A STUDY OF TEMPO TRACKING ALGORITHMS FROM POLYPHONIC MUSIC SIGNALS

M. Alonso, B. David, G. Richard

École Nationale Supérieure des Télécommunications (ENST)
Dépt. Traitement du Signal et des Images
46, Rue Barrault
75634 Paris, Cedex 13
Email:miguel.alonso,bertrand.david,gael.richard@enst.fr

ABSTRACT

In this paper, algorithms for tempo tracking from polyphonic music signals are introduced. These new methods are based on the association of a filter bank with robust pitch detection algorithms such as the spectral sum or spectral product. These algorithms are further improved by using an onset detector in each band. These algorithms are then compared to two of the most reliable methods of the literature using a small manually annotated database of short sequences of music signals. It is shown that, despite their simplicity, these new approaches are very efficient and outperform the tested methods.

1. INTRODUCTION

The enormous amount of unstructured multimedia data available nowadays and the spread of its use as a data source in many applications are introducing new challenges to researchers in information and signal processing. The continuously growing amount of this digital multimedia information increases the difficulty of its access and management, thus hampering its practical usefulness. As a consequence, there is a clear need for content-based multimedia data indexing, processing and retrieval techniques.

If multimedia and multimodal approaches represent essential challenges, the more classical approach consisting in building new analysis or indexing technologies on a given media are still needed to overcome the current limitations of today approaches.

For example in audio, numerous problems still exist to extract high level descriptors directly from polyphonic music signals. The tempo (or beat) is one of the most important descriptor since many applications can be derived from the automatic recognition of the rhythmic structure of a music signal:

- automatic rhythmic alignement of multiple instruments, channels or musical pieces (for mixing or *karaoké*)
- automatic indexing, segmentation and style classification of music databases,
- beat driven computer graphics (virtual dancers, etc..)

Tempo or beat analysis of musical signals is a domain of research that receives a growing interest as shown by the variety of recent publications [8],[10],[1], [12], [6]. This problem is apparently simple (most people even without any musical knowledge have no difficulties to find the beat of a musical performance). However, automatic recognition is more complex especially for music styles that do not include strong rhythmic patterns (such as classical or jazz music, for example).

If earlier approaches focused on MIDI signals (or simple real audio signals such as purely percussive signals [9]), today approaches are directly dealing with polyphonic music. Scheirer [8] proposed a method associating a filterbank with a set of comb-filters. Simpler methods were introduced by by Seppännen [10] using sound onset detection or by Tzanetakis [12] in the context of musical genre classification. Another approach was also proposed by Goto [2] to infer the hierarchical beat structure. In most of these works, the rhythm detection (or periodicity) is based on a simple inter-onset time detection or on the traditional autocorrelation method.

In this paper, several algorithms for tempo tracking are introduced. These methods are based on the association of a filter bank with robust pitch detection algorithms such as the spectral sum or spectral products. The performances of these algorithms are evaluated against some of the most reliable algorithms of the literature using a small manually annotated database of short sequences of music signals.

The paper is organized as follows: the next section describes our new algorithms including some minor improvements of the original Scheirer's algorithm. Results of these algorithms compared to two methods of the literature are given in section 3. Finally, in section 4 we suggest some conclusions.

2. TEMPO ESTIMATION ALGORITHMS

Most of the algorithms designed to estimate the tempo of musical pieces [8, 7, 3, 12] are based on the same basic steps. In particular, they all process several frequency bands separately, combining the results in the end. According to the experimental results found in [8], this assumption preserves rhythm perception for most music signals, and have prove to be efficient for many kinds of frequency bands decomposition. These basic steps are described below:

1. subband decomposition of the signal, provided by a filterbank,
2. onset detection in each subband.
3. estimation of the periodicity in each subband.
4. combination of the results to obtain the general tempo.

The differences between the algorithms found in the litterature rely on the implementations of those steps. For instance, the well-established Scheirer algorithm uses a six band IIR filter bank for the first stage. Nevertheless we also found a eight band filter bank in [7] and a 21 nearly critical band filter bank in [4]. There are also different techniques used to detect the onset times: based on a half wave or full wave rectification, using the enveloppe or its squared value, deriving the difference function or the relative difference function, applying thresholding or not.

According to these descriptions, the structure of the whole system would appear as sketched on the flow diagram of the figure 1.

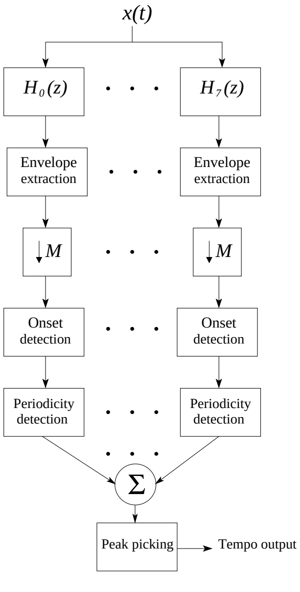

Fig. 1. Proposed tempo estimation system flow diagram.

Our tempo estimation approach uses the general psychoacoustic simplification proposed by [8] and also adopted by Paulus [7]. The input signal to the tempo estimation system is first divided into eight non-overlapping frequency bands using a filterbank of sixth-order butterworth filters. The lowest band is obtained by lowpass filtering with a cutoff frequency of 100 Hz, the seven higher bands are logarithmically distributed in frequency between 100 Hz and half the sampling frequency (8000 Hz for our experiments), as suggested by Paulus in [7]. We obtain the subband signals $x_k(t)$, where $k = 0, \dots, 7$.

Next, extraction of the signal's envelope is carried out. This is a fundamental step aiming at precisely finding the sound onset points provided to the tempo estimation algorithms. This task is accomplished using a system mostly based on the onset detector proposed by Klapuri in [3, 4], a flow diagram is presented in Fig. 2.

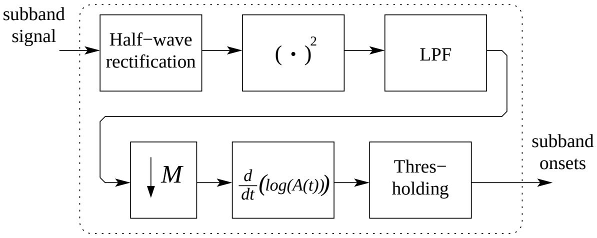

Fig. 2. Envelope extraction and onset detector flow diagram.

At each subband, the input signal is first half-wave rectified and squared. Then, amplitude envelopes, $A_k(t)$, at each frequency channel are calculated by convolving the subband signals with a 100 ms descending half-Hanning window (linear phase lowpass filter) and then the output of each band is decimated in order to reduce the computational burden of the following stages, the decimation factor being 16. For the bandwise onset detection we use the first order *relative difference function* $W_k(t)$, which gives the amount of change in the signal in relation to its absolute level. This is equivalent to differentiating the logarithm of the amplitude envelope, as given by Eq. (1).

$$W_n(t) = \frac{d}{dt}(\log(A_n(t))) \tag{1}$$

The relative difference function is a psychoacoustically relevant measure, since the perceived increase in signal amplitude is in relation to its level, the same amount of increase is more prominent in a quiet signal [3, 4]. Hence, we detect onset components by a peak picking operation, which looks for peaks above a given threshold. The threshold value was found experimentally to be around $1.5\sigma_{W_k}$, where σ_{W_k} stands for the standard deviation of the signal $W_k(t)$.

Until this point, two of the proposed tempo estimation systems (spectral and summary autocovariance function) follow the same principle. From here on they will be treated separately, thus one of them takes place in the frequency domain while the other in the time domain.

2.1. Spectral methods

Due to their strong relationship, two different spectral tempo estimation methods are presented: the *harmonic spectral* *sum* and the *harmonic spectral product*. Both of these methods come from traditional pitch determination techniques. At the output of the onset detection block, subband signals have the appearance of a quasi-periodic train pulse. In order to find the bandwise fundamental frequency of such train pulses, the Fourier transform of the subband signals, $X_k(e^{j\omega_n})$, is calculated. Prior to the FFT calculation, subband signals are zero padded to have a size l_x given by:

$$l_x = 2^{\lfloor \log_2(\text{length}(x_k(t))) \rfloor + 2} \qquad (2)$$

where $\lfloor \cdot \rfloor$ stands for *the integer part of*.

2.1.1. Spectral Sum

The spectral sum is a reliable pitch determination technique. It's principle lies on the assumption that the power spectrum of the signal is formed of strong harmonics located at integer multiples of the signal's fundamental frequency. For the purpose of finding this frequency, the power spectrum is compressed by a factor l, then the obtained spectra are added. In normalized frequency, this is indicated by Eq. (3).

$$S_k(e^{j\omega_n}) = \sum_{l=1}^{M} |X_k(e^{jl\omega_n})|^2 \quad \text{for} \quad \omega_n < \frac{\pi}{M} \qquad (3)$$

Consequently, the signal's fundamental frequency is strongly reinforced. In addition, all subband spectral sums are added together, this strengthens even more the fundamental frequency which is shown in the form of an easily detectable prominent peak, as depicted in Fig. 3 for a particular music signal. There, we can clearly see the most salient peak located roughly at a frequency of 1.05 Hz, which corresponds to a beat rate of 63 Beat Per Minute (BPM).

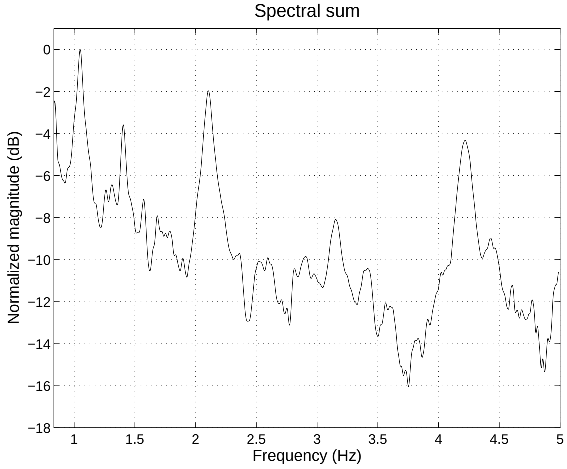

Fig. 3. Spectral sum of a music signal.

In our implementation M was set to 6 and the fundamental frequency search was carried out in the interval ranging from $5/6$ to 5 Hz, which corresponds to a beat rate between 50 and 300 BPM.

2.1.2. Spectral Product

As briefly mentioned, another method for pitch estimation closely related to the spectral sum is the spectral product, in normalized frequency it is defined by Eq. (4).

$$S_k(e^{j\omega_n}) = \prod_{l=1}^{M} |X_k(e^{jl\omega_n})|^2 \quad \text{for} \quad \omega_n < \frac{\pi}{M} \qquad (4)$$

In a similar way to the preceeding method, for the spectral product implementation M was set to 6 and the fundamental frequency search was performed within the same frequency interval. Fig. (4) depicts the result for the aforesaid signal. Once again, we can see the most salient peak located at about the same frequency of 1.05 Hz. Note, however, that the spectral product method shows a much higher prominent peak relatively to the other secondary peaks compared to the spectral sum method.

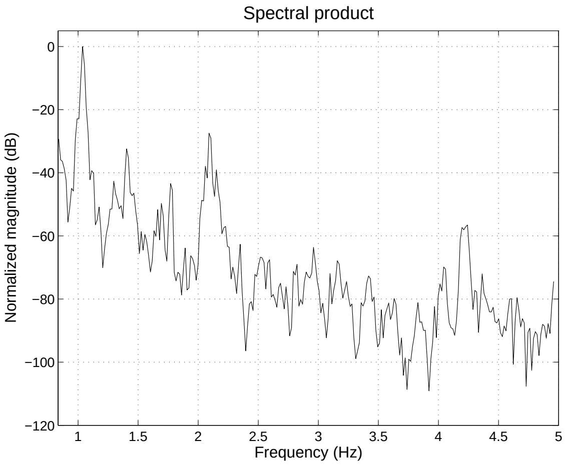

Fig. 4. Spectral product of a music signal.

2.2. Summary Autocovariance Function

The conception of this method was suggested by the work done in the multipitch detection field by [11]. The bandwise train pulse like signals at the output of the sound onset detector are convolved with a 100 ms odd length Hanning window. Then, the autocovariance function, $\Gamma_k(\tau)$, of the subband signals is computed, as given by Eq. (5).

$$\Gamma_k(\tau) = \sum_t [x_k(t+\tau) - \overline{x}_k][x_k(t) - \overline{x}_k] \qquad (5)$$

Then, the bandwise autocovariance functions are added together to form the *summary autocovariance function*, SACVF(τ), obtained by Eq. (6).

$$\text{SACVF}(\tau) = \sum_{k=0}^{7} \Gamma_k(\tau) \qquad (6)$$

As in the spectral methods case, we're only interested in detecting the periodicities of the subband signals $x_k(t)$ who correspond to the most salient peaks in SACVF(τ). In addition, we are only concerned about finding a beat rate between 50 and 300 BPM. Thus, the time lag τ varies only within the range of 200 to 1250 ms. For our SACVF implementation this result is shown in Fig. (5), where the foresaid signal has a dominant periodicity, depicted by the most salient peak, at lag of approximately 950 ms, which nearly corresponds to a fundamental frequency of 1.05 Hz.

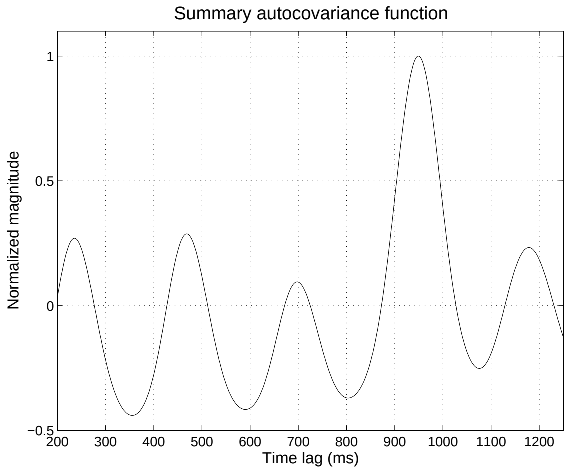

Fig. 5. Summary autocovariance funcion of a music signal.

Finally, to ensure a better tempo extraction, the three most salient peaks are detected and a relation of multiplicity between them is searched via a simple numerical algorithm. For instance, in Fig. (5) the first, second and fourth peaks are the most prominent and they bear a strong relationship between them. The second peak is located at lag practically twice that of first one and the fourth peak is located at a lag nearly four times that of the first one. This allows to have a more robust decision. If no relation of multiplicity is found among the detected peaks, the most salient one is taken as the right tempo.

2.3. Bank of resonators

An alternate method dedicated to the determination of the tempo is given in [8] and [5]. In order to estimate the onset periodicity (pulse) in each subband, a so-called bank of resonators is used. These resonators are simply oversampled versions of autoregressive filters of order 1. For instance, the k$^{\mathrm{th}}$ filter has a z-transfer function of the kind:

$$H(z) = \frac{\beta_k}{1 - \alpha_k z^{-T_k}} \qquad (7)$$

The principle of the method is the following. The impulse response of this filter is zero unless for the time indices multiples of the oversampling factor T_k. When it receives a periodic pulse train of period T_k, the output samples cumulate and the output level is increased.

The factors T_k are set as integer numbers of samples in order to cover the whole range of the analysed tempi, from T_1 to T_M. According to the principle described above, the β_k and α_k coefficients are chosen constant for all filters in order to be able to compare the non-zero outputs. For a T_k-periodic pulse train, the transient time of the k$^{\mathrm{th}}$ filter is thus of the order of

$$\tau_k = -\frac{3T_k}{\ln \alpha}$$

In our implementation, α is set to ensure the maximum transient time τ_M to be far less than the length of the analysis window. The tempo is determined following the main steps:

1. computation of the responses $y_k(t)$ of each filter to the centered amplitude envelopes,
2. estimation the mean power σ_k^2 of y_k,
3. extraction of the time indices $t_i, i = 1, \ldots, N_k$ such as $y_k(t_i) > 3\sigma_k$, and computation of the mean power $P_k = 1/N_k \sum_{i=1}^{N_k} y_k(t_i)^2$,
4. Extraction of the tempo corresponding to the factor T_u with $u = \arg_k \max P_k$

3. SIMULATION RESULTS

3.1. Sound database

The database used for evaluation is constituted of 55 short segments of musical signals (each of 10 seconds long). The short musical excerpts were chosen in order to represent different styles : Classical music (23 % of the database), Rock or modern pop music (33 %), traditional songs (12 %), Latin/cuban music (12%) and jazz (20 %). All signals are sampled at 16kHz. This sound database has been manually annotated by skilled musicians. The procedure for manually estimating the tempo is the following:

- the musician listens to a musical segment using headphones (sometimes several times in a row to be accustomed with the tempi),
- while listening, he/she taps the tempo,
- the tapping signal is recorded and tempo is automatically extracted from it,
- all tempo are finally manually checked, however due to the impulsiveness nature of the tapping signals, no errors was found after the automatic extraction.

Method	Pourcentage of correct estimation
Scheirer	76
Scheirer Modified	85
Autocovariance	87
Paulus	74
Spectral Sum	87
Spectral prod	89

Table 1. Performances obtained for several tempo tracking algorithms

3.2. *Results*

This section gives the results of several algorithms on the database described in the previous section. Despite the limited size of our database, this test gives good indication of the performances of each algorithm. The estimation provided by an algorithm is labelled as correct when it differs from less than 5% from the original tempo without counting errors of doubling or halving. An estimated tempo T_e is then labelled correct if:

$$0.95\ \alpha T < T_e < 1.05\ \alpha T \tag{8}$$

where $\alpha = 0.5$ or $\alpha = 2$, and where T is the valid (manually measures) tempo.

The table 1 gives the results obtained for five algorithm: the original Scheirer algorithm [8], a modified version of it taking into account a new resonator bank, (see section 2.3), the approach introduced by Paulus [7], the autocovariance method, the spectral sum and the spectral product methods.

If there is no significant differences between the four best methods, the approaches introduced in this paper outperform the classical approach of Scheirer and the recently introduced method by Paulus. It is though important to notice that the database used is small and that it will be important to conduct complementary tests on an extended database to confirm these initial results.

4. CONCLUSION

This paper has proposed several new algorithms for tempo tracking and has compared them to a number of methods described in the literature. Although the dataset used for evaluation is rather limited, it is seen that the methods introduced are very accurate. Future work will include an evaluation of these algorithms on a larger dataset, the extension of our best algorithm for real-time tracking of tempo and its adaptation to small size analysis windows. Finally retrieval experiments will be conducted based solely on the rhythmic pattern.

5. REFERENCES

[1] S.E. Dixon. A beat tracking system for audio signals. *Austrian Research Institute for Artificial Intelligence. Vienne.*, pages 311–320, Avr. 2000.

[2] M. Goto and Y. Muraoka. An audiobased realtime beat tracking system and its applications. *Proceedings of the International Computer Music Conference*, 1998.

[3] Anssi Klapuri. Automatic transcription of music. Master's thesis, Department of Information Technology, Tampere University of Technology, 1998.

[4] Anssi Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pages pp. 3089–3092, 1999.

[5] Edward W. Large and John F. Kolen. Resonance and the perception of musical metter. *Connection Science*, Vol. 6:pp. 177–208, 1994.

[6] J. Laroche. Estimating tempo, swing and beat locations in audio recordings. *Proc. Int. Workshop on applications of Signal Processing to Audio and Acoustics*, WAASPA:pp. 131–135, 2001.

[7] Jouni Paulus and Anssi Klapuri. Measuring the similarity of rhythmic patterns. In *3rd. International Conference on Music Information Retrieval (ISMIR)*, 2002.

[8] Eric D. Scheirer. Tempo and beat analysis of acoustic music signals. *J. Acoust. Soc. Am.*, Vol. 103(1):pp. 588–601, Jan. 1998.

[9] A. Schloss. On the automatic transcription of percussive music, 1985.

[10] Jarno Seppänen. Tatum grids analysis of musical signals. *New Paltz, New York*, pages 21–24, Oct. 2001.

[11] Tero Tolonen and Matti Karjalainen. A computationally efficient multipitch analysis model. *IEEE Trans. Speech Audio Processing*, Vol. 8(6):pp. 708–716, Nov. 2000.

[12] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Processing*, Vol. 10(5):pp. 293–301, Jul. 2002.