Multipitch Estimation Applied to Single-Channel Audio Source Separation: Relevant Techniques and Challenges

Alejandro Delgado Castro and John E. Szymanski

Audio Lab, Department of Electronics, University of York
Heslington, York, North Yorkshire, YO10 5DD. United Kingdom
adc533@york.ac.uk
john.szymanski@york.ac.uk

Abstract. The estimation of melody trajectories in single-channel polyphonic signals is a major field of study in digital signal processing, with principal applications in audio source separation, automatic music transcription and musical genre identification. The YIN algorithm and one multiple fundamental frequency estimator are tested using two different mixtures of harmonic sounds in order to identify advantages and limitations. A strategy is presented as a way to overcome these limitations and hence improve the accuracy of the estimated pitch trajectories.

Keywords: Single-Channel Audio Source Separation, Fundamental Frequency Estimation, Multipitch Estimators, Principal Melody Extraction, Pitch Trajectories, YIN algorithm.

1 Introduction

In general, when most western musical instruments are excited to produce a note, when viewed in the frequency domain the result is not an isolated frequency but is rather a series of energy peaks having different frequencies and amplitudes. The spectral envelope characterized by these components is one of the cues that human brain uses to distinguish between different instruments, while the musical note itself can be characterized by its fundamental frequency or perceived pitch[1]. Within this harmonic distribution of energy, the fundamental frequency works as the reference point for generating the rest of the harmonically related components of the note. However, the fundamental frequency does not always coincide with the highest peak in magnitude.

The number of fundamental frequencies present in a mixture of sounds is probably one of the most important parameters that can be used to estimate the number of sources, and it can also be used to further characterize those sources individually. Melody trajectories are usually generated when the pitch

[1] Despite the fact that *Fundamental Frequency* is usually considered as feature of the signal, while *Pitch* refers more to a perceptual measure, the terms will be used interchangeable across the article.

of one or more identified sources is tracked across time. In order to implement an accurate source separation system for harmonic sounds, a reliable fundamental frequency estimator has to be used to generate a reliable pitch trajectory for every source.

Many fundamental frequency estimation methods have been presented [2, 4, 14], both for monophonic and polyphonic signals, and some of them have proven to be an important preprocessing stage for audio source separation algorithms, in particular, for those methods concentrated in isolating or extracting harmonic sounds from single-channel music recordings.

The aim of this article is to give a general review of some established fundamental frequency estimators, namely the YIN algorithm [2] and one multipitch estimator by Duan et. al [4], focussing on their principal features and limitations. Several tests are presented here, using individual and mixed sounds, in order to reveal interesting characteristics and evaluate their performance. The contribution of this paper is to present some proposals on how to improve the accuracy of these algorithms.

2 Fundamental Frequency Estimation in Monophonic Signals

The problem of estimating fundamental frequencies in audio signals was first studied in monophonic recordings and was then applied to speech processing [11]. Since then, many other methods emerged and were specifically designed for music signals. The vast majority of existing algorithms divide the entire time-domain representation of the input signal into short portions, called frames, and then present fundamental frequency estimates for every frame. The so-generated sequence of pitches can be considered as an appropriate representation of the melody trajectory [5].

2.1 Classification

Pitch estimators can be grouped into several types according to the main principle or function that is used to approximate the set of fundamental frequencies. The most common types are listed and explained briefly below.

- **Zero-Crossing Rate.** This is probably the simplest and most inexpensive type of pitch estimator, and consists of counting the number of times the input signal crosses the reference axis (zero level) in order to detect periodicity. Although the method is simple, its results are unreliable when applied to noisy signals. It also struggles to deliver accurate results for those harmonic signals where the fundamental partial is not the strongest.
- **Autocorrelation.** Some of the most frequently used algorithms in pitch detection are based on autocorrelation functions. Periodicity in this case is indicated by the maximum of this function. Therefore, autocorrelation methods select the highest non-zero lag to compute the estimated period [14].

Algorithms in this category have proven to be relatively robust against issues caused by noise, but sensitive to particularities in the spectral characteristics of sounds [5].

- **Cepstrum Analysis.** Cepstrum-based pitch detectors were the first methods to be realizable through digital computation and were used as reference for other algorithms [5]. The cepstrum of the input signal is the inverse Fourier transform of the logarithm of its power spectrum. A peak-picking schema is used to find strong peaks within this function which indicates the underlying periodicity. Cepstral methods normally perform poorly in noisy environments and frequent octave errors have also been reported. However, good results have been obtained when dealing with formants in speech processing [5].
- **Harmonic Matching Methods.** This type of algorithm identifies a fundamental period by analysing a pattern of spectral peaks in the magnitude spectrum. When all the relevant peaks are identified, the most likely fundamental frequency is found that is consistent with the pattern of the observed peaks. Some related drawbacks usually emerge when the levels of noise are high, or when the deviation between ideal harmonics and real partials is also high.
- **Wavelet-based Algorithms.** Multiresolution and multi-scale analysis are techniques that have been also applied to pitch detection. In contrast to Fourier Analysis, the Wavelet Transform utilizes different resolutions to analyse high and low frequency regions. Hence, it can be considered as a form of constant Q frequency analysis.

2.2 The YIN Algorithm

One method that has been widely used in several application areas of pitch detection, is the so-called YIN Algorithm, developed by De Cheveigné et. al. [2]. It is based on the autocorrelation function and incorporates a number of modifications that are combined to prevent errors. According to [2], the most relevant features of the YIN algorithm are presented as follows.

- The error rates were reported to be three times lower than other competing methods.
- It can be applied to either speech or music signals.
- It is suitable for high-pitched voices or music since there is no upper limit on the frequency search range.
- The algorithm has a small number of parameters and any fine tuning of them is not required.

When the YIN algorithm is applied to a non-stationary monophonic signal, the result is a series of frequency estimates that can be used to construct the melody trajectory of the sound. Figure 1 shows the estimated melody trajectory of three different audio signals: a short excerpt of an aria for soprano, a modern viola playing the note A4 without vibrato, and a clarinet playing five different musical notes. The original audio recordings were taken from the Open Air Anechoic Audio Database [9].

The depicted results show that the estimation of fundamental frequencies is highly accurate during the sustain of every note. However, some problems arise during note transitions or silences, where the algorithm produces spurious peaks of short duration. Considering the pitch trajectory for the soprano voice, the spurious peaks can be spotted easily, and they correspond to short silences between different notes. For the clarinet sound, the first and fourth transitions are also problematic in this sense. On the other hand, it can also be observed that the algorithm was able to track the vibrato in those sung notes, which is advantageous if the melody trajectory is going to be used for source separation.

If the YIN algorithm is applied to a mixture of different sounds, i.e. a polyphonic signal, additional problems will appear and the estimated trajectory will not always be accurate. The reason for this failure is that YIN does not perform multipitch estimation, so the algorithm always assumes that only one source is present in every frame. Multipitch estimators will be introduced in the following section.

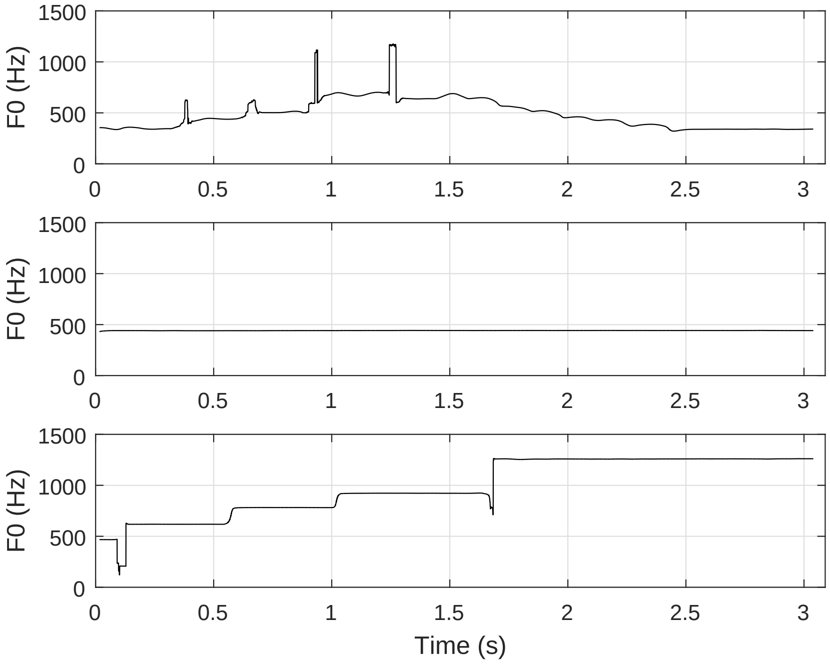

Fig. 1. Estimated melody trajectories for different audio signals, using YIN Algorithm [2]. Singing voice (Upper Chart), Modern viola playing the note A4 (Middle Chart), Clarinet playing the notes A♯4, D♯5, G5, A♯5, and D♯6 (Lower Chart).

3 Fundamental Frequency Estimation for Polyphonic Signals

Multiple fundamental frequency estimation algorithms assume that the input signal $x(t)$ is a mixture of two or more harmonic sources $\tilde{x}(t)$ plus a residual non-harmonic signal $z(t)$. Hence, the model imposed on the input signal can be expressed by the following equation, after [14].

$$x(t) = \tilde{x}(t) + z(t) \approx \sum_{m=1}^{M} \sum_{h=1}^{H_m} A_{m,h} \cos(h\omega_m t + \phi_{m,h}) + z(t) \tag{1}$$

Where H_m is the total number of harmonics to be considered, M is the number of sources and $A_{m,h}$ is the amplitude of the sinusoidal component associated with the *h-th* harmonic of the *m-th* source. In this way, the multipitch estimation problem consists of estimating the number of sources present and their fundamental frequencies [14]. The way in which the multipitch estimator handles overlapping harmonics, transients and reverberation, significantly improves or degrades the accuracy of the results.

3.1 Brief Review of Multipitch Estimation Algorithms

Many multipitch estimation algorithms are used as preprocessing stages for audio source separation. Some of those most commonly used are discussed below.

In 2003, Klapuri presented an iterative method for multiple fundamental frequency estimation based on bandpass filtering and spectral cancellation [7]. The pitch associated with the most prominent sound was estimated first and then its harmonic structure was subtracted from the original mixture, via a spectral smoothness principle. The cycle was repeated using the residual spectrum in order to extract a second pitch and continued until the estimated number of sources in the mixture was reached.

In 2008, a refined approach by Klapuri was presented [8], based on the human auditory system. The model upon which the method was designed can be described as a filter bank that decomposes the original signal into a fixed number of subbands. To achieve such a decomposition, a Gammatone filter was used. The obtained results were reported to be satisfactory when compared to other competing algorithms. The proposed system inspired parallel approaches in source separation techniques, such as [12], where multipitch estimation was used to guide a set of spectral filters to extract harmonic sources from single-channel recordings.

Other methods have concentrated on using probabilistic approaches as a way to select the fundamental frequencies that better explain a given distribution of partials. The method proposed in [4] models the original spectrum as the combination of two regions: spectral peaks and non-peaks. The signal is broken into frames and the maximum-likelihood approach is used to estimate the pitches of the detected harmonic sources. A polyphony estimation method and additional refinement stages were also presented. This particular algorithm will be explored further in Section 3.2.

Bayesian harmonic models were also used in [13] for separating harmonic sources in single-channel recordings. Multipitch estimation was used as a preprocessing stage, and it was carried out in two stages. During the first stage, all fundamental frequencies without octave relations were estimated. Then, the undetected pitches were resolved within the second stage, based on continuity of the magnitudes of harmonic partials.

A different approach was presented in [10] where the Continuous Complex Wavelet Transform (CCWT) was used as a time-frequency representation of the mixture. A Morlet wavelet was used as a filter bank to decompose the original signal into several sets of wavelet coefficients, and a novel type of scalogram was built from the data. According to a set of rules, the most relevant peaks were chosen and used to estimate the fundamental frequency candidates.

The regularized least-squares was used in [6] as a solution for simultaneous sparse source selection and parameter estimation. By exploring the block sparsity, the algorithm allows the estimation of fundamental frequencies to track a set of identified sources, without *a priori* assumptions of the number of harmonics for each source. The addition of a Bayesian prior probability distribution and regularization coefficients was considered, in order to efficiently incorporate both earlier and future blocks in the tracking of frequency estimates.

3.2 Exploring a Multipitch Estimator: A Case Study

The multipitch estimator proposed in [4] has been tested using two types of audio mixtures. First, a combination of two sounds produced by harmonic instruments, and then second, some background voices and drums were also incorporated into the mixture. Here, results are presented and discussed. Moreover, they are the basis for future possible improvements that are proposed in the following section.

The algorithm by Duan et. al. [4] was previously introduced as a probabilistic approach for fundamental frequency estimation. An implementation of this algorithm in Matlab is available for research from the author's webpage [3]. The supplied code was used during the tests in order to evaluate its accuracy and reliability as a melody extraction system.

The first test applied Duan's multipitch estimator to a mixture of two harmonic sounds. Melody trajectories for these two sounds were individually obtained in previous sections, using the YIN algorithm [2], and presented in Figure 1 (the last two plots). To compare the results of Duan's multipitch estimator, the YIN algorithm [2] was also applied to the audio mixture. Final results for the first test are presented in Figure 2.

Observing the output trajectory generated by YIN, it can be inferred that the method was unable to completely estimate the pitch associated with the modern viola. Also, YIN did not detect the pitch trajectory of the clarinet. The Duan multipitch estimator [3][4], on the other hand, correctly identified the two sources and their pitch trajectories.

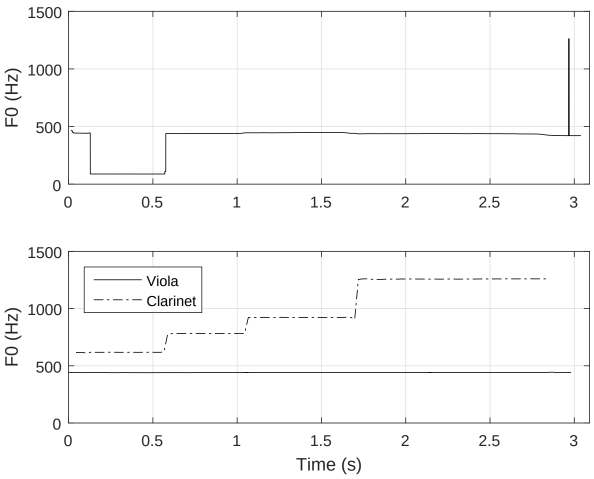

Fig. 2. Pitch trajectories extracted from a polyphonic audio signal comprising two musical instruments: Modern Viola and Clarinet. Using YIN Algorithm [2] (Upper Chart). Using Duan's Multipitch Estimator [3][4] (Lower Chart).

At this point, the good performance of the multipitch estimator is not really surprising, because the mixture was a simple mix of two individual harmonic sounds, and their pitches were selected so they were not in an octave relation. Furthermore, considering that both instruments were recorded under anechoic conditions, the levels of reverberation and noise are negligible. Also, the non-existence of additional background instrumentation reduces significantly the risk of delivering misleading results.

In order to test the algorithm under more realistic conditions, a second test was conducted. The same mixture of two harmonic sounds was combined with background instrumentation consisting of voices and drums. These additional sounds were taken from the Sixth Community-Based Signal Separation Evaluation Campaign 2015 database [1]. Multipitch estimation, using Duan's algorithm [3][4], was applied to the resulting mixture and the extracted pitch trajectories are shown in Figure 3.

The new estimated melody trajectories indicate that the multipitch estimator produced misleading estimates for some specific time-domain frames. There are sections where the algorithm swapped the correct frequency estimates between the two sources. Further, in other frames, the estimator was unable to detect correctly the fundamental frequencies of the second instrument.

Unfortunately, most commercial recordings also have significant levels of background instrumentation and noise, so that, overcoming such issues is an essential step in enlarging the range of audio recordings that can be processed, and improving the quality of the estimated sources.

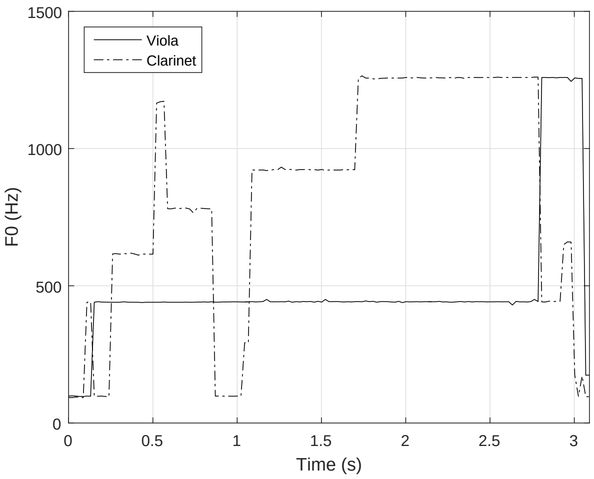

Fig. 3. Pitch trajectories extracted from a polyphonic audio signal comprising two musical instruments plus background voices and drums, using Duan's algorithm [3][4].

4 Challenges and Future Improvements

A possible way to overcome some of the limitations described in the previous section is by introducing additional stages into the system to assist the multipitch estimation algorithm in observing the input signal in an effective way. The rationale for this is that, *a priori*, there is no knowledge regarding the signal content, so that the existing algorithms can be misled by volume differences between different harmonic and non-harmonic sources. Hence, in theory, it is necessary to have more than one observation of the same input mixture.

Since multiple observations are not available the only way to provide additional input is to produce modified versions of the original signal. This should help in selecting the most reliable frequency estimates for every pitch trajectory. The stages of the proposed strategy are explained below.

- The original mixed signal is passed through a set of digital high-pass filters, each one having a different cutoff frequency. As a result, several different versions of the original audio mixture are generated.
- Multipitch estimation is applied to every filtered version of the input signal in order to estimate a set of pitch trajectory candidates.
- These trajectories are arranged in a data structure that has the form of a matrix in which the columns follow the frame number and the rows correspond to the filter used to generate that particular pitch trajectory.
- To evaluate the fundamental frequency estimates in every frame, a measurement of salience is proposed. The process assumes that those frequencies associated with a real pitch trajectory are supposed to exhibit a clear and

structured harmonic pattern, while spurious non-harmonic frequency estimates are not suppose to have any structured distribution in the magnitude spectrum. Therefore, the salience can be measure by calculating the energy of the first few partials associated with every frequency estimate in the frame. These result can be used to organize the fundamental frequency candidates, and those having the highest energy content are considered as the most reliable.

- A continuity-based approach can be also considered for error correction during melody trajectory estimation. If the multipitch estimator is using a short frame, it can be assumed that a normal note has to be present across several adjacent frames. If some unusually rapid change in pitch occurs and it is shorter than the minimum expected note duration, the related frames can be labelled as misleading and replaced with values from the data structure, that preserve the continuity of that particular melody trajectory.

This iterative structure, in which multipitch estimation is applied to different versions of the input audio mixture, is a promising alternative to extract pitch trajectories for harmonic sources in those cases where significant levels of background instrumentation or noise are present.

The proposed strategy requires control parameters to be defined, for example, the number of filters that will be used and their corresponding cutoff frequencies. Also important are the number of partials that will be considered for the salience measurement and the minimum accepted duration for musical notes. Assigning adequate values for these parameters will require further research and tests. Hence, establishing the impact of these parameters on the overall performance of the system is the essential next step in the work.

5 Conclusions

Multipitch estimation in single-channel recordings has been addressed and some relevant algorithms were described and tested using different types of audio inputs. The YIN algorithm and one multipitch estimator by Duan were evaluated for robustness in principal melody extraction tasks, using recordings of harmonic instruments, background voices and drums.

The results obtained confirmed the stability of the YIN algorithm as a melody extraction system for individual harmonic sounds, while Duan's multipitch estimator outperformed the YIN algorithm and delivered positive results for mixtures of two different harmonic sounds with unrelated fundamental frequencies. When background voices and drums were added to the mixture, the estimated melody trajectories showed swapping errors in some frames, while other sections were misleadingly estimated.

A strategy was proposed as a possible way to overcome these limitations and improve melody extraction systems for polyphonic signals. The incorporation of a set of high-pass digital filters, energy-based evaluation of pitch candidates, and continuity, are aspects that will be investigated as possible ways of producing more reliable multipitch estimation for audio source separation systems and many other applications.

Acknowledgements

The authors would like to thank the University of Costa Rica (UCR), and the Costa Rican Ministry of Science, Technology and Telecommunications (MICITT) for their support in funding this research.

References

1. Sixth Community-Based Signal Separation Evaluation Campaign, https://sisec.inria.fr/sisec-2015/2015-underdetermined-speech-and-music-mixtures/
2. De Cheveigné, A., Kawahara, H.: YIN: A Fundamental Frequency Estimator for Speech and Music. The Journal of the Acoustical Society of America 111(4), 1917–1930 (2002)
3. Duan, Z.: Multi-Pitch Analysis, http://www.ece.rochester.edu/~zduan/multipitch/multipitch.html
4. Duan, Z., Pardo, B., Zhang, C.: 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)
5. Gomez, E., Klapuri, A., Meudic, B.: Melody Description and Extraction in the Context of Music Content Processing. Journal of New Music Research (2003)
6. Karimian Azari, S., Jakobsson, A., Jensen, J.R., Christensen, M.G.: Multi-Pitch Estimation and Tracking Using Bayesian inference in Block Sparsity. In: 23rd European Signal Processing Conference (EUSIPCO). pp. 16–20. No. 2, IEEE (2015)
7. Klapuri, A.: Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness. IEEE Transactions on Speech and Audio Processing 11(6), 804–816 (2003)
8. Klapuri, A.: Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model. IEEE Transactions on Audio, Speech, and Language Processing 16(2), 255–266 (2008)
9. Murphy, D., Shelley, S.: Open Air Anechoic Audio Database, http://www.openairlib.net/
10. Ponce de Leon Vazquez, J., Beltrán, F., Beltrán, J.R.: A Complex Wavelet Based Fundamental Frequency Estimator in Single-Channel Polyphonic Signals. Proc. Digital Audio Effects (3), 1–8 (2013)
11. Salamon, J., Gomez, E., Ellis, D.P.W., Richard, G.: Melody Extraction from Polyphonic Music Signals: Approaches, Applications, and Challenges. IEEE Signal Processing Magazine 31(February), 118–134 (2014)
12. Siamantas, G.: An Iterative, Residual-Based Approach to Unsupervised Musical Source Separation in Single-Channel Mixtures. Phd, University of York (2009)
13. Wang, Y., Wang, H., Zhu, B., Wang, X.: Single-Channel Polyphonic Signal Separation Based on a Novel Multi-F0 Estimation Method. In: 14th International Conference on Communication Technology. pp. 1334–1338. IEEE (2012)
14. Yeh, C.: Multiple Fundamental Frequency Estimation. Phd, University of Paris VI (2008)