SCORE-INFORMED ESTIMATION OF PERFORMANCE PARAMETERS FROM POLYPHONIC AUDIO USING AMPACT

Johanna Devaney
Ohio State University
School of Music

Michael I. Mandel
Brooklyn College, CUNY
Computer and Information Science

ABSTRACT

A musical performance can convey both the musicians' interpretation of the written score as well as emphasize, or even manipulate, the emotional content of the music through small variations in timing, dynamics, and tuning. This paper describes the latest developments in a suite of automatic software tools for quantitatively analyzing musical performances for which a corresponding musical score is available, entitled the Automatic Music Performance Analysis and Comparison Toolkit (AMPACT). AMPACT uses a score-informed approach to estimate timing, pitch, and loudness parameters from both monophonic and, now, polyphonic audio. Robust extraction of higher-level timing, pitch, and loudness performance descriptors requires precise frame-level estimation of note onsets and offsets, fundamental frequency, and power. This paper describes the score-informed approaches implemented in AMPACT for this frame-level estimation in polyphonic audio.

1. INTRODUCTION

Precise, frame-level estimation of signal properties is a necessary first step in empirically measuring musical performance parameters. While numerous solutions exist for extracting this type of information from monophonic audio, e.g., [3], estimating it from polyphonic audio remains an unsolved problem. Score-guided approaches offer a means of reducing the complexity of the problem that blind transcription methods face, by providing an indication of which time-frequency regions of the signal are associated with each musical note once the score has been aligned to the audio. This paper describes the score-informed approaches implemented in the current version of the Automatic Music Performance Analysis and Comparison Toolkit (AMPACT) [1] [6] for estimating notes onsets and offsets, fundamental frequency (f_0), and power.

[1] http://www.ampact.org

2. PERFORMANCE PARAMETERS

2.1 Timing

AMPACT uses a hybrid dynamic time warping (DTW) / hidden Markov model (HMM) approach to estimate note onset and offset locations, includes asynchronies between notes marked as simultaneities in the musical score. The use of the DTW alignment removes the need to encode information about the score in the HMM. Specially, by assuming that the DTW alignment is roughly correct, we do not need to rely excessively on noisy f_0 estimates in the HMM. DTW is used in the first pass to obtain a rough estimate of the note locations, but rather than running the HMM on the entire signal, the HMM polyphonic algorithm refines the offset-onset transitions between groups of "simultaneous" notes in the DTW alignment in order to estimate the location of the onsets and offsets for each voice. The HMM assumes that the DTW is roughly correct and only looks at the audio 125 ms before and after the onset identified by the DTW alignment, thus it is only able to correct errors in the DTW alignment by a maximum of that amount. A visual representation of the DTW alignment allows for detection of gross errors, which can be manually corrected. The details of the algorithm are available in [4].

2.2 Fundamental Frequency

In order to obtain f_0 estimates for each note we first extract observations close to the expected frequencies of the harmonics of the fundamental (including the fundamental itself) based on the initial f_0 value from the aligned score. The simplest approach is to use the central bin values of the discrete Fourier transform (DFT). In this DFT approach, we convert these frequencies to the frequency of the corresponding fundamental by dividing by the harmonic number of the closest harmonic, and then take the mean of these frequencies weighted by their respective magnitudes. Mathematically,

$$\hat{f}_0 = \frac{\sum_n \sum_{\omega_i \in \mathcal{N}(nf_0)} \frac{\omega_i}{n} x(\omega_i)}{\sum_n \sum_{\omega_i \in \mathcal{N}(nf_0)} x(\omega_i)}, \quad (1)$$

where $x(\omega_i)$ is the cube root of the magnitude at frequency ω_i and $\mathcal{N}(nf_0)$ is the set of frequencies in the neighborhood of nf_0, the nth harmonic, here five DFT bins. Because the output of this process is a more refined estimate of f_0, we can use this new estimate as the basis for performing the same procedure again, leading to a further refined estimate. Through experimentation we found that

this process tends to converge to a stable estimate after 5–10 iterations, so we use 10 iterations in our calculations.

The DFT runs into problems however for signals composed of sinusoids that are spaced farther apart than the DFT frequency samples. In this case, several consecutive frequencies will be dominated by a single sinusoid, but treated by the DFT as separate sinusoids. An instantaneous frequency (IF) approach, on the other hand, will correctly identify the frequency of this sinusoid in all of them. In the IF approach, the frequency values are estimated from the time derivative of the phase spectrum according to [1, 2], as implemented in MATLAB by Dan Ellis [2] . The weights are still the cube root of the DFT magnitudes at the corresponding points. The instantaneous frequency provides a modified estimate of the frequency of the dominant sinusoid in each DFT bin, ω_i in (1). The IF features use a neighborhood size, $\mathcal{N}(nf_0)$ of 27 Hz, the equivalent of two DFT bins, below and above the predicted frequency.

2.3 Power

The power estimates were derived from the same data that were used for the f_0 estimates in (1), except that instead of using cube root compressed magnitudes, they used squared magnitudes, designated $\tilde{x}(\omega_i)$. In particular, for a given estimated f_0, the power was estimated as

$$\hat{p}(f_0) = \sum_n \sum_{\omega_i \in \mathcal{N}(nf_0)} \tilde{x}(\omega_i). \quad (2)$$

By using a neighborhood larger than a single observation, this method is very unlikely to miss any target energy, but could include additional energy from simultaneous notes.

3. VALIDATION OF f_0 AND POWER ESTIMATES

3.1 Test Data

We validated this approach using the Bach 10 dataset, which contains 10 four-part Bach chorales recorded by violin, clarinet, saxophone and bassoon for a 330 seconds of annotated multi-part audio [7], and a 40 second excerpt from the opening of "Kyrie" from Machaut's four-part Messe de Notre Dame recorded by soprano, alto, tenor, and bass [5]. The Bach10 dataset consists of hand annotated onset estimates for each notated simultaneity, while the Machaut recordings consist of hand annotated onsets and offsets for each individual monophonic line (thus accounting for timing asynchronies between musical lines). In the experiment, the available hand annotated timings were used instead of MIDI alignment to avoid propagating error from the onset/offset estimation step to the f_0 and power estimation. In order to account for discrepancies in the temporal annotations, particularly the absence of offsets in the Bach10 dataset, only the central 80% of the frames for each note were used in the evaluation. MIDI note information corresponding to each onset was also provided in both datasets. The combination of timing and MIDI note information was used to specify time-frequency regions of interest in the signal used by the algorithms described above.

[2] http://labrosa.ee.columbia.edu/projects/coversongs/ifgram.m.html

3.2 Ground Truth

The score-guided estimates of frame-wise f_0 from the polyphonic mixture of four voices were evaluated against the frame-wise f_0 estimates calculated on the original monophonic tracks using a MATLAB implementation [3] of the YIN algorithm [3]. The window size was set adaptively by dividing the sampling rate (sr) by the minimum f_0 specified. Both a minimum and maximum f_0 estimate were set adaptively to the expected frequency of the note minus and plus two semitones. The hop size was set to $32/sr$. The error between the estimates and the ground truth was measured in cents. To combine these errors across all of the frames of a note, a weighted sum was computed

$$E = \sqrt{\frac{\sum_n \left(\hat{f}_0(n) - f_0(n)\right)^2 w_n}{\sum_n w_n}} \quad (3)$$

where w_n is the weighting applied to the nth frame. This weighting was computed from YIN's estimate of the magnitude of the pitched component of the monophonic signal. Specifically, for a given time frame, YIN computes the total power at that frame, p_n and and estimate of the proportion of that power that is due to aperiodic components, a_n. We compute the weights as

$$w_n = \sqrt{(1 - a_n)p_n}. \quad (4)$$

The main motivation for using a weighting like this was to decrease the importance of low-energy regions, such as breaths and transitions between notes, where there is no true f_0 to speak of and all estimates are noisy.

3.3 Results

Overall the DFT and the IF approaches performed comparably against the YIN estimates for f_0 and power. Both approaches ran about 3x slower than YIN (2.82x for the DFT approach and 2.99x for the IF approach). For the f_0 estimates, the median error was 23 cents for the DFT approach and 21 cents for the IF approach, and for both 100% of thef_0 estimates were within 50 cents of the YIN estimates. The IF approach had an advantage, however, as it had both the lowest variability (as measured by the spread between the 25th and 75th percentiles) and did not smooth vibrato sections as the DFT approach did. In the power estimates, th IF marginally outperforms the DFT in terms of having a lower median error (2.2 dB vs 2.3 dB) against the YIN estimates and a smaller spread in its 25th and 75th percentiles.

4. CONCLUSIONS

These approaches to extracting the lower-level frame-wise descriptors of note onsets and offsets, f_0, and power from polyphonic audio have been implemented in the current version of AMPACT and integrated into the existing algorithms in AMPACT for describing higher-level note-wise descriptors of timing, pitch, and loudness.

[3] http://audition.ens.fr/adc/sw/yin.zip

5. ACKNOWLEDGEMENTS

This work was supported by the National Endowment for the Humanities' Office of Digital Humanities [grant number HD-228966-15].

6. REFERENCES

[1] Toshihiko Abe, Takao Kobayashi, and Satoshi Imai. Harmonics tracking and pitch extraction based on instantaneous frequency. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 756–759, 1995.

[2] Toshihiko Abe, Takao Kobayashi, and Satoshi Imai. The if spectrogram: a new spectral representation. In *International Symposium on Simulation, Visualization and Auralization for Acoustics Research and Education*, pages 423–430, 1997.

[3] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[4] Johanna Devaney. Estimating onset and offset asynchronies in polyphonic score-audio alignment. *Journal of New Music Research*, 43(3):266–275, 2014.

[5] Johanna Devaney and Daniel P. W. Ellis. An empirical approach to studying intonation tendencies in polyphonic vocal performances. *Journal of Interdisciplinary Music Studies*, 2(1-2):156, 2008.

[6] Johanna Devaney, Michael I. Mandel, and Ichiro Fujinaga. A study of intonation in three-part singing using the automatic music performance analysis and comparison toolkit (ampact). In *International Society for Music Information Retrieval Conference*, pages 511–6, 2012.

[7] Zhiyao Duan, Bryan Pardo, and Changshui Zhang. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2121–2133, 2010.