ESTIMATING THE ERROR DISTRIBUTION OF A TAP SEQUENCE WITHOUT GROUND TRUTH[1]

Roger B. Dannenberg
Carnegie Mellon University
School of Computer Science

Larry Wasserman
Carnegie Mellon University
Department of Statistics

ABSTRACT

Detecting beats, estimating tempo, aligning scores to audio, and detecting onsets are all interesting problems in the field of music information retrieval. In much of this research, it is convenient to think of beats as occuring at precise time points. However, anyone who has attempted to label beats by hand soon realizes that precise annotation of music audio is not possible. A common method of beat annotation is simply to tap along with audio and record the tap times. This raises the question: How accurate are the taps? It may seem that an answer to this question would require knowledge of "true" beat times. However, tap times can be characterized as a random distribution around true beat times. Multiple independent taps can be used to estimate not only the location of the true beat time, but also the statistical distribution of measured tap times around the true beat time. Thus, without knowledge of true beat times, and without even requiring the existence of precise beat times, we can estimate the uncertainty of tap times. This characterization of tapping can be useful for estimating tempo variation and evaluating alternative annotation methods.

1. INTRODUCTION

Tempo estimation and beat tracking are considered to be fundamental tasks of automatic music analysis and understanding. To evaluate machine performance in these sorts of tasks, it is useful to have audio annotated with beat times. We often assume that beat times are obvious and easily measured, usually through manual annotation. In some sense this is a fair assumption. Humans are good at detecting beats, especially in popular music, and human performance is generally better than machine performance. Most research simply accepts human-generated data as correct.

In cases where the goal is simply to get close to "true" beat times, or to estimate tempo (which can be a long-term average), ignoring potential tapping errors might be reasonable. However, it is troubling to assume errors do not matter without any way to test this assumption. Furthermore, there are some cases where automated methods can deliver quite precise results. For example, onset detection and beat detection in piano music can rely on fast onsets to obtain precise times in the millisecond range automatically. It seems unlikely that humans can tap or otherwise annotate beat times with this degree of precision, so how can we evaluate automatic labels?

The main goal of this work is to characterize the quality of human beat and tempo estimates in prerecorded audio data. A simple approach to this problem is to synthesize music from known control data such as MIDI, using control timing as the "ground truth" for beat times. This approach offers a clear connection to an underlying sequence of precise times, and after a human taps along with the music, some simple statistics can describe the distribution of actual tap times relative to the "true" beats. The problem here is that "real" music seems more complicated: Musicians are somewhat independent, adding their own timing variations, both intentional and unintentional. Musicians play instruments with varying attack times and they sometimes place their note onsets systematically earlier or later than the "true beat" times. How can we know that tapping to carefully controlled synthesized music is indicative of tapping to music in general?

We present an alternative approach in which multiple independent taps to beat-based music are used to estimate a distribution around the underlying "true" beat time. We assume that a true but hidden beat time exists and that observed tap times are clustered around these true times. In addition, we assume "all beats are the same" in the sense that observed tap times for one beat have the same distribution as observed tap times around any other beat. (This assumption will be discussed later.)

It should be apparent that different forms of tapping (tapping with different kinds of audio feedback, tapping with hands or feet, tapping by or while performing a musical instrument) will have subtle implications for the positioning and distribution of the tap times. Our techniques enable us to explore these differences but say nothing about whether one is more correct than another. In other words, there may be different implied "true" beat times for different tapping conditions.

[1] Originally published as: Roger B. Dannenberg and Larry Wasserman, "Estimating the Error Distribution of a Tap Sequence Without Ground Truth," in *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009),* (October 2009), pp. 297-302.

In addition to estimating the distribution of tap times from multiple independent taps, our technique can estimate the distribution of another source of tap times. For example, we will show how a single set of foot tap times captured in a live performance can be used to estimate the accuracy of foot tapping, again without any ground truth. Our technique is interesting because it does not require any manual time estimation using visual editing, ground truth, or acoustical analysis, yet it gives us the ability to describe any sequence of estimated beat times as a probabilistic distribution around the underlying "true" beats.

By collecting data from real music audio examples, we can get a sense not only of the location of beats but the uncertainty of these locations. Since studies of expressive timing and tempo are normally based on beat time estimates, it is important to characterize uncertainty. In real-time computer music performance, estimating tempo and predicting the time of the next beat is an important problem. A good model of tapping and uncertainty can help to clarify the problem and analyze proposed solutions. There is also the potential to apply our model to the evaluation of automated beat tracking systems and to compare their performance to human tapping. Finally, models of timing and tempo change can help to build better beat tracking systems, which must reconcile prediction from past beat estimates using a steady-tempo hypothesis with new but uncertain beat estimates allowing the system to adapt to tempo change.

2. RELATED WORK

Previous studies have looked directly at tapping and synchronization. Michon [1] studied synchronization to sequences of clicks, and Mecca [2] studied human accompanists and how they adapt to tempo change. Wright [3] studied perceptual attack time, the perceived time or distribution of times at which a tone is perceived to begin. Dixon et al. [4] studied tapping to a short musical excerpt with expressive timing. There is a substantial literature on the perception of beats and rhythmic grouping [5]. The automatic detection of beats and tempo also has a long history of study [6, 7]. The Mazurka project [8] has published beat times estimated using acoustic data from expressive performances.

Computer accompaniment [9] is a popular topic in the computer music literature and this work is closely related to ours. Tempo change in computer accompaniment has been modeled using Bayesian belief networks [10]. Our study of beat estimation and tempo in fact addresses shortcomings of existing computer accompaniment systems. In particular, computer accompaniment is usually based on score following, which assumes that a score exists and that there are audio signals to be matched to the score [9]. In reality, popular music often involves improvisation and other deviations from the score (if any), so the computer system must be "aware" of beats, measures, and cues in order to perform effectively with live players [11].

Conducting is another means for synchronizing computers to live performers and another example of human indication of beats. Various conducting systems have been created using traditional conducting gestures as well as simple tapping interfaces [12]. These studies are closely related to our work because any conducting system must sense beat times and make predictions about the tempo and the next beat time. Our work extends previous work by measuring human performance in tapping along to music. The sequential drum [13] and radio drum [14] of Max Mathews are also in the category of conducting systems. These emphasize expressive timing and multidimensional gestural control.

The "virtual orchestra" concept [15, 16] is also related. Virtual orchestras have been created to accompany dance, opera, and musical theater. Most if not all of this work is commercial and proprietary, so it is not known what techniques are used or how this work could be replicated, making any comparative studies impractical. Certainly, a better understanding of beat uncertainty and tempo estimation could contribute to the performance of these systems.

3. THE MODEL AND ASSUMPTIONS

We are interested in characterizing information obtained from tapping to music audio. In an ideal world, we would first label the audio with precise beat times. For example, we might ask a subject to tap by hand many times along with the music, measure the tap times, and compute the mean tap time $\widehat{\theta}_1, \widehat{\theta}_2, \ldots$ for each beat. Presumably, these mean tap times estimate and converge to a precise underlying or "hidden" time θ_i for each beat. In this way, beat times can be estimated with arbitrary precision given sufficient data. Once beat times are estimated, we can study other tap sequences. For example, given a sequence of foot tap times F_i we might like to estimate the distribution of timing errors: $\Delta_i = F_i - \theta_i$. If we ignore the difference between $\widehat{\theta}_i$ and θ_i, it is simple to compute the mean and standard deviation of Δ_i or simply to plot a histogram to characterize the distribution.

It should be noted that the outcome (the distribution of Δ_i) is a distribution over timing errors throughout the entire piece, not a distribution for a particular beat. Timing errors and the distributions of individual beats might be interesting things to study, but these are not considered by our model.

Unfortunately, tapping along to music requires much time, care, and concentration. We want to achieve the same results without tapping along to music many times. In fact, if we make a few assumptions about Δ_i, we only need to tap twice. Then, given a measured sequence of times F_i, we can estimate the corresponding distribution Δ_i.

The assumptions are that, first, Δ_i is normal. We will show some evidence that Δ_i obtained from actual tap data is in fact approximately normal. The second assumption is that the sequence of true beat times θ_i is well defined and the same for all tap sequences. So for example, if we want to compare foot taps to hand taps, we need to assume that the underlying "true" beats for each sequence are the same. Alternatively, if we want to measure the tap distribution of several subjects, we must assume they all share the same

true beats.

In practice, we are seldom concerned about absolute shifts (subject A always perceives beats 10ms earlier than subject B). But introducing a time offset to a collection of tap times, say from subject A, generally increases the estimated variance of the tap times. If we believe that an offset may reflect individual differences, sensors, or calibration problems, then we can simply estimate and subtract off the offset. (Details will follow.) In that case the only assumption is that the "true" beat times for any two sequences of taps are the same except for a constant (but unknown) time offset.

4. ESTIMATING THE DISTRIBUTION

To estimate the distribution of actual tap times, let $\theta_1, \theta_2, \ldots$ denote the true beat times. (These will remain unknown.) Also, we will collect two sets of hand taps at times H_i^1, H_i^2. We assume that these times are normally distributed around the true beat times:

$$H_i^1, H_i^2 \sim \text{Normal}(\theta_i, \sigma^2). \tag{1}$$

An unbiased estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n}(H_i^1 - H_i^2)^2. \tag{2}$$

Thus, with only two sets of taps generated under the same conditions, we can estimate the distribution of the taps relative to the true beat times. It should be mentioned that H_i^1 and H_i^2 must correspond to the same true beat. If, for example, one tap is missing from H^1, then some differences $(H_i^1 - H_i^2)$ will be increased by one beat. In practice, taps rarely differ by more than 150ms and beats are typically separated by 500ms or more (taps can be every 2, 3, or 4 beats if the tempo is faster), so errors are simple to find and correct.

What if H^2 has a constant offset relative to H^1? Since we assume the distribution around the true beat should be the same for both sequences, the mean of their differences $\bar{d}$:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n}(H_i^1 - H_i^2) \tag{3}$$

should be zero. We can "correct" any constant offset (estimated by $\bar{d}$) by replacing H_i^2 by $(H_i^2 + \bar{d})$.

Now suppose we have another set of tap times generated by a different source, for example foot taps or taps from another subject. What is the distribution of these taps? Given H_i^1 and H_i^2, we only need one set of taps (one tap per beat) from the new source.

Let F_i be the new set of tap times, and let $\Delta_i = F_i - \theta_i$. The problem is to estimate the distribution of the Δ_i's. Let us begin by defining

$$\widehat{\Delta}_i = F_i - \widehat{\theta}_i \tag{4}$$

where $\widehat{\theta}_i$ is an estimate of θ_i. For these estimates, we will use

$$\widehat{\theta}_i = \frac{H_i^1 + H_i^2}{2}. \tag{5}$$

From the assumption that F_i is normal, $F_i \sim N(\theta_i, \tau^2)$, it follows that

$$\widehat{\Delta}_i \sim N\left(0, \tau^2 + \frac{\sigma^2}{2}\right) \tag{6}$$

Here, τ^2 is due to random variation in F_i and $\frac{\sigma^2}{2}$ is due to uncertainty in our estimates of the true beat times. Now, if we let s^2 be the expected sample variance of the $\widehat{\Delta}_i$, we obtain

$$s^2 = \tau^2 + \frac{\sigma^2}{2} \tag{7}$$

and hence

$$\tau^2 = s^2 - \frac{\sigma^2}{2} \tag{8}$$

Thus,

$$\Delta_i \sim N\left(0, s^2 - \frac{\sigma^2}{2}\right) \tag{9}$$

We already have an estimate of σ^2, and we can estimate s^2 using the sample variance $\widehat{s}^2$ of $\widehat{\Delta}_i$. Substituting $\widehat{\sigma}^2$ for σ^2 and $\widehat{s}^2$ for s^2, we can estimate the distribution of Δ_i and thus the accuracy of taps from the new source even without any ground truth for beat times. All we need are two additional sets of times obtained by tapping along with the music.

5. GENERALIZATION TO N SEQUENCES

This approach can be generalized to multiple tap sequences. For example, taps from many different subjects might be combined. Suppose that N tap sequences, $H_i^1, \ldots, H_i^N$ are normally distributed with means θ_i and variance σ^2. We estimate means and variance as follows:

$$\widehat{\theta}_i = \frac{1}{N}\sum_{j=1}^{N} H_i^j \tag{10}$$

and

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2 \tag{11}$$

where

$$s_i^2 = \frac{1}{N-1}\sum_{j=1}^{N}(H_i^j - \widehat{\theta}_i)^2. \tag{12}$$

Defining $\widehat{\Delta}_i$ again as in (4), we generalize (6) to

$$\widehat{\Delta}_i \sim N(0, \tau^2 + \frac{\sigma^2}{N}). \tag{13}$$

Letting S^2 be the expected sample variance of the $\widehat{\Delta}_i$,

$$\Delta_i \sim N(0, \tau^2) = N\left(0, S^2 - \frac{\sigma^2}{N}\right). \tag{14}$$

Again, we can estimate S^2 using the sample variance $\widehat{S}^2$ of $\widehat{\Delta}_i$ and estimate the variance of Δ_i as $\widehat{S}^2 - \frac{\widehat{\sigma}^2}{N}$.

6. IS Δ_I NORMAL?

Our analysis assumes that the distribution of Δ_i is normal. We collected some taps to music synthesized from MIDI with note onsets quantized to exact beat times and smoothly varying but mostly constant tempo. Figure 1 shows a histogram of differences between the 117 "true" beats and hand-tapped (by one of the authors) beats, corresponding to Δ_i. To characterize the error, we use the mean of the absolute difference (MAD) between tapped beats and true beats after adjusting an absolute time offset to obtain a mean difference of zero. For this condition, the MAD is 16.13ms. Although the extreme values of +/-60ms seem quite large, the MAD value of 16.13ms compares favorably to the typical value of 10ms cited as the just noticeable difference (JND) for timing deviation [17]. (Since our goal is to describe a representation and its theory, we show only a couple of typical examples from data collected from a growing collection of songs, performers, and tappers.)

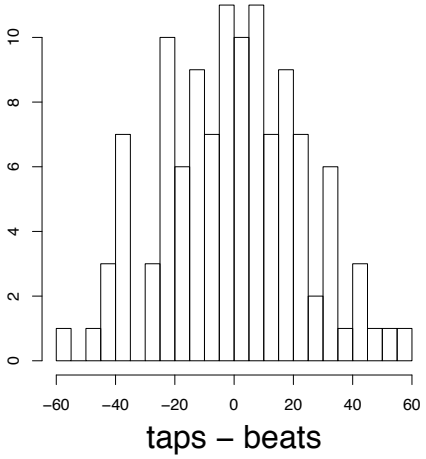

Figure 1. *Histogram of deviations (in ms) of hand tap times from "true" (MIDI) beat times.*

Using live acoustic music, two sets of hand tap times were collected, and Figure 2 shows differences between corresponding hand tap times. In this example, the music was from a big-band jazz rehearsal. Again, the data is from one of the authors, but it is typical of other data we have collected. This differs from Figure 1 in that the time differences are between two tap times to acoustic music rather than between a tap time and a known beat time in synthesized music. The standard deviation is 26ms. As with the MIDI-related data, the general shape of the histogram appears to be Gaussian, so the Normality assumption is at least reasonable. A Shapiro-Wilks test of Normality on data in Figures 1 and 2 yields values of $W = 0.9829$ (p-value = .6445), and $W = 0.9882$ (p-value = .2625), suggesting again that Normality is reasonable.

7. EXAMPLE

We are interested in characterizing foot tapping as an indicator of beat times. For our data used in Figure 2, $\widehat{\sigma} =$ 32.77ms (standard error 2.006ms).

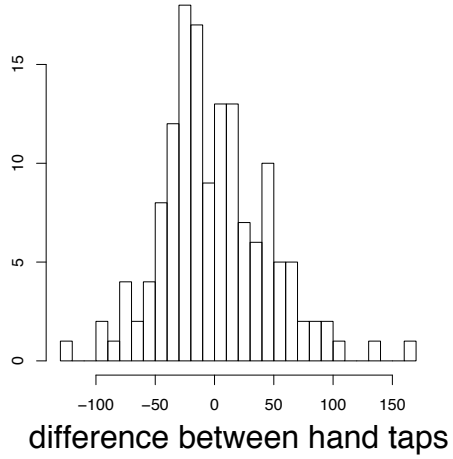

Figure 2. *Histogram of differences (in ms) between two sets of hand tap times to audio recording of a live performance.*

Even before collecting hand tap times, we recorded audio and foot tap times from a live performance. The foot taps are sensed by a custom pedal that uses a force-sensitive resistor (FSR) to control the frequency of a low-power analog oscillator [15]. The audio output from the pedal can be sent to one channel of a stereo recording in synchrony with the live music on the other channel. Later, the "foot pedal channel" can be analyzed to detect foot taps with precise synchronization to the music audio. We then used Sonic Visualizer [18] to record hand taps (twice) while listening to the music channel of the recording.

Finally, using the analysis described in Section 4, we obtain a mean of 0 and a standard deviation of 37.2ms. This number reflects a particular condition involving the type of music, the other players involved, the interference task of performing, and possibly individual differences. Thus, we are not suggesting that one can give meaningful numbers for the accuracy of hand-tapped or foot-tapped beats in general, only that for any given situation, the taps can be accurately and efficiently characterized without a ground truth for beat times.

This example is interesting because it is impossible to obtain more than one set of foot taps or ground truth from a live performance, yet our technique still provides an estimate of the foot tap error distribution.

8. DISCUSSION

Because beats are hidden and perceptual, there are multiple ways to characterize beats. Just as there are differences between pitch (a percept) and fundamental frequency (a physical attribute), a distinction can be made between perceived beat times and acoustic event times. Some research relies on acoustic events to estimate beat times. While objective and often precise, these times are subject to various influences including random errors and physical characteristics of the instrument [19], so even acoustic times are the result of human perception, cognition, and action. After all, performing within an ensemble requires a perception

of the beat and precise timing, so it is not all that different from tapping.

Furthermore, polyphony creates ambiguity because note onsets are often not synchronized. In fact, there is good evidence that note onsets are deliberately *not* placed on "the beat," at least in some important cases [20]. Therefore, this work attempts to identify and characterize *perceptual* beat times through tapping.

Even this approach has limitations. As seen in our example data, beat times are characterized as distributions rather than precise times, reflecting the limited information available from a small number of tap sequences. Moreover, all distributions are assumed to have the same variance. In music with a steady beat, this seems to be a reasonable assumption. In music with expressive timing, rubato, etc., one would expect some beats to be more accurately tapped than others. Learning from repeated listening can affect tapping times [4]. We suspect learning is a bigger factor in music with expressive timing where subjects might learn to anticipate timing variations. In music with a steadier tempo, any learning effect should be minimal.

The "meaning" of variance (τ^2) merits discussion. One interpretation is that the perceived beat time is very precise, but there are limitations in motor control that give rise to variation in tap times. Another interpretation is that the perception of beat times is not consistent from one listening to the next, resulting in different tap times. If different subjects tap, variance could arise from a difference between subjects. Ultimately, τ^2 models real data, so a more detailed model may not be relevant. On the other hand, experiments might be able isolate and characterize different influences on tap timing.

Using a limited amount of "field recording" data, we observed that foot tap timing can be approximated by a normal (Gaussian) random distribution around the "true" beat time. This is suggested by histograms as well as a Shapiro-Wilks test of Normality. The observed variance is almost certainly dependent upon the clarity of the beat, the steadiness of tempo, the skill of the tapper, interference tasks including playing an instrument while tapping, and other factors. The good news is that the method is practical and inexpensive, and the method can be used to study all of these factors.

Many studies in Computer Music, Music Information Retrieval, and Music Perception depend upon estimates of beat times and tempo variation. The techniques described here offer a principled way to go about characterizing the uncertainty of beat times obtained by tapping.

9. APPLICATIONS AND FUTURE WORK

The goal of this paper is to describe a representation of beat timing, the underlying estimation theory, and a practical way to use this representation. Current work is examining data from many sources with the goal of understanding the range of uncertainty (τ^2) observed under different conditions, and perhaps factors that account for differences. Also, experiments could study the degree to which tap time variance results from perceptual uncertainty vs motor control.

One of our goals is to create music systems that perform with live musicians using techniques based on work in Music Information Retrieval. Beat tracking, gesture sensing, analysis of mood, and other aspects of a performance all provide important input to an automated music performer.

In the area of beats and tempo, the techniques presented here are being used to analyze data from a variety of performances. For synchronization to live performers, the data will help us to tune systems that accurately predict the next beat time, allowing an artificial performer to play accurately on the beat. Beat timing variation implies tempo change. Modeling tap times probabilistically can help to distinguish between random timing errors and true tempo change. For example, preliminary analysis has shown that, depending upon the amount of tempo variation in a piece of music, estimating tempo using the previous 6 to 18 beats gives the best prediction of the next beat time. This work is closely related to beat tracking systems where smoothing over beats can help the system stay on track, but smoothing over too many beats makes the system unable to follow tempo changes.

Another application is in the construction and evaluation of score-to-audio alignment systems. While scores have precise beat times, audio recordings do not. By substituting alignment times for foot tap times (F_i in (4)), we can measure score alignment quality without any ground truth.

Audio labeling is another application. We might like to compare beat labels based on audio features to perceptual beat times. Since tap times might have a large variance, one is tempted to conclude that precise audio-based labels are more reliable. With our techniques, this can be tested. Another issue with labeling is the reliability of hand-labeled audio using an audio editor. This is a very difficult task where one might expect to see individual differences among human labelers. The lack of ground truth makes it difficult to evaluate different labelers. Our method might be useful because it does not need the ground truth to provide an analysis.

Finally, it is interesting to study tempo in the abstract. In live performances we have tapped to, we have found substantial tempo changes (on the order of 10%) during solos with a rhythm section where the tempo is nominally steady. As with live synchronization, one must be careful to avoid attributing tempo change to jitter in tap times, and a characterization of the tap time distribution helps to identify true tempo changes.

10. CONCLUSION

Our work concerns the analysis of beat times in music with a fairly steady beat. Our live data collection and analysis indicate that foot tap timing can be modeled well as a Gaussian distribution around a "true" but unknown beat time. We have introduced a new technique for estimating tapping accuracy that *does not require the accurate identification of underlying beats.* By comparing foot tap data

(or data from other sources) to multiple hand taps on the same music, we are able to estimate the standard deviation and thus characterize the uncertainty in the tapping data. A major strength of this approach is that a one-time, irreproducible sequence of taps such as from a live performance can be analyzed in terms of accuracy without ground truth for "true" beat times.

11. ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of Nathaniel Anozie to the initial exploration and analysis of this data, and Cosma Shalizi for helpful consultation and discussions. This work was supported in part by Microsoft Research through the Computational Thinking Center at Carnegie Mellon.

12. REFERENCES

[1] J. A. Michon. *Timing in Temporal Tracking*. Van Gorcum, Assen, NL, 1967.

[2] M. Mecca. Tempo following behavior in musical accompaniment. Carnegie Mellon University, Department of Philosophy, Pittsburgh, PA, USA (Masters Thesis), May 1993.

[3] Matt Wright. *Computer-Based Music Theory and Acoustics*. PhD thesis, Stanford University, CA, USA, March 2008.

[4] S. Dixon, W. Goebl, and E. Cambouropoulos. Perceptual smoothness of tempo in expressively performed music. *Music Perception*, 23(3):195–21, 2006.

[5] P. Fraisse. Rhythm and tempo. In D. Deutsch, editor, *The Psychology of Music*, pages 149–80. Academic Press, New York, 1st edition edition, 1982.

[6] F. Gouyon and S. Dixon. A review of automatic rhythm description systems. *Computer Music Journal*, 29(1):34–54, 2005.

[7] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1832–1844, September 2006.

[8] Craig. Sapp. Comparative analysis of multiple musical performances. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07)*, pages 497–500, 2007.

[9] R. B. Dannenberg and C. Raphael. Music score alignment and computer accompaniment. *Communications of the ACM*, 49(8):39–43, August 2006.

[10] C. Raphael. Synthesizing musical accompaniments with bayesian belief networks. *Journal of New Music Research*, 30:59–67, 2000.

[11] A. Robertson and M. D. Plumbley. B-keeper: A beat tracker for real time synchronisation within performance. In *Proceedings of New Interfaces for Musical Expression (NIME 2007)*, pages 234–237, 2007.

[12] R. B. Dannenberg and K. Bookstein. Practical aspects of a midi conducting program. In *ICMC Montreal 1991 Proceedings*, pages 537–540, San Francisco, 1991. International Computer Music Association.

[13] M. V. Mathews and C. Abbot. The sequential drum. *Computer Music Journal*, 4(4):45–59, Winter 1980.

[14] M. V. Mathews and W. A. Schloss. The radio drum as a synthesizer controller. In *Proceedings of the 1989 International Computer Music Conference*, San Francisco, 1989. Computer Music Association.

[15] Roger B. Dannenberg. New interfaces for popular music performance. In *NIME '07: Proceedings of the 7th International Conference on New Interfaces for Musical Expression*, pages 130–135, New York, NY, USA, 2007. ACM.

[16] Gregory M. Lamb. Robo-music gives musicians the jitters. *The Christian Science Monitor*, December 14, 2006.

[17] A. Friberg and J. Sundberg. Perception of just noticeable time displacement of a tone presented in a metrical sequence at different tempos. Technical report, STL-QPSR, Vol. 34, No. 2-3, pp. 49–56, 1993.

[18] C. Cannam, C. Landone, M. Sandler, and J. P. Bello. The sonic visualizer: A visualization platform for semantic descriptors from musical signals. In *ISMIR 2006, 7th International Conference on Music Information Retrieval*, pages 324–327, 2006.

[19] Patrik N. Juslin. Five facets of musical expression: A psychologist's perspective on music performance. *Psychology of Music*, 31(3):273–302, 2003.

[20] A. Friberg and A. Sundstrom. Swing ratios and ensemble timing in jazz performance: Evidence for a common rhythmic pattern. *Music Perception*, 19(3):333–349, 2002.