Analysis and Improvements of the Cepstrum Method for Fundamental Frequency Estimation in Music Signals

Johannes Gauer
Institute of Communication Acoustics
Ruhr-Universität Bochum
Bochum, Germany
johannes.gauer@rub.de

Diana Kleingarn
Institute of Communication Acoustics†
Ruhr-Universität Bochum
Bochum, Germany
diana.kleingarn@rub.de

Rainer Martin
Institute of Communication Acoustics
Ruhr-Universität Bochum
Bochum, Germany
rainer.martin@rub.de

***Index Terms*—pitch, fundamental frequency, cepstrum, speech, music**

I. Abstract

Fundamental frequency estimation has many applications in speech and music processing. Among other methods, peak tracking in the cepstrum domain has been established as a robust and successful method for speech signals. Its elegance stems from the fact that a harmonic pattern in the log-spectral domain is mapped onto a single bin in the cepstral domain. However, the cepstrum method had not been thoroughly analyzed in the context of music signals. In this work we introduce a novel method for detecting and compensating octave errors to enhance the fundamental frequency search in the cepstrum and combine it with a fine search based on a least squares approximation in the time domain. The performance of the method is evaluated with monophonic music signals across a wide range of fundamental frequencies. In addition, the achievable frequency resolution and estimation error and their dependence on general signal parameters are analyzed.

II. Introduction

The fundamental frequency (F_0) is an often used quantity in speech and music processing, e.g., for the analysis and synthesis of prosody in speech or melody in music signals. Thus, a variety of F_0 estimation algorithms for speech and music signals have been developed over the years. They are based e.g. on the autocorrelation function [1]–[3], on Summation of Residual Harmonics (SRH) [4], [5], or on the harmonic model [6]–[8]. An overview of different methods can be found, e.g., in [5] and [9]. With the advancement of deep learning, also neural network-based approaches have been proposed (e.g. [10]).

Another class of algorithms employs the cepstrum [11], [12] which transforms the log-magnitude spectrum into the cepstral domain via an inverse DFT or DCT. The cepstrum method was initially developed and tested for narrowband speech signals [13]. Its appeal resides in the elegant mapping of harmonic spectral patterns onto a few cepstral bins and in the fact that relative estimation errors do not vary much on the Cent (¢) scale. Thus, by searching for maxima in the cepstral domain the fundamental frequency may be identified with moderate computational effort. It has been used for instance for vibrato analysis and synthesis in [14], however only for singing voices.

Frequently, the estimation of F_0 is organized in two consecutive steps: firstly, a relatively coarse search yields F_0 candidates which are then further refined in a local search step (e.g., [15]). This first step can be implemented by any of the well-known methods mentioned above. In the second step, either a fine-spaced grid search using the least squares method [15], an interpolation using a quadratic function, or a dynamic model using an HMM [3] or a Kalman filter [8] are used to improve the resolution and tracking capabilities.

For signals with small sampling rates and relatively low fundamental frequencies (e.g. speech signals) F_0 can be estimated with a high accuracy using the cepstral approach. It is, however, less clear how the cepstrum method would perform on music signals which requires the consideration of a much wider range of fundamental frequencies. Since the cepstrum method maps harmonic patterns onto single bins it enables a fast search across a wide range of fundamental frequencies.

The remainder of this paper is organized as follows: In Section III we present the cepstrum-based method for fundamental frequency estimation. We then investigate the resolution properties of the cepstrum method in Section IV. In Section V we describe our experimental setup to evaluate the method and summarize the results in Section VI.

III. Method

In the proposed method, we integrate a heuristic octave error compensation into the cepstrum-based F_0 estimation and combine it with a subsequent least squares estimator of a harmonic model in the time domain. Figure 1 shows a flowchart of the proposed method.

†Diana Kleingarn is now with Robotics Research Institute, Technische Universität Dortmund, Dortmund, Germany

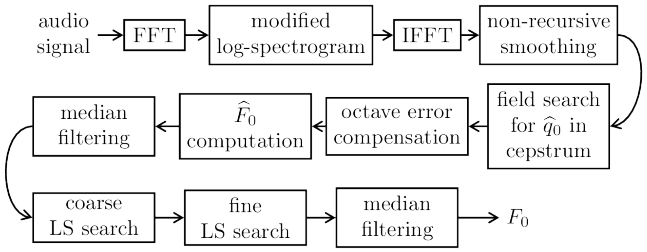

Figure 1: Flowchart of the cepstrum-based method

A. Cepstrum

The core idea of F_0 estimation in the cepstral domain [13] is to reliably find the cepstral peak $\hat{q}_0$ that represents the fundamental frequency F_0 and is located at *quefrency* (q) bin

$$q_0 = \frac{1/F_0}{1/f_s} = \frac{f_s}{F_0}. \tag{1}$$

The discrete-time input signal $x(n)$, $n \in \mathbb{N}^0$, is first divided into Λ overlapping frames $\mathbf{x}^{(\lambda)} = \{x(\lambda R + 1), ..., x(\lambda R + L)\}$ of length L with frame index $\lambda \in \{0, \dots, \Lambda - 1\}$ and frame shift $R < L$. Applying the Fourier transform $\mathbf{X}^{(\lambda)} = \mathcal{F}\{\mathbf{x}^{(\lambda)}\}$ to each frame λ yields the short-time Fourier transform (STFT). The real cepstrum is obtained by applying the inverse Fourier transform to the log-spectrogram and removing the symmetric part ($q > q_{\max} = L/2 + 1$) [11]. However, inspired by [12], we slightly deviate from the standard cepstrum definition and employ a modified version $\mathbf{S}_{\text{mod}}^{(\lambda)}$ of the log-spectrogram to obtain a meaningful and robust cepstral representation $\mathbf{C}^{(\lambda)}$:

$$\mathbf{S}_{\text{mod}}^{(\lambda)} = 20 \log_{10}(1 + |\mathbf{X}^{(\lambda)}|) \tag{2}$$

$$\mathbf{C}^{(\lambda)} = \mathcal{F}^{-1}\{\mathbf{S}_{\text{mod}}^{(\lambda)}\} = \left[C^{(\lambda)}(1), \dots, C^{(\lambda)}(q_{\max})\right] \tag{3}$$

In (2), the use of $1 + |\mathbf{X}^{(\lambda)}|$ instead of $|\mathbf{X}^{(\lambda)}|$ prevents small values of $|\mathbf{X}^{(\lambda)}|$ turning into large negative values while the general shape of the logarithmic spectra is maintained.

Prior to peak search in the cepstrum, a number of additional preprocessing steps are applied: First, we employ local non-recursive smoothing with a range of two *quefrency* bins in each frame. Also, only cepstral peaks beyond a threshold $q_{0,\min}$ can be reasonably interpreted as a fundamental frequency as the first cepstral bins ($q < q_{0,\min}$) carry mostly information on signal energy and timbre. Therefore, we attenuate the lowest q-bins by applying a high-pass soft-mask function

$$C_{\text{HP}}^{(\lambda)}(q) = C^{(\lambda)}(q) \cdot \left(\frac{q}{q + q_{\text{HP}}}\right), \; q \in \{1; q_{\max}\} \tag{4}$$

to the cepstrum. The cepstral peak search range is further restricted following the rationale that a salient peak in the cepstrum can be expected to be preceded by at least one local minimum $q_{\min}$. The index $\hat{q}(\lambda)$ of the global cepstral maximum serves as a preliminary candidate for $\hat{q}_0$:

$$\hat{q}(\lambda) = \underset{q \in \{q_{\min}; q_{\max}\}}{\arg\max} \; C^{(\lambda)}(q). \tag{5}$$

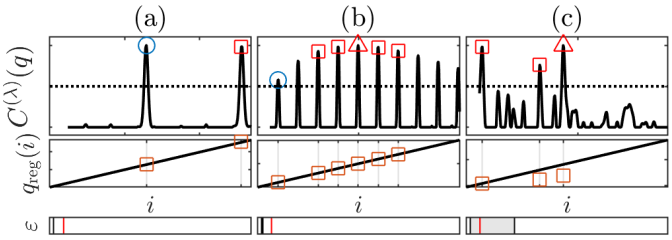

Figure 2: Octave error compensation using the regression approach. Top row: the cepstra $C(q)$ with their most salient peaks $q_p(i)$ (□) and the minimum peak height $C_{\min}$ (dashed line). Mid row: the regression of the peak locations $q_p(i)$. Bottom row: error ε_{reg} and threshold c_{thr} (red line). In (a) the cepstral maximum $\hat{q}$ (◯) coincides with the leftmost cepstral peak (△). In (b) $\hat{q}_0 = q_p(1)$ is chosen as the peaks $q_p(i)$ (□) are on a regular grid of subharmonics of $\hat{q}$ so they all lie close to the regression line $q_{\text{reg}}(i)$. In (c) the peaks (□) do not lie on a regular grid of multiples, thus $\varepsilon_{\text{reg}} > c_{\text{thr}}$ and no compensation is applied.

However, due to the periodic structure of harmonic signals, this maximum might as well be located within the neighborhood of one of the first I integer multiples $\hat{q} = i \cdot q_0$, $i \in \mathbb{N}^{\leq I}$ of q_0. This leads to so-called sub-octave errors as the corresponding frequency estimates $\hat{F} = f_s/\hat{q} = f_s/(i \cdot q_0) = F_0/i$ relate to the subharmonic series of the actual fundamental frequency F_0 (cf. Figure 5).

B. Octave error compensation

To probe the cepstrum for possible sub-octave errors, the N_{p} most salient peak locations $[q_{\text{p}}(1), \dots, q_{\text{p}}(N_p)]$ are collected. This search is constrained by a minimal distance of ε_q and a minimum peak height $C_{\min} = C^{(\lambda)}(\hat{q})/2$ within a range limited by $q'_{\max} = 2\hat{q} + \varepsilon_q$. The locations $q_{\text{p}}(i)$, $i \in \mathbb{N}^{\leq I}$ are approximated as integer multiples of the median peak distance $\tilde{d}_{\text{p}}$ and used as supports for a linear regression with slope a_{reg} and intercept q_{int}:

$$q_{\text{p}}(i) \approx i \cdot \tilde{d}_{\text{p}} \tag{6}$$

$$q_{\text{reg}}(i) = a_{\text{reg}} \cdot i + q_{\text{int}}. \tag{7}$$

If the peak locations $q_{\text{p}}(i)$ are located close to a linear grid of (sub)-harmonics of $\hat{q}$, they are well approximated by $q_{\text{reg}}(i)$ so the relative regression errors decrease (cp. Figure 2).

In the case of a regular grid and a well-fitting regression the location $q_{\text{p}}(1)$ of the leftmost salient cepstral peak refers to the smallest *rahmonic* period of the particular frame. It is chosen as candidate $\hat{q}_0(\lambda)$ if the root mean square approximation error

$$\varepsilon_{\text{reg}} = \sqrt{\frac{1}{I}\sum_i^I \left(1 - q_{\text{reg}}(i)/q_{\text{p}}(i)\right)^2} \tag{8}$$

is below the heuristically determined threshold $c_{\text{thr}} = 0.05$.

We denominate this heuristics as *harmComp*. It proved to be rather robust with regard to different cepstral shapes. A narrow median filter with a width of 3 frames smoothens the estimated track $\hat{F}_0(\lambda) = f_s/\hat{q}_0(\lambda)$ and reduces single outliers.

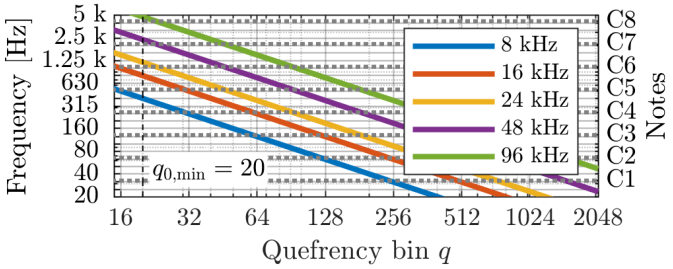

Figure 3: Relation between *quefrency* and frequency for different sampling rates f_s and frame lengths up to $L = 4096$.

C. Time-domain based refinement

To improve the F_0 estimation accuracy beyond the grid of integer *quefrency* bins q we propose a least squares (LS) harmonic approximation [16] within a defined range of frequency candidates in the vicinity of the preliminary F_0 estimate $\hat{F}_0(\lambda) = f_s/\hat{q}_0(\lambda)$. For each signal frame λ the time domain signal $\mathbf{x}^{(\lambda)}$ is approximated by a sinusoidal model with a_k, $\Omega_k = 2\pi f_k/f_s$ and ψ_k denoting the amplitudes, normalized frequencies and phases for each of its K components.

$$h(n) = a_0 + \sum_{k=1}^{K} a_k \cos\left(\Omega_k n + \psi_k\right) \tag{9}$$

Searching for the fundamental frequency F_0 of $\mathbf{x}^{(\lambda)}$, one can assume that the components of $h(n)$ are harmonic ($k \in \mathbb{N}^{\leq K}$), thus $\Omega_k = k\Omega_0 = k(2\pi F_0/f_s)$. With eq. (9), the best harmonic approximation $\mathbf{y}^{(\lambda)}(\Omega_0)$ in the sense of the least squares method can then be found as described in [15]. For each signal frame λ it minimizes the error function

$$\varepsilon_{\mathrm{LS}}(\lambda, \Omega_0) = \|\mathbf{x}^{(\lambda)} - \mathbf{y}^{(\lambda)}(\Omega_0)\|_2^2. \tag{10}$$

The refined fundamental frequency $\tilde{F}_0$ is found by picking the frequency candidate with the minimum error $\varepsilon_{\mathrm{LS}}(\lambda, \Omega_0)$. As the amplitude and phase components a_k and ψ_k of the harmonic model are not required for the fundamental frequency estimation task, their computation can be omitted.

In our default approach the refinement step is performed twice: first with a coarse spacing of $\Delta F_{0,c} = 10\,\mathrm{Cent}$ within the search range of $\hat{F}_0 \pm 50\,\mathrm{Cent}$ and afterwards with a finer spacing of $\Delta F_{0,f} = 2\,\mathrm{Cent}$ within the range of $\hat{F}_0 \pm 10\,\mathrm{Cent}$.

IV. Analysis of F_0 Resolution

Frequency analysis in the cepstral domain benefits from the fact that a uniform cepstral grid results in a high resolution for low frequencies and a relatively sparse representation for higher frequencies. As the resolution in the frequency domain depends on the sampling rate f_s (cf. (1)), the relation between frequency and *quefrency* is shown in Figure 3 for seven different sampling frequencies. Interestingly, (1) does not depend directly on the DFT length L. In fact, the DFT length determines the lower bound of fundamental frequencies $F_{0,\min} = 2f_s/L$ that can be extracted. Note that this boundary coincides with the spectral resolution of the DFT when a rectangular analysis window is used. For other window functions $F_{0,\min}$ will be larger.

In order to further quantify the resolution we investigate the relative fundamental frequency estimation error $\Delta F_0 = \hat{F}_0 - F_0$ in relation to the error $\Delta q_0 = \hat{q}_0 - q_0$ in the cepstral domain and find

$$\frac{\Delta F_0}{F_0} = \frac{f_s}{F_0}\left(\frac{\hat{q}_0 - q_0}{\hat{q}_0 q_0}\right) = \frac{\Delta q_0 F_0}{f_s + \Delta q_0 F_0}. \tag{11}$$

Assuming that peak picking results in errors in the cepstral domain in the order of one q-bin ($\Delta q_0 \leq 1$) we find $\frac{\Delta F_0}{F_0} \leq \frac{F_0}{f_s + F_0}$. With a sampling rate of $f_s = 44.1\,\mathrm{kHz}$ and fundamental frequencies between $F_{0,\min} = 50\,\mathrm{Hz}$ ($\widehat{\approx}\,\mathrm{G1}$) and $F_{0,\max} = 5\,\mathrm{kHz}$ ($\widehat{\approx}\,\mathrm{D}^{\#}8$) the deviation is in the range of $5\,\mathrm{Cent}$ to $94\,\mathrm{Cent}$ and thus smaller than one semitone. This error can be further reduced by the subsequent time-domain based refinement step described in Section III-C.

V. Experimental Evaluation

In a first step we systematically investigated the influence of different instruments and pitches[1] on the fundamental frequency estimation by applying the proposed method to the McGill University Master Samples (MUMS) library [17]. This dataset contains up to 6000 sound samples of a wide range of (pitched) musical instruments playing single notes across their respective tonal range, so an analysis both across different instruments at fixed pitch and across the particular tone range of each instrument is possible.

However, as the MUMS database exhibits a certain number of errors like tuning error, octave errors, and erratic labeling [18], it is less suited to evaluate the overall performance of a fundamental frequency estimation algorithm. Therefore, in the second step of evaluation we applied our method to the MDB-melody-synth database [19]. It contains 65 songs taken from the MedleyDB database [20] where the melody tracks have been resynthesized using a sinusoidal analysis/synthesis framework [21]. Similar to signals synthesized from MIDI data this dataset provides perfect fundamental frequency annotations while timbre and dynamics of the synthesized tracks are very close to the original recordings.

The Gross Pitch Error (GPE), Fine Pitch Error (FPE), Voicing Decision Error (VDE) and F0 Frame Error (FFE) error metrics [9], [22] have been used as evaluation criteria:

$$\frac{\mathrm{GPE}}{100\,\%} = \frac{N_{\mathrm{GPE}}}{N_{\mathrm{VV}}} \tag{12}$$

$$\frac{\mathrm{FPE}}{\mathrm{Cent}} = \sqrt{\mathrm{Var}\left(1200\log_2\left(\frac{\hat{F}_0(\lambda)}{F_0(\lambda)}\right)\right)},\ \lambda \in \overline{\Lambda_{\mathrm{V,GPE}}} \tag{13}$$

$$\frac{\mathrm{VDE}}{100\,\%} = \frac{N_{\mathrm{VU}} + N_{\mathrm{UV}}}{N_{\mathrm{total}}} \tag{14}$$

$$\frac{\mathrm{FFE}}{100\,\%} = \frac{N_{\mathrm{VU}} + N_{\mathrm{UV}} + N_{\mathrm{GPE}}}{N_{\mathrm{total}}} = \frac{\frac{N_{\mathrm{VV}}}{N_{\mathrm{total}}}\,\mathrm{GPE} + \mathrm{VDE}}{100\,\%} \tag{15}$$

[1] Simplified, the *pitch* denotes the human percept of fundamental frequencies. In a musical context, it assigns musical tones to relative positions on a musical scale. Throughout this paper, the term *pitch* is used where fundamental frequency is assessed from a musical point of view.

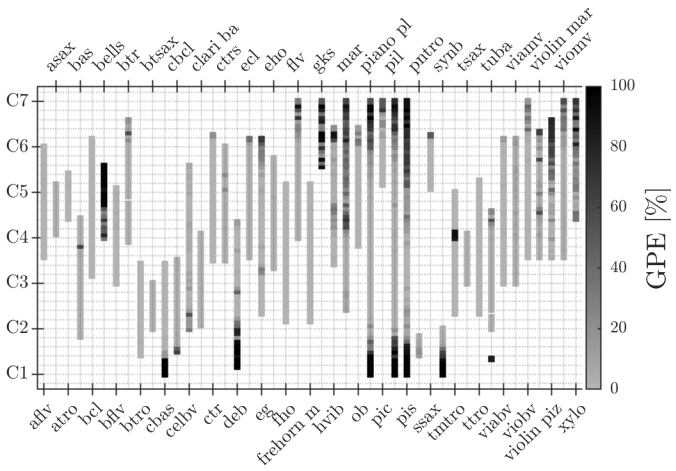

Figure 4: Gross Pitch Error (GPE) with method Ceps6 for several instruments from the MUMS database

The set of frames detected as voiced but with a Gross Pitch Error is defined as $\Lambda_{\mathrm{V,GPE}} = \{\lambda : \left|1200 \log_2 \left(\frac{\hat{F}_0(\lambda)}{F_0(\lambda)}\right)\right| > \Delta F_{\mathrm{GPE}}\}$ whereas $\overline{\Lambda_{\mathrm{V,GPE}}}$ denotes the set of frames within the GPE tolerance ΔF_{GPE}. $N_{\mathrm{GPE}} = |\Lambda_{\mathrm{V,GPE}}|$ denotes the number of GPE frames, N_{VU} and N_{UV} indicate the numbers of true voiced frames detected as unvoiced and vice versa, and N_{total} denotes the total number of evaluated frames.

VI. Results

To assess the impact of the particular refinement steps, we denote the following methods (cf. Figure 1): Ceps1 comprises only the cepstral peak search, in Ceps2 the *harmComp* heuristics (cf. Section III-B) is added and in Ceps3 a subsequent median filter with a width of 3 frames is applied to prevent single outliers. Subsequently, the coarse and fine LS searches described in Section III-C are considered in Ceps4 and Ceps5, followed by another median filter in method Ceps6.

Although the MUMS dataset is not perfect for an instrumental evaluation of F_0 estimation performance (see Section V), the GPE results for method Ceps6 depicted in Figure 4 show that the proposed enhanced cepstral method performs reliably for a wide range of instruments and frequencies. Notably, only for the extreme ends of the practically used F_0 range and for idiophone instruments like tubular bells (*bells*), glockenspiel (*gks*), marimba (*mar*) and xylophone (*xylo*) a poorer fundamental frequency estimation performance is obtained. This can be attributed to the inharmonic distribution of partial tones in these instruments. The performance in the low frequency range can be further improved by increasing the transformation length L at the cost of computational effort (cf. Section IV).

Table I shows the error metrics yielded for the MDB-melody-synth database. For better comparison with other methods we applied two different GPE tolerances of $\Delta F_{\mathrm{GPE}} = 50\,\mathrm{Cent}$ (corresponding to one semitone) and $\Delta F_{\mathrm{GPE}} = 100\,\mathrm{Cent}$ as in e.g. [3], [9]. As a benchmark we refer to the well-established methods PYIN [3] (obtained with the Vamp plugin for *Sonic Annotator* [23]) and CREPE [10].

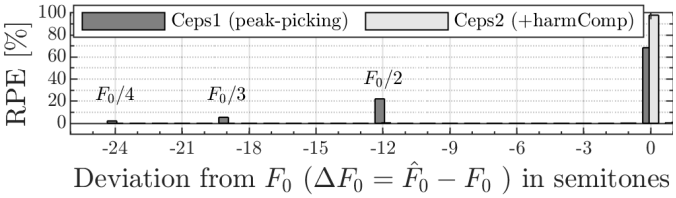

Figure 5: Distribution of Raw Pitch Errors (RPE) for all voiced frames from the MDB-melody-synth database.

	GPE [%]		FPE [¢]		VDE [%]		FFE [%]	
	50 ¢	100 ¢	50 ¢	100 ¢	50 ¢	100 ¢	50 ¢	100 ¢
Ceps1	31.55	31.09	7.69	9.50	2.01	2.01	17.92	17.69
Ceps2	2.03	1.16	9.95	11.55	2.01	2.01	3.03	2.59
Ceps3	1.98	1.02	10.26	12.01	1.02	1.02	2.02	1.53
Ceps4	1.91	1.04	9.29	11.14	1.02	1.02	1.98	1.54
Ceps5	2.05	1.06	8.72	10.96	1.02	1.02	2.05	1.55
Ceps6	1.98	1.05	8.50	10.71	0.03	0.03	1.02	0.56
PYIN	6.22	5.33	6.78	9.57	0.96	0.96	4.09	3.64
CREPE	1.45	0.76	7.38	9.30	3.15	3.15	3.88	3.54

Table I: Pitch Error Scores for MDB-melody-synth database with GPE tolerances ΔF_{GPE} of 50 Cent and 100 Cent.

The Raw Pitch Error (RPE) describes the deviation of more than one semitone from the true F_0. As depicted in Figure 5, sub-octave errors ($\hat{F}_0 < F_0$) account for almost 30% of all false F_0 estimations across the MDB-melody-synth dataset if only the maximum cepstral peak is considered (Ceps1). Raw Pitch Errors caused by higher harmonics ($\hat{F}_0 > F_0$) could not be observed throughout the complete dataset. Thus, Figure 5 and the results in Table I illustrate that the *harmComp* heuristics introduced in method Ceps2 (cf. Section III-B) significantly improves the estimation performance such that the PYIN method is outperformed in terms of GPE and VDE on this dataset. As the additional refinement steps in methods Ceps3 to Ceps6 further reduce the FPE and VDE measures, we propose method Ceps6, although method Ceps3 exhibits a slightly better GPE.

The proposed method has been implemented and evaluated using MATLAB v9.5 (R2018b) on a standard PC with Intel Core i7-4790 CPU. For the most elaborate and computationally demanding method Ceps6 an average real-time ratio of $r_{RT} = T_{\mathrm{comp}}/T_V = 0.33$ was measured with $L = 2048$ and $R = 512$ on the MDB-melody-synth dataset sampled at $f_s = 44.1\,\mathrm{kHz}$. T_{comp} and T_V denote the overall computation time and the total duration of all voiced frames in the database.

VII. Conclusions

In this paper we presented a fundamental frequency estimation method for music signals that combines a search for plausible F_0 candidates in the cepstral domain supplemented by a regression-based heuristics to avoid (sub-)octave errors and a subsequent fine search based on a least squares approximation of a harmonic time domain model.

A preceding analysis of the F_0 resolution (cf. Section IV) demonstrated that the expected F_0 estimation error for simple peak-picking in the cepstrum depends on the fundamental

frequency but is below $100\,\mathrm{Cent}$ and thus smaller than one semitone. Especially for high-pitched instruments we therefore advise to also use the resolution refinement described in Section III-C. Then, the resolution and computational complexity demands can be well balanced for a wide range of notes.

An experimental evaluation showed that with the proposed cepstrum-based fundamental frequency estimation method an F_0 accuracy comparable to other established methods can be obtained. It also proved to be robust to a wide range of different instrumental sounds and pitches and shows a low computational complexity.

Acknowledgment

This work was funded by DFG Collaborative Research Center SFB 823, subproject B3.

References

[1] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, vol. 17, Amsterdam, 1993, pp. 97–110.

[2] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[3] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 659–663.

[4] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding Synth.*, vol. 495, p. 518, 1995.

[5] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.

[6] M. G. Christensen, P. Stoica, A. Jakobsson, and S. Holdt Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88, no. 4, pp. 972–983, Apr. 1, 2008.

[7] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "A Robust and Computationally Efficient Subspace-Based Fundamental Frequency Estimator," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 3, pp. 487–497, Mar. 2010.

[8] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "A Kalman-based fundamental frequency estimation algorithm," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2017, pp. 314–318.

[9] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Vancouver, Canada, May 2013, pp. 7815–7819.

[10] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A Convolutional Representation for Pitch Estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 161–165.

[11] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking," in *Proc. of the Symposium on Time Series Analysis*, vol. 15, New York: Wiley, 1963, pp. 209–243.

[12] P. McLeod, "Fast, Accurate Pitch Detection Tools for Music Analysis," PhD thesis, University of Otago, Otago, New Zealand, May 30, 2008.

[13] A. M. Noll, "Cepstrum Pitch Determination," *J. Acoust. Soc. Am.*, vol. 41, no. 2, pp. 293–309, Feb. 1, 1967.

[14] D. Arfib and N. Delprat, "Alteration of the vibrato of a recorded voice," *Int. Comput. Music Conf. Proc.*, vol. 1999, 1999.

[15] A. Bánhalmi, K. Kovács, A. Kocsor, and L. Tóth, "Fundamental frequency estimation by least-squares harmonic model fitting," in *Proc. Annu. Conf. Int. Speech Communic. Assoc. (INTERSPEECH)*, Lisbon, Portugal, Sep. 4–8, 2005, pp. 305–308.

[16] B. G. Quinn and P. J. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78, no. 1, pp. 65–74, Mar. 1, 1991.

[17] F. Opolko and J. Wapnick, *McGill University master samples (3 CDs)*, 1987.

[18] T. Eerola and R. Ferrer, "Instrument Library (MUMS) Revised," *Music Perception*, vol. 25, no. 3, pp. 253–255, Feb. 1, 2008.

[19] J. Salamon, R. Bittner, J. Bonada, J. J. Bosch, E. Gómez, and J. P. Bello, *MDB-melody-synth*, Nov. 8, 2018.

[20] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Taipei, Taiwan, Oct. 28, 2014.

[21] J. Salamon, R. M. Bittner, J. Bonada, J. J. Bosch, E. Gómez Gutiérrez, and J. P. Bello, "An Analysis/synthesis framework for automatic F0 annotation of multitrack datasets," 2017.

[22] W. Chu and A. Alwan, "Reducing F0 Frame Error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Apr. 2009, pp. 3969–3972.

[23] C. Cannam, M. Sandler, M. O. Jewell, C. Rhodes, and M. d'Inverno, "Linked Data and You: Bringing Music Research Software into the Semantic Web," *J. New Music Res.*, vol. 39, no. 4, pp. 313–325, Dec. 1, 2010.