

ISE575/CSCI575

Week 6  
February 17, 2005

Mohammad Anwar Hossain  
Student ID: 8813190224

## **Automatic Transcription of Musical Recordings**

*By - Anssi Klapuri, Tuomas Virtanen, Antti Eronen, Jarno Seppänen*

### **In one sentence, what the paper is about?**

A brief description of a music transcription system of real world music by adding a non-harmonic noise suppressor and number-of-voice estimation system to the author's earlier presented algorithm.

### **What is Music Transcription and Why we need it?**

Transcription of music is the process of analyzing an acoustic musical signal so as to write down the parameters of the sounds that constitute the piece of music in question. Traditionally, written music uses note symbols to indicate the pitch, onset time, and duration of each sound to be played. However, written music is primarily a performance instruction, rather than a representation of music.

Music transcription allows musicians to reproduce and modify the original performance. Some other application could be - structured audio coding, searching musical information, Music analysis etc.

### **Background on the topics and the previous efforts:**

Until these days, music transcription has fallen clearly behind humans in accuracy and flexibility. Some attempts made in this field are - Martin proposed a system that utilized musical knowledge in transcribing four voice piano compositions, Kashino et al. describe a model which was able to handle several different instruments, Goto's system was particularly designed to extract melody and bass lines from real-world musical recordings and so on.

### **Discussion on Proposed Methods:**

The focus of this paper is the 2 extensions to some previous algorithms (e.g. onset detection and multipitch estimation) that the authors presented earlier. For the onset detection they used algorithm using differentiation of the logarithm of the amplitude envelopes at each band. In this case, oscillations in the amplitude envelope do not matter too much after the sound has set on. The multipitch estimation algorithm consists of 2 parts - *Estimate # of voices* and *Subtraction of sound from mixture* which are applied in iterative succession.

The first of 2 extensions is Noise detection & suppression. Noise detection in music is quite different from that of speech. In music recording, practically there is no continuous noise that could be estimated over a long period of time. Instead non-harmonic parts are due to drums and percussive instruments which are transient-like in nature and short in duration. Thus the authors proposed an algorithm which estimates and removes noise independently in each analysis frame. Even though the authors tried different methods to remove additive and convolutive noise but they found that removing both types of noise simultaneously tends to work better as mentioned RASTA spectral processing.

The second of 2 extensions is Estimating the # of concurrent voices. The authors took a statistical approach. Random mixtures from zero to six concurrent harmonic sounds were generated by allotting sounds from 26 musical instruments. The mixtures were then contaminated with pink noise or random drum sounds, signal-to-noise ratios (SNR) varying between 23 dB and -2 dB. It turned out to be necessary to perform polyphony estimation using two different models. The first detects voicing, i.e. if there are any harmonic sounds in the input, and the second estimates the number of concurrent voices, if any.

Drum sounds turned out to be the biggest problem in voicing detection. Approximately half of the acoustic energy of the sound of bass drums, snares and tom-toms is harmonic, resulting from the drum

membrane which vibrates at mode frequencies. This tends to mislead a voicing detector. On the contrary for pink noise alone, the voicing detector can be designed to work almost perfectly, even though the harmonic sounds themselves vary from the double bass to the transverse flute.

In the case that a section in music has been determined to be voiced, another model is used to control the stopping of the iterative multipitch estimation system, i.e., to estimate the number of sounds to be extracted. The likelihood  $L_i$  of a sound detected by the predominant pitch estimator at iteration  $i$  was again a single best feature for controlling the iteration stopping.

The last step was sound separation and stream formation. Provided that the correct sounds are detected by the multipitch estimator, and that drums do not dominate a musical signal too badly, separation works rather well. A preliminary attempt towards stream formation from the separated notes was performed by utilizing acoustic features used in musical instrument recognition research. Mel frequency cepstral coefficients, the fundamental frequency, the spectral centroid, and features describing the modulation properties of notes were used to form 17 dimensional feature vectors, which were then k-means clustered. Based on the observations, stream formation according to sources is possible provided that the timbres of the sound sources are different enough, and that the distinctive characteristics do not get lost in the separation process.

### **Conclusion:**

This paper was very brief and seems like rushed piece of work. At the time of writing this paper, the transcription system suffered from defects and that's why authors did not publish the result for signal synthesized from MIDI. Another defect of the process is that long-duration sounds are detected several times at successive onsets. This results in insertion errors. The authors under-estimated the skills of human to detect concurrent voice by saying they tend to miss some voices. But, when the authors described their result they came back to the problem that overestimation of voice is not a good thing and can cause disturbing sound. So if human actually miss any voice that is actually fine since they don't hear it anyhow so why bother.