AN EVALUATION OF SCORE-INFORMED METHODS FOR ESTIMATING FUNDAMENTAL FREQUENCY AND POWER FROM POLYPHONIC AUDIO

Johanna Devaney*
Ohio State University
School of Music
Columbus, OH

Michael Mandel
Brooklyn College, CUNY
Computer and Information Science
Brooklyn, NY

ABSTRACT

Robust extraction of performance data from polyphonic musical performances requires precise frame-level estimation of fundamental frequency ($f_0$) and power. This paper evaluates a new score-guided approach to $f_0$ and power estimation in polyphonic audio and compares the use of four different input features: the central bin frequencies of the spectrogram, the instantaneous frequency, and two variants of a high resolution spectral analysis. These four features were evaluated on four-part multi-track ensemble recordings, consisting of either four vocalists or bassoon, clarinet, saxophone, and violin (the Bach10 data set) created from polyphonic mixes of the monophonic tracks both with and without artificial reverberation. Score information was used to identify time-frequency regions of interest in the polyphonic mixes for each note in a corresponding aligned score, from which $f_0$ and power estimates were made. The approach was able to recover ground truth $f_0$ within 20 cents on average in reverberation and power within 5 dB for anechoic mixtures, but only within 10 dB for reverberant.

Index Terms— fundamental frequency estimation, power estimation, score-informed signal processing

1. INTRODUCTION

Precise, frame-level estimation of performance parameters from audio 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., [1], estimating this information 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 the time-frequency regions of the signal associated with each musical note.

This paper describes the evaluation of a new score-guided approach to fundamental frequency ($f_0$) and power estimation in polyphonic audio and compares four different input features: the discrete Fourier transform, the instantaneous frequency, and two versions of high resolution spectral analysis. The polyphonic audio-based $f_0$ and power estimates can successfully estimate $f_0$ and power measurements generated by the pYIN algorithm on corresponding monophonic audio tracks.

*Thanks to National Endowment for the Humanities Office of Digital Humanities (Grant number HD-228966-15).

2. BACKGROUND

Alignment between audio and a corresponding musical score can be used for estimating note onsets and offsets and generally performs with more precision than blind estimation algorithms. The use of a digitized version of a musical score, typically in MIDI, as a guide to facilitate transcription was first described in [2]. While earlier work provided a single time estimate for notated simultaneities in the musical score [3–7], more recent work has addressed the issue of identifying asynchronies between notes marked as simultaneities in the score [8–12], which allows for more precise guiding of signal processing algorithms.

2.1. Score-Guided Estimation of $f_0$

Polyphonic, frame-wise $f_0$ estimation and note tracking (the determination of frames-wise $f_0$ estimates that should be grouped together into notes) are extremely challenging. Overall, there has been an observable ceiling of 70% accuracy for the state-of-the-art methods, regardless of their approach [13]. This has been demonstrated in the recent history of the MIREX Multiple Fundamental Frequency Estimation & Note Tracking task, where submitted systems are evaluated on their ability to make frame-wise $f_0$ estimates (within 50 cents of the ground truth) as well as track notes and timbre: in 2016, the top system’s accuracy was 53.7% [14] on the MIREX data set; in 2015, it was 65.4% [15]; and in 2014, it was 72.3% [16]. A recent summary article of the state-of-the-art in music transcription identifies score-and perceptually-informed approaches as a potential way of breaking the current performance ceiling [13].

2.2. Score-Guided Estimation of Power

To date, there has been only been limited work on power estimation in polyphonic audio. [17] estimated note intensities in polyphonic piano recordings using a corresponding score. The score was aligned to the audio and used to create a parameterized model spectrogram, which achieved 83% accuracy. This was evaluated against the calculation of intensity from the energy in the spectrogram bins corresponding to the fundamental and first five harmonics of each note in the aligned score, which only achieved 66% accuracy.

3. EVALUATION OF SCORE-INFORMED APPROACH

3.1. $f_0$ Estimates

The evaluated approach takes features of a single time frame and an initial estimate of the $f_0$ from the aligned score as input. The input features consist of a set of magnitude measurements at a potentially
irregularly sampled set of frequencies, $x(\omega_i)$. These measurements come from a standard discrete Fourier transform, an instantaneous frequency-grain, the high resolution spectral estimator, and the same high resolution method applied after comb filtering at harmonics of the initial $f_0$ estimate\footnote{Code available at: http://www.ampact.org}. For any measurement type, we identify 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. We then 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 this can be expressed as

$$\hat{f}_0 = \frac{\sum_i \sum_{\omega \in \mathcal{N}(nf_0)} \omega x(\omega_i)}{\sum_i \sum_{\omega \in \mathcal{N}(nf_0)} x(\omega_i)},$$

where $\mathcal{N}(nf_0)$ is the set of frequencies in the neighborhood of $nf_0$, the $n$th harmonic. 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. In our experiments, this process tended to converge to a stable estimate after 5–10 iterations, so we use 10.

When applied to the discrete Fourier transform (DFT), the $\omega_i$ frequency values are uniformly spaced between DC and Nyquist. For the observations, we took the cube roots of the DFT magnitudes in order to apply a non-negative perceptual weighting\footnote{http://labrosa.ee.columbia.edu/projects/coversongs/ifgram.m.html} that can be used in a weighted sum. While the $\omega_i$ grid itself is fixed for a given DFT length and sampling frequency, the weighting in (1) can interpolate $f_0$ estimates between grid points.

The instantaneous frequency (IF) replaces the fixed frequency grid of the DFT with frequency values estimated from the time derivative of the phase spectrum according to [19, 20], as implemented in MATLAB by Dan Ellis\footnote{http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html}. The weights are still the cube root of the DFT magnitudes at the corresponding points. The instantaneous frequency essentially provides a modified estimate of the frequency of the dominant sinusoid in each DFT bin. For a signal composed of sinusoids that are spaced farther apart than the spacing of DFT frequency samples, several consecutive frequencies will be dominated by a single sinusoid and the instantaneous frequency will correctly identify the frequency of this sinusoid in all of them. 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.

We also use features from the high resolution (HR) sinusoidal estimation method of [21]. It uses a generalized version of the ESPRIT method to estimate mixtures of complex exponentials modulated by polynomials. By assuming signal components are sinusoidal, these estimation methods can achieve much more precise estimates of their frequencies. We then measure the frequency and cube root of the amplitude of each modulated complex exponential as $\omega_j$ and $x(\omega_j)$ for (1). The HR features use a slightly wider neighborhood, $\mathcal{N}(nf_0)$, of 40 Hz, because it uses fewer frequency bands than the IF analysis.

The HR tended to produce noisy estimates at the beginnings and ends of notes, where the signal tended to be noisier. To reduce this variability, we introduced a novel pre-processing step for the high resolution method (HR-C), which performed comb filtering using the sum of a bank of zero-delay, constant-Q, one-zero gammatone filters\footnote{http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html} with center frequencies placed at the harmonics of the estimated fundamental. The purpose of such a filter is to remove energy outside of the target frequencies, but to allow for larger error in the higher harmonics, where small changes in fundamental are amplified. See Figure 1 for an example filter response and the filters from which it was composed. We used the gammatone filter implementation from the Voicebox toolkit\footnote{http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html}, with third-order IIR gammatone filters, a Q-factor of 12, and the first 40 harmonics. The signal was filtered forward in time and then backward in time to achieve zero delay.

3.2. Power Estimates

The power estimates were derived from the same data as 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

$$\tilde{p}(f_0) = \sum_{n} \sum_{\omega_i \in \mathcal{N}(nf_0)} \tilde{x}(\omega_i).$$

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

4. EXPERIMENTS

4.1. Materials

The first part of the test set is 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[23]. The second part is 40 seconds from the opening of “Kyrie” from Machaut’s four-part Messe de Notre Dame recorded by soprano, alto, tenor, and bass [24]. 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 addition, note boundaries were further refined by keeping only the time samples for each note corresponding to the longest contiguous segment having a monophonic YIN periodicity estimate of at least 95%. MIDI note information corresponding to each onset was also provided in both datasets. The combination of
The use of the comb filter is quite effective at reducing these spurious estimates. In this experiment, we use the pYIN VAMP plugin\(^4\) with a window size of 2048 and the hop size of 256. The error between the estimates and the ground truth was measured in cents. This is based on the ground truth approach used in the MIREX evaluation, except that in that evaluation octave-errors made by YIN were hand-corrected. To combine these errors across all of the frames of a note, the root mean square error was computed

\[ E = \sqrt{\sum_n (f_0(n) - \hat{f}_0(n))^2}. \] (3)

The ground truth for the power estimates was calculated in a similar way, using the RMS estimates produced by the pYIN VAMP plugin. We noticed that many of the errors in estimation happened around the annotated onsets and offsets of notes, even with the use of YIN’s periodicity estimate to identify the periodic portion of each note. We thus provide results both on the entire notes and on the central 80% of each note, discarding the 10% of frames at the beginning and end.

### 4.3. Results

Figure 2 compares the estimates of the four score-informed algorithms on the polyphonic mixes against the pYIN estimates on the corresponding anechoic, monophonic tracks as a function of reverberation. It shows these comparisons for both the entire notes and the central 80% of frames of each note, trimming off the ends that contain most of the erroneous estimates due to transitions in and out of the notes. The figure shows that for \(f_0\), the high resolution (HR) estimates have the largest error, caused by numerous spurious estimates. The use of the comb filter is quite effective at reducing these spurious estimates.
peaks, by focusing the high resolution method on the neighborhood of the harmonics of the approximate fundamental. The IF and DFT estimates both perform better than the HR-C features, with the DFT slightly, but not significantly, outperforming the IF method. When the ends of the notes are trimmed off, the IF and DFT results are not affected much, but the HR and HR-C features perform much better. In this condition, the HR-C features perform comparably to the IF and DFT features, showing that many of the errors in the HR-C estimates occur at the beginnings and ends of notes. When reverberation is introduced, it tends to harm the IF estimates the most, causing the error to increase with increasing source-mic distance. This is likely caused by reverberant speech violating the assumption that there is a single sinusoid in each DFT bin.

For power estimation, shown in the bottom row of Figure 2, the IF and DFT features are equivalently accurate, followed by the HR-C and then HR features. The equivalence between IF and DFT makes sense, as they use the same magnitude estimates. They could differ in their inclusion or exclusion of bins in the weighted sum of 1 due to differences in frequency estimates, but these results suggest that they do not differ much. The results also show that the comb filtering operation is able to make the power estimation more accurate by removing energy from spurious frequency estimates. Estimating the original anechoic power of a signal in reverberant conditions is difficult. All of the features tested here perform worse for power estimation in reverberation, but maintain the same relative performance.

In order to use a similar metric to the MIREX evaluation, we also calculated the number of frames that were more than one semitone (50 cents) from the ground truth. For the anechoic version of the signals the DFT features had 2.6% frames beyond this threshold, IF 1.1%, HR 8.0%, and HR-C 1.5%. These results deteriorated with the introduction of reverberation, where the DFT features had 3.4%, IF 3.2%, HR 14.0%, and HR-C 2.6%. These values are all for the entire notes, with no trimming.

The run time of the DFT and IF features are comparable, with the IF running at 1.06 times the speed of the DFT. Extracting HR features is much slower, with the simple HR method running 14.49 times slower than the DFT and HR-C running 17.83 times slower.

A closer evaluation of the performance of the algorithms is shown in Figure 3. In the figure, (a) visualizes the $f_0$ trace of a four-note segment estimated by each of the algorithms (pYIN, DFT, IF, HR, and HR-C). The most notable differences between the $f_0$ traces is the larger number of spurious estimates from the HR algorithm in the first note (0.2–0.8 s) and the smoothing of the vibrato by the DFT in the fourth note (2.1–3.2 s). While (b) shows the power estimate trace of the same four-note segment as Figure 3(a). All of the estimates follow the same trend as each other, particularly for the first two notes (0.2–1.5 s). Although the HR without comb filtering has more spurious estimates. For the third note (1.5–2.0 s), pYIN’s estimate is smoother than those based on the polyphonic recording. This discrepancy may be due to activities in the other voices. In the fourth note (2.1–3.2 s), HR and HR-C consistently over-estimate the power compared to pYIN, IF, and DFT. This is likely an illustration of the larger trend exhibited in the frame-level results.

5. CONCLUSIONS

Overall, the proposed score-guided approach is able to estimate $f_0$ and power for individual voices from a reverberant polyphonic mixture. The DFT and instantaneous frequency features performed best in terms of having the lowest RMS error for both $f_0$ and power estimates. When omitting onset and offset portions of notes, the comb filtered high resolution features performed comparably as well. The use of the comb filter as pre-processing for the HR analysis significantly improves the performance of the algorithm for both $f_0$ and power estimation. The addition of reverberation caused some challenges with power estimation, but did not have much effect on $f_0$ estimation. The performance of the IF features deteriorated slightly in reverberation, but not significantly.

The next step in this project is to move from frame-wise measurements to note-wise estimates informed by perceptual models. Perceptually-informed note-wise estimates are important for applying these algorithms to the study of expressive musical performance. This is fairly straightforward in the case of perceived pitch, where a model need only be applied to the $f_0$ estimates, e.g., [26]. Perceived loudness estimation is more complicated because a cochlear model would need access to frame-wise estimates of all partials, e.g., [27]. In the future we will also investigate alternatives to using pYIN to generate ground truth, such as synthesized tracks using high-quality vocal models.
6. REFERENCES


