HIGH TIME-RESOLUTION ESTIMATION OF MULTIPLE FUNDAMENTAL FREQUENCIES

Jayme Garcia Arnal Barbedo[1,2], Amauri Lopes[2], and Patrick J. Wolfe[1]

[1] School of Engineering and Applied Sciences, Harvard University
33 Oxford Street, Cambridge, MA 02138-2901, USA
{jbarbedo, patrick}@seas.harvard.edu

[2] School of Electrical and Computer Engineering, State University of Campinas
Cidade Universitária Zeferino Vaz, C.P. 6101, CEP: 13083-970, Campinas, SP, Brazil
{jgab, amauri}@decom.fee.unicamp.br

ABSTRACT

This paper presents a high time-resolution strategy to estimate multiple fundamental frequencies in musical signals. The signal is first divided into overlapping blocks, and a high-resolution estimate made of the short-term spectrum. The resulting spectrum is modified such that only the most relevant spectral components are considered, and an iterative algorithm based on earlier work by Klapuri is used to identify candidate fundamental frequencies. Finally, a context-based rule is used to improve the accuracy of fundamental frequency estimates. The performance of this technique is investigated under both noiseless and noisy conditions, and its accuracy is examined in cases where the polyphony is known and unknown a priori.

1 INTRODUCTION

The estimation of fundamental frequencies (F0) of mixtures of several sound sources is a problem whose solution can benefit several areas of digital audio processing, including automatic music transcription, music information retrieval, and sound source separation, among others.

Early work relating to this problem aimed to solve the problem of transcribing polyphonic music [1, 2]; however, such methods worked well only under very restrictive constraints. A new phase in multiple-F0 estimation started with the work of Meddis and Hewitt [3], which has provided the foundation for most approaches used in more recent methods. Cheveigné explored the model proposed in [3] to develop an iterative procedure in which the sound component corresponding to a particular estimated F0 is removed, and a new round of F0 estimation then proceeds using the residual [4]. Tolonen and Karjalainen simplified the approach of [3] to create a strategy reported to be both accurate and computationally efficient [5], and statistical inference was used by Davy and Godsill to estimate multiple fundamental frequencies [6]. Klapuri proposed a method based in the harmonicity and spectral smoothness of the signals [7], as well as a more recent perceptually motivated strategy that uses an iterative estimation-cancellation approach [8].

Like other approaches, the method proposed here begins with the division of the musical signal into overlapping short-time frames. The high-resolution spectral estimation proposed in [9] is then applied to each frame in turn, in order to allow a finer analysis of the spectral structure of the signal. The resulting spectrum is modified in such a way only the most relevant frequency components are considered. Additionally, the remaining components are quantized into only two levels, making the data more homogeneous. The modified spectrum is then analyzed using an iterative algorithm based on the procedure presented in [8], but which also introduces some further processing intended to refine the selection of the correct F0. If the polyphony is known a priori, the iterations stop as soon as the number of sound sources has been reached. If the polyphony is unknown, the iterations are interrupted if one out a set of rules is fulfilled. After the fundamental frequencies have been determined for all frames, a further procedure is applied to improve the estimates. All frames contained in a segment between two events [1] are analyzed, and the estimates are all made the same according to majority rules. In the cases where the polyphony is known a priori, this last procedure only changes the values of F0, but in cases where the polyphony is unknown at the outset, there can be changes in the estimated number of sound sources. Those context-corrected F0 estimates comprise the output of the algorithm.

The remainder of this paper describes this algorithm and its application in more detail, and is organized as follows. Section 2 presents a description of the algorithm. The analysis of the results is presented in Section 3, and Section 4 presents the conclusions and final remarks.

[1] An event, in the context of this work, is any change in the number of sound sources and/or fundamental frequencies present in the signal.

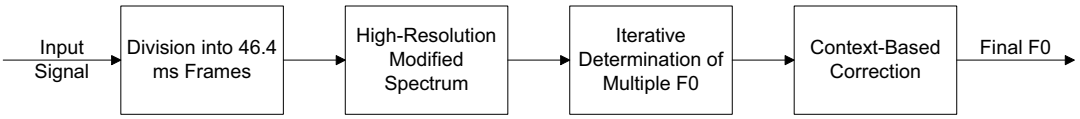

Figure 1. Block diagram of the estimation strategy

2 ALGORITHM DESCRIPTION

Figure 1 shows the basic structure of the method. As can be seen, there are four basic procedures, which will be described in the following.

2.1 Division into Frames

The first step of the algorithm is the division of the signal into frames. Assuming a sample rate $f_s = 44100$ kHz, the length of the frames is taken to be 2048 samples, corresponding to 46.4 ms. This value was chosen in order to provide a good temporal resolution to the analysis—thereby minimizing, as much as possible, the estimation problems that can arise when a new event occurs in the middle of a frame. (This also provides better estimates when new events occur in short time intervals.) In spectral analysis, better time resolution comes at the expense of worse frequency resolution. To minimize this problem, a high-resolution estimate of the spectrum was employed [9], as described in Section 2.2 below.

2.2 Computing High-Resolution Modified Spectrum

The tight compromise between time and frequency resolutions in spectral analysis has motivated the use of an algorithm that makes a high-resolution estimate of the spectrum [9]. The Matlab code and more details can be found in [10]. More specifically, a spectrum with 44100 points (i.e., 1 Hz resolution) is estimated based on a frame length of 2048 samples. The resultant estimates are reasonably accurate, and were observed to provide much better results than using the regular Discrete Fourier Transform (DFT).

Besides providing a high-resolution spectral estimate, the method of [9] also reduces greatly the amplitude of sidelobes that appear in the computation of regular DFT. This is achieved by extrapolating the data, and properly exploring the autocorrelation function of the extended data. (The mathematical details of the procedure can be found in [9].) The only major shortcoming of this strategy is the computational burden, which is much higher than that required by the Fast Fourier Transform (FFT) algorithm.

The high-resolution spectrum is then modified to consider only relevant components. First, all spectral components that do not represent a local magnitude peak are set to zero. The component with greatest magnitude is identified and taken as a reference value X_m. Then, the following assignment is applied for $k \in \{1, 2, \ldots, \lfloor f_s/2 \rfloor\}$:

$$\begin{cases} S_m[k] = 0 & X[k] < 0.01X_m \\ S_m[k] = 100 & 0.01X_m \leq X[k] < 0.1X_m \\ S_m[k] = 1000 & X[k] \geq 0.1X_m \end{cases},$$

where X represents the magnitude spectrum, k is the frequency index, and X_m is the value of the component of largest magnitude.

The resulting modified spectrum S_m is then used to determine the fundamental frequencies present, according to the procedures described in Section 2.3 below.

2.3 Iterative Determination of Multiple F0

Part of the iterative procedure to determine the fundamental frequencies was based on part of the strategy presented in [8]. This procedure is performed according to the following steps:

1. A residual spectrum S_r is initialized and set equal to the modified spectrum S_m.

2. A simplified version of the fundamental period estimation presented in [8] is applied, given by

$$\lambda(\tau) = \left(\frac{f_s}{\sqrt{\tau}}\right) \sum_{j=1}^{\tau/2} \max_{k \in \boldsymbol{k}_{j,\tau}} \left(S_r(k)\right), \qquad (1)$$

where f_s is the sampling frequency, τ is the candidate fundamental period, and $\boldsymbol{k}_{j,\tau}$ is a set that defines a range of frequency bins in the neighborhood of the harmonic partials of the k'th F0 candidate, given by

$$\boldsymbol{k}_{j,\tau} = \left[k_{j,\tau}^{\min} \quad \cdots \quad k' \quad \cdots \quad k_{j,\tau}^{\max}\right],$$

where

$$k_{j,\tau}^{\min} = \lfloor jK/(\tau+1) \rfloor + 1$$
$$k_{j,\tau}^{\max} = \max\left(\lfloor jK/(\tau-1) \rfloor + 1, k_{j,\tau}^{\min}\right)$$

and $K = f_s = 44100$ is the total number of spectral bins of the modified spectrum. As can be seen, there are some differences between the procedure adopted here and that described in [8]. Particularly, the weighting factor that simulates the bandwidth of the auditory filters was not used here, nor the balancing operation over λ. This is because such operations were not seen to improve the results of the method proposed here. Additionally, $\sqrt{\tau}$ is used in (1), instead of τ itself; this decision was taken as a result of optimization tests performed with the method over large databases, as described in Section 3 below.

3. The candidate fundamental frequency is given by

$$f_c = f_s / \max(\lambda).$$

4. In the next step, the partials corresponding to f_c are removed from the residual spectrum S_r according to

$$S_r(P_m) = 0,$$

where for $m \in \{1, 2, \ldots, \lfloor \frac{f_s}{2\tilde{f}_c} \rfloor\}$, P_m is defined as

$$P_m = \underset{n \in \mathbb{N} \cap [\alpha_1, \alpha_2]}{\operatorname{argmax}} S_r(n);$$
$$\alpha_1 = \lfloor (0.975 - a) f_c \rfloor, \quad \alpha_2 = \lfloor (0.975 + a) f_c \rfloor,$$
$$a = \max\Big((\tilde{f}_c - 1), 0\Big) \cdot 10^{(\tilde{f}_c - 25)/5},$$

with $\tilde{f}_c$ the candidate F0 in kHz. As can be seen, the interval around each partial grows for large frequencies in order to account for deviations caused by inharmonicity.

5. If the polyphony is known a priori, Steps 1–4 are repeated until the number of estimated fundamentals coincides with the number of concurrent sounds. Otherwise, some stopping criteria must be applied. In particular, if at iteration i the stopping criteria

$$\max\left(S_r^i\right) \leq 100 \text{ and } \sum_k S_r^i(k) < 200$$

are met, then the candidate fundamental $f_c(i)$ will be accepted and the algorithm will proceed to Step 6 below. Alternatively, define at iteration i the criteria

$$f_c(i) \geq 500 \text{ and } \sum_k S_r^{i-1}(k) - \sum_k S_r^i(k) < 200$$
$$f_c(i) < 500 \text{ and } \sum_k S_r^{i-1}(k) - \sum_k S_r^i(k) < 300,$$

where S_r^{i-1} and S_r^{i-1} are, respectively, the residual spectra before and after removing the partials of the current F0 estimate, and $S_r^0 = S_m$. If either of these criteria are met, then the candidate $f_c(i)$ will be rejected and the algorithm will proceed to Step 6 below. Such rules interrupt the iterations if there are too few significant spectral components remaining, or if the current fundamental frequency was estimated using too few spectral components. The rules are tighter for low F0, because more significant spectral components are expected to be present in such situations.

6. Finally, the first estimated F0 is checked to verify that it is the lowest of all estimated F0. If so, the estimates resulting from Steps 1–5 are the final output of this stage of the algorithm. If not, these procedures are repeated, this time forcing the lowest detected frequency to be the first to be considered in the iterative process. This method is employed as in many cases an overtone partial of the actual F0 is taken as the estimated F0, leading to an estimation error. In that case, the correct F0 is normally also detected as a potential F0 in a subsequent iteration. Moreover, in such cases it was often observed that the correct F0 would be the lowest among all estimated F0. Forcing the lowest partial to be considered first greatly reduces the frequency of this problem.

2.4 Context-Based Correction

The last stage of the algorithm is a context-based correction of the estimated F0. The homogeneous segments between two events normally will contain more than one frame and, since it is expected that all frames of the segment present the same results, a procedure to homogenize the estimates is applied. If the polyphony N is known a priori, the set of estimated F0 for all frames in the segment will comprise the N fundamental frequencies that appear most frequently among these frames. If the polyphony is unknown, then a given F0 is included in the set of estimates of all frames in the segment only if it appears in at least 50% of the frames; otherwise it is removed. This simple procedure improves the overall algorithm performance by about 25%, as will be seen in the next section.

Poly-	Context Correction		Klapuri Method [8]	
phony	Without	With	46 ms	92 ms
1	1.5	1.1	7.2	2.1
2	5.2	3.9	12.0	7.0
3	7.5	6.0	21.3	10.2
4	13.0	9.9	26.9	12.8
5	19.8	15.3	35.5	17.1
6	29.0	24.2	42.4	21.3

Table 1. Estimation error (%) with known polyphony

3 RESULTS

The database used for testing is composed of 1200 mixtures of one to six concurrent sounds, taken from both the RWC Music Instrument Sound database [11] and from the University of Iowa Musical Instrument Samples database (http://theremin.music.uiowa.edu/MIS.html). These mixtures have lengths between 0.05 and 1 second, with an average of about 250 ms. Sounds from 30 instruments were used, and each mixture was the result of a random selection among all instruments and respective note ranges. Calibration was performed using 200 mixtures, and the remaining 1000 signals were used in the tests.

Table 1 shows the percentage of mis-estimates obtained for the method in the case where the polyphony is known, with and without the context correction. The results are compared with the 46-ms and 92-ms frame versions of the method proposed by Klapuri, implemented according to the guidelines presented in [8]. As can be seen in Table 1, the strategy performs very well for few concurrent sounds, and the performance begins to degrade rapidly when more sounds are present. This is due, firstly, to the tendency of any method to lose reliability when too many spectral components are present. Additionally, the elimination of spectral components considered irrelevant (see Section 2.2) sometimes removes partials of actual F0, leading to an error. However, the method performs better with this "clean" spectrum than without any component selection. Table 1 also reveals that the context-based correction reduces errors by approximately 25%. It is important to note that the effects of correction are more effective when the segment between events is longer; as described above, the results of Table 1 were obtained for segments with lengths between 50 ms and 1 s (250 ms on average).

Poly-	Without Correction			With Correction		
phony	Corr.	Miss	False	Corr.	Miss	False
1	98.5	1.5	3.4	98.9	1.1	1.2
2	94.1	5.9	6.6	95.6	4.4	4.2
3	91.0	9.0	8.8	92.5	7.5	6.6
4	84.7	15.3	12.3	87.1	12.9	10.3
5	77.9	22.1	18.0	80.0	20.0	15.9
6	68.1	31.9	23.4	71.8	28.2	23.0

Table 2. Estimation error (%) with unknown polyphony

Table 2 shows the performance of the technique for the case in which polyphony is unknown. Its columns show the percentage, with respect to the number of actual F0, of the correct estimates, missed F0, and incorrectly detected F0. As can be seen, the results exhibit similar behavior to that observed in Table 1, but since the polyphony is unknown, the results are (as expected) slightly worse. It also can be seen that there is a balance between the number of missed F0, and false F0. The method was calibrated in this way, but simple changes in the algorithm can change the compromise between missed and false detections.

The results shown in Tables 1 and 2 were obtained from mixtures where the levels of the sounds were the same. If the relative levels between the sounds change, the accuracy of the technique tends to be reduced. To test the influence of level in algorithm performance, all mixtures of three sounds were taken and each estimation procedure was repeated, with the level of one of the sounds being reduced at a time. Figure 2 shows the percentage of F0 mis-identifications in the level-reduced sound, as a function of the reduction factor. As can be seen, the technique is quite robust to mild variations in the level of the sounds, but it quickly begins to lose reliability if the target sound is more than 5 dB below the levels of the others.

4 CONCLUSIONS

This paper presented a new method to estimate multiple fundamental frequencies of concurrent sounds. It uses a modified spectrum as input to an iterative algorithm that estimates a candidate set of F0. A set of rules is applied to improve these estimates, and an additional context-based correction procedure provides the final F0 estimates. The method performs well in cases where the polyphony is known or unknown a priori, and is robust to mild differences in the levels of the sounds. The main shortcoming of the technique is its high computational complexity; future work will search for solutions to this problem. New procedures and statistical models to replace current rule-based heuristics will be investigated, in an attempt to improve estimation performance and quantify F0 uncertainty.

Acknowledgments: *Special thanks are extended to FAPESP and Capes for supporting this work under Grants 04/08281-0 and 03/09858-6 (FAPESP), and Grant 2234/06-8 (Capes).*

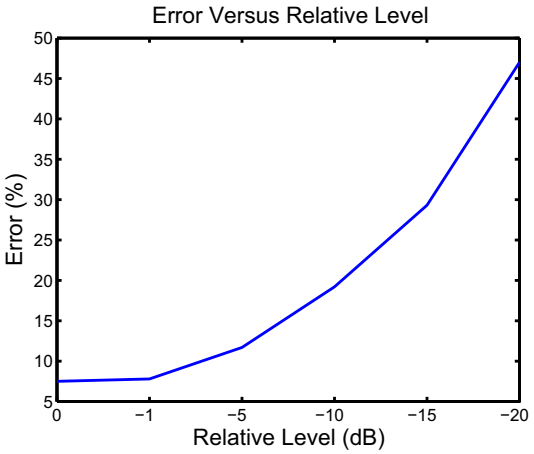

Figure 2. Estimation error (%) for varying relative level

5 REFERENCES

[1] J. A. Moorer, "On the transcription of musical sound by computer," *Comput. Mus. J.*, vol. 1, no. 4, pp. 32–38, 1977.

[2] C. Chafe and D. Jaffe, "Source separation and note identification in polyphonic music," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1986, pp. 1289–1292.

[3] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery I: Pitch identification," *J. Acoust. Soc. Am.*, vol. 89, no. 6, pp. 2866–2882, 1991.

[4] A. de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *J. Acoust. Soc. Am.*, vol. 93, no. 6, pp. 3271–3290, 1993.

[5] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 11, pp. 708–716, 2000.

[6] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of Western tonal music," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2498–2517, 2006.

[7] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–815, 2003.

[8] A. P. Klapuri, "A perceptually motivated multiple-F0 estimation method," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2005.

[9] V. Y. Liepin'sh, "An algorithm for evaluating a discrete Fourier transform for incomplete data," *Autom. Control Comput. Sci.*, vol. 30, no. 3, pp. 27–40, 1996.

[10] V. Y. Liepin'sh, "Extended DFT," Available: http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=11020objectType=File

[11] M. Goto, "Development of the RWC music database," in *Proc. 18th Int. Congr. Acoust.*, 2004, pp. 553–556.