MUSIC PITCH REPRESENTATION BY PERIODICITY MEASURES BASED ON COMBINED TEMPORAL AND SPECTRAL REPRESENTATIONS

Geoffroy Peeters

Ircam - CNRS (Sound Analysis/Synthesis Team)
1, pl. Igor Stranvinsky - 75004 Paris - France
geoffroy.peeters@ircam.fr

ABSTRACT

Periodicity estimation of an audio signal, for applications such as pitch, multiple pitch or tempo estimation is often problematic due to the presence of multiple harmonics in the audio signal producing octave errors. While pitch models or rhythm models can be used, they remain often dedicated to a specific problem. In this paper, we propose a straightforward approach for periodicity estimation based on the combination of a spectral representation and a temporal representation. This method allows a better emphasis on the frequencies corresponding to the various pitches. We show the ability of this representation to adequately estimate pitch and visualize signals with multiple pitch content.

1. INTRODUCTION

Because of its numerous applications (transcription, separation, front-end for further sound analysis), periodicity of a musical audio signal is one of the most important information. Numerous proposals have been made for this task, either based on novel signal representations ([2] [3] [4]), or on novel periodicity models ([5] [6] [7] [8] [9]). In this paper, we propose a straightforward approach for periodicity estimation based on the combination of a spectral representation (amplitude of the discrete Fourier transform, amplitude of the frequency-reassigned DFT or autocorrelation of the previous) and a temporal representation (autocorrelation function or real-cepstrum). This method allows a better emphasis on the frequencies corresponding to the various pitches.

The paper is organized as follows. In part 2, we give an overview of the spectral and temporal periodicity representations that will be used for our method. In part 3, we propose our method for periodicity estimation based on the combination of spectral and temporal representations. Five different functions are proposed. In part 4, we evaluate the proposed functions for two different tasks: pitch estimation of musical instrument and multiple pitch signal visualization.

2. PERIODICITY REPRESENTATION

2.1. Spectral representation

The **Discrete Fourier Transform of a temporal signal** (DFT) is one of the most used representation for periodicity estimation. There exist numerous methods to estimate the periodicity from the amplitude of the DFT, $X(k)$, (maximum likelihood [5], two-way mismatch [6], ...). A naive method would be to take the first peak of the DFT amplitude (within a certain range). This would make the underlying assumption that energy exist at the pitch frequency and that the higher harmonics of the spectrum do not bring significant extra-information. Since the DFT of a periodic impulse signal is a set of harmonically related components, a peak also exists at "twice the pitch" frequency. An octave error is therefore likely to occur.

The spectral resolution of the DFT can be improved using the **Frequency Reassignment** (REAS) based on the reassigned spectrogram proposed by [4]. It consists of reallocating the energy of the "bins" of a DFT to the frequency ω_r corresponding to their center of gravity. It is based on the computation of the instantaneous frequency (time derivative of the phase) which can be efficiently computed by:

$$\omega_r(x,\omega_k) = \omega_k - \Im\left\{\frac{DFT_{dh}(x,\omega_k)}{DFT_h(x,\omega_k)}\right\} \quad (1)$$

where $\Im$ stands for the imaginary part, h stands for the analysis window and dh stands for the time derivative of the window: $\partial h(t)/\partial t$. Each bin ω_k of the DFT is then reassigned to its center of gravity ω_r. The values are accumulated over frequency. We note $X_{re}(k)$ the resulting amplitude spectrum.

The **Auto-Correlation Function of the amplitude of the DFT** (ACFofDFT) measures the periodicity of the DFT amplitude peak positions using an auto-correlation function [10] [11]. The autocorrelation of $X(k)$ at a frequency k is defined as

$$\hat{R}(k) = \frac{1}{N-k}\sum_{K=0}^{N-k-1} X(K)X(K+k) \quad (2)$$

$\hat{R}(k)$ does not make any assumption related to the presence/absence of energy at the pitch frequency. It is therefore robust to missing fundamental. However, as it is the case for the DFT, a peak exists at "twice the pitch" frequency and octave errors are likely to occur. Moreover, in the presence of multiple pitch (f_1,f_2), $\hat{R}(k)$ will be simultaneously influenced by the inter-distance between the harmonic positions of each single pitch separately (nf_1,nf_2) and by the inter-distance between the harmonics of the various pitches ($nf_1 - n'f_2$). Also, for a spectrum reducing to a single component (like some flute sounds), the estimation of the autocorrelation of $X(k)$ is not appropriate. In the following we will use the autocorrelation of $X(k)$ but also of $X_{re}(k)$.

2.2. Temporal representation

The **Auto-Correlation Function of the temporal signal** (ACF) [12] measures the correlation of the signal with a time delayed version

Part of this work was conducted in the context of the European IST project Semantic HIFI [1] (http://shf.ircam.fr). Many thanks to Xavier Rodet for the fruitful discussions.

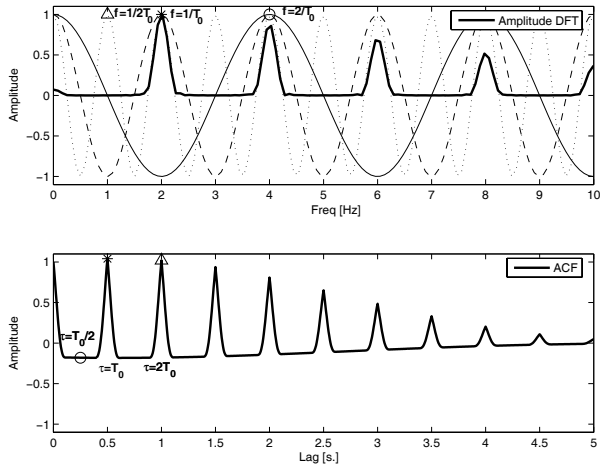

Fig. 1. [top] Amplitude of the DFT of the signal; superimposed: $g_{\tau=T_0/2}(f)$ (continuous line) and $f = \frac{2}{T_0}$ (circle mark); $g_{\tau=T_0}(f)$ (dashed line) and $f = \frac{1}{T_0}$ (star mark); $g_{\tau=2T_0}(f)$ (dotted line) and $f = \frac{1}{2T_0}$ (triangle mark); [bottom] autocorrelation function; superimposed: $\tau = \frac{T_0}{2}$, $\tau = T_0$, $\tau = 2T_0$ positions; on a periodic impulse signal at f_0=$\frac{1}{T_0}$=2Hz.

of itself. There exist numerous methods to estimate the periodicity from the ACF ([7]). The simplest one is to take the first positive peak of the ACF (within a certain lag range). However, since the ACF of a periodic impulse signal is a set of periodically related components, a peak exists at "twice the period" ("half the pitch") lag. An octave error is therefore likely to occur. The ACF can be computed efficiently using the inverse Fourier transform of the power spectrum. Since the power spectrum is real and symmetric, its inverse Fourier transform reduces to the real part. Therefore we can consider $\hat{r}(l)$ as the projection of the power spectrum $X^2(f)$ on a set of cosine functions (with $\phi(f = 0) = 0$) with frequencies equal to the lag $\tau_l = l/\text{sr}$[1]:

$$\hat{r}(l) = \frac{1}{N-l} \sum_k |X(k)|^2 cos\left(2\pi k \frac{l}{N}\right) \quad (3)$$

$\hat{r}(l)$ measures the periodicity of the peak positions of the power spectrum. In Fig. 1, we illustrate this for a periodic impulse signal at f_0=$\frac{1}{T_0}$=2Hz, the power spectrum of which is a set of harmonically related peaks. We note $g_\tau(f) = cos(2\pi f\tau)$ the cosine function. Positive values of $\hat{r}(\tau)$ occur for τ values for which only the positive parts of $g_\tau(f)$ coincide with the power spectrum peaks. This is the case for $\tau = kT_0, k \in \mathbf{N}$ in the figure. Non-positive values when positives parts and negative parts of $g_\tau(f)$ in turn coincide with the power spectrum peaks. This is the case for $\tau = T_0/k, k > 1, k \in \mathbf{N}$ in the figure. This explains why the ACF is likely to produce "twice the period" ("half the pitch") octave errors but not "half the period" ("twice the pitch"). However, a major drawback of the autocorrelation function is its sensitiveness to the spectral envelope.

Apart from its homomorphic properties, the **Real-Cepstrum of the temporal signal (CEP)** [13] [14] can also be considered as a projection of the spectrum on a set of cosine functions. In the case of the real-cepstrum, the amplitude spectrum is expressed in the log-domain before taking its inverse Fourier transform. The log-scale allows a compression on the amplitude range, and therefore reduces the amplitude variations due to the spectral envelope. This is why the real-cepstrum is often preferred over the ACF for pitch estimation. The real-cepstrum is expressed as:

$$\hat{c}(l) = \frac{1}{N-l} \sum_k \log(|X(k)|) cos\left(2\pi k \frac{l}{N}\right) \quad (4)$$

However the log-scale emphasizes the noise part of the spectrum. In the following, when using the real-cepstrum, a threshold is applied to $\log(|X(k)|)$ in order to reduce the influence of noise.

3. COMBINING TEMPORAL AND SPECTRAL REPRESENTATION

In the rest of the paper, we note $X(k)$ as the amplitude of the DFT, $X_{re}(k)$ the amplitude of the Frequency Reassigned DFT, $\hat{R}(k)$ the autocorrelation of $X(k)$, $\hat{R}_{re}(k)$ the autocorrelation of $X_{re}(k)$, $\hat{r}(l)$ the autocorrelation function of the signal and $\hat{c}(l)$ the real-cepstrum function of the signal.

For the signal of Fig. 1, it is easy to see that only for $\tau = T_0$ we have simultaneously a maximum of the projection of the power spectrum on $g_\tau(f)$ and an energy peak in the power spectrum at $f = 1/\tau$. The same is true when using $X_{re}(k)$, $\hat{R}(k)$ or $\hat{R}_{re}(k)$ instead of $X(k)$ and when using $\hat{c}(l)$ instead of $\hat{r}(l)$. We therefore note $S(k)$ either $X(k)$, $X_{re}(k)$, $\hat{R}(k)$ or $\hat{R}_{re}(k)$ and $t(l)$ either $\hat{r}(l)$ or $\hat{c}(l)$.

The proposed periodicity representation takes advantage of the inverse octave errors of $S(k)$ and $t(l)$ by combining the two functions into a single function, therefore reducing the octave ambiguity. The combination is done by computing the product of $S(k)$ with a frequency-mapped version of $t(l)$.

The computation is done as follows:

1. $S(k)$ and $t(l)$ are computed for the same signal frame with a window of size N.
2. The value of $t(l)$ at lag l is considered as a measure of the periodicity at the frequency $f_l = \text{sr}/l$. Each lag $l > 0$ can be mapped to the frequency domain. We call Frequency-Mapped the function $t(l)$ mapped to the frequency domain.
3. We would like $t(l)$ to measure the periodicities at the same frequencies f_k as $S(k)$. We therefore interpolate the values of $t(l)$ at the positions $l = \text{sr}/f_k = N/k$. The frequency-mapped function is noted $t_S(k)$. One easily see that, since $t(l)$ has a constant temporal resolution, $t_S(k)$ has a decreasing frequency resolution. A solution would be to use a variable scale transform such as a constant-Q transform ([2]) instead of a FFT/IFFT algorithm.
4. The proposed periodicity function is obtained by computing the product of $S(k)$ and $t_S(k)$. Depending on the choice of $S(k)$ and $t(l)$ eight different functions can be estimated. We consider the five following:
   - $Y_{X\hat{r}}(k) = X(k) \cdot \hat{r}_S(k)$,
   - $Y_{X\hat{c}}(k) = X(k) \cdot \hat{c}_S(k)$,
   - $Y_{\hat{R}\hat{r}}(k) = \hat{R}(k) \cdot \hat{r}_S(k)$,
   - $Y_{\hat{R}\hat{c}}(k) = \hat{R}(k) \cdot \hat{c}_S(k)$.
   - $Y_{\hat{R}_{re}\hat{c}}(k) = \hat{R}_{re}(k) \cdot \hat{c}_S(k)$.

[1] sr stands for sampling rate.

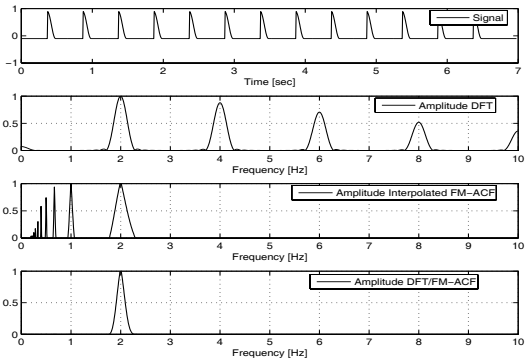

Fig. 2. From top to bottom [1] Signal [2] Amplitude of the DFT [3] ACF function mapped to the frequency domain [4] $Y_{X\hat{r}}(k)$.

$Y(k)$ constitutes a periodicity measure at the frequency k. Because of the use of the two transforms, $Y(k)$ is less sensitive to octave errors than any of the above mentioned representations. Of course autocorrelation and spectral methods are identical (by doing a matrix multiply) but the non-linearities of the frequency mapping and product function matter. In Fig. 2 we illustrate the computation of the product function $Y_{X\hat{r}}(k)$ with a simple example (a periodic impulse signal at 2Hz) . In this case, the sub-harmonic of the ACF and over-harmonic of the DFT cancel each other. Only one peak remains at the 2Hz periodicity frequency.

4. APPLICATIONS

We present here two applications of the periodicity function $Y(k)$: pitch estimation and visualization of multiple pitch signals.

4.1. Pitch estimation

The objective is not to propose yet another pitch estimation method but rather to test if the value of $Y(k)$ which corresponds to the pitch frequency can be sufficiently discriminated from the other frequencies so that just taking the frequency corresponding to the maximum value of $Y(k)$ provides a good approximation of the exact pitch. If it is the case this approximation could be used as an initialization (bootstrap) for other pitch estimation algorithms.

We've set up a large test set of 5371 musical instrument sounds coming from various databases: Studio-On-Line Ircam database, Iowa university database, McGill university database, a Microsoft database, and two private databases vi and pro. Each sound contains a single constant pitch note. The pitches range from 27.5 Hz (A0) to 7900 Hz (B8). 27 instruments are represented from piano (slightly inharmonic sound), pizzicato string (fast damping), bowed strings, brasses, single reeds, double reeds instruments, ... For each sound, a straightforward method is used to derive the pitch. We compute $Y(k,t)$ on a frame basis. In order to be able to discriminate adjacent pitches in the lower part of the spectrum, we have used a large window size of 250ms (with a 66% overlap between frame). We then compute the energy weighted time average of $Y(k,t)$. Finally, the pitch is chosen as the frequency corresponding to the maximum of $Y(k)$. Five versions of $Y(k)$ are tested: $Y_{X\hat{r}}(k)$ (DFT/ ACF), $Y_{X\hat{c}}(k)$ (DFT/ CEP), $Y_{\hat{R}\hat{r}}(k)$ (ACFofDFT/ ACF), $Y_{\hat{R}\hat{c}}(k)$) (ACFofDFT/ CEP) and $Y_{\hat{R}_{re}\hat{c}}(k)$ (ACFofREAS/ CEP). For comparison, we have also tested the performance of the well-known Yin algorithm [7] (reference code given by the author) with parameters set

	accuracy 1	accuracy 2
DFT / ACF	81,6	91,7
DFT / CEP	91,4	95,8
ACFofDFT / ACF	95	96,1
ACFofDFT / CEP	**97**	**97,6**
ACFofREAS / CEP	97	97,3
Yin	94,9	95,5

Table 1. Recognition rate of pitch estimation for the various algorithms

to the minimum frequency 27.5 Hz and maximum frequency 7900 Hz. The process described above is also applied here to derive a single pitch estimate for each sound. Two performance measures are used: *accuracy 1*: the recognition rate of the correct note, *accuracy 2*: the recognition rate of the correct chroma (note position inside the octave).

The results are indicated in Table 1. The best results are obtained with $Y_{\hat{R}\hat{c}}(k)$ (ACFofDFT/CEP): 97% for accuracy1, 97.6% for accuracy2. This is not surprising since $Y_{\hat{R}\hat{c}}(k)$ benefits from the use of the autocorrelation of the DFT (allowing to solve the missing fundamental problem), and from the real-cepstrum (allowing to reduce the spectral envelope influence). The small difference between accuracy 1 and 2 indicates that octave errors remain low. The use of the Frequency Reassigned DFT (ACFofREAS/CEP) does not improve the performances. For comparison, the results obtained with the reference Yin algorithm are lower: 94.9% for accuracy1, 95.5% for accuracy2. In Fig. 3, we indicate the detailed analysis of accuracy 1/2 by pitch and instrument class for the $Y_{\hat{R}\hat{c}}(k)$ case. Most errors occurred for the piano (inharmonicity), pizzicato strings (fast damping) and recorder/piccolo (single component spectrum not appropriate for $\hat{R}(k)$)[2]. The recognition rate is also lower for the pitch extremes (very low-pitch and very high-pitch).

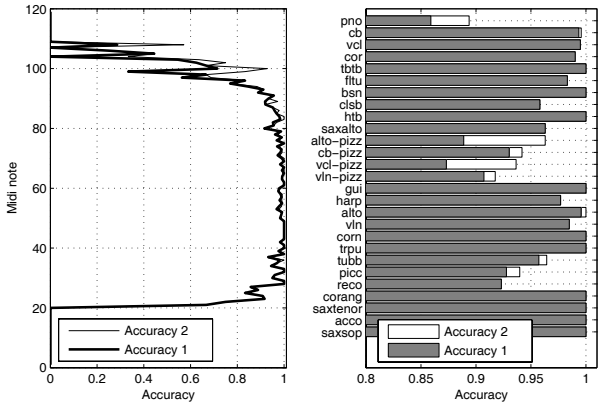

Fig. 3. Accuracy 1/2 using ACFofDFT/CEP algorithm [left] by pitch class (in midi note number) [right] by instrument class

4.2. Multiple pitch visualization

$Y(k)$ can be used efficiently for visual representation of multiple pitch signals. For multiple pitch signals, $Y_{X\hat{c}}(k)$ is preferred over $Y_{\hat{R}\hat{c}}(k)$, since, as mentioned before, autocorrelation of the spectrum

[2]piano=pno, pizzicato strings=vln-pizz/ vcl-pizz/ cb-pizz/ alto-pizz, reco=recorder, piccolo=picc.

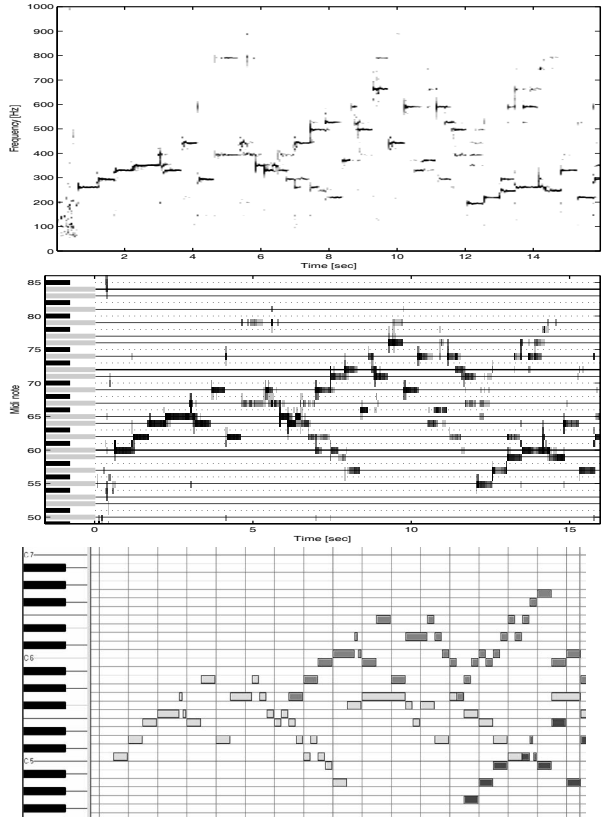

Fig. 4. [top] $Y_{X_{re}\hat{c}}(k,t)$ over time represented in a spectrogram way [middle] mapping of $Y_{X_{re}\hat{c}}(k,t)$ frequencies to midi notes [bottom] manual midi note transcription on J. S. BACH, Well-Tempered Clavier I, 02 Fugue in CM, window size=100 ms.

does not provide meaningful information in case of multiple pitch. For time-varying signals, a shorter analysis window is used: 100 ms. In order to increase the spectral resolution $X_{re}(k)$ is used instead of $X(k)$. In Fig. 4 upper part, we draw the values of $Y_{X_{re}\hat{c}}(k,t)$ over time in a spectrogram way. The signal corresponds to the first 16s of the Well-Tempered Clavier I, 02 Fugue in CM from J. S. BACH. The middle part represents a mapping of $Y_{X_{re}\hat{c}}(k,t)$ frequencies to the logarithmic midi note scale. It is represented as a piano-roll (X-axis = times, Y-axis = midi note). For comparison, the lower part represents the manual midi note transcription. Except an octave error present around time 5s, $Y_{X_{re}\hat{c}}(k,t)$ allows to visualize clearly the various multiple pitch of the piano performance. Although deep tests remain to be done, informal tests allow to believe that $Y_{X_{re}\hat{c}}(k,t)$ could be used as an efficient initialization (bootstrap) for multiple pitch estimation algorithms.

4.3. Tempo estimation

In [15], we have used $Y_{X\hat{r}}(k)$ to estimate the periodicities related to the rhythm of an audio signal . In this case three periodicities were to be estimated: the *measure* periodicity (lowest frequency), the *beat* periodicity (middle frequency) and the *tatum* periodicity (smaller time unit subdivision of the score, highest frequency). An onset-energy function is first extracted from the audio signal. $Y_{X\hat{r}}(k)$ is then used to emphasize the three main frequencies of this function.

5. CONCLUSION

In this paper, we've proposed a straightforward approach for periodicity estimation of musical audio signals based on the combination of a temporal representation and a spectral representation. Various time and spectral representations have been considered. For single pitch estimation, the best results were obtained when combining the autocorrelation of the amplitude of the DFT with a frequency-mapped real-cepstrum. For multiple pitch visualization, the best results were obtained when combining the amplitude of the frequency reassigned DFT with a frequency mapped real-cepstrum. Further works will concentrate on exploiting the capability of this method for initialization of pitch and multiple pitch estimation algorithms. It is worth to note that any two pitch estimation models with inverse octave errors could be combined in the way as proposed.

6. REFERENCES

[1] H. Vinet, "The semantic hifi project," in *ICMC*, Barcelona, Spain, 2005, pp. 503–506.

[2] J. Brown, "Calculation of a constant q spectral transform," *JASA*, vol. 89, no. 1, pp. 425–434, 1991.

[3] J. Brown and M. Puckette, "Calculation of a "narrowed" autocorrelation function," *JASA*, vol. 85, no. 4, pp. 1595–1601, 1989.

[4] P. Flandrin, *Time-Frequency/Time-Scale Analysis*, Academic Press, San Diego, California, 1999.

[5] B. Doval and X. Rodet, "Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and hmms," in *ICASSP*, Minneapolis, 1993, vol. 1, pp. 221–224.

[6] R. Maher and J. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way lismatch procedure," *JASA*, vol. 95, no. 4, pp. 2254–2263, 1994.

[7] A. de Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, 2002.

[8] A. Klapuri, *Automatic Transcription of Music*, Phd thesis, Tampere University of Technology, 1997.

[9] C. Yeh, A. Roebel, and X. Rodet, "Multiple fundamental frequency estimation of polyphonic music signals," in *ICASSP*, Philadelphia, PA, USA, 2005, vol. 3, pp. 225–228.

[10] M. Lahat, R. Niederjohn, and D. Krubsack, "A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech," *IEEE Trans. on Speech and Audio Processing*, vol. 35, no. 6, 1987.

[11] N. Kunieda, T. Shimamura, and J. Suzuki, "Robust method of measurement of fundamental frequency by aclos:autocorrelation of log spectrum," in *ICASSP*, Atlanta, GA, USA, 1996.

[12] L. Rabiner and B. Juang, *Fundamentals of speech recognition*, Prentice-Hall, New-York, 1993.

[13] A.V. Oppenheim and R.W. Schafer, *Discrete-time signal processing*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[14] A. Oppenheim, "A speech analysis-synthesis system based on homomorphic filtering," *JASA*, vol. 45, pp. 458–465, 1969.

[15] G. Peeters, "Time variable tempo detection and beat marking," in *ICMC*, Barcelona, Spain, 2005, pp. 539–542.