

# Engineering Approaches to Music Perception & Cognition

Baharak Zali  
8679928115

Homework # 5  
Feb. 17, 2005

## Review #2

---

### **Automatic Transcription of Musical Recordings, by Anssi Klapuri, Tumas Virtanen, Antti Ernen, Jarno Seppanen**

The goal of the research was automatic transcription of real-world musical recordings. Transcription is defined as the act of listening to a piece of music and writing down the musical scores. Until recent years, the automatic systems worked worse than human transcriptionists but during the past years such systems has become more advanced.

This paper suggests an extension to the authors' earlier work: *signal processing methods for detecting the beginning of discrete acoustic events in musical signals*, and *estimating the multiple pitches of concurrent musical sounds*. The system included three consecutive parts: *temporal segmentation*, *multi-pitch estimation*, and *sound separation and stream formation*. The multi-pitch estimation part itself, comprised two processes: *predominant pitch estimation*, and *removal of the detected pitch from the mixture*. In the predominant pitch estimation, the pitch of the most prominent sound is estimated in presence of other harmonic and noisy sounds and this is done by utilizing the harmonic concordance of simultaneous spectral components. Removal of the detected pitch from the mixture uses the fact that the spectral envelopes of real sound sources tend to be continuous. After each pitch is detected, it is extracted from the signal and the whole process is repeated for the residual signal.

In their new work, two processes are added to the multi-pitch segmentation, to reduce the error rate and improve the performance of the transcription system; those are *noise suppression*, and *iterative voice number estimation*. With these extensions, multi-pitch estimation includes four steps now. The first step is noise suppression in which different types of noise existing in the signal should be suppressed. It is been experienced that if the *additive* and *convolutive noise* are removed from the signal consecutively, the result would not be acceptable; only when both additive and convolutive noise are removed simultaneously the result is satisfiable. To get the best results, the additive noise is estimated and subtracted in the power spectrum, and for the convolutive noise, the logarithm is taken and an estimate of the convolutive noise is subtracted in logarithmic magnitude. The next step is predominant pitch estimation, then the removal of the detected pitch, followed by estimation of the number of pitches, which will be run iteratively until all the pitches are detected and stored. The second extended step is estimation of the number of pitches that requires two models to do the polyphony estimation. These two models are, voice detection, and estimation of the number of sounds. At first, system should detect whether a voice, defined as a harmonic sound, exist in the input signal, and then if there was any, it should estimate the number of voices.

Sound Separation and Stream Formation is the last step in the process, which includes a 2-stage model: first, finding initial sound parameters by applying multi-pitch estimator, and second, estimating more accurate and time-varying sinusoidal parameters. Stream formation from separate notes was attempted by using acoustic features in musical instrument recognition research. Stream formation is a difficult job and it is possible only if timbers of the sound sources are different enough and also if distinctive characteristics do not get lost in separation process.

For the experiments, the authors used artificially created harmonic signals, created by mixing 0-6 random number of harmonic sounds, from 26 musical instruments. Then they added pink noise and random drum sounds to the mixture. The signal to noise ratio, SNR, plays an

# Engineering Approaches to Music Perception & Cognition

Baharak Zali  
8679928115

Homework # 5  
Feb. 17, 2005

---

## Review #2

---

important role in the whole process and five different SNRs between 23 dB and -2 dB were used in the experiments. Their results showed that for voicing detection experiments with presence of drum noise, both 93 ms frame size, and 193 ms frame size, had the same *extraneous voicing* error rate (1.8%), and *undetected voicing* error rate was 6.1% for the former and 1.6% for the latter.

The results of the estimation of the number of concurrent voices showed that up to 3 concurrent voices are estimated almost perfectly, but the system has difficulty with detecting the concurrent voices when their number passes four. Even human listener has a difficulty in recognizing all the concurrent voices when they are more than four. They have tested the estimation performance for two different frame sizes, 93 ms and 190 ms, and with two different types of noise, pink noise and drum noise. Among all four combinations, I believe the results for input signals with 190 ms frame size and pink noise was the best: perfect estimation for 1-3 concurrent voices but not very accurate when the number of concurrent voices were five, and six. They did not provide any results for the experiments on real-world musical recordings, since as they mention, the synthesized music, MIDI, has the scores in it while comparing those for real music is not easy.

In detection or estimation of the pitches there are three types of errors: first type is *deletion error*, in which a pitch is not detected and it is the least disturbing error. The second one is *insertion error* that happens when a non-existing voice is detected due to errors somewhere in the path. This type of error is the most disturbing error and its rate should be kept low. The last error type is erroneous error, which occurs when a pitch is detected incorrectly.

Simulation results are averaged over 3 SNRs: 23dB, 13 dB, & 3dB. To calculate the error results, the insertion, deletion and erroneous errors are summed and divided by the sum by the number of notes. The final results showed that about 66% of the errors were insertion errors, about 1% deletion error and the rest erroneous ones. The error rates were almost twice in 93 ms frames compared to 190 ms frames and they dramatically increased by increasing in the number of polyphony. The minimum error was seen in 190 ms when the number of polyphonies was 1 (6.9%) and the highest was seen in 93 ms frame with 6 polyphonies (61%). A disturbing defect that was observed was that long-duration sounds are detected several time at successive onsets.

The system still has a long way to become accurate with high number of polyphonies. It has not been tested on the noisy operating environments, and it is not tested on the real-world music.