MORPHING TECHNIQUES FOR ENHANCED SCAT SINGING

Jordi Janer

Technology Department
Universitat Pompeu Fabra, Barcelona, Spain
jordi.janer@iua.upf.es

Alex Loscos

Technology Department
Universitat Pompeu Fabra, Barcelona, Spain
alex.loscos@iua.upf.es

ABSTRACT

In jazz, scat singing is a phonetic improvisation that imitates instrumental sounds. In this paper, we propose a system that aims to transform singing voice into real instrument sounds, extending the possibilities for scat singers. Analysis algorithms in the spectral domain extract voice parameters, which drive the resulting instrument sound. A small database contains real instrument samples that have been spectrally analyzed offline. Two different prototypes are introduced, reproducing a trumpet and a bass guitar respectively.

1. INTRODUCTION

1.1. History of scat singing

Scat is defined as a vocal improvisation using phonetic sounds similar to the instrumental sounds of jazz. Although, scat is believed to have its origins in African American ring shout ceremonies, it was first popularized by Louis Armstrong in a 1926 recording, when he dropped the lyrics and spontaneously substituted *scat* for the words. This particular singing style was adopted by other great jazz singers such as Ella Fitzgerald or Sarah Vaughan. Today, *scat* maintains its relevance with performers like Bobby McFerrin, who exploits his vocal register producing a wide range of sounds.

1.2. Transforming the voice – Voice driven synthesis

In the context of spectral audio processing, *Morphing* [10] appeared some years ago as a very exciting technique for generating new and never heard sounds. Actually, the concept of *Morphing* comes from the field of video processing, where an image is transformed sequentially into another. In audio processing, attempts for synthesizing sounds, whose acoustic characteristics lie between two real instruments sounds, were carried out. However, in the presented approach, we make use of *Morphing* techniques for transforming the input singing voice into a musical instrument sound.

A particularity of such as system is that it can be understood by means of two different perspectives: as a voice transformation effect, or as a voice-driven synthesizer. This issue can motivate a more thorough study that is beyond the scope of this article.

2. VOICE ANALYSIS

Several techniques have been developed over the years to study the human voice. Most research is related to the field of speech processing, but the study of the singing voice has brought also additional approaches. Basically, all these methods, though, rely on the source/filter approach, in which the phonatory system is seen as a coupled excitation (modulated air-flow) and a resonator (vocal tract cavity).

Time domain techniques were widely used, for instance, to estimate the pitch of a periodic signal by means of its correlation signal. With the rapid evolution of processor velocity, spectral techniques were also affordable, and implied a major progress in the audio processing field. In this work we use a Phase-Vocoder based method for voice analysis and also for the spectral morph synthesis.

The Phase-Vocoder was first developed by Flanagan [1] in 1966 at Bell Laboratories, and was used for speech processing. Later, in its FFT form, it was brought to musical applications by J.A. Moore and M. Dolson. Primarily, this technique was used for modifying the time scale or the pitch of incoming voice or monophonic signals. A problem that arouse with the Phase-Vocoder algorithm is the loss of presence, or artificial reverberation after processing the sound. In order to overcome this, M. Puckette proposed the Phase-Locked Vocoder[2], which introduces phase-locking between adjacent pairs of FFT channels. Puckette's article gives a more thorough description about this technique.

A further step in the Phase-vocoder development is the new algorithm proposed by J. Laroche [3]. This technique allows direct manipulation of the signal in the frequency domain. Some of the application are pitch-shifting, chorusing and harmonizing. The underlying idea behind the algorithm is to identify peaks in the Sort-Time Fourier Transform, and translate them to new arbitrary frequencies of the spectrum. Here, the spectrum is divided in several regions around peaks.

This technique was integrated in the *Spectral Peak Processing* (SPP) framework [4] by Bonada and Loscos. It performs a frame based spectral analysis of the audio, giving as output of the STFT, the harmonic peaks and the pitch. Here, the harmonic peaks are calculated by parabolic approximation. For the pitch detection, we used the technique developed by Cano et al. [5]. Basically, the SPP considers the spectrum as a set of regions, each of which belongs to one harmonic peak and its surroundings. The main goal of such technique is to preserve the local convolution of the analysis window after transposition and equalization transformations. In SPP, any transformation applies on all bins within an harmonic region. Common transformations include pitch-shift and equalization (*timbre modification*).

3. TRUMPET MORPHING

When writing a song with a sequencer, adding convincing wind instrument tracks is easier said than done if the composer can not use a wind controller, whether because he lacks it or he lacks the skills to use it. When so, either the composer records the score using a MIDI keyboard; either writes it by clicking the mouse.

Clearly none of these solutions give intuitive and meaningful control over the final synthesis result and they usually require of a tedious test and trial tuning.

The hypothesis in which the so-called Larynxophone application stands is that voice can be successful in specifying main characteristics of a melody but also the expressivity nuances (such as attack sharpness / smoothness or dynamic envelope) of a wind instrument performance.

Similar to [6], the Larynxophone processes can be decomposed in non real time processes, which are the ones that take place as a prelude, before the application runs, and the processes that take place in real time, which are the ones that occur while the user is performing.

The analysis used for both the wind instrument samples and the voice signal captured by the microphone is based is frame-based and uses spectral domain techniques that stand on the rigid phase-locked vocoder [7].

3.1. Non-real time processes: instrument database creation

The non real time processes focus on the wind instrument database creation. That is, on recording real performances, editing and cutting them into notes, analyzing and labeling them, and storing them as an instrument database.

In the current implementation, the database is made out of three trumpet notes at A3, A#5, and C5. For each sample the database contains a binary file in which the necessary data resulting from the analysis is stored. These data is pitch and energy envelope.

3.2. Real time processes: Voice Analysis, Transformations and Synthesis

The real-time processes start with a frame based spectral analysis of the input voice signal out of which the system extracts a set of voice features. This voice feature vector is first used to decide which sample to fetch from the database, and later, as input for the cross-synthesis between the instrument sample and the voice signal.

The criterion that decides the trumpet sample to use at each voice frame is: take the nearest sample in pitch. From that sample, the trumpet frame is chosen sequentially taking into account loops. The frame is transformed so to fit the user's energy and tuning note specification, for which energy correction and transposition with spectral shape preservation is applied with similar techniques to those described in [3].

Finally, the synthesis is in charge of concatenating the synthesis frames by inverse frequency transformation and the necessary window overlap-add related processes.

Regarding the mapping between voice and synthesis parameters, some considerations need to be done. If we think about controlling a tone generator, a sampler, or a synthesizer, it instantly comes to mind using MIDI. In the case we are dealing with this implies using a voice to midi conversion to turn voice expression attributes into MIDI control and use these to drive the synthesis. However, there is an alternative solution in which voice directly drives the wind instrument synthesis using real-time morph / cross-synthesis techniques.

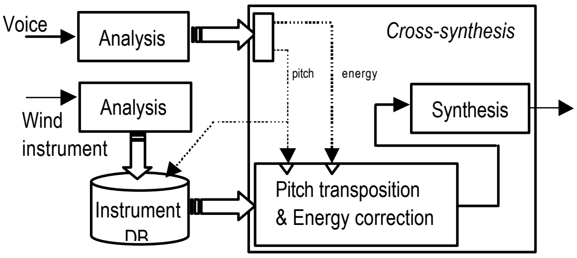

Figure 1: *Larynxophone block diagram*

For the voice to midi converters, there already exist several hardware and software commercial voice-to-midi converters. Among diverse softwares that offer real-time conversion we can find: MidiVoicer (www.e-xpressor.com/m_voicer.html), and Digital Ear (www.digital-ear.com/digital-ear/info.htm). Some hardware solutions are: Vocal to MIDI (www.geocities.com/vocal2midi/), and MidiVox (www.healingmusic.net).

Figure 2: *MidiVox general view*

Some of these converters might be useful and give reasonable results in some cases but they generally lack of robustness. The problem comes from the singing voice real-time note onset detection. Such critical process is in charge of deciding at each analysis frame time if the current voice data belongs to a new note or if it has to be considered part of the note that was already being performed. This decision has to be taken with no frame delay and with no prior knowledge on the final melody (not even key or/and scale). The considerable complexity of this problem makes it nearly impossible to avoid the converter outcoming false notes.

Regardless the modest suitability of using a voice to MIDI converter, it is important and probably a key factor to decide how to map voice attributes to MIDI messages.

Basically, driving a synthesis with voice utterances requires from a mapping between pitch / dynamic related voice attributes and MIDI messages: note on / note off (key number and key velocity), polyphonic key pressure (aftertouch), and pitch bend change. Of course neither these voice attributes fulfill vocal expressivity space nor the MIDI messages are able to reproduce all possible wind instrument expressive nuances; however, when adequately mapped will allow a useful basic control.

The proposal is to use the pitch envelope obtained in the fundamental frequency analysis to fill the key number and its associated modulation along the note, the pitch bend; and to use the Excitation plus Residual (EpR) Voice Model excitation parameters [4] to fill the key velocity and its associated modulation, the aftertouch.

The goal of the voice to MIDI conversion would be to outcome each note of the fragment, and for each one, the pitch envelope would define:
-key note: the mean of the pitch envelope over the note region quantized to the 12-tone equal temperament scale.
-pitch bend: the difference between the pitch envelope and the key note frequency value.

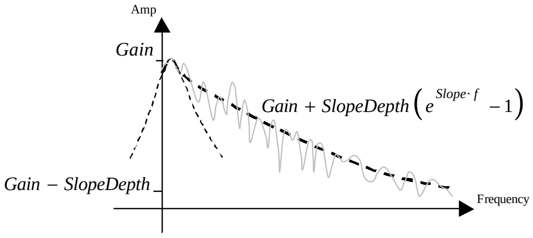

Figure 3: *EpR model voice excitation representation*

The mapping between dynamic related attributes and key velocity and aftertouch would be applied just the same. In this case, dynamic envelope is calculated out of the three EpR excitation parameters, that is, excitation gain, excitation slope, and excitation slope (see figure 3).
Obviously though, this process has to run on the fly and this means once a frame has been detected as the onset of a note, the converter takes the current pitch and dynamic values (possibly averaged along some short past history) as the mean over the note. Thus, all following frames that are considered part of that note define aftertouch and pitch bend messages from the difference between its current values and the onset frame values.

The cross-synthesis between wind instrument samples and voice utterances is a shortcut technique that avoids intermediate MIDI conversions. We define cross synthesis as the technique by which elements of one or more sounds combine to create a new one with hybrid properties. Taking profit of this technique capabilities, we can extend the voice control further than pitch, dynamics and their associated modulations and set off continuous control over, for example, the sharpness of the attack or the instrument timbre modulations.
In this cross-synthesis, wind instrument will take synthesis pitch and dynamics directly from the pitch and dynamics attributes of the input voice. More complex controls such as the ones previously mentioned can achieved by using, for example, voice excitation gain first derivative or spectral tilt voice attribute to drive the wind synthesis.

3.3. Additional Transformations

The prototype incorporates some additional transformations. One of them is a one octave up / down transposition. When turned on, it is applied to the voice analysis data before using it to fetch the trumpet frame.

Another additional transformation is pitch quantification, implemented as in [8] with optional specifications of key and chord.

There is also an 'Extract Unvoiced' switch which can mute the unvoiced parts of the input. This allows the musician to use unvoiced allophones in the performance.
Finally, the prototype includes a morph feature by which the resulting synthesis timbre can defined as a balanced interpolation of both voice and trumpet timbre. Timbres are defined by means of the harmonic peaks magnitude data. The morph slider in the interface defines the morph interpolation value.

4. BASS GUITAR MORPHING

Since the introduction of the jazz music, the bass, originally a double bass, has become an indispensable instrument of any rhythm section in popular music. Jointly with the drums, it provides the basis upon the other instruments play. The role of the bass continued evolving in new musical styles such as *rock* or *funk*, after the introduction of the electric bass by the Fender company in the 50's. Furthermore, other playing styles appeared, such as the "slap'n'pop" that was invented by Larry Graham in the 60's and often found in funk productions. The sound of a plucked bass guitar consists of impulsive notes that decay exponentially in time. Due to the long note decay time, bass players usually damp the vibrating string with the finger before plucking a new note. Usually, both the double bass and the electric bass have four strings tuned at E1 (41.3*Hz*), A1, D2 and G2. Therefore, the fundamental frequency remains below other instruments range, attaining a good timbre balance when playing together.

4.1. Voice features

Essentially, vocal gestures can be of three different kind in the voice production mechanism, depending on whether its origin is either in the breathing system, in the glottis, or in the vocal tract. In this approach, the chosen features are classified based on control aspects: Excitation, Vocal Tract, Voice Quality and Context.

A more detailed description of the feature extraction algorithms, which derive from the *SPP Analysis*, appears in an article by the author [9]. *Excitation* descriptors are probably the most elemental and intuitive for the user, since they are related to the instantaneous sung energy and fundamental frequency. In a further step, we find the voice color or voice timbre. For voiced sounds, the timbre is associated to a particular vowel. We assume that a vowel can be approximately determined by its two first formant frequencies. A very simple algorithm based on spectral centroid, *Dual-Band Centroid Estimation* [9], will estimate these frequencies approximately. In addition to the introduced features, we argue that the singing voice comprises other characteristics that might be also controllable by the singer, and thus useful in this kind of systems. Algorithms for estimating timbrical aspects (*Voice Quality*) such as *roughness* and *breathiness* were developed. Although these algorithms need further research, it is a good start-point for controlling other aspects of the transformed sound.
Finally, we include the *Context* features, which give us information of the note's state. We included the *Attack unvoiceness* descriptor. Assuming that a note consists of a pitched sound, this descriptor attempts to determine whether the attack of the note was unvoiced or not. The formula derives from the accumulative zero-crossing rate of the signal weighted with the energy, which is evaluated just before a note onset. The initial motivation for this descriptor is to use it for determining the harshness of the synthesized note's attack. In our case of a bass guitar, it might be related to the differences between a soft fingered and a sharp slap electric bass.

4.2. Mapping

In our case, from the *Energy* feature, we defined a note onset signal that triggers a note depending on the energy's derivative. The *Pitch* feature is transposed one octave lower and passed as continuous parameter, thus allowing pitch bend actions. A quantization function for simulating the influence of the frets is foreseen, but not yet integrated in the current implementation. In addition, the two formant frequencies, which are related to the timbre, are also passed though as continuous parameters to the synthesis technique dependant layer. In the following sections, we address the mapping particularities of each synthesis technique.

4.3. Synthesis

Our Spectral Model approach combines a sample-based synthesis with transformations, based in the Spectral Peak Processing (see section 2). A particularity of our sample-based algorithm is that it works in the frequency domain. Basically, depending on the input voice's parameters, a sound template track is selected from the database. Each template track contains all spectral frames from a single note. Then, we read periodically the spectral frames and transform some characteristics using the SPP framework. Finally, the processed spectral frames are converted to time domain through the inverse Fourier transform.

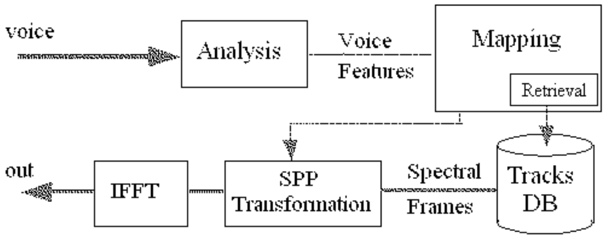

Figure 4: *System's overview.*

For our bass synthesizer implementation, we set up a very small database consisting of 12 template tracks (containing one note sample). The tracks are analyzed off-line in the spectral domain, and stored in the database in form of binary files containing spectral data (complex spectrum, harmonic peaks, estimated pitch, etc.). In a further step, all tracks have been labeled by hand according to its characteristics. Currently, only three features were annotated: *Pitch*, *Dynamics* and *Attack Type*. The pitch values are specified in Hz, *Dynamics* and *Attack Type* range is [*0..1*]. In the case of Dynamics, *0* corresponds to a *pp* sound, and *1* to a *ff*. The attack type is a novel concept that we defined, and it is instrument dependant. Concerning bass sounds, we decided to classify two types of sounds depending on the plucking technique pick-plucked or fingered, whose sounds are primarily related to the attack.

A retrieval method calculates the minimum Euclidean distance between the Input Vector (Pitch, Loudness and Attack Harshness) and the database elements. This outputs an ID corresponding to the selected track. Then, in a further step, we start reading the spectral frame of the track. Since we are dealing with a very small database, few combinations of loudness and pitches are available. Currently, the transformation process is reduced to a transposition.

Another factor that must be taken into account is the timing. The pre-analyzed template tracks have a certain duration. For each track, this corresponds to a certain number of spectral frames. In our system, though, the performer's voice controls the synthesized sound. Hence, is the input voice that decides the output duration. We analyze the track by hand and set two loop points (*sustained* and *release*) within a steady region of the original sound. When a note is triggered, we start reading the template track frame by frame. When the point sustained is reached, we repeat this frame continuously keeping the phase coherence in the generated spectrum. When the performer releases the current sung note, the last frames are played until the note's end.

5. ACKNOWLEDGEMENTS

This research has been partially funded by the EU-FP6-IST-507913 project SemanticHIFI.

6. REFERENCES

[1] Flanagan, J.L. and Golden, R.M., "Phase Vocoder", Bell Systems Technology Journal, vol 45, 1966

[2] Puckette, M.S. "Phase-locked Vocoder" Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1995.

[3] Laroche, J. and Dolson, M. "New Phase-Vocoder Techniques for Pitch-Shifting, Harmonizing and other exotic Effects." Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 1999

[4] Bonada, J. and Loscos, A., "Sample-based singing voice synthesizer by spectral concatenation", Proceedings of Stockholm Music Acoustics Conference 2003, Stockholm, Sweden, 2003.

[5] Cano, P. "Fundamental Frequency Estimation in the SMS analysis", Proceedings of COST G6 Conference on Digital Audio Effects, Barcelona, 1998.

[6] Cano P., Loscos A., Bonada J., de Boer M., Serra X. "Voice Morphing System for Impersonating in Karaoke Applications", *Proceedings of International Computer Music Conference*, Berlin, Germany, 2000.

[7] Puckette M. S. "Phase-locked vocoder", Proceedings of IEEE Conference on Applications of Signal Processing to Audio and Acoustics, Mohonk, USA, 1995.

[8] Zolzer U. DAFX – Digital Audio Effects. Wiley, John & Sons, March, 2002.

[9] Janer, J. "Feature Extraction for Voice-driven Synthesis", Proceedings of the 118thAES Convention, Barcelona, 2005.

[10] Cook, P.R., "Toward the Perfect Audio Morph? Singing Voice Synthesis and Processing", Proceedings of the 1st. International Conference on Digital Audio Effects (DAFX), Barcelona, 1998.