# KALEIVOICECOPE: VOICE TRANSFORMATION FROM INTERACTIVE INSTALLATIONS TO VIDEO-GAMES

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

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

A Real-time Voice Transformation Technology and its applications are presented in this article. The technology allows the transformation of the human voice, such as changing gender from male to female, or transforming a teenager to an old woman. More exotic transformations are also possible, for instance robotizing the voice or giving the voice an alien character as if it was taken from a science fiction film. The technology has been already used for real-time installations in museums and in post-production applications. Now, it's being adapted to interactive videogames to transform the voice of the user or any of the game characters.

## INTRODUCTION

KaleiVoiceCope is the name of a real-time voice transformation technology developed at the Music Technology Group of the Universitat Pompeu Fabra. 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.

## 1 TECHNOLOGY

The KaleiVoiceCope Technology consists of a Software Library containing state of the art frequency domain signal processing algorithms that convert and modify the human voice, based in a set of meaningful controls that preserve its natural quality. These meaningful transformations include adding vibrato, changing fundamental frequency or amplitude, controlling spectral and physical characteristics of the voice or modifying timbre. High-Level Presets based on these transformations can be easily created to allow, for instance, gender change transformations, converting to operatic voice or robotizer effect.

This Voice Transformation Library can be easily integrated in post-production studios or software FX plug-ins to transform isolated singing or speech voices from a song, film or advertisement or even generate new voices for virtual characters from existing ones.

### 1.1 Voice Processing

Our system implements the Wide-Band Harmonic Sinusoidal Modeling (WBHSM) technique [1][2]. 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 [3]. 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, as shown in Figure 2.

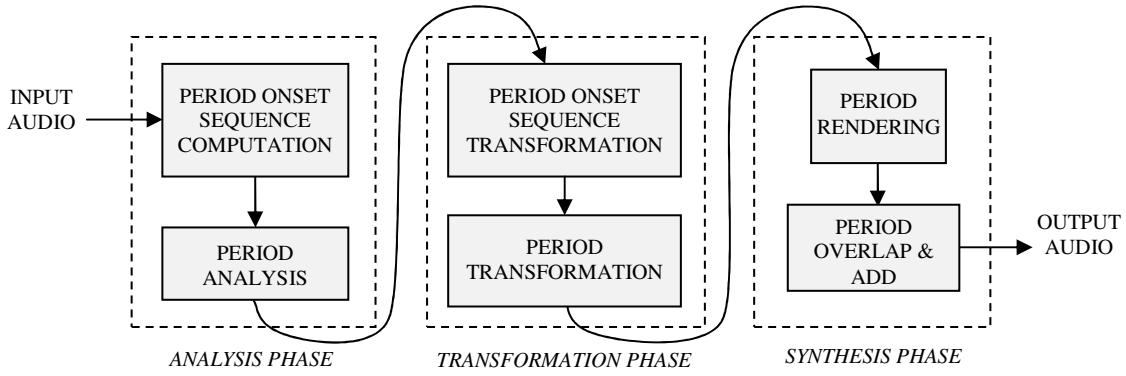


Figure 1 *Processing framework*

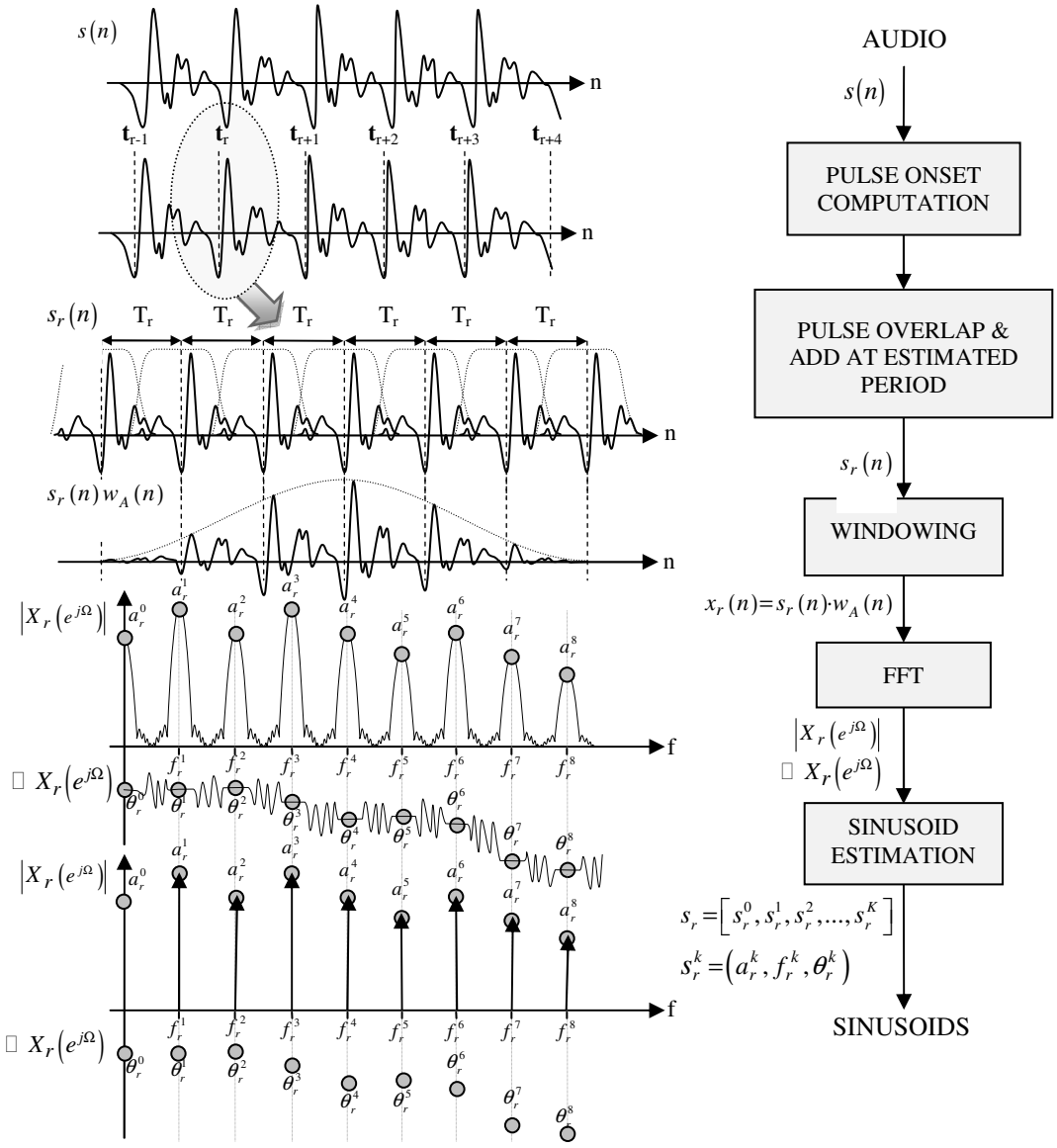


Figure 2 *Block diagram of the analysis phase*

There are two main types of transformations, the ones related to the period onset sequence and the ones related to each individual period, as depicted in Figure 1.

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.

## 2 CONTROLLABLE PARAMETERS FOR VOICE TRANSFORMATION

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

### 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.

### 2.2 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.

### 2.3 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.

Remove Unvoiced: controls if the unvoiced consonants are synthesized or not after transforming the voice.

### 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.

### 2.5 Main Parameters

Output Gain: controls the output gain of the output transformed sound.

Panorama: assigns a location of the transformed voice in the stereo panorama.

### 2.6 Real-time Control and Preset Files

All the above transformations can be controlled in real-time for each frame of the input vice to generate the output voice. The user can control these parameters through a Graphical User Interface using sliders, buttons and other controls because the API of the software library allows controlling these parameters in an analysis/synthesis frame based process. For offline processing of the voice, the values for each parameter can be specified using an easy to edit ASCII preset file.

The next figure 3 illustrates an example of preset transformation file to transform a female voice into a male voice.

```
NumberOfVoices 1
VoiceNumber 0
VoiceActive 1
LowPitch 0
PitchTransposition 0.250000
TonalityQuantization 0
PitchQuantization 0
VibratoDepth 0.000000
VibratoFreq 0.000000
VibTremolo 0.000000
Gain 0.800000
Panorama 0.500000
FrequencyStretch 0.000000
FrequencyShift 0.000000
OddEvenHarmonics 0.000000
Roughness 0.000000
Breathiness 0.000000
Whisper 0
RemoveUnvoiced 0
Robotizer 0
Alienator 0
EnvSize 6
TimbreMappingEnv000 0.000000 1.000000
TimbreMappingEnv001 0.065719 0.901274
TimbreMappingEnv002 0.156306 0.270701
TimbreMappingEnv003 0.324468 0.160256
TimbreMappingEnv004 0.632979 0.134615
TimbreMappingEnv005 1.000000 0.000000
TMFormantShift 0.500000
TMFormantStretch 0.500000
TMGamma 0.750000
DrawTimbreMapping 1
```

Figure 3: Example of parameter file used for Female to Male preset transformation.

## 3 SPEAKER CLASSIFICATION FOR CONTENT-ADAPTED TRANSFORMATION

As described in the previous sections, several parameters of the transformation can be adjusted to match the desired result. However, there are some applications (e.g. see section 4.3), where the user selects a preset of the target voice (male, female, child, elder, etc.). In these cases, a fixed transformation is applied regardless of the qualities of the input voice. It might produce an unnatural transformed sound depending on the characteristics of the input voice. For instance, applying the child preset to a child input voice would result in an extreme childish voice. Although, it might be interesting as an audio effect, in this section we introduce a method that adapts the transformation to the characteristics of the speaker, which shall improve the naturality in applications in which fixed presets are used.

### 3.1 Method description

This method relies on, first, defining multiple *sub-presets* for a given target voice preset (e.g. child). Each sub-preset is adjusted in order to sound natural for one specific template speaker. Second, prior to transforming a new input, the input voice signal is analyzed and classified according to the speaker templates. Then, the voice signal is transformed using the most appropriate sub-preset. Current approach uses four template speakers, two male and two female.

Concerning the classification procedure, it is based on a Gaussian Mixture Model (GMM) for each template speaker. GMM is a classic parametric model typically used in speaker recognition or voice conversion [4]. The speaker's data is a set of spectral vectors $\{\mathbf{x}_t, t=1..N\}$, where each spectral vector $\mathbf{x}_t$ contains the discrete mel-frequency cepstrum coefficients (MFCC's) describing the spectral envelope. The GMM assumes that the speaker data can be parameterized as:

$$p(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i N(\mathbf{x}; \mu_i, \Sigma_i)$$

where $N(\mathbf{x}; \mu_i, \Sigma_i)$ is a multidimensional normal distribution with mean vector $\mu_i$ and covariance matrix $\Sigma_i$. To compute the parameters of the GMM from the input spectral vectors $\mathbf{x}_t$, the expectation maximization algorithm is used. In this implementation, we modelled four speakers, who recorded English phrases suggested by ITU-T recommendation P.800 for determining a MOS (Mean Opinion Score): *You will have to be very quiet, There was nothing to be seen, They worshipped wooden idols, I want a minute with the inspector, Did he need any money?*. All phrases were recorded twice in three different pitches (low, mid and high), generating around one thousand of $\mathbf{x}_t$ vectors per speaker, containing 12 MFCC's with a filterbank of 20 filters at a sampling rate of 8 kHz. Due to the limited size of the data size, we suggest to use a reduced number of Gaussian components as explained in section 3.2.

Finally, to select a sub-preset for an unknown input voice $\mathbf{y}_t$, a scoring function $s(\mathbf{y}_t)$ for all four GMM models $P_n$, is computed:

$$s(\mathbf{y}_t) = \arg\max_n pdf(\mathbf{y}_t, P_n)$$

This scoring function returns the index of the best model, which corresponds to the template speaker model that best match the input speaker.

### 3.2 Results

In order to evaluate the speaker classification method, we used a subset of the TIMIT speech database, consisting of 200 files from 24 speakers. We compare different configurations of a genre classifier (male/female). Figure 4 shows the accuracy of four configurations of the classifier for different number of Gaussian components. We observe that better results are obtained without using the MFCC derivatives, where there are no significant differences when using a diagonal covariance matrix.

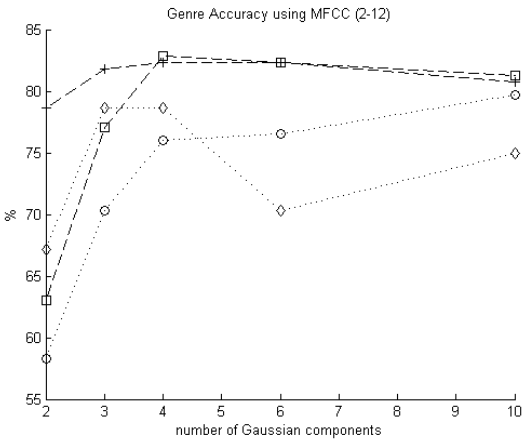


Figure 4: Classifier accuracy using MFCC (dashed), MFCC+DMFCC (dotted); and with full covariance matrix (cross and diamond) and with diagonal covariance matrix (square and circle).

In another evaluation, we compared the previous results (using four models, two male and two female) to a classifier that uses two models (male and female), combining training data from two speakers. The best result with two models was *55.20%,* compared to an accuracy of *82.8%* when using four models.

Besides, it is still remaining an evaluation of the actual improvement that represents the usage of the speaker classification within the voice transformation. This evaluation is only possible by means of a perceptual test. This evaluation will be carried out gathering feedback form user of the web-service (see section 5).

## 4 APLICATIONS

The KaleiVoiceCope technology can be employed in a wide range of applications from professional audio effects software for voice manipulation to entertainment, being also applicable for exhibitions or videogames.

### 4.1 Plug-in application

The KaleiVoiceCope technology has been integrated as audio plug-ins aimed for a wide range of users from amateurs to professionals. One example is ComboVox (see figure 5), a plug-in included in the Bonus DVD Pack of the Pinnacle Studio 10 software which allowed the user to transform a voice in an audio track of a video using a set of predefined presets including human based modifications such as gender change or age change but also fiction transformation like a robotizer effect, alien effect, ogre effect, and others. With this Plug-In even unskilled users can very easily transform a kid voice into an elderly woman or give a man's voice a special robot sound or monster effect.

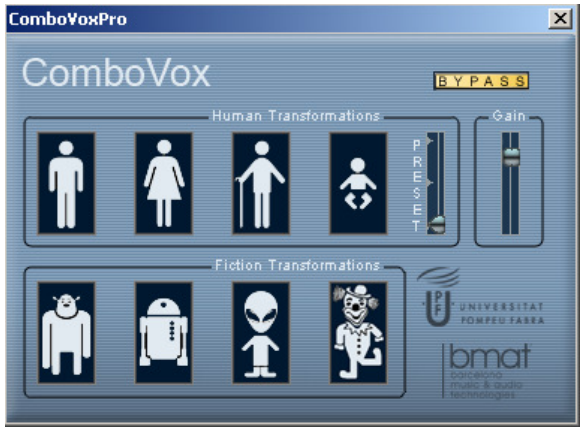


Figure 5: ComboVox Plug-in.

Another example of integration of these voice transformation techniques is the VST plug-in [5] developed in the context of the SALERO European project. It allows transforming voices in real-time, based in a set of meaningful controls like tuning (pitch and vibrato), sinusoidal controls like frequency shift & stretch, amplitude modulation, excitation controls like roughness, whisper, breathiness or robotizer and timbre modification. Harmonizer controls allow transforming one voice to many voices in real-time controlling each voice independently setting the desired control parameters and panning for each one of them separately. The Plug-In offers the possibility to create presets based on the control parameters and store it for later use. Also,

it provides visual feedback of the input and output sound, showing in/out pitch (piano roll visualization), in/out spectrum and in/out waveform (see figure 6).

This prototype has been already developed in C++ and runs in real-time under any compatible VST hosts. All transformation algorithms are based on the spectral techniques described in section 1 that analyze and modify the content in the frequency domain. This plug-in is aimed for advanced users to design new voices in a post processing studio but also offers basic functionalities for an amateur user, that wants to transform a voice easily using predefined transformations.


Figure 6: Voice Transform VST Plug-in.

### 4.2 Real-time installations for museums

The high quality and robustness of the KaleiVoiceCope technology makes it suitable for Real-time public installations in museums where the visitor speaks to a microphone, selects the desired voice transformation to be applied from a set of presets and is able to listen and visualize in real-time some parameters of the transformed voice.

#### *4.2.1 The Voice Kaleidoscope*

The Voice Kaleidoscope is an installation composed by an interactive kiosk with a 19" touch display for selecting the transformations, a big 40" screen for in/out voice parameters visualization, a microphone to capture the input voice and a set of speakers for reproducing the output transformed voice. In the touch screen, the allowed voice transformations are represented by an image icon (male to female, female to male, elder, child, monster, robot, alien or cartoon). The user presses one of these icons to select the desired transformation.

Depending on the image selected, the system analyzes the sound coming from the microphone near the kiosk, transforms it to sound like pressed icon and reproduces it in real-time through a set of speakers situated in front of the user. Some extracted parameters from the voice in the analysis process are represented in the big display so the user can view in real-time some characteristics of the input voice and the output transformed one, like the waveform, spectrum and pitch.

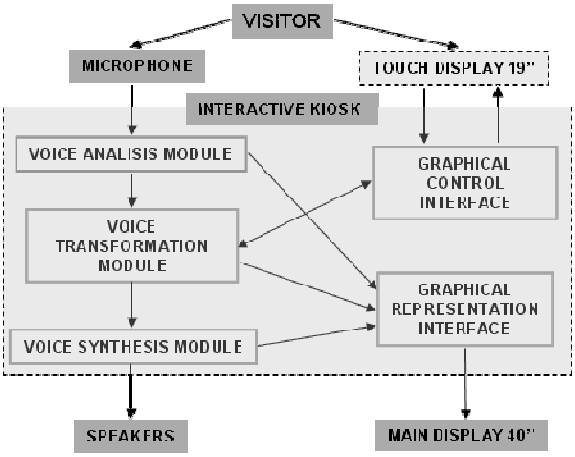


Figure 7: Block diagram of The Voice Kaleidoscope Installation.

This installation is being used by thousands of visitors daily and is gathering positive feedback, proving that the technology is reliable and robust in 24/7 situations.


Figure 8: The Voice Kaleidoscope Installation.

#### *4.2.2 My Voice Produces Waves*

In this installation, the user (children from 7 to 9 years) talks to a microphone and pressing some buttons is able to transform and listen to his/her voice in real-time. The waveform of the user voice and the transformed voice are drawn in real-time in a panoramic display so children can understand that the voice produces sound waves and that the transformation of the voice also

transforms the shape and periodicity of the sound waves. Every transformation corresponds to a different light button and an image icon identifies each transformation, the last button pressed remains illuminated to indicate the current transformation that is being applied.

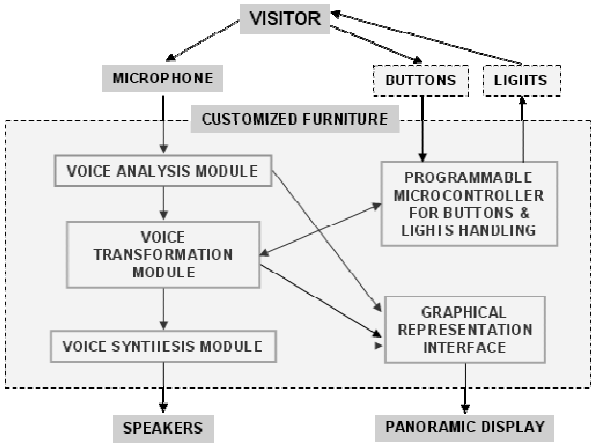


Figure 9: Block diagram of My Voice Produces Waves Installation.

This installation is being used by hundreds of children visitors daily, mainly by organized school groups. The feedback gathered from monitors and instructors demonstrates the success of such installation, and the reliability and robustness of the technology, allowing for instance 7 year-old kids' voice to sound like their parents.

Figure 10: My Voice Produces Waves Installation.

### 4.3 Web Application

Another application of the Voice Transformation technology is to transform audio files in an offline process using a web service application where the client uploads a file to the server and selects a transformation and the server returns a URL with the transformed output file. This application allows us to carry out a perceptual experiment with an extense number of users. Users fill a short questionnaire about voice quality, naturalness and plausibility to efficiently evaluate the technology making use of the opinion of hundreds of users. This application is targeted for example to post-production studios or as web service for Text-To-Speech systems.

### 4.4 Video games

This technology can be easily applied in entertainment software and has a wide range of applications. For instance, allowing to transforming the user's voice in an online multiplayer game with real voice chat. Voice chats are a new add-on to new generation consoles with internet connections as the existing ones in the Nintendo Wi-Fi Connection or the Xbox Live service. The idea is that the user could choose a transformation so other online gamers will listen to him/her with the voice transformed; it can be funny to sound like a robot in an online voice chat, or an application for children to sound like their parents.

The KaleiVoiceCope technology would also allow the gamer to customize the voice of the characters of the videogame, in the same way that modern games allow to select/change the appearance of the virtual characters. A set of sliders and buttons would be enough to change the voice of the characters in the game, offering a wide range of possibilities.

Other 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. The user can perform a song and then go to the in-game virtual post-production studio to improve your recorded voice to sound like a professional singer, or apply funny transformations to it.

This voice transformation technology can also be used as an offline tool to help videogame sound designers to create new fiction voices for the characters in a videogame, allowing real-voices to sound as if they were synthetic like robots, aliens or monsters. An example of this application has been done in the context of the SALERO European project and can be found in a series of Flash based games called MyTinyPlanets [5] which incorporate the KaleiVoiceCope technology to modify the virtual character voices created using a Text-to-Speech synthesis system and some other voices recorded by a real speaker. From an original female voice, applying transformations, we can create voices for a male, a child, an alien, a robot and other characters that appear in the videogame.


Figure 11: MyTinyPlanets game

## 5 EVALUATION

In order to evaluate the KaleiVoiceCope technology, in terms of efficiency, it performs twice faster than real-time in a state of the art computer, latency between input and output audio due to processing is about 50 ms, this is mainly related to the frequency domain analysis process where we need some frames in advance for computing some descriptors before we can synthesize an output frame. Nevertheless, this latency added to a low latency audio driver is sufficiently small to allow integrating the voice transformation in real-time exhibitions as it is stated in section 4.2.

In terms of quality, naturalness and plausibility, this technology needs to be evaluated doing perceptual tests. For human-like transformations including transforming any voice to female, male, child or elder, a questionnaire has been created where users are asked to listen some original and transformed sounds mixed and score them using the MOS (Mean Opinion Score) [6] which provides a numerical indication of the perceived quality in terms of naturalness and plausibility, expressed as a single number in the range 1 (bad) to 5 (excellent). For conducting this hearing tests English phrases suggested by ITU-T recommendation P.800 for determining a MOS where used as it's explained in section 3.1. For fiction-like transformations when transforming human voices to robot, alien, monster and clown voices, the users are asked to rate the quality of the voices in terms of plausibility, intelligibility and an overall quality question asking if the user likes or dislikes the transformation applied to the voice.

Preliminary results of the evaluation suggest that the quality of the technology is ready for professional applications where it is already being used.

## 6 CONCLUSIONS

In this article we have presented a frequency domain voice transformation technology and its current and future applications. The technology combines some of the main pros of both time and frequency-domain methods while avoiding some of their main drawbacks. The applications where it's already been applied, including professional software audio plug-ins, real-time installations for museums and video-games demonstrate that the technology has enough quality to be reliable and robust in 24/7 situations.

## 7 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. The interactive installations for Museums have been funded by "La Caixa" foundation.

## REFERENCES

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

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

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

[4] Y Stylianou, O Cappe, E Moulines, "Continuous probabilistic transform for voice conversion", *IEEE Transactions on speech and audio processing*, 1998.

[5] MyTinyPlanets on-line game . http://www.mytinyplanets.com

[6] John H. L. Hansen, Bryan L. Pellom. "An effective quality evaluation protocol for speech enhancement algorithms", *Proceedings of the International Conference on Speech and Language Processing*, Sydney, Australia (1998).