Polyphonic Transcription with Deep Layered Learning

Anders Elowsson **Anders Friberg**
KTH Royal Institute of Technology
elov@kth.se afriberg@kth.se

ABSTRACT

This short informal abstract introduces a polyphonic transcription system submitted to MIREX 2014. Here we will simply give an introduction to our main ideas concerning MIR and present some results from the MIREX homepage, whereas details concerning the implementation is reserved for a potential future publication. This is necessary, as the implementation is an ongoing project, where some steps are likely to be added and some will be removed.

1. INTRODUCTION

In our research at KTH Royal Institute of Technology we are trying to develop perceptually relevant representations of music audio [1-2]. One idea is that intermediate layers will allow us to solve many Music Information Retrieval (MIR) tasks, as in [3]. Music engages us at an emotional level while at the same time featuring a complex hierarchy of components interrelated through the time dimension. We believe that to accurately model music we must:

- Increase our ability to unweave the layers of information that it consists of.
- Also attain some intermediate representations without human supervision (i.e. Deep Learning).

One of the layers that we are seeking to discern is that of fundamental frequency in music. It seems reasonable to assume that this task could (just as emotional response [2]) be disentangled by an automatic (or supervised) layer-wise decomposition of the underlying perceptual phenomena.

2. METHOD

We have used multiple spectrograms with different time/frequency-resolutions (e.g. 1024 samples, 2048 samples, 4096 samples) and are using a 256 samples resolution in the time direction. We are using a Hanning window for the FFT and phase information from the spectrograms is being discarded at this stage in the implementation process. The range for the transcription was set to midi pitch 26-105 (drop D bass to high pitched human whistling) and frequencies in the spectrogram above 14 kHz were discarded. However, no frequency bins in the bass was discarded. No restrictions regarding polyphony level were made and no musicological assumptions were used.

We have specifically been focused on the detection of onsets with a corresponding pitch, as we believe that they are an important feature in music perception. Spectral information is transformed into higher-level representations within a machine-learning framework, from which onsets and offsets are extracted. A frame-based output is also supported, as it is possible to utilize information of onset and offset as well as pitch characteristics. Further details are TBA.

3. RESULTS

The system achieves high results as presented in Table 1.

Task 2	Precision	Recall	Ave. F-Measure
All	0.843	0.807	0.821
Piano	0.845	0.764	0.802

Table 1. The results for note tracking (onsets only) in the MIREX multiple fundamental frequency estimation tasks.

One important factor seems to be the system's ability to avoid false detections of incorrect onsets, measured as the *Precision*. The system's ability to find correct onsets (*Recall*) is also quite high. The *F-Measure* is the harmonic mean of *Precision* and *Recall*.

It is hard concretize the results without a proper comparison. How does the system compare with other submissions? Results are presented in relation to the other submissions of the multiple fundamental frequency estimation tasks, since the present set of music excerpts were established in 2009. In Figure 1 the main result (*Ave. F-Measure* for all excerpts) is presented.

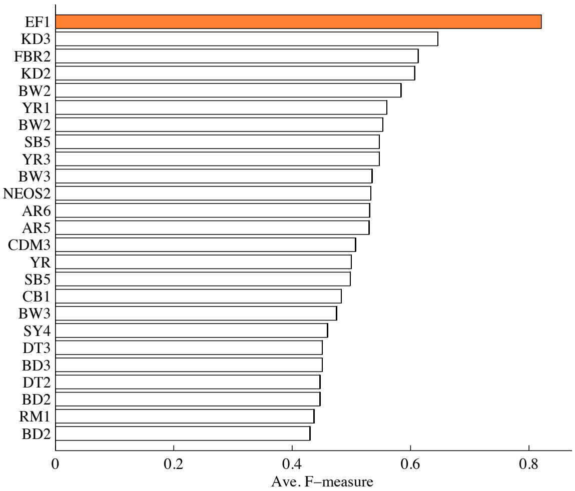

Figure 1. The average *F-Measure* (onsets detection) for the 25 highest performing contributions (of 52). Our system is represented by the orange bar.

Another statistic reported is the ability to identify both onsets and offsets. This is an error prone task as perceptual offsets are hard to define in a consistent manner, and we will simply conclude that our system seems to handle the task well.

Many systems are focused on piano music transcription, so a set of piano recordings is evaluated separately as presented in Table 1. How does our system compare with submissions specialized for the task? As shown in Figure 2, it performs well.

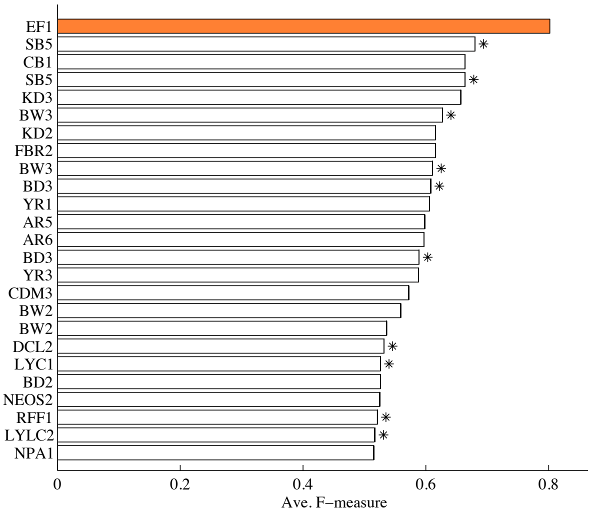

Figure 2. The average *F-Measure* (onset detection) for the 25 highest performing contributions (of 53) for piano recordings. Our system is represented by the orange bar. The systems that were designed with the purpose to transcribe piano-music are highlighted with an asterisk (*).

Results for the ability to find the active fundamental frequencies in each frame are shown in Table 2.

	Precision	**Recall**	**Accuracy**
Task 1	0.857	0.775	0.723

Table 2. The results of the frame-based evaluation in the MIREX multiple fundamental frequency estimation tasks.

Here *Accuracy* is used instead of *F-Measure* and it represents the number of correct detections in relation to the number of annotations + the number of erroneously given detections. We note that the system is achieving the highest *Accuracy* reported so far, among 51 contributions.

One drawback of our submission is that it did not have real-time properties due to an excessive feature calculation step. This step will be omitted in the final implementation and we are at this point expecting the run-time to be around 0.7-0.9 of the audio length.

4. CONCLUSIONS

Our submission performed well for polyphonic transcription. Both onsets and offsets seem to be located accurately. The system seems to generalize well for different types of instrumentations (we only explored a few piano excerpts in the development of the system, but achieved high results in the piano task in relation to systems specialized for that instrument). It is also convenient that a frame-based output is supported by the system. We expect the system to be useful in our quest to decipher emotion in music.

5. DISCUSSION

For next year's MIREX we hope to extend this system to handle Automatic Melody Extraction (AME) as we believe that (AME) is a key factor for our understanding of musical meaning.

It is our understanding that this task have some risks of explicit or "implicit" overfitting from knowledge of e.g. style, timbre and onset characteristics of some of the instruments in the test sets. We and many others have been exploring e.g. timbral properties of instruments in some of the recordings in the publically available Bach10 set, which seems to have some examples also featured here. We could define a better protocol or similar in the task page[1] if it is of importance. Or we could develop a new hidden set where neither instrumentation/timbre nor polyphony level are known to the participants. If any authors would like to join us in a discussion of how we could finance the development of such a test set (via a third party), please contact the first author. We have not seen any discussion of this although it should be relevant.

6. REFERENCES

[1] A. Elowsson and A. Friberg: "Modelling the Speed of Music Using Features from Harmonic/Percussive Separated Audio," In Proc. of ISMIR, pp. 481-486, 2013.

[2] A. Friberg, E. Schoonderwaldt, A. Hedblad, M. Fabiani, and A. Elowsson. "Using listener-based perceptual features as intermediate representations in music information retrieval," The Journal of the Acoustical Society of America, 136(4), 1951-1963, 2014.

[3] A. Elowsson and A. Friberg: "Tempo Estimation by Modelling Perceptual Speed," In MIREX Audio Tempo Estimation task, 2013.

[1] http://www.music-ir.org/mirex/wiki/2014:Multiple_Fundamental_Frequency_Estimation_%26_Tracking