CURRENT CHALLENGES IN THE EVALUATION OF PREDOMINANT MELODY EXTRACTION ALGORITHMS

Justin Salamon
Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
justin.salamon@upf.edu

Julián Urbano
Department of Computer Science
University Carlos III of Madrid, Leganés, Spain
jurbano@inf.uc3m.es

ABSTRACT

In this paper we analyze the reliability of the evaluation of Audio Melody Extraction algorithms. We focus on the procedures and collections currently used as part of the annual Music Information Retrieval Evaluation eXchange (MIREX), which has become the de-facto benchmark for evaluating and comparing melody extraction algorithms. We study several factors: the duration of the audio clips, time offsets in the ground truth annotations, and the size and musical content of the collection. The results show that the clips currently used are too short to predict performance on full songs, highlighting the paramount need to use complete musical pieces. Concerning the ground truth, we show how a minor error, specifically a time offset between the annotation and the audio, can have a dramatic effect on the results, emphasizing the importance of establishing a common protocol for ground truth annotation and system output. We also show that results based on the small ADC04, MIREX05 and INDIAN08 collections are unreliable, while the MIREX09 collections are larger than necessary. This evidences the need for new and larger collections containing realistic music material, for reliable and meaningful evaluation of Audio Melody Extraction.

1. INTRODUCTION

The task of melody extraction has received growing attention from the research community in recent years [4–7, 10–12]. Also referred to as Audio Melody Extraction, Predominant Melody Extraction, Predominant Melody Estimation or Predominant Fundamental Frequency (F0) Estimation, the task involves automatically obtaining a sequence of frequency values representing the pitch of the main melodic line from the audio signal of a polyphonic piece of music. As the number of researchers working on the task grew, so did the need for proper means of evaluating and comparing the performance of different algorithms. In 2004, the first Audio Description Contest (ADC) was hosted by the Music Technology Group at Universitat Pompeu Fabra in Barcelona, Spain. This initiative later evolved into the Music Information Retrieval Evaluation eXchange (MIREX) [3], which is held annually in conjunction with the ISMIR conference.

MIREX has become the de-facto benchmark for evaluating and comparing the performance of melody extraction algorithms, with over 50 algorithms evaluated since the first run in ADC 2004. Whilst this is without doubt an indication of the formalization of the topic as an established research area, it has recently been argued that some of the evaluation procedures employed by the Music Information Retrieval (MIR) research community still lack the rigor found in other disciplines such as Text IR [13]. In this paper we examine the evaluation of melody extraction algorithms, as currently carried out in the MIREX Audio Melody Extraction (AME) task. We focus on three aspects of the evaluation: first, we examine the annotation procedure used for generating a ground truth for evaluation. Specifically, we study the influence of a systematic error in the annotations, in the form of a fixed time offset between the ground truth annotation and the output of the algorithms. This issue is particularly relevant, as such an error has actually been detected in past MIREX AME evaluations. Next, we consider the duration of the audio excerpts (clips) used for evaluation. Currently all collections used for evaluation are comprised of short excerpts taken from full songs. The underlying assumption is that performance on a short clip is a good predictor for performance on a full song. However to date this assumption has neither been confirmed nor confuted. Finally, we consider the aspect of collection size. Currently, the size of most collections used for AME evaluation is relatively small compared to collections used in other IR tasks, and so we assess whether this presents any problems or not. Through these factors, we aim to assess the reliability of the evaluation procedure, as well as the meaningfulness of the results and the conclusions that are drawn from them.

The remainder of the paper is as follows. In Section 2 we explain the current evaluation procedure for AME algorithms. Section 3 takes a closer look at the annotation procedure, assessing the potential influence of a systematic error in the annotation process. In Section 4 we study the relationship between system performance and clip duration. In Section 5 we consider the influence of the size of the music collection used for evaluation. Then, in Section 6 we provide further insight into the results obtained in the previous sections, and finally we present the conclusions in Section 7.

2. MELODY EXTRACTION EVALUATION

We start by describing the current procedure for evaluating melody extraction algorithms, as carried out in the yearly MIREX AME evaluation.

2.1 Ground Truth Annotation

The ground truth for each audio excerpt is generated using the following procedure: first, the annotator must acquire the audio track containing just the melody of the excerpt. This is done by using multitrack recordings for which the separate tracks are available. Given the melody track, the pitch of the melody is estimated using a monophonic pitch tracker with a graphical user interface such as SMSTools [1] or WaveSurfer [2], producing an estimate of the fundamental frequency (F0) of the melody in every frame. This annotation is then manually inspected and corrected in cases of octave errors (double or half frequency) or when pitch is detected in frames where the melody is not present (unvoiced frames). Finally, the estimated frequency sequence is saved into a file with two columns - the first containing the time-stamp of every frame, starting from time 0, and the second the value of the fundamental frequency in Hertz. In ADC 2004 a hop size of 5.8 ms was used for the annotation, and since 2005 a hop size of 10 ms between frames is used. Frames in which there is no melody present are labelled with 0 Hz.

2.2 Evaluation Measures

An algorithm's output for a single excerpt is evaluated by comparing it to the ground truth annotation on a frame-by-frame basis, and computing five measures which summarize its performance for the complete excerpt. For a full music collection, these five measures are computed per excerpt and then averaged over the entire collection. To facilitate the evaluation, algorithms are required to provide the output in the same format as the ground truth. The only difference between the algorithm's output and the ground truth annotation is that for frames estimated as unvoiced (i.e. no melody present) by the algorithm, the algorithm may return either 0 Hz (as in the ground truth) or a negative frequency value. The negative value represents the algorithm's pitch estimation in case its voicing estimation is wrong and the melody is actually present in that frame. This allows us to separate two different aspects in the evaluation - the algorithm's voicing estimation (determining when the melody is present and when it is not) and the algorithm's pitch estimation (determining the F0 of the melody). The five evaluation measures currently employed in MIREX, as defined in [11], are summarized in Table 1.

2.3 Music Collections

Over the years, efforts by different researchers/groups have been made to generate annotated music collections for AME evaluation. The combination of the limited amount of multitrack recordings freely available, and the time-consuming annotation process, means most of these collections are quite small compared to those used in other MIR disciplines. In Table 2 we provide a summary of the music collections used in MIREX for AME evaluation since 2009.

[1] http://mtg.upf.edu/technologies/sms

[2] http://www.speech.kth.se/wavesurfer/

Voicing Recall Rate: the proportion of frames labeled as voiced in the ground truth that are estimated as voiced by the algorithm.
Voicing False Alarm Rate: the proportion of unvoiced frames in the ground truth that are estimated as voiced by the algorithm.
Raw Pitch Accuracy: the proportion of voiced frames in the ground truth for which the F0 estimated by the algorithm is within $\pm\frac{1}{4}$ tone (50 cents) of the ground truth annotation.
Raw Chroma Accuracy: same as the raw pitch accuracy, except that both the estimated and ground truth F0 sequences are mapped into a single octave, in this way ignoring octave errors in the estimation.
Overall Accuracy: combines the performance of the pitch estimation and voicing detection to give an overall performance score. Defined as the proportion of frames (out of the entire excerpt) correctly estimated by the algorithm, i.e. unvoiced frames that are labeled as unvoiced and voiced frames with a correct pitch estimate.

Table 1. AME evaluation measures used in MIREX.

Collection	Description
ADC2004	20 excerpts of roughly 20s in the genres of pop, jazz and opera. Includes real recordings, synthesized singing and audio generated from MIDI files. Total play time: 369s.
MIREX05	25 excerpts of 10-40s duration in the genres of rock, R&B, pop, jazz and solo classical piano. Includes real recordings and audio generated from MIDI files. Total play time: 686s.
INDIAN08	Four 1 minute long excerpts from north Indian classical vocal performances. There are two mixes per excerpt with differing amounts of accompaniment resulting in a total of 8 audio clips. Total play time: 501s.
MIREX09	374 Karaoke recordings of Chinese songs (i.e. recorded singing with karaoke accompaniment). Each recording is mixed at three different levels of signal-to-accompaniment ratio {-5dB, 0dB, +5dB} resulting in a total of 1,122 audio clips. Total play time: 10,022s.

Table 2. Test collections for AME evaluation in MIREX.

3. GROUND TRUTH ANNOTATION OFFSET

In this section we study the influence of a specific type of systematic error in the annotation on the results. Whilst there are other aspects of the annotation process that are also worth consideration, we find this issue to be of particular interest, since it was actually identified recently in one of the music collections used for Audio Melody Extraction evaluation in MIREX.

As explained in the previous section, all AME evaluation measures are based on a frame-by-frame comparison of the algorithm's output to the ground truth annotation. Hence, if there is a time offset between the algorithm's output and the ground truth annotation, this will cause a mismatch in all frames. Since melody pitch tends to be continuous, a very small time offset may not be noticed. However, as we increase the offset between the two sequences, we expect it to have an increasingly detrimental effect on the results.

To evaluate the effect of such an offset, we compiled a collection of 30 music clips from publicly available MIREX training sets: 10 from ADC 2004, 9 similar to MIREX05 and 11 similar to MIREX09. We used the ground truth annotations generated by the original authors of each collection, and ensured that the first frame of each annota-

tion was centered on time 0. For evaluation, we use the output of six different melody extraction algorithms that were kindly provided by their authors: KD [4], DR [3] [5], FL [6], HJ [7], RP [9] and SG [12]. For each algorithm, we computed the mean raw pitch and overall accuracy for the entire collection, as a function of a fixed time offset introduced in the ground truth annotation, from -50 ms to 50 ms using 1 ms steps. To emulate offsets smaller than the hop size of the annotation (10 ms), the ground truth was upsampled using linear interpolation.

3.1 Results

In Figure 1 we display the results of the evaluation, where we have subtracted from all values the score at offset 0. In this way, the graph reflects the absolute difference between the score at a given offset and the optimal score of the algorithm (assuming it is centered on time 0). Plot (a) contains the results for the raw pitch measure, and plot (b) for the overall accuracy.

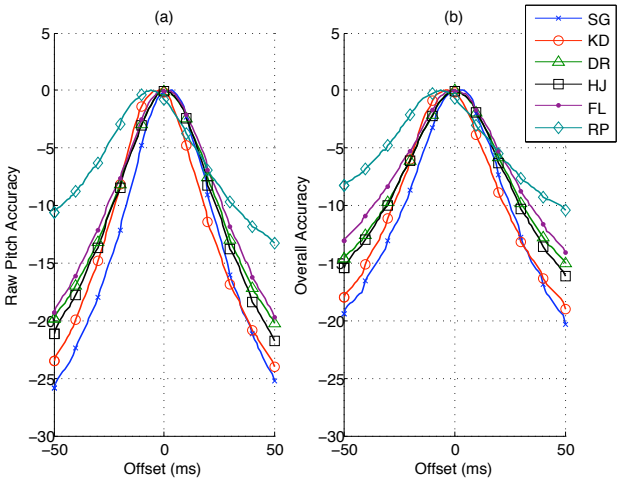

Figure 1. Absolute performance drop versus annotation offset: (a) raw pitch accuracy, (b) overall accuracy.

As can be seen, the effect of the offset is quite dramatic, causing an absolute drop of up to 25% in the raw pitch accuracy and 20% in the overall accuracy for the most extreme offset evaluated (50 ms). Though a 50 ms offset is perhaps an exaggerated case, in 2011 it was discovered that one of the MIREX collections had a 20ms offset. In our evaluation, a 20 ms offset would cause the most affected algorithms to loose 17% in raw pitch accuracy, and 13% in overall accuracy. Another interesting observation is that some algorithms do not perform best at offset 0 (most visibly RP, whose peak performance is at -6 ms). This emphasizes the fact that it does not suffice for the annotation to be centered on time 0, but rather, that there must be a strict convention to which both the annotations and algorithms adhere. Finally, we found there is a correlation between absolute performance and the effect of annotation offset: the higher the absolute performance of the algorithm, the more sensitive it is to an offset in the annotation. This is particularly important, since it suggests that the best algorithms are those who will be most affected by this type of systematic error.

[3] The output was computed using a different implementation than that of the paper, available at: https://github.com/wslihgt/separateLeadStereo

4. CLIP DURATION

A common criticism of evaluation in MIR, and particularly in MIREX, is the use of clips instead of full songs. One might argue that the use of clips is unrealistic and that observed performance on those clips may be very different from performance on full songs [13]. The collections used in the AME evaluation contain some very short excerpts, some only 10 seconds long. The use of such small clips is especially striking in AME: these clips contain primarily voiced frames, and so the generalization of the results to full songs should be questioned. We designed an experiment to assess the effect of clip duration on the reliability of the AME evaluations.

For each of the 30 clips used in the previous experiment (referred to as the *x1* clips), we created a series of subclips: 2 subclips of half the duration, 3 subclips of one third of the duration, and 4 subclips of one forth of the duration (referred to as the *x1/2*, *x1/3* and *x1/4* subclips). Note that the *x1/4* subclips can also be considered as *x1/2* versions of the *x1/2* subclips. This gives us 180 *x1/2* subclips, 90 *x1/3* subclips and 120 *x1/4* subclips, all of which were used to evaluate the six algorithms. We computed the performance difference between all subclips and their corresponding *x1* versions, leading to a grand total of 2340 data-points.

4.1 Results

In Figure 2 we show the log-scaled distribution of relative performance differences. Mean differences vary between 13% and 21% for overall accuracy and raw pitch, while for voicing false-alarm the means are around 50%. We note that there is a large amount of outliers in the distributions. However, these outliers were not found to correspond to particular songs or algorithms (they are rather randomly distributed). There seems to be a clear correlation: the shorter the subclips, the larger the performance differences (all significant by a 1-tailed Wilcoxon test, α=0.01). In principle, therefore, one would want the clips used for evaluation to be as long as possible; ideally, the full songs.

In Figure 3 we plot the log-scaled relative performance differences in overall accuracy, this time as a function of the log-scaled actual subclip duration (other measures produce very similar plots). We see that the negative correlation between subclip duration and performance difference appears to be independent of the duration of the *x1* clip. We fitted a non-linear model of the form $diff = a \cdot duration^b$, where a and b are the parameters to fit, to the results of each of the relative durations (*x1/2*, *x1/3*, *x1/4*), and as the plot shows, they are very similar. In fact, an ANCOVA analysis revealed no significant difference between them. This suggests that the error decreases as the clip duration increases, regardless of the duration of the full song.

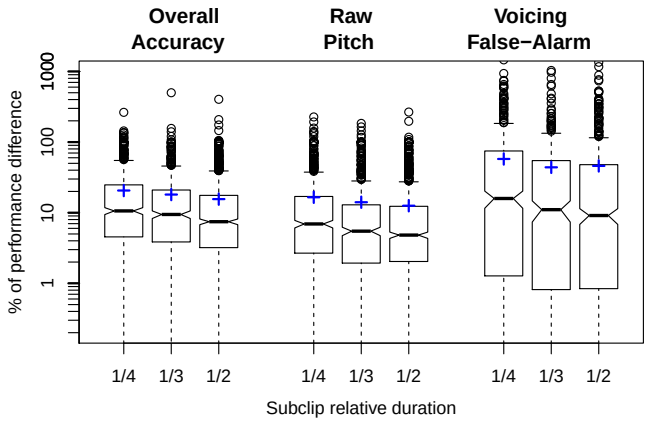

Figure 2. Releative performance differences between subclips and their corresponding *x1* clips. Blue crosses mark the means of the distributions.

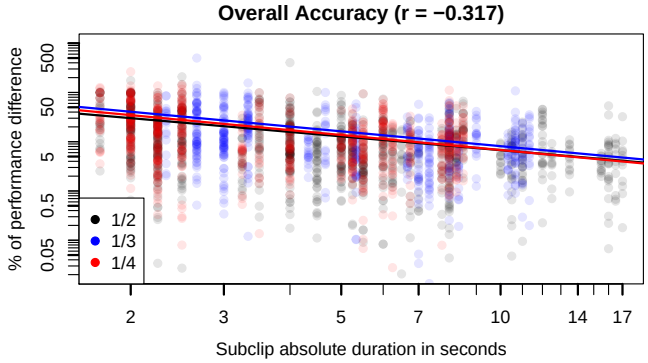

Figure 3. Relative performance differences with subclips as a function of subclip actual duration.

5. COLLECTION SIZE

Regardless of the effectiveness measure used, an AME experiment consists of evaluating a set of algorithms $\mathcal{A}$ using a set of songs $\mathcal{S}$. Such an evaluation experiment can be viewed as fitting the following model:

$$y_{as} = \overline{y} + \overline{y}_a + \overline{y}_s + \varepsilon_{as} \quad (1)$$

where y_{as} is the score of algorithm a for song s, $\overline{y}$ is the grand average score of all possible algorithms over all possible songs, $\overline{y}_a$ is the algorithm effect (the average deviation of algorithm a from the grand average $\overline{y}$), $\overline{y}_s$ is the song effect and ε_{as} is a residual modeling the particular deviation of algorithm a for song s. In our case, where we do not consider other effects such as annotators, this ε_{as} residual actually models the algorithm-song interaction effect: some algorithms are particularly better (or worse) for particular songs.

When a researcher carries out an AME evaluation experiment, they evaluate how well an algorithm performs for the set $\mathcal{S}$ of songs, but ideally they want to generalize from the performance of that specific experiment to the average score the algorithm would obtain for the population of all songs represented by the sample $\mathcal{S}$, not just the sample itself. The reliability when drawing such general conclusions based on the observations on samples (test collections) can be measured with Generalizability Theory (GT) [1, 2].

From the model in Eq. 1 we can identify two sources of variability in the observed scores: actual performance differences among algorithms and difficulty differences among songs. Ideally, we want most of the variability in y_{as} to be due to the algorithm effect, that is, the observed effectiveness differences to be due to actual differences between algorithms and not due to other sources of variability such as songs, annotators, or specific algorithm-song interactions. Note that this does not mean a collection should not contain varied musical content. Ideally, we want an algorithm to work well for all types of musical material, and hence a varied collection in terms of content does not necessarily imply large performance variability due to the song effect. However, a small collection that contains songs with a great degree of variability (in terms of difficulty) is likely to result in performance variability that is dominated by the song effect and possibly by algorithm-song interactions (e.g. algorithm X is especially good for jazz but poor for rock), thus reducing our ability to claim that the observed differences between the algorithms can be generalized to the universe of all songs. Using GT [1, 2], we can measure the proportion of observed variability that is due to actual differences between the algorithms. This proportion reflects the stability of the evaluation, and as such it is also a measure of efficiency: the higher the stability, the fewer the songs necessary to reliably evaluate algorithms [1, 8]. GT does not only help evaluate the stability of past collections, but also estimate the reliability of yet-to-be created collections as a function of their size. However, the results of GT only hold if the original data used for the analysis is *representative* of the wider population of songs to which we want to generalize in the future.

5.1 Variance Analysis and Collection Stability

In the model in Eq. 1, the grand mean $\overline{y}$ is a constant, and the other effects can be modeled as random variables with their own expectation and variance. As such, the variance of the observed scores is modeled as the sum of these variance components:

$$\sigma^2 = \sigma_a^2 + \sigma_s^2 + \sigma_{as}^2 \quad (2)$$

where σ_a^2 is the variance due to the algorithm effect, σ_s^2 is the variance due to the song effect, and σ_{as}^2 is the variance due to the algorithm-song interaction effect (the residual). This variance decomposition can be estimated by fitting a fully-crossed ANOVA model for Eq. 1:

$$\widehat{\sigma}_{as}^2 = \mathrm{E}MS_{as} = \mathrm{E}MS_{residual}$$
$$\widehat{\sigma}_a^2 = \frac{EMS_a - \widehat{\sigma}_{as}^2}{|\mathcal{S}|} \quad , \quad \widehat{\sigma}_s^2 = \frac{\mathrm{E}MS_s - \widehat{\sigma}_{as}^2}{|\mathcal{A}|} \quad (3)$$

where $\mathrm{E}MS_x$ is the expected Mean Square of component x. In practice, $\mathrm{E}MS_x$ is approximated by the Mean Square of component x as computed with the ANOVA model [1, 2]. Using the estimates in Eq. 3 we can estimate the proportion of variability due to the algorithm effect as per Eq. 2. The stability of the evaluation can then be quantified with the dependability index Φ:

	Overall Accuracy				Raw Pitch				Voicing False-Alarm			
	$\widehat{\sigma}_a^2$	$\widehat{\sigma}_s^2$	$\widehat{\sigma}_{as}^2$	$\widehat{\Phi}$	$\widehat{\sigma}_a^2$	$\widehat{\sigma}_s^2$	$\widehat{\sigma}_{as}^2$	$\widehat{\Phi}$	$\widehat{\sigma}_a^2$	$\widehat{\sigma}_s^2$	$\widehat{\sigma}_{as}^2$	$\widehat{\Phi}$
ADC04	27%	27%	46%	.879	23%	28%	49%	.859	55%	21%	23%	.961
MIREX05	11%	47%	42%	.758	15%	54%	31%	.817	57%	20%	23%	.971
INDIAN08	16%	50%	34%	.600	24%	57%	19%	.721	70%	13%	16%	.950
04 + 05 + 08	16%	39%	45%	.909	16%	43%	41%	.912	56%	21%	23%	.986
MIREX09 0dB	52%	20%	28%	.998	50%	20%	31%	.997	81%	5%	14%	.999
MIREX09 -5dB	40%	23%	37%	.996	40%	24%	35%	.996	82%	5%	13%	.999
MIREX09 +5dB	58%	17%	26%	.998	48%	18%	34%	.997	83%	4%	14%	.999

Table 3. Variance components and $\widehat{\Phi}$ score for all three measures and all six collections plus the joint 04+05+08 collection.

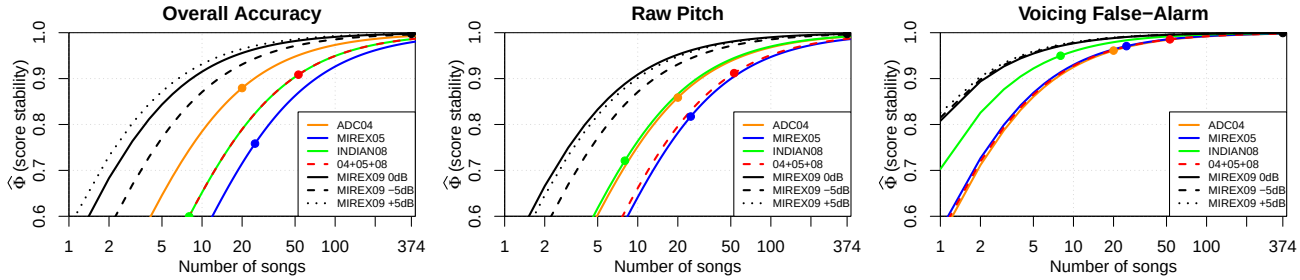

Figure 4. Dependability index as a function of the number of songs for Overall Accuracy (left), Raw Pitch (middle) and Voicing False-Alarm (right). The points mark the actual number of songs per collection.

$$\Phi = \frac{\sigma_a^2}{\sigma_a^2 + \frac{\sigma_s^2 + \sigma_{as}^2}{|\mathcal{S}|}} \tag{4}$$

which measures the ratio between algorithm variance and the variance in absolute effectiveness scores (total variance) [1, 2]. This measure increases with the song set size (i.e. with an infinite number of songs all the observed variability would be due to algorithm differences) [8].

5.2 Results

In Table 3 we show the estimated proportion of variability due to the algorithm, song and algorithm-song interaction effects. For these calculations we used the results of the MIREX campaign directly, combining the results of the five algorithms from MIREX 2010 and ten algorithms from MIREX 2011. In both years the same six test-collections were used for evaluation, so we can consider the grouping of algorithms from both years as a single larger evaluation round leading to a fully crossed experimental design. We also joined the three smaller collections into a single larger one referred to as "04+05+08", discussed in Section 6.

In general, it can be seen that the estimated variance due to the algorithm effect is much larger in the MIREX09 collections. For overall accuracy, the average is 50%, while for the earlier collections it is just 18%, and as low as 11% for MIREX05. These differences show that generalizations of results based on the earlier collections are not very reliable, especially in the case of the MIREX05 and INDIAN08 collections, because a large part of the variability in the scores is due to the song characteristics rather than differences between the algorithms.

Figure 4 shows the estimated dependability index as a function of the number of songs used (log scaled). The points mark the value of $\widehat{\Phi}$ for the actual number of songs in each collection (cf. Table 3). Again we observe that the MIREX09 collections are considerably more stable than the earlier collections, especially MIREX05 and INDIAN08, where $\widehat{\Phi}$ is as low as 0.6. More interesting is the fact that the dependability index in the MIREX09 collections rapidly converges to 1, and there is virtually no appreciable difference between using all 374 songs in the collection or just 100: $\widehat{\Phi}$ would only drop from an average of 0.997 to 0.990, showing that most of the variability in performance scores would still be attributable to the algorithm effect. However, we must also consider the content validity of this collection (i.e. whether it is representative or not) [13]. We discuss this in the next section.

6. DISCUSSION

Starting with the annotation offset issue, we note that there are two crucial parameters that must be fixed in order to prevent this problem: the precise time of the first frame, and the hop size. Since 2005, all the annotations use a hop size of 10 ms, and all algorithms are required to use this hop size for their output. However, the exact time of the first frame has not been explicitly agreed upon by the community. When the short-time Fourier transform (or any other transform which segments the audio signal into short frames) is used, it is common practice to consider the time-stamp of each frame to be the time exactly at the middle of the frame. Thus, for the first frame to start exactly at time zero, it must be centered on the first sample of the audio (filling the first half of the frame with zeros). Nonetheless, while this is common practice, it is not strictly imposed, meaning algorithms and annotators might, rather than center the first frame on the first sample, start the frame at this sample. In this case, the frame will not be centered on time zero, but rather on an arbitrary time which depends on the length of the frame. Since different algorithms and annotations use different frame sizes, this scenario could lead to a different fixed offset between every algorithm and every

annotation, leading to a systematic error in the evaluation.

In terms of clip duration, we saw that there is a clear correlation between the relative duration of the clip (compared to the full song) and evaluation error, suggesting that performance based on clips might not really predict performance on full songs. However, Figure 3 suggests that this correlation is independent of the actual duration of the full song. That is, there might be a duration threshold of x seconds for which observed performance on clips does predict performance on full songs (within some error rate), no matter how long they are. While counter-intuitive at first, this result does somehow agree with general statistical theory. How large a sample needs to be in order to reliably estimate unknown parameters of the underlying population, is independent of how large the population actually is, as long as the sample is *representative* of the population. This usually requires to sample randomly or follow other techniques such as systematic or stratified sampling. For AME evaluation it does not make sense to randomly sample frames of a song, but the results suggest that there might be a sampling technique such that audio clips, if selected appropriately, can be representative of the full songs.

Regarding the collection size, we observed that the earlier ADC04, MIREX05 and INDIAN08 collections are unstable because a larger proportion of the variability in the observed performance scores is due to song difficulty differences rather than algorithm differences. As such, results from these collections alone are expected to be unstable, and therefore evaluations that rely solely on *one* of these collections are not very reliable. In Table 3 (and Figure 4) we see that by joining these collections into a single larger one ("04+05+08") the evaluation results are considerably more stable ($\widehat{\Phi} > 0.9$ for all three measures), and so we recommend fusing them into a single collection for future evaluations. On the other hand, we saw that the MIREX09 collections are in fact much larger than necessary: about 25% of the current songs would suffice for results to be highly stable and therefore generalize to a wider population of songs. However, all MIREX09 music material consists of Chinese karaoke songs with non-professional singers, and therefore we should expect the results to generalize to *this* population of songs, but not to the general universe of *all* songs (essentially everything that is not karaoke). Therefore, the AME community is found in the situation where the collections with sufficiently varied music material are too small to be reliable, while the ones that are reliable contain very biased music material.

7. CONCLUSION

In this paper we analyzed the reliability of the evaluation of Audio Melody Extraction algorithms, as performed in MIREX. Three main factors were studied: ground truth annotations, clip duration and collection size. We demonstrated how an offset between the ground truth and an algorithm's output can significantly degrade the results, the solution to which is the definition and adherence to a strict protocol for annotation. Next, it was shown that the clips currently used are too short to predict performance on full songs, stressing the need to use complete musical pieces. It was also shown that results based on one of the ADC04, MIREX05 or INDIAN08 collections alone are not reliable due to their small size, while the MIREX09 collection, though more reliable, does not reflect real-world musical content. The above demonstrates that whilst the MIREX AME evaluation task is an important initiative, it currently suffers from problems which require urgent attention. As a solution, we propose the creation of a new and open test collection through a joint effort of the research community. If the collection is carefully compiled and annotated, keeping in mind the issues mentioned here, it should, in theory, solve all of the aforementioned problems that current AME evaluation suffers from. Furthermore, we could consider the application of low-cost evaluation methodologies that dramatically reduce the annotation effort required [14]. Finally, in the future it would also be worth studying the appropriateness of the evaluation measures themselves, the accuracy of the manual ground truth annotations and further investigate the effect of clip duration.

8. ACKNOWLEDGMENTS

We would like to thank the authors of the melody extraction algorithms for their contribution to our experiments. This work was supported by the Programa de Formación del Profesorado Universitario (FPU) and grant TIN2010-22145-C02-02 of the Spanish Government.

9. REFERENCES

[1] D. Bodoff. Test theory for evaluating reliability of IR test collections. *Inf. Process. Manage.*, 44(3), 2008.

[2] R. L. Brennan. *Generalizability Theory*. Springer, 2001.

[3] J. Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 2008.

[4] K. Dressler. Audio melody extraction for mirex 2009. In *Music Inform. Retrieval Evaluation eXchange (MIREX)*, 2009.

[5] J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE TASLP*, 18(3), 2010.

[6] B. Fuentes, A. Liutkus, R. Badeau, and G. Richard. Probabilistic model for main melody extraction using constant-Q transform. In *IEEE ICASSP*, 2012.

[7] C. Hsu, D. Wang, and J. Jang. A trend estimation algorithm for singing pitch detection in musical recordings. In *IEEE ICASSP*, 2011.

[8] E. Kanoulas and J. Aslam. Empirical justification of the gain and discount function for nDCG. In *ACM CIKM*, 2009.

[9] R. P. Paiva. *Melody Detection in Polyphonic Audio*. PhD thesis, University of Coimbra, Portugal, 2007.

[10] R. P. Paiva, T. Mendes, and A. Cardoso. Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness. *Comput. Music J.*, 2006.

[11] G. E. Poliner, D. P. W. Ellis, F. Ehmann, E. Gómez, S. Steich, and B. Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE TASLP*, 15(4), 2007.

[12] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE TASLP*, 20(6), 2012.

[13] J. Urbano. Information retrieval meta-evaluation: Challenges and opportunities in the music domain. In *ISMIR*, 2011.

[14] J. Urbano and M. Schedl. Towards Minimal Test Collections for Evaluation of Audio Music Similarity and Retrieval. In *WWW Workshop on Advances in MIR*, 2012.