TUNING FREQUENCY ESTIMATION USING CIRCULAR STATISTICS

Karin Dressler
Fraunhofer IDMT
Langewiesener Str. 22
98693 Ilmenau, Germany
dresslkn@idmt.fraunhofer.de

Sebastian Streich
Music Technology Group
Pompeu Fabra University
Barcelona, Spain
sstreich@beat.yamaha.co.jp

ABSTRACT

In this document a new approach on tuning frequency estimation based on circular statistics is presented. Two methods are introduced: the calculation of the tuning frequency over an entire audio piece, and the estimation of an adapting reference frequency for a single voice.

The results for the tuning frequency estimation look very good for audio pieces where the dominant voices are tuned close to the equal-temperament scale and exhibit only moderate frequency dynamics. For the analysis of popular western music, the method does not achieve very robust results due to the strong frequency dynamics of the human singing voice. Nevertheless, the method could be improved by excluding the singing voice from the calculation taking only the accompaniment into account.

The main advantage of the proposed method lies especially in the easy computation of an adaptive reference frequency using an exponential moving average. This adaptive reference can for example be used in the quantization of the singing voice into a note representation.

1 INTRODUCTION

The tuning frequency estimation for musical audio signals by itself is not a very popular research topic in the MIR domain. If addressed at all, it is mostly regarded as a minor preprocessing step in a larger system performing for example key or chord detection. However, results from key detection experiments with real world recordings show [1] that the tuning frequency estimation might have some impact on the overall system performance. In general, any system that at some point quantizes frequencies from the continuous scale into pitch classes or semitone intervals needs a specified reference frequency. Since in most practical cases it is not possible to make a safe assumption about the underlying tuning frequency of a music recording the only possibility is to estimate this value from the data itself.

Several methods reported in the literature make use of pitch histograms (e.g. [2],[3] and [4]). These approaches share the disadvantage that the pitch values already need to be quantized for the preprocessing in order to construct the histograms. Others rely on iterative optimization mechanisms (e.g. [5]). Here we want to propose two different approaches (either as a global or as an adaptive estimation) originating from the domain of circular statistics that we consider more elegant and advantageous in several ways. The methods do not require a quantization, they are efficiently computable, consume only minimal storage, and produce as a byproduct a confidence measure for the reference frequency estimate.

2 METHOD

If we assume a twelve-tone equal temperament scale, each semitone lies exactly on a 100 cent grid. The cent is a logarithmic unit of measure used for musical intervals. One octave equals 1200 cents. For the frequency to cent conversion, we compute cent values relative to the standard tuning frequency of 440 Hz:

$$c = 1200 \cdot \log_2 \left(\frac{f}{440\,\mathrm{Hz}} \right). \tag{1}$$

When frequencies have to be quantized to the nearest semitone, all frequencies within ± 50 cent distance are assigned to the same tone. The absolute cent deviation may not become bigger than 50 cent, because in this case simply the next semitone will be considered.

For the estimation of the tuning frequency only the cent deviation from the 100 cent grid is of importance, no matter to which semitone in particular a frequency is assigned. If a quantity ”wraps around” after reaching a certain value we may speak of circular data. The simple calculation of the arithmetic mean is not an appropriate statistic for such data. In this case histogram techniques can be used to exploit a certain trend in the data.

Another way to deal with circular or directional quantities is circular statistics, a subdiscipline of statistics that is for example applied when directions or periodic time measurements (e.g. day, week, month) have to be evaluated[6]. Such data are often best handled not as scalars, but as (unit) vectors in the complex plane. Each cent value is treated as a unit vector $\hat{u}$ whose angle ϕ is the appropriate fraction of a full circle:

$$\hat{u} = 1 \cdot e^{j\phi} \tag{2}$$

with

$$\phi = \frac{2\pi}{100} \cdot c.$$

2.1 Overall Estimate

In order to determine the tuning frequency of the entire audio piece the sum of all "cent" vectors is computed and then divided by the number of values N to get a mean $\bar{z}$ of the circular quantities:

$$\begin{aligned} \mathrm{Re}(\bar{z}) &= \frac{\sum_i \cos(\phi_i)}{N} \\ \mathrm{Im}(\bar{z}) &= \frac{\sum_i \sin(\phi_i)}{N}. \end{aligned} \tag{3}$$

The mean $\bar{z}$ is again a vector in the complex plane which means it can be decomposed into a magnitude and a phase value. The magnitude $|\bar{z}|$ lies in the interval $[0, 1]$. It depends on the amount of variation in the vectors that are averaged. The more the cent values are scattered, the smaller the magnitude of the complex number $\bar{z}$ will be, whereas if $\bar{z}$ has a magnitude close to 1 that would imply a significant tendency in the data. We can therefore see $|\bar{z}|$ as a confidence measure for the tuning frequency estimate.

The phase angle $\bar{\phi}$ of $\bar{z}$ is converted back into cents to obtain the deviation from the standard tuning frequency of 440 Hz in cents:

$$\Delta c = \frac{100}{2\pi} \arg(\bar{z}). \tag{4}$$

With help of the estimated cent deviation Δc the reference frequency can be calculated as

$$f_{ref} = 2^{\Delta c/1200} \cdot 440\,\mathrm{Hz}. \tag{5}$$

The results of the calculation can be improved if each cent vector is weighted by a certain weighting factor r_i:

$$\begin{aligned} \mathrm{Re}(\bar{z}) &= \frac{\sum_i r_i \cos(\phi_i)}{\sum_i r_i} \\ \mathrm{Im}(\bar{z}) &= \frac{\sum_i r_i \sin(\phi_i)}{\sum_i r_i} \end{aligned} \tag{6}$$

Three different approaches using equation (6) were examined:

Spectral peaks: The instantaneous frequencies and magnitudes of salient spectral peaks were computed using the spectral analysis frontend described in [7]. The cent vectors calculated from the instantaneous frequencies are weighted with the respective peak magnitudes to avoid a high impact of noise peaks.

Melody pitch: In the second approach, only the estimated pitch frequencies of the melody voice are taken into account. The resulting cent vectors are weighted by the pitch magnitude. The pitch contour and the pitch magnitude are computed by the melody extraction algorithm described in [8].

Stable melody pitch: There are many instruments that allow a substantial frequency variation within one single tone. The human singing voice is a prominent example – a sung vibrato can easily exceed more than one semitone in either direction from the note itself. Nevertheless the perceived tone height is to a certain degree stable. In our third approach, we successively assign stable tone heights to the identified melody tone in order to get a more reliable reference frequency for the melody voice.

2.2 Exponential Moving Average

The exponential moving average (EMA) is a technique used to analyze time series data. The EMA applies weighting factors which decrease exponentially in time – giving more importance to recent observations while still not discarding older observations entirely. If the pitch of a tone needs to be quantized into the equal-temperament scale, it has to be evaluated relative to the current reference frequency, which might change over time. At each time instance a reference frequency estimate is computed by the following recursive formula:

$$z'_i = (1-\alpha) \cdot z'_{i-1} + \alpha \cdot r_i \, e^{j\phi_i} \quad \text{with} \quad z'_0 = 0. \tag{7}$$

Each cent vector $e^{j\phi_i}$ is again weighted with a magnitude measure r_i. The decay of older values is determined by a constant smoothing factor α

$$\alpha = 1 - 0.5^{\Delta t/T_{1/2}}. \tag{8}$$

The parameter Δt specifies the time advance between two observations, the parameter $T_{1/2}$ is the half-life period. In case of an STFT spectrogram Δt is given by

$$\Delta t = \frac{L}{f_s}, \tag{9}$$

where L is the hop-size in samples and f_s determines the sampling rate of the audio data.

Finally, z'_i is normalized, so that the magnitude of $\overline{z'_i}$ lies in the interval $[0, 1]$:

$$\overline{z'_i} = \frac{z'_i}{r'_i} \tag{10}$$

with

$$r'_i = (1-\alpha) \cdot r'_{i-1} + \alpha \cdot r_i$$

For each frame, the argument of $\overline{z'_i}$ corresponds to a reference frequency estimate which can be calculated as given in equations (4) and (5). Again the magnitude of the normalized complex $\overline{z'_i}$ may be used as a confidence measure for the estimate. Of course, meaningful results can be expected only after an initial settling phase has passed. Two different values of the half-life period $T_{1/2}$ are used for the evaluation: $T_{1/2} = 2$s gives more stable results, while $T_{1/2} = 0.5$s allows for very quick adaption.

	1.) spectral peaks			2.) melody pitch			3.) stable melody pitch		
file	$\lvert\bar{z}\rvert$	Δc	f_{ref}	$\lvert\bar{z}\rvert$	Δc	f_{ref}	$\lvert\bar{z}\rvert$	Δc	f_{ref}
daisy1	0.54	1.7	440.4	0.77	0.8	440.2	0.83	0.7	440.2
daisy4	0.70	0.2	440.1	0.87	0.5	440.1	0.95	0.5	440.1
jazz1	0.43	-2.1	439.5	0.67	-2.2	439.5	0.71	-1.1	439.7
jazz4	0.37	-0.5	439.9	0.83	0.3	440.1	0.84	1.4	440.4
midi3	0.82	-4.4	438.9	0.96	-4.7	438.8	0.99	-4.7	438.8
midi4	0.39	-2.9	439.3	0.93	-0.7	439.8	0.94	-0.8	439.8
opera4	0.28	8.5	442.3	0.12	12.8	443.3	0.30	-6.9	438.3
opera5	0.13	15.4	443.9	0.11	10.9	442.8	0.48	25.8	446.6
pop2	0.31	-4.4	438.9	0.35	-10.5	437.3	0.45	-13.1	436.7
pop3	0.21	-2.4	439.4	0.41	3.2	440.8	0.50	3.2	440.8

Table 1. Overall tuning frequency estimation of the ISMIR2004 data set

3 RESULTS

3.1 Data Set

For the ISMIR 2004 Audio Description Contest, the Music Technology Group of the Pompeu Fabra University assembled a diverse set of 20 polyphonic musical audio pieces and corresponding melody transcriptions including MIDI, Jazz, Pop and Opera music as well as audio pieces with a synthesized voice (daisy examples) [1] . The small data set allows only a rather limited statistical analysis. Yet, it was intentionally chosen in order to demonstrate the significant influence of the music style on the tuning frequency estimation. We would also like to point out that it is almost impossible to obtain valid ground truth data for a time-varying tuning frequency other than for completely synthesized music. We therefore focus on a qualitative analysis here rather than providing actual detection accuracies.

3.2 Overall Estimate

Table 1 shows the evaluation results of the overall tuning frequency calculation for each audio piece. A very good algorithm performance with high confidence measures can be noted with artificially generated data (midi, daisy) and pieces with a rather stable frequency throughout a tone like in the jazz examples. This is especially true for the third approach using stable melody pitches. The first approach (spectral peaks) has a lower confidence measure for these examples, because noisy peaks and the deviation from the equal-temperament scale of some higher harmonics affect the result[4].

Clearly, the results look very different for the pop and opera examples. The dominant human voice with high frequency dynamics has a strong impact on the tuning frequency estimation. This fact becomes obvious especially in the opera examples, which contain a voice with a very strong vibrato. The expressivity of the voice lays in the fundamental frequency variations, eg. glissandi and vibrato.

We tried to compensate such effects by assigning stable pitch estimates before calculating the reference frequency (stable melody pitch approach). However, the singer is not exactly in tune with the accompaniment and the microtuning of the voice slightly changes during the performance. That is why the spectral peak approach surprisingly yields comparable confidence values in the opera examples, because the harmonics of the accompaniment are taken into account.

Evidently, the proposed method struggles to accurately estimate the overall tuning frequency as long as the human singing voice ”corrupts” the data. We expect to markedly increase reliability by isolating the instruments that have stable pitches.

We conclude that an evaluation of tuning frequency estimators on artificially generated audio does not necessarily show the efficiency of the method with e.g. popular western music.

3.3 Exponential Moving Average

The figures 1 and 2 show two examples for an adapting reference frequency calculated as exponential moving average. We see the computed cent deviation and the confidence measure over time for two different values of the half-life period parameter $T_{1/2}$. The short-term estimates with $T_{1/2} = 0.5$s adapt very quickly to a shifting reference frequency, while the long-term reference frequency estimate has a more stable progression. Of course one will not expect the reference frequency to shift that quickly in a professional recording – otherwise it would not be a reference. Still, even the short-term estimate may be of value if you have to analyze a very mistuned voice of an untalented singer. In this case, if the pitch being quantized repeatedly fits badly to the grid defined by the long-term reference frequency, the algorithm could switch to the short-term reference.

Even if the overall tuning frequency estimate of the jazz1 example does not differ much from the standard tuning, the local reference frequency of the saxophone deviates slightly with the played melody because the saxophone is not tuned in equal-temperament.

The reference frequency of the male voice in the opera5 recording cannot be determined with a high confidence

[1] The data set can be found on the contest web page `http://ismir2004.ismir.net/melody_contest/results.html`

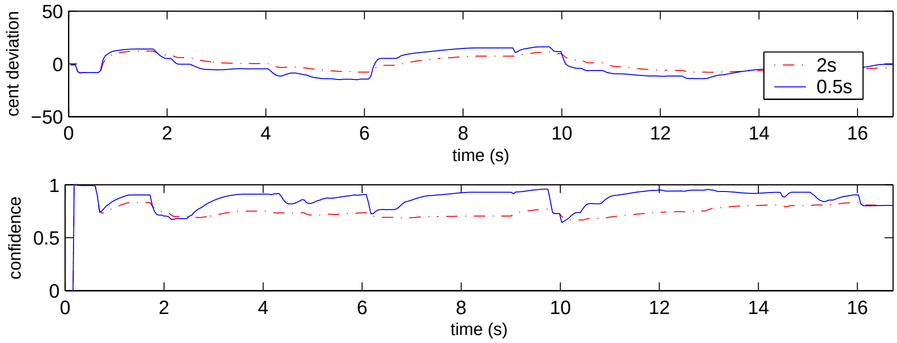

Figure 1. Short-term and long-term estimates of the cent deviation and the confidence for the jazz1 recording

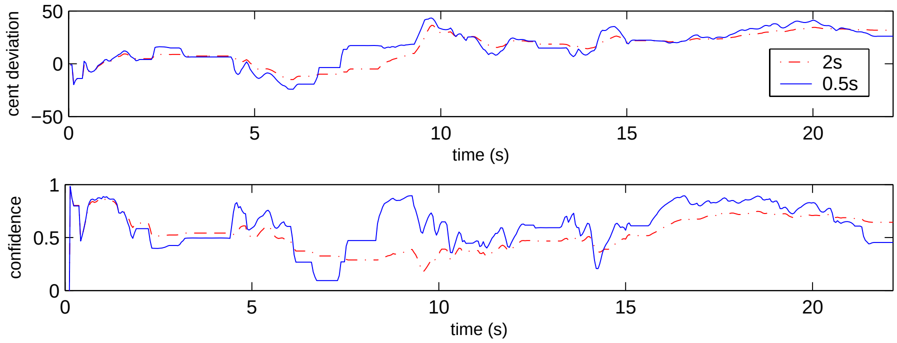

Figure 2. Short-term and long-term estimates of the cent deviation and the confidence for the opera5 recording

measure, as there is no ”ground truth” tuning frequency in the human voice. It can be noted that the reference frequency changes over time: Towards the end of the tune the opera singer is singing sharp while increasing the volume of his voice.

4 CONCLUSION

In this document, we presented a new approach to tuning frequency estimation based on circular statistics. Two methods have been introduced: The calculation of the tuning frequency over an entire audio piece, and the estimation of an adapting reference frequency for a single voice.

The estimation of the tuning frequency over the entire audio piece is a convenient way to estimate the tuning frequency, as long as one can assume that the tuning stays the same throughout the music piece and the dominant voices exhibit only moderate frequency modulation. Yet, especially the human singing voice often does not fulfill these requirements. The method could be improved by the identification of the accompaniment with stable pitch and the exclusion of the human voice from the calculation.

The estimation of the adaptive reference frequency can be used to study the evolution of the tuning over time, which may be useful to monitor the tuning of instruments or singers/choirs singing a capella. The EMA approach also proved useful for the quantization of a sung melody into a note representation, for example in a query by humming system or a melody extraction algorithm. In the easy computation of an adapting reference frequency in different time scales, we see a decisive advantage of the proposed method over histogram techniques.

5 REFERENCES

[1] A. Lerch, “On the requirement of automatic tuning frequency estimation,” in *Proc. of the 7th Int. Conf. on Music Information Retrieval*, 2006, pp. 212–215.

[2] M. P. Ryynänen, “Probabilistic modelling of note events in the transcription of monophonic melodies,” M.S. thesis, Tampere University of Technology, 2004.

[3] C. A. Harte and M. B. Sandler, “Automatic chord identification using a quantised chromagram,” in *Proc. of the 118th AES Convention*, 2005, number 6412.

[4] E. Gómez, *Tonal Description of Music Audio Signals*, Ph.D. thesis, UPF, Barcelona, Spain, 2006.

[5] S. Dixon, “Multiphonic note identification,” in *Australian Computer Science Communications, 18, 1, Proc. of the 19th Australasian Comp. Sc. Conf.*, 1996, pp. 318–323.

[6] E. Batschelet, *Circular Statistics in Biology*, Mathematics in Biology. Academic Press, 1981.

[7] K. Dressler, “Sinusoidal extraction using an efficient implementation of a multi-resolution FFT,” in *Proc. of DAFx-06*, 2006, pp. 247–252.

[8] K. Dressler, “An auditory streaming approach on melody extraction,” in *2nd Music Information Retrieval Evaluation eXchange (MIREX)*, 2006