ESTIMATING TEMPO, SWING AND BEAT LOCATIONS IN AUDIO RECORDINGS

Jean Laroche jeanl@creative-atc.com

Creative Advanced Technology Center
1500 Green Hills Road
Scotts Valley, CA95067

ABSTRACT

The problem of estimating the tempo of audio recordings (the number of beats per minute, or BPM) has received an increasing amount of attention in the past few years. Applications include the synchronization of multiple audio tracks for simultaneous playback, "tempo-synchronous" audio effects, automatic looping of audio tracks... This article presents techniques for estimating the tempo and the swing, and locating the beats in audio recordings, under the assumption that the tempo is constant. The techniques rely on a preliminary transient detection stage where note onsets/offsets, percussion hits and other time-localized events are detected. This first step is followed by a maximum likelihood estimation of the tempo, swing and downbeat. Suggestions are given to minimize the computation load of the methods.

1. INTRODUCTION

Estimating the tempo of a musical piece is a complex problem that has received an increasing amount of attention in the past few years. The problems consists of estimating the number of beats per minutes or BPM, at which the music is played, and possibly to identify exactly when these beats occur. Commercial devices already exist, that attempt to extract a "MIDI[1] clock" from an audio signal, indicating both the tempo and the actual location of the beat. The MIDI clock can then be used to synchronize other devices, such as a drum machine, to the audio source. The systems commercially available tend to be fairly unsophisticated, as they seem to rely mostly on the presence of a strong and regular bass-drum kick at every beat, an assumption that tends to hold with modern musical genres such as "techno" or "drums and bass". For music with a less "pronounced" tempo, such techniques fail miserably and more sophisticated algorithms are needed. As mentioned above, tempo estimation and beat detection can be used to synchronize MIDI instruments or effects to audio signals. Beat detection can also help in the usually tedious process of manipulating audio material in audio editing software: Cut and paste operations are made considerably easier if "markers" are positioned at the each beat or at bar boundaries. Looping a drum track over two bars becomes trivial once the location of the beats is known. Finally, tempo estimation can be a very useful tool for music education. The ability to obtain a precise estimate of the time-evolution of the tempo from a musical performance (be it recorded or live) can help a musicologist in his/her analysis, or enable a performer to fine tune his art.

A very comprehensive bibliography on tempo, beat or rhythm estimation is given by Scheirer in [1]. In this paper, we investigate the much simpler case where the tempo of the audio track can be assumed constant. The techniques described in this paper rely on a preliminary transient analysis stage in which significant transients (percussion hits, note onsets etc) are detected in the audio track. The actual tempo, swing and beat estimation stage makes use of the information obtained in the first stage, and is based on a maximum likelihood approach, using the transient times, or the inter-transient elapsed times as observations.

[1] MIDI: Musical Instrument Digital Interface.

2. ALGORITHM

2.1. Transient analysis

For simplicity, we use a rather loose definition of what transients are. During the transient analysis stage, we attempt to detect times at which the energy of the signal in some frequency band increases sharply. This could be caused for example by a percussion hit, or by a note onset, or by a rapid spectral change in the signal. Since percussion hits, and note changes tend to occur on the beat, or on a subdivision of the beat in a wide variety of musical genres, it is reasonable to assume that the tempo information can still be extracted from the mere knowledge of transient times and amplitudes. In fact, an informal test of this hypothesis consists of replacing the original audio by a series of sharp clicks whose time-locations and amplitudes correspond to the transients detected in the audio. Interestingly, it is still possible for a human to correctly tap his/her foot to this click track, in spite of the drastic reduction of the information available to the listener.

The transient detection algorithm described in this paper relies on a fairly standard approach similar to that described in [1] (even though transients are not actually extracted in that paper). The idea consists of running a short-time Fourier transform on the original signal $x(n)$,

$$X(f,t_i) = \sum_{n=0}^{N-1} h(n)x(n+t_i)e^{-2j\pi fn}$$

end expressing the spectral energy $|X(f,t_i)|$ in a non-linear scale (e.g. in dB). $h(n)$ is an analysis window (e.g., a Hanning window), and t_i denotes the time at the short-time Fourier transform frame i. The spectral energy is then summed over a predefined frequency range, yielding some sort of a time-varying "energy curve" $E(t_i)$.

$$E(t_i) = \int_{f_{\min}}^{f_{\max}} C(|X(f,t_i)|)df$$

Here, $C(z)$ denotes the non-linear compressed scale, and the integration range is the interval $\{f_{\min} \;\; f_{\max}\}$. In actual implementations, the Fourier transform will most likely be a discrete Fourier transform and the integral above will become a sum over the DFT bins. In practice, one can use $C(z) = \log(z)$ for a dB-like scale, but problems arise for values of z close to 0. Another possibility is to use $C(z) = \sqrt{z}$ which compresses the dynamic range but behaves nicely[2] around 0. A rectified first-order difference is then calculated on the energy curve, resulting in a signal $\Delta E(t_i)$ exhibiting large positive peaks at transient times.

$$\Delta E(t_i) = E(t_i) - E(t_{i-1})$$

A fairly simple peak extraction stage follows, in which the most significant peaks are selected. The peak times and amplitudes are saved as a series of transient times t_i, to be later used in the tempo and beat detection stage. The result of such an analysis is presented in Fig. 1 for a percussion track. The detector picks up the most visible transients but also less "visible" ones, such as the fourth one from the left, which corresponds to a high-frequency hit.

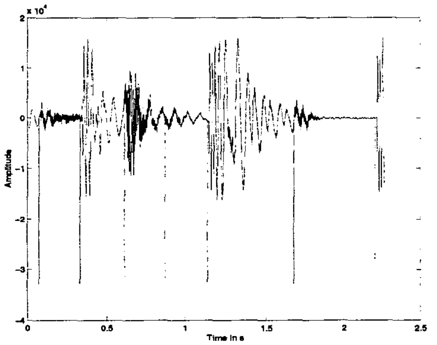

Figure 1: Example of transient detection for a percussion track. The vertical lines on top of the original signal denote the detected transient times.

2.2. Defining beat, tempo and swing

We will use the following (fuzzy) definitions: beats are time-instants distributed somewhat regularly along the time-axis. Beats are placed in such a way that most of the musical events (note onsets, transients) fall on beats or on subdivisions of the beat (e.g. half or quarter beats). The tempo, measured in BPM (beats per minutes) counts the number of beats per minute. Our definition of the beat is not precise enough that tempo can be measure unambiguously: if a track has a tempo of 60 beats per minute, one could also say the tempo is 120 beats per minutes if half-beats were counted as beats (which our definition does not preclude). To avoid this problem, we will constrain the tempo to lie between, say, 70 and 140. Swing is a musical attribute that is commonly found in jazz, but also in rock music. In this paper, swing is defined as a slight delay of the second and fourth quarter-beats (see Fig.2), measured as a percent of the quarter-beat duration. Including swing in the tempo estimation will prove to be important if the audio track has a strong swing (failing to do so can lead to poor tempo or beat estimations).

2.3. Writing the likelihood

The hypothesis that the tempo is constant over the duration of the audio track greatly simplifies the estimating task, because only three parameters must be estimated: the tempo in beats per minute, the swing in percent and the location of the first beat. To estimate these three parameters, we use a maximum likelihood technique [2]. First, given a tempo T in beats per minutes, the location of the first beat b_0, the value of the swing S in percent and the duration of the track, we need to define a probability to observe a transient at time t_i. We do this in a fairly ad-hoc manner. The tempo, swing and first beat define a series of quarter-beat times b_i at which there is a maximum probability to observe a transient. Because the tempo is constant, the transient times can be expressed modulo P (the beat period), see Fig. 3. One could subdivide the beat into two half-beats, or 8 eighth-beats, but half-beats are too coarse (many musical events occur on quarter beats for BPMs between 70 and 140) and eighth-beats are too fine (fewer relevant musical events occur on eighth-beats). Note that one can also subdivide the beat into 3 third-beats, which would be required for certain time-signatures such as 6/8. This choice does not affect the rest of the discussion. A very simple-minded PDF is given by:

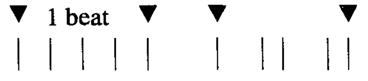

Figure 2: Left: one beat is subdivided into four equal quarter-beats. Right: swing measures the delay in quarter-beats 2 and 4.

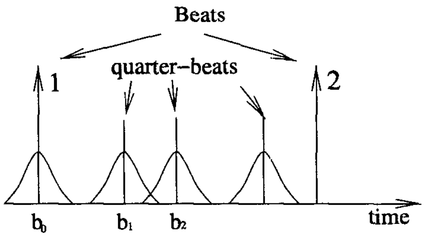

Figure 3: Simple probability density function for observing a transient at time t modulo the beat period P.

$$p_t(t) = \frac{1}{4} \sum_{i=0}^{3} G([t]_T - b_i)$$

where $[t]_T$ represents t modulo T, b_i represents the four quarter-beats and $G(x)$ is a zero-mean, symmetrical template PDF (e.g. gaussian). The dependence on tempo T, swing S and first beat b_0 is through the location of the subbeats b_i. This simple PDF expresses the fact that transients should occur most likely around subbeat times. There is a problem, however, because the PDF is periodic in t with a period of one half-beat (if there is swing) or one quarter beat (if there is no swing), which means we will only be able to determine the location of the beat up to a half or quarter beat. To remove this ambiguity, we can arbitrarily introduce an asymmetry between the four quarter-beats: we can make it more likely for a transient to occur on the first quarter-beat, a bit less likely on the third quarter-beat, and even less likely on the second and fourth quarter-beats. The assumption is that transients will tend to occur on the first quarter-beat (the "bottom of the beat" in musical terms), or on the third one (the "up-beat"), but will less often occur on the other quarter-beats. To demonstrate that this assumption is somewhat founded, we calculated a histogram of transient locations for a techno audio track of known tempo and beat location. Fig. 4 shows that our assumption is valid, at least for

[2] Thanks to Miller Puckette for this trick.

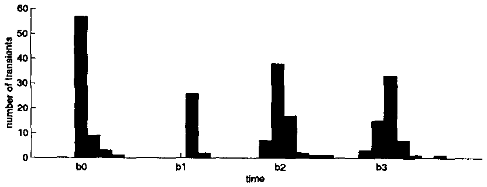

Figure 4: Histogram of transient location relative to the quarter-beats for a real audio track. Notice the slight swing delaying the second and fourth quarter-beats.

this track. We will see that the assumption is false for other musical genres. The modified PDF, given in Fig. 5 can be expressed as

$$p_t(t) = \sum_{i=0}^{3} p_i G([t]_T - b_i)$$

where p_i expresses the overall probability of a transient falling around subbeat i, with $\sum p_i = 1$. In addition to this PDF, we assume that transients times are independent from each other, such that the probability of observing N transients at times t_i is simply the products of $p_t(t_i)$. This assumption is difficult to test in practice, but is a reasonable one.

The maximum-likelihood estimate of the tempo T, the swing S and the beat b_0 is simply obtained by maximizing the log-likelihood $L_t(T, S, b_0)$ with respect to T, S and b_0, given the N observed transient times t_i:

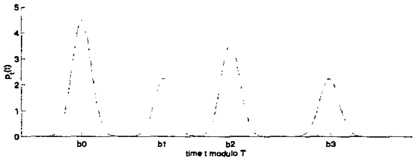

Figure 5: A better probability density function with preference on quarter-beat 1 (b_0) and 3 (b_2).

$$L_t(T, S, b_0) = \log\left(\prod_{i=0}^{N-1} p_t(t_i)\right) = \sum_{i=0}^{N-1} \log p_t(t_i) \quad (1)$$

Unfortunately, this cannot be done analytically, because the dependence of L on T, S and b_0 is too complicated. A number of non-linear optimization techniques could be used to maximize $L_t(T, S, b_0)$ but the likelihood exhibits a large number of local maxima, a characteristic that renders non-exhaustive optimization techniques unreliable. Because the search space is somewhat limited, we can afford an exhaustive search, provided our unknowns are discretized appropriately. For that purpose, we will test combinations of T, S and b_0 taken from sets of discrete values T^k, S^l and b_0^m, calculate the corresponding likelihood $L_t(T^k, S^l, b_0^m)$ and look for the triplet (k, l, m) that maximizes L. We now investigate the practical details of the algorithm implementation.

2.4. Practical details

Selecting the pdf: First, we need to select $G(x)$ which represents the probability that a transient will occur at time x for a subbeat positioned at time $t = 0$ and the weight factors p_i. The best solution for that would be do derive the PDF from actual observations, i.e., transient locations for tracks whose tempo, swing and beat location are known. This is a chicken/egg problem though, because determining the tempo, swing and beat location by hand is not easy, so we revert to an ad-hoc choice, with the possibility to adjust the PDF once the proper tempo has been estimated. It was found in practice that the type of PDF did not matter much as long as it was symmetrical and localized around $t = 0$. A Gaussian distribution is an appropriate choice, but the variance must be adjusted as a function of the tempo T so the PDF does not "spill" into adjacent quarter-beats, but does not become too small too fast either. A PDF that is too narrow would make it extremely unlikely for a transient to occur anywhere else than very near a quarter-beat, potentially giving too much weight to mis-estimated transient locations. In practice, if Δt is the time interval between two quarter-beats, $\sigma^2 = 0.05\Delta t$ is an appropriate value. In addition, we set $p_0 = .4, p_2 = .3, p_1 = p_3 = .15$.

Discretizing the variables: Values of swing can be discretized in 10% increments from 0 to 40%. If a more accurate estimate of the swing must be obtained, the coarse estimate can be refined in a subsequent stage. For a given tempo, the location of the first beat b_0 can be constrained to lie between time $t = 0$ and $t = P$ where P is the beat period. In practice, it is sufficient to divide this period into 32 evenly spaced test beat locations $b_0^l = Pl/32$. Discretizing the tempo T is more tricky, because a small delta in the tempo translates into a large variation of transient locations modulo P $[t_i]_P$ for the transients lcoated at the end of the track. Even though we might only be interested in knowing T with a precision of $\pm.5$BPM, we must search through a much finer grid to make sure the beats have a chance to align with the true beats in the track. Specifically, denoting D the duration of the track, there are approximately D/P beats in the track, and if the first beat is fixed, the last $[t_i]_P$ will vary by $\Delta[t_i]_P = D\Delta P/P$ if the beat period P is varied by ΔP. To keep $\Delta[t_i]_P$ smaller than about one eighth of the beat period requires that $\Delta P < \frac{P^2}{8D}$ or $\Delta T < \frac{60}{8D}$. For a 1mn audio track, this means that the tempo must be searched every $1/8$BPM, a very fine grid resulting in an expensive search. In the next section, we will see how this problem can be mitigated.

Doing the search: Now that the discrete search space is defined, the likelihood L is calculated for every possible triplet (T^k, S^l, b_0^m) according to Eq. 1, and the triplet that maximize L yields the maximum likelihood estimate of the tempo, the swing and the first beat.

2.5. Speeding things up

The algorithm above can be fairly expensive, especially for long tracks where the tempo must be searched on a very fine grid. For example, for a 4mn track, we end up with about 6 values for the swing, 32 values for the beat location, and $70/(1/8.4) = 2240$ values for the tempo. The search requires estimating L over 430000 times, and the cost increases linearly with the track duration. This suggests a very simple manner in which the search can be sped up.

Using a small audio segment: Since the tempo is assumed to be constant, we can run the search on a small portion of the track, (say 5 to 15 seconds), get a first estimate of the tempo, beat location and swing, then refine these values by doing a second search on the whole track, with a smaller search space centered around the rough estimates. The first search will be fast because the audio segment is of small duration, and the second will also be fast because the search range has been considerably restricted. In the same vein, we can also do a coarser search on the swing and beat location, then refine in a subsequent stage.

Sampling the likelihood: Rather than calculating the likelihood according to Eq. 1 for every transient t_i, it is possible to precompute a sampled version of it, for each swing value, and store it in a multidimensional table. Estimating the likelihood is simply done by a table-lookup, much cheaper than a direct computation.

Using inter-transient times: Another way to speed up the algorithm consists of eliminating one search variable altogether. The beat position b_0 can be eliminated if we use the time separating transients rather than the transient times themselves. In other word, if instead of using t_i we use $e_i = t_i - t_{i-1}$, and come up with a probability density function for it, this PDF will be independent from the position of the first beat, and will only be a function of the tempo T and the swing S, yielding a two-dimensional search rather than a 3-D one. In the example above where 32 discrete values of b_0 were used, the search would be 32 times faster. If we have a PDF for the transient times t_i $p_t(t)$ and if they are supposed to be independent, then it is a standard result [3] that the PDF $p_e(e)$ for e_i is obtained by the following convolution:

$$p_e(e) = p_t(e) \star p_t(-e) = \int_{-\infty}^{\infty} p_t(\tau) p_t(\tau - e) d\tau$$

Fig.6 presents such a PDF when e_i is expressed modulo P, for a 0% and a 30% swing. Strictly speaking, the e_i are no longer independent from each-other, but we can still force the assumption of independence, to make the PDF for the set of e_i the product of the individual PDFs. Fig. 7 presents a visualization of the likelihood as a function of the swing and tempo for a given audio track. The likelihood is maximized over the tempo T and the swing S, and the beat location b_0 can be estimated by use of the transient-time based likelihood L_t (Eq.1). In that final stage, the search could be restricted to b_0 but should ideally include a refinement of the values of T and S estimated with $L_e(T, S)$. Note that the tempo T can be discretized much more coarsely than in the previous case, because a small variation of T does not translate into a larger variation of the elapsed time between beats (the likelihood in Fig. 7 is not very peaky). The resulting search is considerably faster.

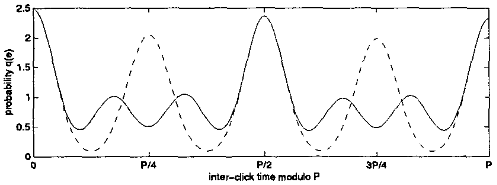

Figure 6: PDF for $e_i \mod P$ in the absence of swing (solid line) and for a 30% swing (dashed line).

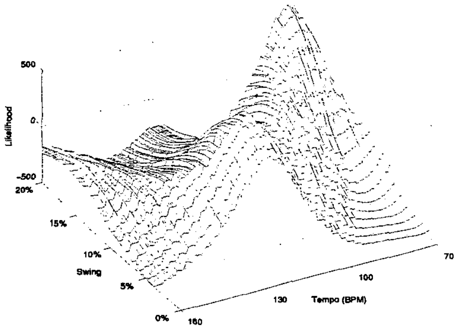

Figure 7: 3D representation of the likelihood for the set of e_i for an actual audio track.

3. RESULTS

For tracks whose tempo is actually constant, the results of the tempo-analysis techniques are surprisingly good. The simple hypothesis of quarter-beat asymmetry (Fig. 5) allows the algorithm to fairly consistently identify the true location of the beat, for most modern muscial genres. There are musical styles however that clearly violate this assumption, e.g. latin music where percussions tend to occur on the "up-beat", for which the algorithms will place the beat at the wrong (but consistent) location. The algorithms can also be run with the (cheaper) assumption of a 0% swing, which will work well on most tracks, but fail on tracks that have a fairly strong swing or a "ternary" time-signature (i.e. 3/8, usually treated as a 33% swing by the algorithm). The failure will usually translate into a tempo that is 33% slower than the true tempo (see Fig. 8). Other audio tracks that will trick the tempo detection algorithms are tracks that do not contain transients or sharp attacks. A violin solo played legato, for example, will be hard to analyze because the transient detection stage might have a hard time detecting anything.

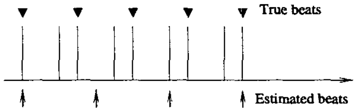

Figure 8: Tempo mis-estimation for a 3/8 time-signature. The true beats (black triangles) occur every three third-beats. Under the quarter-beat hypothesis with no swing, the estimated beats occur every four third-beats, yielding an erroneous tempo estimate.

4. CONCLUSION

The techniques described here work fairly well as long as the transient detection stage is able to identify salient features in the audio track, and the constant-tempo assumption is valid. Dealing with time-varying tempi is much more difficult, because many more parameters must be estimated. One approach consists of using the methods described here on small, possibly overlapping portions of the track (say 4 to 5s) and using the previous estimates in the calculation of the current one, possibly in a Bayesian framework.

5. REFERENCES

[1] E.D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, vol. 104, no. 1, pp. 588–601, Jan 1998.

[2] S.M. Kay, *Fundamentals of Statistical Signal Processing - Estimation Theory*, Prentice-Hall, Englewood Cliffs, MJ, 1993.

[3] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 1991.