Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

Multiple F0 tracking in solo recordings of monodic instruments

Chunghsin Yeh[1], Axel Röbel[1], and Xavier Rodet[1]

[1]*IRCAM Analysis Synthesis team, 1 place Igor Stravinsky , Paris, France*

Correspondence should be addressed to Chunghsin Yeh (cyeh@ircam.fr)

ABSTRACT

This article is concerned with the F0 tracking in monodic instrument solo recordings. Due to reverberation, the observed signal is rather polyphonic and single-F0 tracking techniques often give unsatisfying results. The proposed method is based on multiple-F0 estimation and makes use of the a priori knowledge that the observed spectrum is generated by a single monodic instrument. The predominant F0 is tracked first and the secondary F0 tracks are then established. The proposed method is tested on reverberant recordings and show significant improvements compared to single-F0 estimators.

1. INTRODUCTION

Many single-F0(fundamental frequency) estimators have been developed through the years. However, when it comes to analyzing solo recordings of monodic instrument, most of which are recorded in a reverberant environment, the results of most of the single-F0 estimators are not satisfying. This comes from the fact that reverberation extends the note duration and makes the observed spectrum polyphonic. However, a single-F0 estimator assumes that there is only one F0 present in the observed signal. If the algorithm does not make use of instrument models, a single-F0 estimator often tends to favor a subharmonic which explains both the current note and the reverberation of the preceding notes.

Several studies have tried to cope with the reverberation issue in monodic instrument solo recordings. In [1], instrument model priors and duration priors have been included in a Bayesian inference framework. The performance for transcribing solos is promising but requires parameter tuning on prior distributions. In [2] the authors adapt a double-F0 estimator (an extension of YIN [3]) to the task of F0 tracking for monodic instrument recordings and significant improvements for F0 estimation of reverberant sounds have been found. This encourages us to treat this problem as a multiple-F0 tracking task. Under the assumption that there is single monodic instruments playing, the observed short-time spectrum can be modeled by a predominant harmonic

source plus the reverberant parts of the preceding notes and background noise. Therefore, we propose to first decode the predominant F0 track from a set of hypothetical F0 combinations and keep non-dominant F0s for serving as continuity of predominant F0 tracks.

This paper is organized as follows. First, an overview of the proposed method is introduced. In section **2**, a frame-based multiple-F0 estimation is presented. For each analysis frame, multiple-F0 estimation provides a list of hypothetical F0 combinations for the later tracking stage which is explained in section **3**. Lastly, testing examples are shown and conclusions are drawn.

2. SYSTEM OVERVIEW

The proposed F0 tracking system is mainly composed of three stages (Fig. 1). For each analysis frame, multiple-F0 estimation provides the list of the best-ranked hypothetical F0 combinations. F0 tracking can thus be considered as decoding the optimal path through the trellis structure form by the hypothetical F0 combinations across the frames. As the example shown in Fig. 2, multiple-F0 estimation proposes at each frame a pre-fixed number of candidate combinations for each hypothetical number of F0s (denoted as M). Each hypothetical combination is denoted as $\{F0^i_{M,c}\}$ (where M ranges from 1 to the estimated number of F0s, c ranges from 1 to the pre-defined number of candidate combinations to be considered) for the cth top-ranked candidate combination at time i. We propose to decode first the predominant F0 track based on individual F0 probability which is inferred from the multiple-F0 combinatorial properties. Then, the secondary F0s can be tracked by extending the predominant F0 tracks. In this article, we consider the secondary F0s to be the results of the reverberation only.

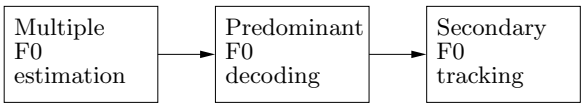

Fig. 1: Overview of the F0 tracking system

3. MULTIPLE F0 ESTIMATION

In [4], we have proposed a frame-based multiple-F0 estimation algorithm based on a generative polyphonic signal model. The inference procedure is similar to the Bayesian model proposed in [5]. However, to prevent the huge computational requirements of numerical likelihood maximization, a more pragmatic approach is proposed to construct and evaluate hypothetical sources, which is guided by three physical principles for nearly-harmonic sounds:

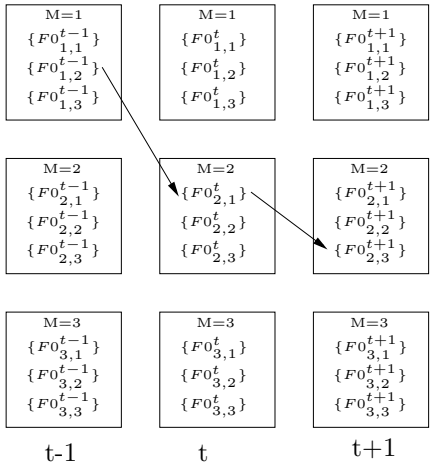

Fig. 2: Decoding the optimal multiple-F0 path

1. Spectral match with low inharmonicity
2. Spectral smoothness
3. Synchronous amplitude evolution within a single source

These principles are formulated as four criteria: harmonicity *HAR*, mean bandwidth *MBW* and centroid *SPC* of Hypothetical Partial Sequences (*HPS*), and the standard deviation of mean time of hypothetical partials *SYNC*. The four criteria together evaluate the plausibility of each F0 combination, which is proportional to the likelihood $p(\mathcal{O}^i|\{F0^i_{M,c}\})$ where $\mathcal{O}^i$ denotes the observed spectrum at instant i. An overview of the proposed multiple-F0 estimation is shown in Fig. 3. The process is listed step by step in the following.

i. Hidden partial extraction:
Extracting hidden partials is essential to increase the accuracy of polyphonic signal analysis since the resolution is necessarily limited. To

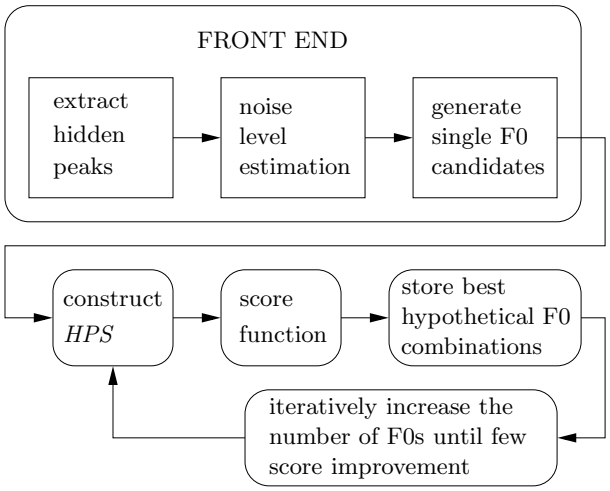

Fig. 3: Multiple-F0 estimation

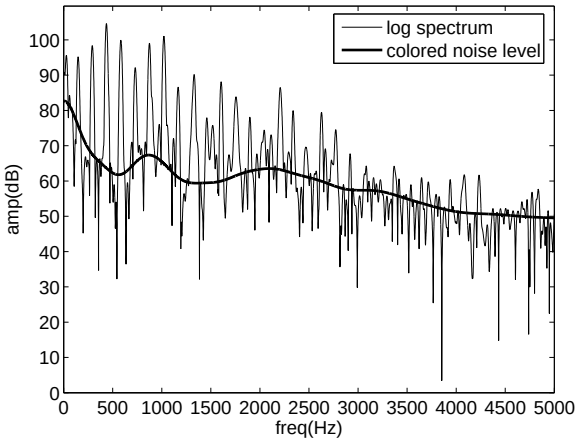

Fig. 4: Noise level estimation

search for the potential spectral collision possibly containing hidden partials, we evaluate the shapes of the observed peaks and their spectral properties using four descriptors [6]. This allows selecting the possibly overlapped partials which are then processed to extract hidden peaks [7].

ii. Noise component estimation:
It is important to identify target components to be explained by the generative nearly-harmonic model and disregard the unwanted components. In [8], we have developed an iterative algorithm to estimate the noise level adapted to the observed spectrum (see the example shown in Fig. 3), by which the noise peaks are classified. During the harmonic matching process in the later stage, matches to noisy peaks are disregarded.

iii. Single-F0 candidate selection:
A harmonic matching technique is used to provide the single-F0 candidate list.

iv. Hypothetical Partial Sequence construction:
Constructing *HPS*s utilizes **Principle 2** and the knowledge of spectral locations where partial overlaps may occur according to the multiple-F0 combination under investigation. We have developed a method for reassigning overlapped partials [4], by which partials in a *HPS* are classified as "effective" and "non-effective". The non-effective partials don't have specified amplitudes and are disregarded.

v. Scoring multiple-F0 combinations:
At this stage, a score function [7] is used to iteratively evaluate the plausibility of the number of F0s starting from one. This iterative search is stopped once the score improvement falls below a threshold. The idea is simple: when one source more than the true source number is added in the model to explain the observed spectrum, the score improvement should be limited.

The score function is defined as the linear combination of four score criteria:

$$D = p_1 \cdot HAR + p_2 \cdot MBW + p_3 \cdot SPC + p_4 \cdot SYNC \quad (1)$$

where $\{p_j\}_{j=1}^{4}$ are the weighting parameters for the four criteria. Here we briefly summarize the score criteria below:

HAR is an indication of harmonicity and totally explained energy. To evaluate the smoothness of the spectral envelope of a hypothetical source, we use the mean bandwidth *MBW* and the centroid *SPC* of a *HPS*. Due to partial overlapping, the "non-effective" partials don't provide specific spectral amplitudes. To evaluate *MBW*, we remove the "non-effective" par-

tials. To evaluate *SPC*, we reconstruct the "non-effective" partial amplitudes by interpolation. To evaluate the synchronicity of the temporal evolution of the hypothetical partials, we rely solely on the "effective" partials.

For each *F0* hypothesis we define *effective weighting* as the sum of linear amplitudes of "effective" partials. Then the individual properties of the last three criteria are weighted by the effective weighting and then summed to define the combinatorial properties.

Notice that *HAR* favors energy explanation of the observed spectrum, while *MBW*, *SPC* and *SYNC* work together as constraints to the hypothetical spectral models. Therefore, the criteria perform in a complementary way and the weighting parameters have been optimized by an evolutionary algorithm to balance the relative contribution of each criterion. To refine precise F0 values, we apply a linear regression of effective partial frequencies. An experimental setup similar to [9] has been carried out, which shows competitive performance [4].

vi. Iterative increase the number of F0s:
We propose an iterative search to infer the plausible hypothetical number of F0s. The true number of F0s is denoted as N, while the inferred hypothesis is denoted as S_M. Starting with S_1, the system iteratively evaluates the score improvements of all possible hypotheses $\{S_1, \ldots, S_M, S_{M+1}\}$, where S_{M+1} is the last hypothesis. S_{M+1} provides a score improvement (w.r.t. the score of S_M) under a threshold δ, which leads to the termination of the iterative evaluation. Then, the hypothesis S_M is considered as the most plausible number of F0s in the current frame. Therefore, the number of F0 is inferred if $S_M = S_N$.

In order to obtain δ, we investigate the score improvement of the correct estimates evaluated on our artificially mixed polyphonic database [4]. The score improvements of iterative F0 search are shown in Fig. 5 for two-note, three-note and four-note mixtures. While $S_M = N+1$, we observe that the score improvements are close to zero. This means that an additional harmonic source does not significantly improve the likelihood of the underlying model. Based on the observed score improvements, we model the improvements of scores from S_M to S_{M+1} by means of Gaussian distributions. This serves as a mechnism to stop the iterative search and meanwhile defines the probability of the most probable state. In the current implementation, the threshold is set to include 85% of the correct estimates from S_4 to S_5 in the four-note mixtures. This might result in some spurious F0s when $N < 4$ but guarantee the inclusion of the correct F0 combinations. The top-five ranked hypothetical F0 combinations from $M = 1$ to $M = 4$ are kept for the later tracking stage. We are currently relating δ to the estimated noise level.

4. TRACKING MULTIPLE-F0 TRAJECTORIES

After evaluating the plausibility of the most probable F0 combinations $\{F0^i_{M,c}\}$s, we start decoding the optimal path for the trellis structure, guided by two principles: local likelihood and temporal continuity. However, it is difficult to define the transition probability between two hypothetical F0 combinations with different Ms. Therefore, we propose to decode the predominant F0 tracks first and the secondary ones, which are assumed to be mostly reverberant parts, can thus be tracked by evaluating their combinatorial probability with the predominant F0s.

4.1. Predominant F0 tracking

For solo recordings of monodic instruments, the predominant F0s clearly relate to the monophonic melody line being played. As long as the reverberation of preceding notes is less dominant than the notes being played, taking the most significant F0 as the predominant F0 is generally accepted.

To track predominant F0s, we rely on the individual scores of F0 candidates. The individual score is defined similarly to eq.(1) with the combinatorial criteria replaced by the individual criteria, that is, for each single-F0 candidate in one combination, the missing information of "non-effective" partials is disregarded. For each hypothetical number of F0s, the individual score of a single F0 candidate is weighted by the combinatorial probability (derived from the relative score in the top-five ranked combinations) to define the average individual probability. Therefore, an F0 candidate appearing in the combinations

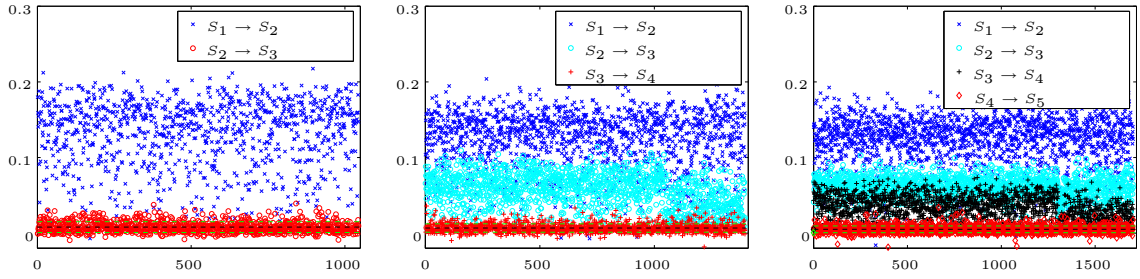

Fig. 5: Score improvement observations for different number of F0s. x-axis represents the wave file number and y-axis represents the score improvement.

with higher score is considered more important. The individual probability is further averaged over different hypothetical number of F0s. In such a way, the plausibility of each F0 candidate among the most probable combinations can be derived.

Given the individual probability as observations, the best state sequence of predominant F0s is going to be inferred. We propose a two-stage tracking method.

I. *Forward connection between frames*
The first stage makes the connection among the F0 candidates between consecutive frames. For each F0 candidate, the connection is allowed for a frequency range of one half tone. For every "pair" of frames, the connection that gives the highest product of individual probability is kept for the next stage.

II. *Track construction*
From the connected single F0s, a track can be defined. However, there are often several "holes" in-between tracks to be taken care of. These holes might be the result of note onsets, where the observed spectrum is disturbed. This is done by linear prediction similar to [10]. To reconstruct the "holes" in-between tracks, a backward/forward linear prediction tracking is applied on the neighboring two tracks. We start by backward linear prediction to find F0 candidates until no match is found. Then forward linear prediction is performed to reconstruct the rest of the missing predominant F0s.

4.2. Secondary F0 tracking

Once the predominant F0 track is decoded, the secondary F0s can be tracked by prolonging the predominant F0s. To track the reverberant parts of the predominant F0 tracks, we search the combination containing the current predominant F0 and the previous predominant F0s. As long as the "effective weighting" of a secondary F0 is larger than 0.01, the reverberant tracks are considered as effective.

5. EXPERIMENTAL RESULTS

To demonstrate the proposed method, we have tested two solo recordings: bassoon and violin. For the bassoon solos, we compare our method with the state-of-art single-F0 estimator "YIN". The F0 search range is set from 50Hz to 2000Hz. As shown in Fig. 6, "YIN" produces subharmonic errors while the reverberant parts of the preceding notes have competitive significance. This shows the complexity of F0 tracking for monodic solo recordings, which can be barely handled by a single-F0 estimator.

In the second example, a violin solo, our proposed method gives promising results for the fast arpeggios of which the reverberant parts are well tracked, too.

6. CONCLUSIONS

We have presented a method using multiple-F0 tracking algorithm for solo recordings of monodic instruments. We propose to decode the predominant F0 track and then the secondary F0s, based on the combinatorial properties of hypothetical F0 combination. Testing examples have shown that a multiple F0 estimation is necessary for automatic transcription of solo recordings. There are several issues to

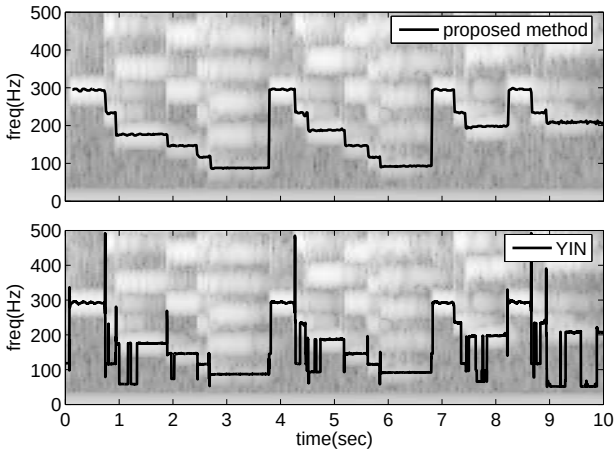

Fig. 6: Comparison of predominant F0 estimation using one Mozart's bassoon solo

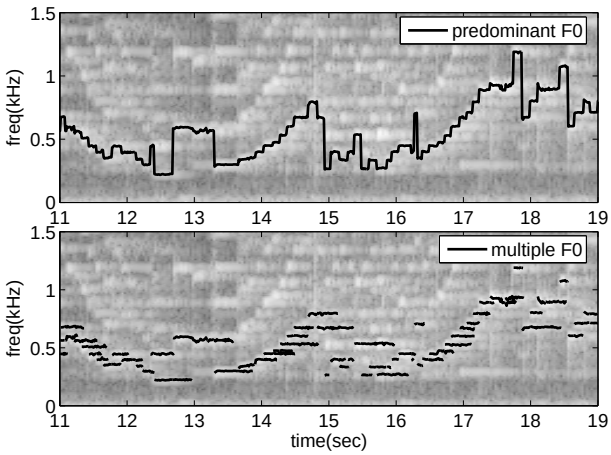

Fig. 7: Multiple F0 tracking tested on one Bach's violin solo

be addressed. If the reverberant parts of the preceding notes are stronger than the following notes (for example, a strongly bowed note followed by left hand pizzicati), our multiple-F0 tracker might favor the reverberation that is more dominant in energy.

7. REFERENCES

[1] E. Vincent, *Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux*, PhD thesis, Université Paris VI, 2004.

[2] A. Baskind and A. de Cheveigné, "Pitch-Tracking of Reverberant Sounds, Application to Spatial Description of Sound Scenes," in *AES 24th International Conference*, Banff Centre, Canada, 2003.

[3] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, pp. 1917–1930, 2002.

[4] C. Yeh, A. Röbel, and X. Rodet, "Multiple fundamental frequency estimation of polyphonic music signals," in *Proc. IEEE, ICASSP'05*, Philadelphia, 2005.

[5] M. Davy and S. Godsill, "Bayesian Harmonic Models for Musical Signal Analysis," in *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, Valencia, Spain, 2003.

[6] A. Röbel and M. Zivanovic, "Signal decomposition by means of classification of spectral peaks," in *Proc. of the International Computer Music Conference (ICMC'04)*, Miami, Florida, 2004.

[7] C. Yeh and A. Röbel, "A new score function for joint evaluation of multiple F0 hypotheses," in *Proc. of the 7th Int. Conf. on Digital Audio Effects (DAFx'04)*, Naples, 2004.

[8] C. Yeh and A. Röbel, "Adaptive noise level estimation," in *Workshop on Computer Music and Audio Technology(WOCMAT'06)*, Taipei, 2006.

[9] Anssi Klapuri, *Signal processing methods for the automatic transcription of music, Ph.D dissertation*, Tampere University of Technology, 2004.

[10] M. Lagrange, S. Marchand, and J.-B. Rault, "Using Linear Prediction to Enhance the Tracking of Partials," in *Proc. IEEE, ICASSP 04*, Montreal, 2004.