Musical Gestures and Audio Effects Processing

Eric Métois, Ph.D
metois@media.mit.edu http://www.media.mit.edu/~metois

Abstract

We introduce the notion of *musical gestures* as time varying measurements which identify the audio input stream's musical skeleton without attempting to implement any involved model of musical understanding. Living comfortably at an intermediate level of abstraction between wave forms and music transcriptions, these musical gestures are used to control the behavior of an audio processing module. The resulting scheme qualifies as an audio effects processing system as it essentially transforms an audio stream into another.

1 Introduction

Audio effects are used extensively not only by producers and mastering engineers but also by musicians, who often consider effects to be critical extensions to their instruments. If a complex combination of compressors, distortions, noise gates, delays, phase modulators, pitch shifters and other modules leads to a deep metamorphosis of the audio input then the resulting system may qualify as a sound-controlled synthetic instrument. A direct functional relationship between the input and the output creates systems that are both responsive and expressive, but the sole reliance on direct wave form transformation as a method of synthesis has clear limitations in terms of the timbral properties of the outcome. For instance, such processing will typically result in sounds that are spectrally richer than the input. Instead of confining itself to the role of a control, the input severely constrains the sonic nature of the synthetic instrument. The limited complexity of the sound produced by a plucked string is the main reason why such wide combinations of effects modules may still offer some musical use to guitar players.

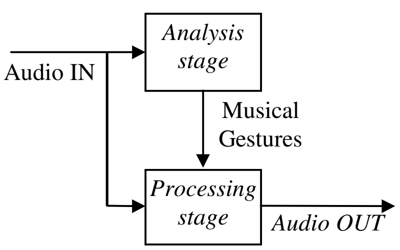

Figure 1. General overview of the proposed framework

Motivated by overcoming such limitations in the design of a sound-controlled synthetic instrument, we suggest a general framework that attempts to unify effects processing, machine listening, and sound synthesis. As illustrated in *Figure 1*, the effects processing module is split into an analysis and a processing stages. Avoiding the common pitfall of audio to MIDI converters, the analysis stage attempts to model the audio input without trying to identify musical intentions. This stage provides a set of *musical gestures* which stands for a minimal carrier of musical intentions. They are fed as control parameters to the processing stage that is responsible for the creation of the audio output. In the context of this work, the processing stage will be some synthesis engine.

As a first step, we will attempt to clarify the general nature of musical gestures. We will then suggest a particular choice for these musical gestures and means to estimate them in the analysis stage. In order to build a complete system, we will discuss briefly the nature of an acceptable synthesis engine and present the result as a proof of principle.

2 Musical Gestures

Music is an artistic medium of ideas and emotions and in an attempt to illustrate that point, *Figure 2* represents a general musical process as a chain of communication. Very much like for language, this chain of communication spans both the cognitive and the physical worlds, requiring both of them in order to make any sense. In this diagram, *sounds* refers literally to the wave form produced by the instrument. By *Low level auditory perception* we refer to the set of features provided by the first stages of our auditory system (external and internal ear). This is not to be taken literally in its physiological sense; we refer hereby to some fairly straightforward signal processing (such as frequency analysis) which might be related to those taking place in our cochlea. This explains why *Low level auditory perception* was excluded from the *cognitive field* in this diagram.

Although this diagram might look somewhat trivial, it is not rare to come across attempts to recover *musical intentions* from sounds via signal processing only, underestimating the role of human perception. This confusion illustrates the obscure boundary between *musical intentions* and *musical gestures*. Neither of these notions have the pretension to be universal. They reflect the author's convictions concerning the boundary between the roles of information theory and of psychology in this chain of communication.

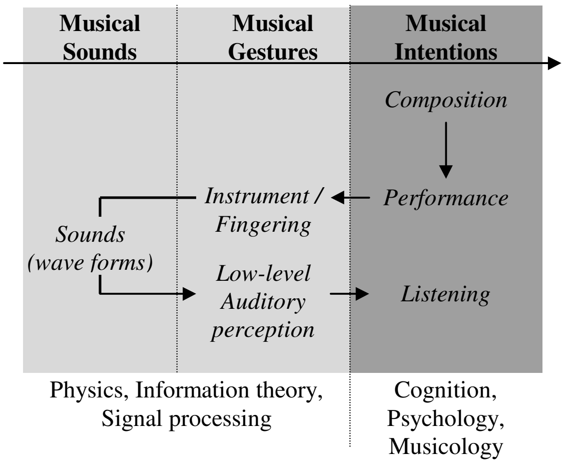

Figure 2. Music as a chain of communication.

Musical intentions are objects that require some knowledge or expectations about what music is supposed to be. Their nature reflects each individual's musical experience and culture [1]. If we keep language as an analogy, these objects are the musical analogs of words, sentences and meaning. They can often be seen as the results of some decision making given the prior knowledge of a context. The lowest-level musical intention is probably a note played in a specific fashion on a specific instrument.

There is a diversified set of objects spanning the gap between the lowest-level musical intention (cognition, psychology, musicology) and a simple wave form (physics). These objects will be referred to as *musical gestures* and they should be seen as the features based on which musical intentions will eventually be recovered through some decision making. Here, the terms "decision making" should be taken fairly loosely as the author does not intend to trivialize music understanding [6]. Back to our analogy with language, these objects would be analogous to formants, phonemes and intonation. Implied by the preceding diagram is also the claim that although the information fed to an instrument and to a listener's brain have different natures, they share similar levels of abstraction. This similarity motivated the author to label them both as musical gestures.

Such a definition by the negative for musical gestures clearly offers a wide range of interpretations and choices. The choice of a specific set of musical gestures will typically be a function of its purpose within a specific system. However, this discussion provides some insight as to an appropriate level of description that these gestures provide. As a representation of the audio stream's musical content, musical gestures should live in an intermediate level between wave forms and musical intentions. They should address basic perceptual and musical concerns such as harmonic structures and modulations in order to capture the musical intentions of the input wave form at a higher level of description. Yet, they should not attempt to describe the musical content of the input at too high a level of abstraction which would require a musical context and understanding [1,6].

For an application such as ours, the representation that is provided by the chosen musical gestures does not need to be complete. That is to say that while they offer valuable information concerning the input stream, they do not need to provide the ability to reproduce a perceptually identical audio stream.

3 Harmonic Structures Likelihood

3.1 Suggestion

In the light of our motivations, we wish for a set of musical gestures which, without pretending to provide an accurate transcription of the input musical stream, will provide relevant information concerning the predominant harmonic structures that are exhibited by the input stream. Such objects are obviously related to the kind of analysis that commonly take place in the first stages of automatic transcription or auditory scene analysis systems [2, 3]. Their estimation will typically involve some frequency analysis followed by some frequency and time grouping. However, the purpose of these analysis stages in our context differs greatly from those systems. Indeed, these musical gestures are to remain nothing more but time varying measurements and they shall not attempt to identify any musical intention at a higher level of abstraction.

We suggest the following as a simple but reasonable set of musical gestures for our purpose. We define a *soft key* as an object that describes both the energy and the pitch of a single harmonic structure whose fundamental happens to live within the range that was assigned to that soft key. The "soft" term comes from the fact that one could visualize this object as a rubber version of a standard key on a keyboard; both its pitch and its volume would be controlled through continuous changes of pressure and position instead of confining itself to onset and offset information. Our set of musical gestures will essentially be the state of a chosen set of soft keys. The energy of each soft key will be assigned accordingly to the input's energy as well as the likelihood of the harmonic structure that the soft key describes. Also, the frequency of each soft key will be finely tuned within the pitch range that was assigned to it. In what follows, we will briefly discuss three stages (tonal component analysis, harmonic grouping, soft key dynamics) that lead to the estimation of the suggested musical gestures.

3.2 Stage 1: Tonal components

The first stage of the analysis examines the frequency content of an instantaneous snapshot of the incoming signal. We compute an FFT-based estimation of the power spectrum distribution (PSD). Let L be the size of this FFT and Fs be the sampling frequency of the input stream. While the frequency resolution of this PSD (Fs/L) might not appear to be sufficient, it should exhibit peaks around the principal tonal components of the input. As shown in *Figure 3*, we retrieve these peaks and compute a higher resolution estimate of their frequency based on the an instantaneous frequency estimation. While one might think that such an estimation should require at least two FFTs applied to two time-shifted version of the snapshot, a simple approximation makes a decent instantaneous frequency estimation possible from the exploitation of aliasing between the frequency bins of a single FFT applied with a rectangular window [4].

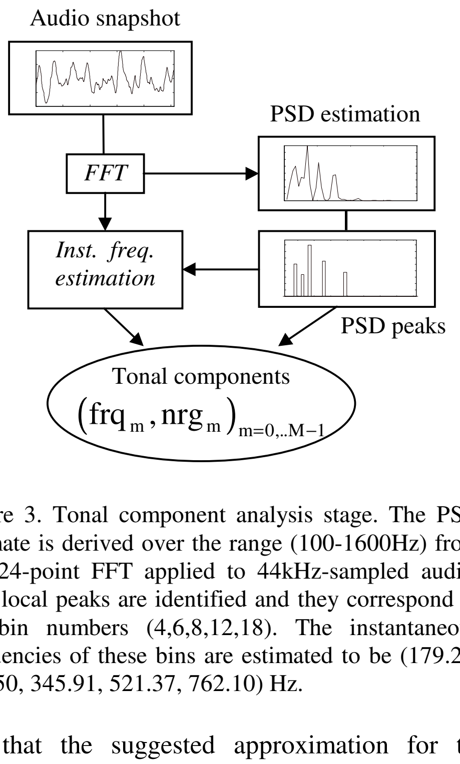

Figure 3. Tonal component analysis stage. The PSD estimate is derived over the range (100-1600Hz) from a 1024-point FFT applied to 44kHz-sampled audio. Five local peaks are identified and they correspond to the bin numbers (4,6,8,12,18). The instantaneous frequencies of these bins are estimated to be (179.24, 268.50, 345.91, 521.37, 762.10) Hz.

Note that the suggested approximation for these instantaneous frequencies will stop being meaningful in the cases where two tonal components are spaced within less than 2 bins (that is 2Fs/L in Hz). This is the criterion according to which the size L of the FFT analysis window will be chosen.

This stage provides a list of tonal components and their associated frequencies (from the instantaneous frequency estimation) and energies (from the PSD estimation).

3.3 Stage 2: Harmonic grouping

We now wish to further summarize the information provided by the previous stage by grouping the estimated tonal components according to their harmonic relationships. This stage can be seen as the decomposition of our tonal component information onto a non-orthogonal basis of harmonic structures. Projecting our tonal components onto each one of these harmonic structures provides a measurement of these structures' likelihoods. It is obviously impractical to attempt this decomposition literally as the vectorial space's dimension is infinite. Fortunately, our vector of tonal component will typically be sparse enough to lead to a fair estimation of this decomposition with minimal computations.

Without going in great details, we first summarize eventual harmonic relationships within the tones via a sparse matrix H as follows.

For each $(\mathrm{i}, \mathrm{j}) \in \{0,..,\mathrm{M}-1\}^2$, we define:

$\mathrm{frac}_{i,j} = \mathrm{frq}_j / \mathrm{frq}_i$; $\mathrm{frac\,int}_{i,j} = \mathrm{round}\left(\mathrm{frac}_{i,j}\right)$; and

$\varepsilon_{i,j} = \left| 1 - \frac{\mathrm{frac\,int}_{i,j}}{\mathrm{frac}_{i,j}} \right|$, which measures the harmonic relationship between the tonal components (i,j) in terms of an error. the matrix H is then define as:

$$H_{i,j} = \begin{cases} \mathrm{frac\,int}_{i,j} & \text{if } \varepsilon_{i,j} < \varepsilon_{th} \\ 0 & \text{otherwise} \end{cases}$$

This matrix is subsequently used to estimate the likelihood of each tonal component as being the fundamental of a harmonic structures within the input signal. This is done via a projection followed by a masking process which makes up of the non-orthogonal basis of harmonic structures. Finally, the fundamental frequencies that are associated with the resulting most likely harmonic structures are derived from weighted averages of the tonal components' frequencies that contribute to these structures.

Following the example of figure 3, the five tonal components (179.24, 268.50, 345.91, 521.37, 762.10) Hz lead to the following value for H:

$$H=\begin{bmatrix}1&0&2&3&0\\0&1&0&2&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix} \text{ (with } \varepsilon_{th} = 0.05)$$

The derived projection, masking and remainder eventually lead to three most likely harmonic structures which frequencies are estimated to be (174.90, 263.66, 762.10) Hz and which associated likelihoods are respectively (0.5401, 0.3240, 0.1359).

3.4 Stage 3: Soft Key Dynamics

Up until now, the measurements that we've gathered have been solemnly relying on the observation of a single snapshot of the input stream. These measurements will not qualify as musical gestures until they are taken within the context of their time evolution through some physically meaningful dynamics. In order to do this, we associate a trivial dynamical system (which is essentially a low-pass filter) to the behavior of each soft key. For each time frame, the estimated most likely harmonic structures are assigned to the appropriate soft keys (pitch range). All the other soft keys are assigned nominal inputs such as an instantaneous likelihood of zero and an instantaneous fundamental frequency equal the center of the pitch range assigned to that soft key. These inputs (both likelihood and frequencies) are then integrated over time through the dynamics of each key.

Our musical gestures will consist in the state of our set of soft keys as they are excited by the instantaneous harmonic structures likelihoods and frequencies that are estimated at each time frame.

4 A Complete System

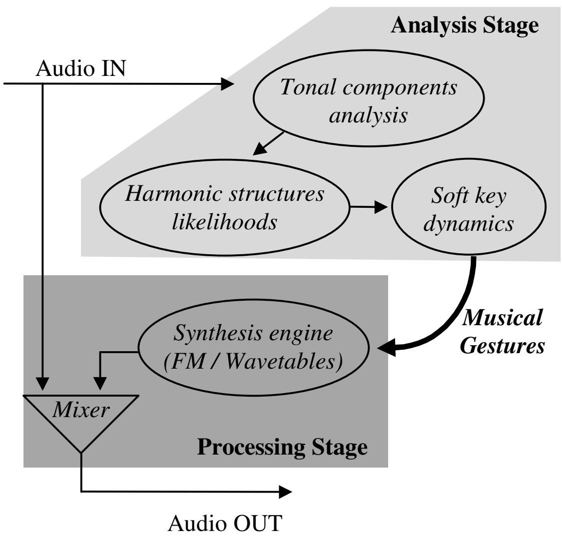

Figure 4. A complete illustration as implemented in C for Win32 and an Analog Devices Shark chip.

The nature and purpose of the processing stage is not a central issue to this paper. The processing stage may chose to use the leverage of the information provided by the musical gestures in many different ways. One of the most ambitious design choices is illustrated in figure 4 with the choice of a synthesis engine for the processing stage. It is an ambitious choice because the resulting audio processing effects is more likely to exhibit the imperfections of our analysis stage. The basic requirement for a valid synthesis engine is dictated by the nature of the musical gestures that will serve to control it. In our case, the control set of the synthesis engine is our set of soft keys. As a proof of principle, we've experimented with simple FM operators [4] and wavetables.

5 Conclusions

As the audio output ceases to be a direct transformation of the input wave-form, the suggested framework leads to a wider timbral freedom. At the same time, the lack of any ambitious musical decision within the analysis stage maintains the responsiveness and expressiveness of more traditional effects-based synthetic instruments. The construction of the system that we've presented should be seen as an exercise that aims to show the benefit of a set of musical gestures as an expressive, yet humble, representation of some relevant musical information. We adopt the philosophy that much expressiveness can be gained when this information is never dissociated from its ambiguity. Rather, we carry this ambiguity throughout our processing and leave it to the listener's brain to resolve it.

The analysis stage that was illustrated in this paper was purposely simplistic and the use of more popular multiscale frequency analysis [7] may reveal to be more appropriate. Also, other sets of musical gestures could be chosen. For instance, harmonic likelihoods may be inappropriate in the context of monophonic sounds or even speech [4,8,9]. Finally, the nature of the audio processing stage is obviously not limited to synthesis engines within the general framework we suggested.

6 References

[1] J. Blacking. "How musical is man?" University of Washington Press, 1973.

[2] D.P.W. Ellis. "Prediction-driven computational auditory scene analysis." Ph.D. dissertation, EECS department, MIT, 1996.

[3] P. Fernández-Cid, F. J. Casajús-Quirós. "Multi-pitch estimation for polyphonic musical signals".

ICASS 1998.

[4] E. Métois. “Musical sound information.” Ph.D. dissertation, MAS department, MIT, 1996.

[5] J. Chowning. “The synthesis of complex audio spectra by means of frequency modulation.” Reprinted “Foundations of Computer Music”. MIT press, 1985.

[6] S. Handel. “Listening.” MIT Press, 1989.

[7] S.N. Levine, T.S. Verma, J.O. Smith III. “Multiresolution sinusoidal modeling for wideband audio with modifications.” ICASS 1998.

[8] W. Oliver, J. Yu, E. Métois. “The singing tree, design of an interactive musical interface.” [http://www.acm.org/sigchi/dis97/dis.htm]

[9] W. Oliver. “Voice analysis.” [http://theremin.media.mit.edu/~woliver/]