions applying different amounts of timbre change and pitch transposition to the original voice. After analyzing the answers in the questionnaire, we can state:

- 15% of the users cannot distinguish between real and transformed voices.
- 40% of the users rate the transformed voices as completely natural and plausible, whilst only 5% rate them as completely unnatural.

The Voice Transformation technology described in this paper is mature enough to be used in real-time for games. It is demonstrated by some existing installations in museums where they run in a 24/7 basis since some years ago having lots of visits each day.

The Flexible Audio technology has only been tested by informal hearing tests and plug-in usability by audio engineering students and has been reported as having better quality than some existing state of the art technologies implemented in commercial time stretch software adding the functionality of the swing effect and real-time processing not present in any of the known commercial systems.

5 CONCLUSIONS

The technologies presented in this paper are mature enough to be integrated in video-games. This is demonstrated by the quality of the technology itself as well as the successful integrations within some games explained in the paper.

ACKNOWLEDGEMENTS

Part of the work presented in this paper has been co-funded with support from the European Union through the IST program under FP-027122, inside the SALERO European project.

REFERENCES

[1] FMOD interactive audio middleware
http://www.fmod.org/

[2] Wwise: WaveWorks Interactive Sound Engine
http://www.audiokinetic.com/en/products/wwise/introduction

[3] XACT: Microsoft Cross-Platform Audio Creation Tool
http://msdn.microsoft.com/en-us/library/bb174772.aspx

[4] S. Farner, A. Röbel, X. Rodet, "Natural transformation of type and nature of the voice for extending vocal repertoire in high-fidelity applications", *35th International AES Conference (Audio for Games),* London (2009).

[5] TC-Helicon Voice Pro Professional Studio: Professional Voice Processor
http://www.tc-helicon.com/products/voicepro/

[6] J. Laroche, "Time and pitch scale modification of audio signals", *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Kluwer, Norwell, MA (1998).

[7] Z-plane Élastique Pro
http://www.zplane.de/index.php?page=description-elastique

[8] Serato's Pitch'n'Time
http://serato.com/pitchntime

[9] G. Thallinger, G. Kienast, O. Mayor, C. Cullen, R. Hackett, J. Jose, "SALERO: Semantic AudiovisuaL Entertainment Reusable Objects". *International Broadcasting Conference, IBC.* Amsterdam, The Netherlands (2010).

[10] O. Mayor, J. Bonada, J. Janer, "KaleiVoiceCope: Voice transformation from interactive installations to video-games". *AES 35th International Conference: Audio for Games*. London, UK (2009).

[11] J. Bonada, "Wide band harmonic sinusoidal modeling". *11th International Conference on Digital Audio Effects DAFx-08*, Espoo, Finland (2008).

[12] J. Bonada, "Audio Signal Transforming". *Patent nº 22679-002001* (2008).

[13] A. Loscos, and J. Bonada, "Emulating rough and growl voice in spectral domain." *7th Int. Conference on Digital Audio Effects DAFx-04*. Naples, Italy (2004).

[14] O. Mayor, "KaleiVoiceKids: Interactive Real-Time Voice Transformation for Children". *The 9th International Conference on Interaction Design and Children.* Barcelona, Spain (2010).

[15] J. Bonada, "Automatic Technique in Frequency Domain for Near-Lossless Time-Scale Modification of Audio". *International Computer Music Conference ICMC2000*, Berlin, Germany (2000).

[16] F. Gouyon, L. Fabig, J. Bonada, "Rhythmic expressiveness transformations of audio recordings swing modifications". *6th Int. Conference on Digital Audio Effects*, London (2003)

[17] J. Janer, J. Bonada, S. Jordà, "Groovator - an implementation of real-time rhythm transformations". *AES 121st convention*, San Francisco, USA (2006).