IMPROVING PERCEPTUAL TEMPO ESTIMATION WITH CROWD-SOURCED ANNOTATIONS

Mark Levy
Last.fm Ltd., Karen House, 1-11 Baches Street, London N1 6DL, United Kingdom
mark@last.fm

ABSTRACT

We report the design and results of a web-based experiment intended to support the development and evaluation of tempo estimation algorithms, in which users tap to music and select descriptive labels. Analysis of the tapping data and labels chosen shows that, while different listeners frequently entrain to different metrical levels for some pieces, they rarely disagree about which pieces are fast and which are slow. We show how this result can be used to improve both the evaluation metrics used for automatic tempo estimation and the estimation algorithms themselves. We also report the relative performance of two recent tempo estimation methods according to a further controlled experiment that does not depend on groundtruth values of any kind.

1. INTRODUCTION

Numerous algorithms for estimating the tempo of music directly from an audio signal have been developed in recent years, motivated by the obvious value of tempo information to automated tools for use in playlisting and DJ mixing [4]. Automatic estimation of the tempo of a track as a simple value measured in beats per minute (bpm) is now regarded as an established technique. Bpm estimation algorithms have gained a place in widely-distributed commercial hardware mixers for DJs, as well as software applications and web service APIs aimed at musicians, recording labels and mobile application developers. Meanwhile the annual MIREX algorithm evaluation competition offers a more formal benchmark for the performance of tempo estimation software. This has led to the proliferation of rival methods: seven different algorithms were submitted to the audio tempo estimation competition during the past year alone.

A common observation made in both informal and formal evaluation of tempo estimation methods is that they frequently suffer from so-called *octave error*, where the machine estimate is some simple multiple or fraction of the perceived tempo [5, 10]. These errors appear to be analogous to a phenomenon observed in studies of human perception of the rhythmic properties of music: humans can also sometimes disagree about the frequency of the main beat of a piece of music. In particular two influential experiments on the perception of tempo attempt to generalise observed variations in human responses into somewhat more formal models of tempo ambiguity [8, 9].

The definition of *tempo ambiguity* proposed in [8] is based on the authors' observation that, while users tend to agree on a bpm value for many tracks, in the remaining cases opinion is divided between two candidates. They quantify ambiguity as the strength of support for the larger of these two candidates, divided by their mean support:

$$A = \frac{2\max(H(T_1), H(T_2))}{H(T_1) + H(T_2)} \tag{1}$$

where $H(T_1)$ and $H(T_2)$ are the number of users who tap at T_1 and T_2, the most and second most commonly observed bpm values, respectively. The study attempts to model the tempo ambiguity of a track in two different ways. Firstly the authors suggest that tempo ambiguity may be related to the mean of $H(T_1)$ and $H(T_2)$; and secondly they investigate how a related *resonance deviation* statistic might be predicted from the value of an acoustic periodicity difference feature computed from the audio signal. The first model was found to be consistent with data collected from a group of 33 listeners for a set of 24 ten-second excerpts, but the result could not be replicated in a second study of 24 subjects who tapped to the beat of 60 thirty-second excerpts. The second model was not supported convincingly by either experiment, although a modified formulation of resonance deviation was found to be correlated with periodicity difference in a third study of 40 subjects [9].

Despite the inconclusive results reported in [8, 9], the studies have been indirectly influential in the research community due to the adoption of their experimental data, and of an evaluation methodology based on their observations, in recent rounds of the MIREX tempo estimation competition. Algorithms entered for the competition have been required to output two different bpm estimates for each track, together with associated weights intended to represent "perceptual strength", credit being given for weights similar to $H(T_1)$ and $H(T_2)$, as well as for estimates close to T_1 and T_2.

A separate line of research has, however, highlighted a negative side-effect of putting tempo ambiguity at the heart of an evaluation metric for machine estimates. Imagine a track that has been submitted for automatic tempo estimation and marked as "70bpm or 140bpm". Is this track appropriate for a playlist of pieces at walking pace, as suggested by an estimate of 70bpm? Is it more suitable for a high energy playlist, as suggested by 140bpm? Will the track really be perceived as slow by some listeners and as fast by others? Or is one of the values simply a poor estimate resulting from a shortcoming of the algorithm, which would be better ignored?

The authors of [5] go so far as to suggest that such uncertainty means that machine bpm estimates are simply not usable in practice for many potential applications, and should be abandoned in favour of categorical labels such as *slow* and *fast*. The study goes on to report extremely high accuracies achieved with a slow-fast classifier trained on a bag of well-known low level audio features, and using social tags as its groundtruth annotations. This suggests that the ambiguity intrinsic to bpm estimation may simply not arise in relation to perceptual tempo categories.

In this paper we attempt to reconcile these apparently conflicting views of perceptual tempo estimation by crowdsourcing a large set of responses through a web-based experiment. The responses include both tapping data and selections from a list of categorical labels. The remainder of the paper is structured as follows: in Section 2 we describe the design of the experiment and give some background about web experiments in general; in Section 3 we report results, in particular exploring the relationship between label selection and human bpm estimates; in Section 4 we outline how categorical labels might be used to improve bpm estimates from existing tempo estimation algorithms; in Section 5 we describe and report the results of a controlled experiment to compare different algorithms without reference to any groundtruth values; and in Section 6 we draw conclusions and outline future work. Last but not least we provide links to our experimental data, making it available for future research and evaluation.

2. EXPERIMENTAL DESIGN

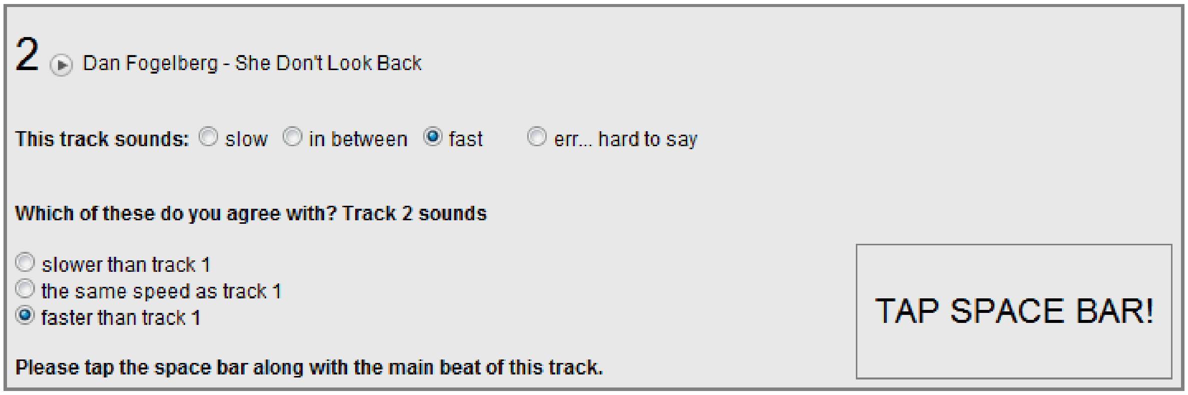
While studies of the perception of music are traditionally carried out under laboratory conditions, in recent years the web has begun to be regarded as a potential source of perceptual data. Social tags, such as those submitted to the music service Last.fm [1], can be seen as an abundant source of perceptual responses although their quality is low: the "experimenter" has no control whatsoever over the circumstances in which a tag is applied, indeed there is no guarantee that the user of a music tagging system has even listened to the music which they are tagging. Social tags have nonetheless been used in several studies intended to capture listeners' characterizations of perceptual characteristics of music [6]. The appeal of tags to researchers is that the cost of acquiring them is essentially zero, and they are often available in sufficient numbers for statistics to be robust even if individual tags are unreliable. Other experiments have been designed as appealing internet games [7]. These games give considerably more control over the circumstances in which data is collected, but require a relatively large investment in design and development.

For this study we opted for a middle course, designing our experiment along the lines of a traditional laboratory questionnaire, but hosting it on the web and simply appealing to visitors to contribute to our research. Besides providing a source of data for the questions at hand, we were particularly interested to find out if visitors would take part in response to such a bald invitation. This approach, if successful, could offer a useful platform for future research, offering considerably more control than social tags, but at much lower cost than an internet game. The web page for the experiment was hosted on the companion labs site to a large music website. Although the main site receives many millions of pageviews per day, traffic to the labs site is several orders of magnitude lower, typically a few thousand pageviews per day.

On each view of the experiment, the web page shows artist and title information, along with an associated set of questions, for thirty-second excerpts of either one or two tracks. The excerpts are chosen at random from a pool of several thousand audio clips, described in more detail in Section 3. The first excerpt starts playing as soon as the page has fully loaded, and the second excerpt starts as soon as the first has finished. As shown in Figure 1, users are first asked to select a speed label for each track, choosing either from a 3-point scale from *slow* to *fast*, or a visually separate category to report cases where they are not sure. On a page displaying two excerpts they are then asked whether the second excerpt sounds *slower*, *the same speed* or *faster* than the first. Finally the visitor is asked to tap along with the main beat of the music.

The sequence in which answers can be provided is not strictly locked down, but highlighting on the page is used to encourage the visitor to answer the questions in order. In particular the large call to action, shown in Figure 1, is displayed only once the preceding question has been answered. Conventional audio play/pause buttons are provided, so it is

[1] e.g. `http://www.last.fm/tag/slow`



2 ▶ Dan Fogelberg - She Don't Look Back

This track sounds: ○ slow ○ in between ● fast ○ err... hard to say

Which of these do you agree with? Track 2 sounds

○ slower than track 1
○ the same speed as track 1
● faster than track 1

Please tap the space bar along with the main beat of this track.

Figure 1. Questions asked on the experiment web page.

also possible to stop and restart the tracks at will, or to listen to them more than once. Once tapping begins, the bpm meter is highlighted in red, changing to green once ten taps have been recorded, to give the visitor an idea of when they have tapped for long enough to allow a reasonable estimate of bpm to be made. If a visitor resumes tapping after a pause of two seconds or more, the bpm meter and its internal counters are reset and the tapping is considered as a new attempt to answer the question. Although not explicitly messaged on the page, this allows users to try again if they are unhappy with their tapping for any particular track. It also imposes a lower limit of 30bpm on the tempo which the experiment can record. When the visitor presses the Save button, their label choices for each track are stored, together with a single bpm value computed simply as the mean interval between their taps.

The web can reasonably be regarded as a hostile environment for perceptual experiments when compared to a laboratory setting, but provided the rules of engagement are understood in advance then it is possible to design reasonable safeguards into the way in which responses are collected. We restrict access to the experiment to logged-in users of the main website, allowing us to associate responses with the users who have submitted them. To attract users to contribute more responses, we award points for each question answered and display total scores for top contributors on a separate leaderboard page, a ploy which unfortunately is also known to encourage cheating. To mitigate the effects of cheating we store at most one set of responses per user for each track. Although organised cheating of course remains possible, it would require a very determined attempt given the relatively low profile of the experiment website, and spurious data associated with any particular set of users can easily be filtered out of any analysis. In practice we discarded only tapping estimates of over 300bpm, which most likely correspond to misunderstanding of the interface.

With these considerations in mind, however, we do ensure that the design of the experiment also allows us to collect data for a more robustly controlled comparison of different tempo estimation algorithms. This is discussed more fully in Section 5 below.

	Listeners	Tracks	Responses
Labels	2141	4006	21444
Bpm estimates	1919	3929	19451
Comparisons	1438	3825	7597

Table 1. Responses received at the time of writing.

3. ANALYSIS OF RESULTS

The experiment continues to be publicly available at `http://playground.last.fm/demo/speedo`. Table 1 summarises the number of responses received at the time of writing. The tracks presented on any given view of the experiment web page are chosen essentially at random, as described in detail in Section 5, and consequently the distribution of responses between tracks is not uniform. Most of the following analysis concentrates on tracks which were annotated by at least five listeners: in particular 1437 tracks received five or more speed labels, while 1263 of those received at least five bpm estimates.

The annotated tracks are predominantly rock, country, pop, soul, funk and rnb, jazz, latin, reggae, disco and rap, but also include music from numerous other genres, including punk, electronic, trance, industrial, house and folk. They range from recent releases back to the 1960s. A full list of tracks used in the experiment is available (see Section 6).

3.1 Ambiguity in perceptual tempo labels

As described in Section 2, visitors to the experiment were asked both to tap along to each excerpt and to describe its

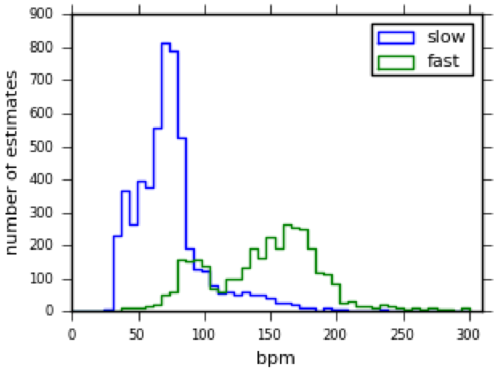


Figure 2. Observed distribution of all bpm estimates by speed category.

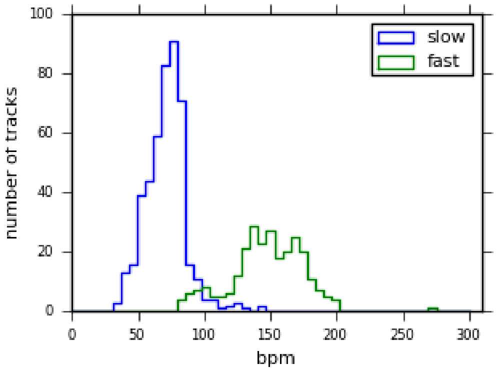


Figure 3. Distribution of peak bpm estimates by speed category.

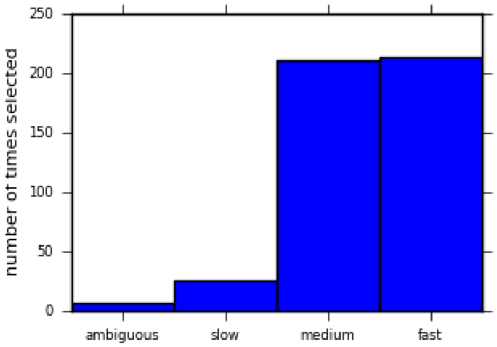


Figure 4. Labels submitted by half-speed tappers for tracks generally considered to be fast.

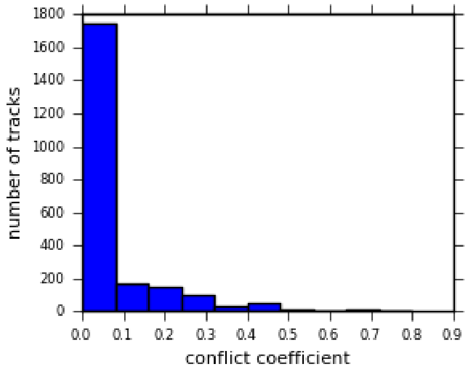


Figure 5. Observed distribution of conflict coefficient C.

speed on a three-point scale of slow, medium and fast, or to indicate if they found it hard to decide. Figure 2 shows distributions of all bpm estimates computed from tapping, for tracks annotated by at least five people of whom a majority described them as slow or fast respectively. Figure 3 shows the corresponding distributions of single *peak* bpm estimates for each track, computed as follows. Individual listeners' estimates are histogrammed into ten bins; the peak estimate is then the median value in the most populated bin. If adjacent bins contain the same number of values they are merged into a single bin before taking the median.

The shape of the distributions in Figure 2 suggests that we can be specific about octave disagreement in human tapping: when listening to tracks generally regarded as fast, some listeners tap half as fast as the majority. Figure 4 shows the distribution of labels supplied for these tracks by "slow tappers": there are cases in which they consider the music to be slow, but they are rare.

To model the extent of disagreement over perceptual slow and fast categories, by analogy with (1) we define the *conflict coefficient* for a track:

$$C = \frac{\min(L_s, L_f)}{\max(L_s, L_f)} \cdot \frac{L_s + L_f}{L} \qquad (2)$$

where L is the total number of labels supplied, of which L_s are slow and L_f are fast. The first term represents the extent to which fast and slow labels conflict, while the second term applies a discount to this if other users have labelled the excerpt as medium. Figure 5 gives the distribution of C over all tracks with at least five labels, showing that, for the huge majority of tracks, listeners do not disagree at all when describing excerpts as either slow or fast. To test whether listeners disagree over perceptual categories in the face of tempo ambiguity, we define ambiguous tracks to be those for which more than 30% of listeners tap at either double or half the peak bpm estimate, allowing a 4% margin of error when comparing bpm values. The mean conflict coefficient for ambiguous tracks is 0.062, slightly higher than the mean of 0.058 for the remaining unambiguous tracks, but the difference is not significant (p = 0.647). We conclude that in general there is no evidence that listeners disagree over which excerpts sound slow, and which sound fast, even when they tap at different metrical levels.

3.2 Evaluating machine bpm estimates

In order to demonstrate the potential value of crowd-sourced annotations in evaluating tempo estimation algorithms, we selected excerpts for the experiment for which bpm estimates were readily available from several sources. We report results here for the following three sources: estimates from the commercial EchoNest API, as distributed with the Million Song Dataset [1]; the BPM List, a published list of bpm values claimed to be computed at least partly by hand, using a variety of commercially-available tools [2]; and, finally, estimates generated using an implementation of methods reported in [3] and distributed as a plugin for the VAMP framework for audio analysis[2].

Some selection of values was necessary for the EchoNest and VAMP sources. The Million Song Dataset was found in a number of cases to contain data for different versions of the same song: we rejected any songs for which the duplicate tempo estimates differed by more than 2%, and otherwise simply used the fist value encountered. The VAMP plugin is designed to produce multiple segment-wise tempo estimates: we selected the estimate associated with the longest segment(s) of audio.

Table 2 shows evaluation results for the three sources relative to peak human estimates. The evaluation is restricted to tracks for which at least five crowd-sourced bpm values were available. In order to observe systematic types of error in the sources, estimates not matching the human reference values are split between six categories, corresponding to six types of octave error, and a final ‘unrelated’ category for estimates that do not match any of the preceding ones. An estimate is considered to match the groundtruth bpm, or one of its related values, if it differs by less than 4% of the reference[3].

The results given in Table 2 show significant differences in the performance of the three sources: the strongest source, the BPM List, is correct some 70% more often than the weakest, the EchoNest. The BPM List also suffers the least from octave error, presumably confirming that humans were involved in the creation of its estimates. While categorised results like Table 2 are useful to understand the strengths and weaknesses of particular methods, a robust single performance value can also easily be computed as a weighted combination of the percentages of estimates classed as correct and as unrelated. The weights can be tuned to reflect the potential harm caused by octave errors for any particular application.

	first faster	same	second faster
EchoNest	39.2	27.3	33.4
Bpm List	33.9	34.7	31.4
VAMP	34.0	34.0	32.0

Table 4. Percentage of answers given when comparing two tracks annotated with the same bpm by a particular source.

4. IMPROVING AUTOMATIC BPM ESTIMATES

Results presented in [5] report that classifiers can be trained to recognise tracks belonging to perceptual slow and fast categories with extremely high accuracy. The separable distributions shown in Figure 3 suggest that we can use the output of such classifiers to remove a great deal of octave error in machine estimates. The following simple algorithm can be used to adjust bpm estimates in cases where they conflict with predicted labels: any estimate of over 100bpm for a track classified as slow should be halved, and vice versa for fast tracks. While evaluating this approach directly remains for future work, Table 3 illustrates the substantial gains possible in the best case, by assuming a classifier that always predicts the label chosen by the majority of humans in the experiment.

5. COMPARING ESTIMATION ALGORITHMS

In addition to allowing data collection for conventional evaluation against a groundtruth, the experiment was designed to contain a controlled experiment enabling the comparison of different sources of bpm estimates without reference to any groundtruth. The experiment holds indexes from the bpm estimates of each source, rounded to the nearest integer, to a list of all tracks for which the source gave that estimate. When a visitor arrives at the experiment web page, the server first chooses a source at random. It then chooses a rounded bpm value, and finally selects two corresponding tracks from the index (or a single track, if the source only annotated one track in the collection with that particular value). This ensures not only that visitors are asked to annotate tracks with a wide range of likely tempo, but in particular that any two tracks presented together are regarded by at least one of the sources as having the same tempo.

Sources can then be compared for consistency by examining responses to the second question shown in Figure 1, in which listeners are asked to say which of the two tracks sounds faster. This is clearly a leading question, likely to

[2] `http://www.vamp-plugins.org`

[3] The MIREX 2010 evaluation allows a relative error of 8%.

	bpm * 4	bpm * 3	bpm * 2	correct	bpm / 2	bpm / 3	bpm / 4	unrelated
EchoNest	0.6	1.7	30.5	40.7	2.4	0.0	0.1	24.0
Bpm List	0.0	0.2	8.2	68.1	5.2	0.1	0.0	18.3
VAMP	0.7	1.6	23.0	58.3	4.0	1.6	0.0	12.3

Table 2. Performance of three sources of bpm estimates relative to peak crowd-sourced value. Numbers in each category are percentages of tracks evaluated for each source.

	bpm * 4	bpm * 3	bpm * 2	correct	bpm / 2	bpm / 3	bpm / 4	unrelated
EchoNest	0.0	0.5	19.5	53.0	1.7	0.0	0.0	25.2
Bpm List	0.0	0.0	5.7	72.8	3.0	0.1	0.0	18.5
VAMP	0.1	0.1	10.9	73.6	1.6	0.0	0.0	13.9

Table 3. Upper bound performance of three sources of bpm estimates after adjustment for label conflict.

cause the listener either to attend to subtle differences between tracks, or to pick faster or slower at random, on the assumption that the question would be unlikely to be posed in relation to two tracks known to be the same speed. Although we cannot know in advance what proportion of listeners will choose each option, we can safely assume that, all things being equal, the proportion will be independent of the source of the bpm estimates.

As the results given in Table 4 illustrate, this method is successful in highlighting differences between the sources, with the EchoNest estimates again shown to be significantly less consistent than either other source.

6. CONCLUSIONS

This study shows how, with a suitable experiment, simple crowd sourcing of annotations can be used to evaluate algorithms such as bpm estimation. Analysis of tens of thousands of responses collected within just a few days leads to the proposal of a straightforward and robust approach to evaluation against a human groundtruth, which is both consonant with perceptions of tempo, and designed to reward the estimates most likely to be useful in practical applications. A second controlled experiment allows validation of these results without reference to any groundtruth values. Finally we outline a method to combine classification with conventional tempo estimation, which promises significant improvements over current methods. Future work includes implementing and evaluating this approach, and extending crowd sourcing to evaluate a wider range of MIR algorithms. Data collected for this study is freely available for research purposes [4] .

[4] http://users.last.fm/~mark/speedo.tgz

7. REFERENCES

[1] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proc. ISMIR*, 2011. (submitted).

[2] D. Brusca. *BPM List: A Music Reference Guide for Mobile DJs*. Lulu.com, 7 edition, 2011.

[3] M. E. P. Davies and M. D. Plumbley. Context-dependent beat tracking of musical audio. *IEEE Trans. ASLP*, 15(3):1009–1020, 2007.

[4] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, and G. Tzanetakis. An experimental comparison of audio tempo induction algorithms. *IEEE Trans. ASLP*, 14(5):1832–1844, 2006.

[5] J. Hockman and I. Fujinaga. Fast vs slow: Learning tempo octaves from user data. In *Proc. ISMIR*, 2010.

[6] C. Laurier, M. Sordo, J. Serrà, and P. Herrera. Music mood representations from social tags. In *Proc. ISMIR*, 2009.

[7] E. Law, K. West, M. Mandel, M. Bay, and J.S. Downie. Evaluation of algorithms using games: The case of music tagging. In *Proc. ISMIR*, 2009.

[8] M. McKinney and D. Moelants. Deviations from the resonance theory of tempo induction. In *Proceedings of the Conference on Interdisciplinary Musicology*, 2004.

[9] D. Moelants and M. McKinney. Tempo perception and musical content: What makes a piece slow, fast, or temporally ambiguous? In *Proc. ICMPC*, 2004.

[10] L. Xiao, A. Tian, W. Li, and J. Zhou. Using a statistic model to capture the association between timbre and perceived tempo. In *Proc. ISMIR*, 2008.