TOWARDS EVALUATING MULTIPLE PREDOMINANT MELODY ANNOTATIONS IN JAZZ RECORDINGS

Stefan Balke[1] **Jonathan Driedger**[1] **Jakob Abeßer**[2]
Christian Dittmar[1] **Meinard Müller**[1]
[1] International Audio Laboratories Erlangen, Germany
[2] Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany
stefan.balke@audiolabs-erlangen.de

ABSTRACT

Melody estimation algorithms are typically evaluated by separately assessing the task of voice activity detection and fundamental frequency estimation. For both subtasks, computed results are typically compared to a single human reference annotation. This is problematic since different human experts may differ in how they specify a predominant melody, thus leading to a pool of equally valid reference annotations. In this paper, we address the problem of evaluating melody extraction algorithms within a jazz music scenario. Using four human and two automatically computed annotations, we discuss the limitations of standard evaluation measures and introduce an adaptation of Fleiss' kappa that can better account for multiple reference annotations. Our experiments not only highlight the behavior of the different evaluation measures, but also give deeper insights into the melody extraction task.

1. INTRODUCTION

Predominant melody extraction is the task of estimating an audio recording's fundamental frequency trajectory values (F0) over time which correspond to the melody. For example in classical jazz recordings, the predominant melody is typically played by a soloist who is accompanied by a rhythm section (e. g., consisting of piano, drums, and bass). When estimating the soloist's F0-trajectory by means of an automated method, one needs to deal with two issues: First, to determine the time instances when the soloist is active. Second, to estimate the course of the soloist's F0 values at active time instances.

A common way to evaluate such an automated approach—as also used in the Music Information Retrieval Evaluation eXchange (MIREX) [5]—is to split the evaluation into the two subtasks of activity detection and F0 estimation. These subtasks are then evaluated by comparing the computed results to a single manually created reference annotation. Such an evaluation, however, is problematic since it assumes the existence of a single ground-truth. In practice, different humans may annotate the same recording in different ways thus leading to a low inter-annotator agreement. Possible reasons are the lack of an exact task specification, the differences in the annotators' experiences, or the usage of different annotation tools [21, 22]. Figure 1 exemplarily illustrates such variations on the basis of three annotations $A_1, \ldots, A_3$ of the same audio recording, where a soloist plays three consecutive notes. A first observation is that A_1 and A_2 have a fine frequency resolution which can capture fluctuations over time (e. g., vibrato effects). In contrast, A_3 is specified on the basis of semitones which is common when considering tasks such as music transcription. Furthermore, one can see that note onsets, note transitions, and durations are annotated inconsistently. Reasons for this might be differences in annotators' familiarity with a given instrument, genre, or a particular playing style. In particular, annotation deviations are likely to occur when notes are connected by slurs or glissandi.

Inter-annotator disagreement is a generally known problem and has previously been discussed in the contexts of audio music similarity [8, 10], music structure analysis [16, 17, 23], and melody extraction [3]. In general, a

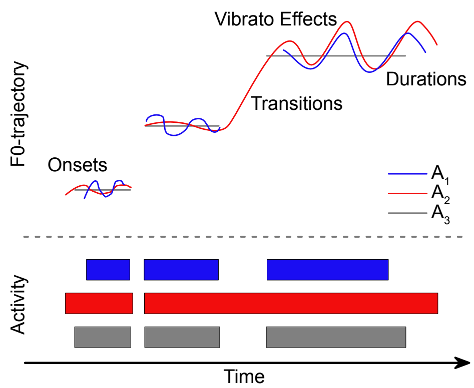

Figure 1. Illustration of different annotations and possible disagreements. A_1 and A_2 are based on a fine frequency resolution. Annotation A_3 is based on a coarser grid of musical pitches.

SoloID	Performer	Title	Instr.	Dur.
`Bech-ST`	Sidney Bechet	Summertime	Sopr. Sax	197
`Brow-JO`	Clifford Brown	Jordu	Trumpet	118
`Brow-JS`	Clifford Brown	Joy Spring	Trumpet	100
`Brow-SD`	Clifford Brown	Sandu	Trumpet	048
`Colt-BT`	John Coltrane	Blue Train	Ten. Sax	168
`Full-BT`	Curtis Fuller	Blue Train	Trombone	112
`Getz-IP`	Stan Getz	The Girl from Ipan.	Ten. Sax	081
`Shor-FP`	Wayne Shorter	Footprints	Ten. Sax	139

Table 1. List of solo excerpts taken from the WJD. The table indicates the performing artist, the title, the solo instrument, and the duration of the solo (given in seconds).

single reference annotation can only reflect a subset of the musically or perceptually valid interpretations for a given music recording, thus rendering the common practice of evaluating against a single annotation questionable.

The contributions of this paper are as follows. First, we report on experiments, where several humans annotate the predominant F0-trajectory for eight jazz recordings. These human annotations are then compared with computed annotations obtained by automated procedures (MELODIA [20] and pYIN [13]) (Section 2). In particular, we consider the scenario of soloist activity detection for jazz recordings (Section 3.1). Afterwards, we adapt and apply an existing measure (Fleiss' Kappa [7]) to our scenario which can account for jointly evaluating multiple annotations (Section 3.2). Note that this paper has an accompanying website at [1] where one can find the annotations which we use in the experiments.

2. EXPERIMENTAL SETUP

In this work, we use a selection of eight jazz recordings from the *Weimar Jazz Database* (WJD) [9, 18]. For each of these eight recordings (see Table 1), we have a pool of seven annotations $\mathcal{A} = \{A_1, \ldots, A_7\}$ which all represent different estimates of the predominant solo instruments' F0-trajectories. In the following, we model an annotation as a discrete-time function $A : [1 : N] \to \mathbb{R} \cup \{*\}$ which assigns to each time index $n \in [1 : N]$ either the solo's F0 at that time instance (given in Hertz), or the symbol '$*$'. The meaning of $A(n) = *$ is that the soloist is inactive at that time instance.

In Table 2, we list the seven annotations. For this work, we manually created three annotations $A_1, \ldots, A_3$ by using a custom graphical user interface as shown in Figure 2 (see also [6]). In addition to standard audio player functionalities, the interface's central element is a salience spectrogram [20]—an enhanced time-frequency representation with a logarithmically-spaced frequency axis. An annotator can indicate the approximate location of F0-trajectories in the salience spectrogram by drawing *constraint regions* (blue rectangles). The tool then automatically uses techniques based on *dynamic programming* [15] to find a plausible trajectory through the specified region. The annotator can then check the annotation by listening to the solo recording, along with a synchronized sonification of the F0-trajectory.

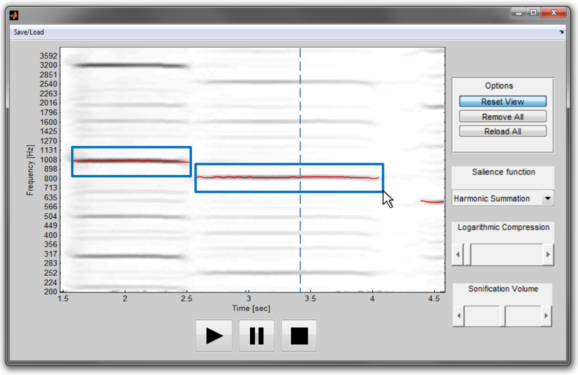

Figure 2. Screenshot of the tool used for the manual annotation of the F0 trajectories.

Annotation	Description
A_1	Human 1, F0-Annotation-Tool
A_2	Human 2, F0-Annotation-Tool
A_3	Human 3, F0-Annotation-Tool
A_4	Human 4, WJD, Sonic Visualiser
A_5	Computed, MELODIA [2, 20]
A_6	Computed, pYIN [13]
A_7	Baseline, all time instances active at 1 kHz

Table 2. Set $\mathcal{A}$ of all annotations with information about their origins.

In addition to the audio recordings, the WJD also includes manually annotated solo transcriptions on the semitone level. These were created and cross-checked by trained jazz musicians using the *Sonic Visualiser* [4]. We use these solo transcriptions to derive A_4 by interpreting the given musical pitches as F0 values by using the pitches' center frequencies.

A_5 and A_6 are created by means of automated methods. A_5 is extracted by using the MELODIA [20] algorithm as implemented in Essentia [2] using the default settings (sample rate = 22050 Hz, hop size = 3 ms, window size = 46 ms). For obtaining A_6, we use the tool Tony [12] (which is based on the pYIN algorithm [13]) with default settings and without any corrections of the F0-trajectory.

As a final annotation, we also consider a baseline $A_7(n) = 1$ kHz for all $n \in [1 : N]$. Intuitively, this baseline assumes the soloist to be always active. All of these annotations are available on this paper's accompanying website [1].

3. SOLOIST ACTIVITY DETECTION

In this section, we focus on the evaluation of the *soloist activity detection* task. This activity is derived from the annotations of the F0-trajectories $A_1, \ldots, A_7$ by only considering active time instances, i. e., $A(n) \neq *$. Figure 3 shows a typical excerpt from the soloist activity annotations for the recording `Brow-JO`. Each row of this matrix shows the annotated activity for one of our annotations from Table 2. Black denotes regions where the soloist is annotated as active and white where the soloist is annotated

Ref. \ Est.	A_1	A_2	A_3	A_4	A_5	A_6	A_7	∅
A_1	–	0.93	0.98	0.92	0.74	0.79	1.00	0.89
A_2	0.92	–	0.97	0.92	0.74	0.79	1.00	0.89
A_3	0.84	0.84	–	0.88	0.69	0.74	1.00	0.83
A_4	0.85	0.86	0.94	–	0.70	0.75	1.00	0.85
A_5	0.84	0.84	0.90	0.85	–	0.77	1.00	0.87
A_6	0.75	0.76	0.81	0.77	0.65	–	1.00	0.79
A_7	0.62	0.62	0.71	0.67	0.55	0.65	–	0.64
∅	0.80	0.81	0.89	0.83	0.68	0.75	1.00	0.82

Table 3. Pairwise evaluation: *Voicing Detection* (VD). The values are obtained by calculating the VD for all possible annotation pairs (Table 2) and all solo recordings (Table 1). These values are then aggregated by using the arithmetic mean.

as inactive. Especially note onsets and durations strongly vary among the annotation, see e. g., the different durations of the note event at second 7.8. Furthermore, a missing note event is noticeable in the annotations A_1 and A_6 at second 7.6. At second 8.2, A_6 found an additional note event which is not visible in the other annotations. This example indicates that the inter-annotator agreement may be low. To further understand the inter-annotator agreement in our dataset, we first use standard evaluation measures (e. g., as used by MIREX for the task of *audio melody extraction* [14]) and discuss the results. Afterwards, we introduce Fleiss' Kappa, an evaluation measure known from psychology, which can account for multiple annotations.

3.1 Standard Evaluation Measures

As discussed in the previous section, an estimated annotation A_e is typically evaluated by comparing it to a reference annotation A_r. For the pair (A_r, A_e), one can count the number of time instances that are *true positives* $\#\mathrm{TP}$ (A_r and A_e both label the soloist as being active), the number of *false positives* $\#\mathrm{FP}$ (only A_e labels the soloist as being active), the number of *true negatives* $\#\mathrm{TN}$ (A_r and A_e both label the soloist as being inactive), and the number *false negatives* $\#\mathrm{FN}$ (only A_e labels the soloist as being inactive).

In previous MIREX campaigns, these numbers are used to derive two evaluation measures for the task of activity detection. *Voicing Detection* (VD) is identical to *Recall* and describes the ratio that a time instance which is annotated as being active is truly active according to the reference annotation:

$$\mathrm{VD} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN}} \,. \tag{1}$$

The second measure is the *Voicing False Alarm* (VFA) and relates the ratio of time instances which are inactive according to the reference annotation but are estimated as being active:

$$\mathrm{VFA} = \frac{\#\mathrm{FP}}{\#\mathrm{TN} + \#\mathrm{FP}} \,. \tag{2}$$

In the following experiments, we assume that all annotations $A_1, \ldots, A_7 \in \mathcal{A}$ have the same status, i. e., each annotation may be regarded as either reference or estimate. Then, we apply the standard measures in a pairwise fashion. For all pairs $(A_r, A_e) \in \mathcal{A} \times \mathcal{A}$ with $A_r \neq A_e$, we extract VD and VFA (using the MIR_EVAL [19] toolbox) for each of the solo recordings listed in Table 1. The mean values over the eight recordings are presented in Table 3 for the VD-measure and in Table 4 for the VFA-measure.

As for the Voicing Detection (Table 3), the values within the human annotators $A_1, \ldots, A_4$ range from 0.84 for the pair (A_3, A_2) to 0.98 for the pair (A_1, A_3). This high variation in VD already shows that the inter-annotator disagreement even within the human annotators is substantial. By taking the human annotators as reference to evaluate the automatic approach A_5, the VD lies in the range of 0.69 for (A_3, A_5) to 0.74 for (A_2, A_5). Analogously, for A_6, we observe values from 0.74 for (A_3, A_6) to 0.79 for (A_1, A_6).

As for the Voicing False Alarm (see Table 4), the values among the human annotations range from 0.05 for (A_3, A_1) to 0.30 for (A_1, A_3). Especially annotation A_3 deviates from the other human annotations, resulting in a very high VFA (having many time instances being set as active).

In conclusion, depending on which human annotation we take as the reference, the evaluated performances of the automated methods vary substantially. Having multiple potential reference annotations, the standard measures

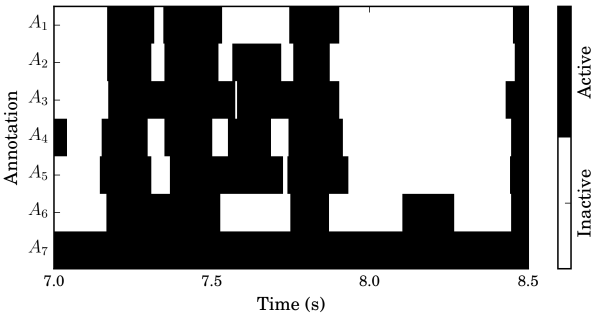

Figure 3. Excerpt from Brow-JO. $A_1, \ldots, A_4$ show the human annotations. A_5 and A_6 are results from automated approaches. A_7 is the baseline annotation which considers all frames as being active.

Ref. \ Est.	A_1	A_2	A_3	A_4	A_5	A_6	A_7	∅
A_1	–	0.13	0.30	0.27	0.22	0.44	1.00	0.39
A_2	0.12	–	0.29	0.26	0.22	0.43	1.00	0.39
A_3	0.05	0.07	–	0.14	0.18	0.43	1.00	0.31
A_4	0.16	0.16	0.27	–	0.24	0.46	1.00	0.38
A_5	0.34	0.35	0.48	0.44	–	0.49	1.00	0.52
A_6	0.38	0.38	0.54	0.49	0.35	–	1.00	0.52
A_7	0.00	0.00	0.00	0.00	0.00	0.00	–	0.00
∅	0.17	0.18	0.31	0.27	0.20	0.38	1.00	0.36

Table 4. Pairwise evaluation: *Voicing False Alarm* (VFA). The values are obtained by calculating the VFA for all possible annotation pairs (Table 2) and all solo recordings (Table 1). These values are then aggregated by using the arithmetic mean.

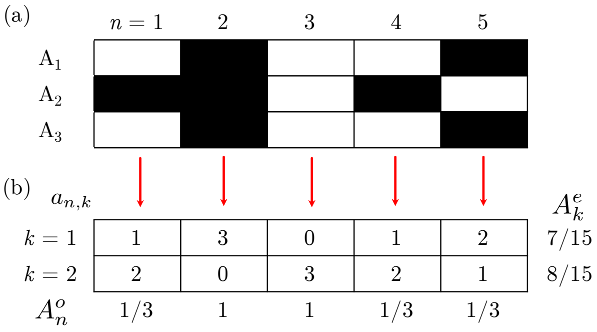

Figure 4. Example of evaluating Fleiss' κ for $K = 2$ categories, $N = 5$ frames, and three different annotations. (a) Annotations. (b) Number of annotations per category and time instance. Combining $A^o = 0.6$ and $A^e = 0.5$ leads to $\kappa = 0.2$.

< 0	$0 - 0.2$	$0.21 - 0.4$	$0.41 - 0.6$	$0.61 - 0.8$	$0.81 - 1$
poor	slight	fair	moderate	substantial	almost perfect

Table 5. Scale for interpreting κ as given by [11].

are not generalizable to take these into account (only by considering a mean over all pairs). Furthermore, although the presented evaluation measures are by design limited to yield values in $[0, 1]$, they can usually not be interpreted without some kind of baseline. For example, considering VD, the pair (A_2, A_3) yields a VD-value of 0.97, suggesting that A_3 can be considered as an "excellent" estimate. However, considering that our uninformed baseline A_7 yields a VD of 1.0, shows that it is meaningless to look at the VD alone. Similarly, an agreement with the trivial annotation A_7 only reflects the statistics on the active and inactive frames, thus being rather uninformative. Next, we introduce an evaluation measure that can overcome some of these problems.

3.2 Fleiss' Kappa

Having to deal with multiple human annotations is common in fields such as medicine or psychology. In these disciplines, measures that can account for multiple annotations have been developed. Furthermore, to compensate for chance-based agreement, a general concept referred to as *Kappa Statistic* [7] is used. In general, a kappa value lies in the range of $[-1, 1]$, where the value 1 means complete agreement among the raters, the value 0 means that the agreement is purely based on chance, and a value below 0 means that agreement is even below chance.

We now adapt *Fleiss' Kappa* to calculate the chance-corrected inter-annotator agreement for the soloist activity detection task. Following [7, 11], Fleiss' Kappa is defined as:

$$\kappa := \frac{A^o - A^e}{1 - A^e} \,. \tag{3}$$

In general, κ compares the mean observed agreement $A^o \in [0, 1]$ to the mean expected agreement $A^e \in [0, 1]$ which is solely based on chance. Table 5 shows a scale for the agreement of annotations with the corresponding range of κ.

SoloID \ Comb.	κ_H	$\kappa_{H,5}$	$\kappa_{H,6}$	ρ_5	ρ_6
`Bech-ST`	0.74	0.60	0.55	0.82	0.75
`Brow-JO`	0.68	0.56	0.59	0.82	0.87
`Brow-JS`	0.61	0.47	0.43	0.78	0.71
`Brow-SD`	0.70	0.61	0.51	0.87	0.73
`Colt-BT`	0.66	0.55	0.49	0.84	0.74
`Full-BT`	0.74	0.66	0.61	0.89	0.83
`Getz-IP`	0.72	0.69	0.64	0.96	0.90
`Shor-FP`	0.82	0.65	0.58	0.80	0.70
$\varnothing$	0.71	0.60	0.55	0.85	0.78

Table 6. κ for all songs and different pools of annotations. κ_H denotes the pool of human annotations $A_1, \ldots, A_4$. These values are then aggregated by using the arithmetic mean.

To give a better feeling for how κ works, we exemplarily calculate κ for the example given in Figure 4(a). In this example, we have $R = 3$ different annotations $A_1, \ldots, A_3$ for $N = 5$ time instances. For each time instance, the annotations belong to either of $K = 2$ categories (*active* or *inactive*). As a first step, for each time instance, we add up the annotations for each category. This yields the number of annotations per category $a_{n,k} \in \mathbb{N}$, $n \in [1 : N]$, $k \in [1 : K]$ which is shown in Figure 4(b). Based on these distributions, we calculate the observed agreement A^o_n for a single time instance $n \in [1 : N]$ as:

$$A^o_n := \frac{1}{R(R-1)} \sum_{k=1}^{K} a_{n,k}(a_{n,k} - 1) \,, \tag{4}$$

which is the fraction of agreeing annotations normalized by the number of possible annotator pairs $R(R-1)$, e. g., for the time instance $n = 2$ in the example, all annotators agree for the frame to be active, thus $A^o_2 = 1$. Taking the arithmetic mean of all observed agreements leads to the mean observed agreement

$$A^o := \frac{1}{N} \sum_{n=1}^{N} A^o_n \,, \tag{5}$$

in our example $A^o = 0.6$. The remaining part for calculating κ is the expected agreement A^e. First, we calculate the distribution of agreements within each category $k \in [1 : K]$, normalized by the number of possible ratings NR:

$$A^e_k := \frac{1}{NR} \sum_{n=1}^{N} a_{n,k} \,, \tag{6}$$

e. g., in our example for $k = 1$ (active) results in $A^e_1 = 7/15$. The expected agreement A_e is defined as [7]

$$A^e := \sum_{k=1}^{K} (A^e_k)^2 \tag{7}$$

which leads to $\kappa = 0.2$ for our example. According to the scale given in Table 5, this is a "slight" agreement.

In Table 6, we show the results for κ calculated for different pools of annotations. First, we calculate κ for the

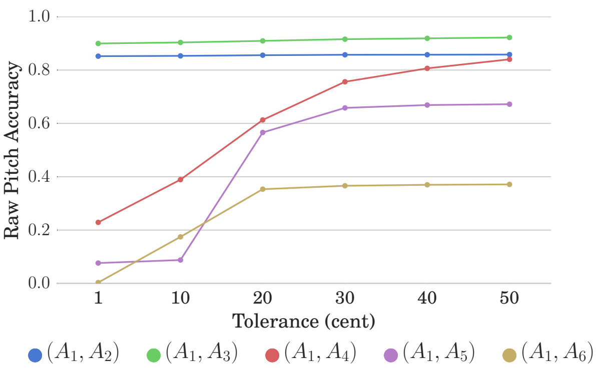

Figure 5. *Raw Pitch Accuracy* (RPA) for different pairs of annotations based on the annotations of the solo recording `Brow-JO`, evaluated on all active frames according to the reference annotation.

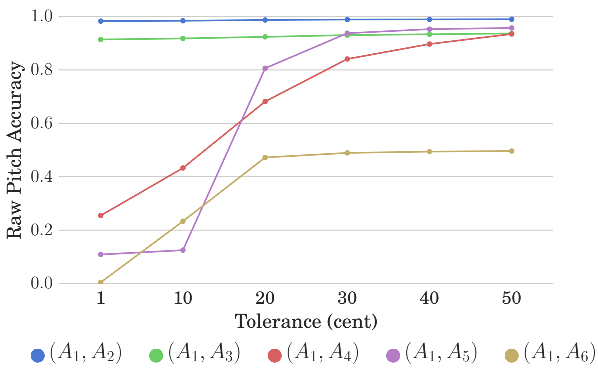

Figure 6. *Modified Raw Pitch Accuracy* for different pairs of annotations based on the annotations of the solo recording `Brow-JO`, evaluated on all active frames according to the *union* of reference and estimate annotation.

pool of human annotations $H := \{1, 2, 3, 4\}$, denoted as κ_H. κ_H yields values ranging from 0.61 to 0.82 which is considered as "substantial" to "almost perfect" agreement according to Table 5.

Now, reverting to our initial task of evaluating an automatically obtained annotation, the idea is to see how the κ-value changes when adding this annotation to the pool of all human annotations. A given automated procedure could then be considered to work correctly if it produces results that are just about as variable as the human annotations. Only if an automated procedure behaves fundamentally different than the human annotations, it will be considered to work incorrectly. In our case, calculating κ for the annotation pool $H \cup \{5\}$ yields values ranging from 0.47 to 0.69, as shown in column $\kappa_{H,5}$ of Table 6. Considering the annotation pool $H \cup \{6\}$, $\kappa_{H,6}$ results in κ-values ranging from 0.43 to 0.64. Considering the average over all individual recordings, we get mean κ-values of 0.60 and 0.55 for $\kappa_{H,5}$ and $\kappa_{H,6}$, respectively. Comparing these mean κ-values for the automated approaches to the respective κ_H, we can consider the method producing the annotation A_5 to be more consistent with the human annotations than A_6.

In order to quantify the agreement of an automatically generated annotation and the human annotations in a single value, we define the proportion $\rho \in \mathbb{R}$ as

$$\rho_5 := \frac{\kappa_{\text{H},5}}{\kappa_{\text{H}}}, \rho_6 := \frac{\kappa_{\text{H},6}}{\kappa_{\text{H}}}. \tag{8}$$

One can interpret ρ as some kind of "normalization" according to the inter-annotator agreement of the humans. For example, solo recording `Brow-JS` obtains the lowest agreement of $\kappa_H = 0.61$ in our test set. The algorithms perform "moderate" with $\kappa_{H,5} = 0.47$ and $\kappa_{H,6} = 0.43$. This moderate performance is partly alleviated when normalizing with the relatively low human agreement, leading to $\rho_5 = 0.78$ and $\rho_6 = 0.71$. On the other hand, for the solo recording `Shor-FP`, the human annotators had an "almost perfect" agreement of $\kappa_{H,6} = 0.82$. While the automated method's approaches were "substantial" with $\kappa_{H,5} = 0.65$ and "moderate" with $\kappa_{H,6} = 0.58$. However, although the automated method's κ-values are higher than for `Brow-JS`, investigating the proportions ρ_5 and ρ_6 reveal that the automated method's relative agreement with the human annotations is actually the same ($\rho_5 = 0.78$ and $\rho_5 = 0.71$ for `Brow-JS` compared to $\rho_5 = 0.80$ and $\rho_5 = 0.70$ for `Shor-FP`). This indicates the ρ-value's potential as an evaluation measure that can account for multiple human reference annotations in a meaningful way.

4. F0 ESTIMATION

One of the used standard measures for the evaluation of the F0 estimation in MIREX is the *Raw Pitch Accuracy* (RPA) which is computed for a pair of annotations (A_r, A_e) consisting of a reference A_r and an estimate annotation A_e. The core concept of this measure is to label an F0 estimate $A_e(n)$ to be correct, if its F0-value deviates from $A_r(n)$ by at most a fixed tolerance $\tau \in \mathbb{R}$ (usually $\tau = 50$ cent). Figure 5 shows the RPA for different annotation pairs and different tolerances $\tau \in \{1, 10, 20, 30, 40, 50\}$ (given in cent) for the solo recording `Brow-JO`, as computed by `MIR_EVAL`. For example, looking at the pair (A_1, A_4), we see that the RPA ascends with increasing value of τ. The reason for this becomes obvious when looking at Figure 7. While A_1 was created with the goal of having fine grained F0-trajectories, annotations A_4 was created with a transcription scenario in mind. Therefore, the RPA is low for very small τ but becomes almost perfect when considering a tolerance of half a semitone ($\tau = 50$ cent).

Another interesting observation in Figure 5 is that the annotation pairs (A_1, A_2) and (A_1, A_3) yield almost constant high RPA-values. This is the case since both annotations were created using the same annotation tool—yielding very similar F0-trajectories. However, it is noteworthy that there seems to be a "glass ceiling" that cannot be exceeded even for high τ-values. The reason for this lies in the exact definition of the RPA as used for MIREX. Let $\mu(A) := \{n \in [1 : N] : A(n) \neq *\}$ be the set of all active time instances of some annotation in $\mathcal{A}$. By definition, the RPA is only evaluated on the reference annotation's active time instances $\mu(A_r)$, where each

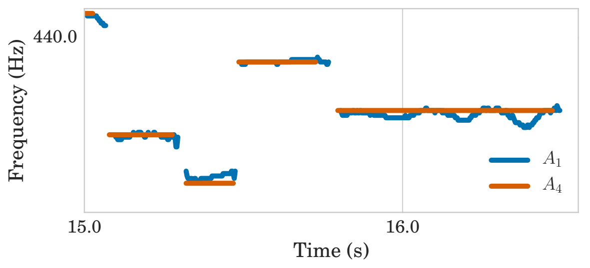

Figure 7. Excerpt from the annotations of the solo `Brow-JO` of A_1 and A_4.

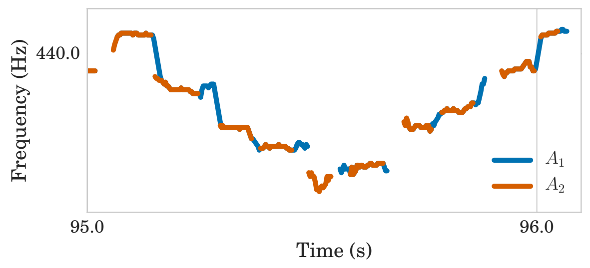

Figure 8. Excerpt from the annotations of the solo `Brow-JO` of A_1 and A_2.

$n \in \mu(A_r) \setminus \mu(A_e)$ is regarded as an incorrect time instance (for any τ). In other words, although the term "Raw Pitch Accuracy" suggests that this measure purely reflects correct F0-estimates, it is implicitly biased by the activity detection of the reference annotation. Figure 8 shows an excerpt of the human annotations A_1 and A_2 for the solo recording `Brow-JO`. While the F0-trajectories are quite similar, they differ in the annotated activity. In A_1, we see that transitions between consecutive notes are often annotated continuously—reflecting glissandi or slurs. This is not the case in A_2, where the annotation rather reflects individual note events. A musically motivated explanation could be that A_1's annotator had a performance analysis scenario in mind where note transitions are an interesting aspect, whereas A_2's annotator could have been more focused on a transcription task. Although both annotations are musically meaningful, when calculating the RPA for (A_1, A_2), all time instances where A_1 is active and A_2 not, are counted as incorrect (independent of τ)—causing the glass ceiling.

As an alternative approach that decouples the activity detection from the F0 estimation, one could evaluate the RPA only on those time instances, where reference *and* estimate annotation are active, i. e., $\mu(A_r) \cup \mu(A_e)$. This leads to the modified RPA-values as shown in Figure 6. Compared to Figure 5, all curves are shifted towards higher RPA-values. In particular, the pair (A_1, A_2) yields modified RPA-values close to one, irrespective of the tolerance τ—now indicating that A_1 and A_2 coincide perfectly in terms of F0 estimation.

However, it is important to note that the modified RPA evaluation measure may not be an expressive measure on its own. For example, in the case that two annotations are almost disjoint in terms of activity, the modified RPA would only be computed on the basis of a very small number of time instances, thus being statistically meaningless. Therefore, to rate a computational approach's performance, it is necessary to consider both, the evaluation of the activity detection as well as the F0 estimation, simultaneously but independent of each other. Both evaluations give valuable perspectives on the computational approach's performance for the task of predominant melody estimation and therefore help to get a better understanding of the underlying problems.

5. CONCLUSION

In this paper, we investigated the evaluation of automatic approaches for the task of predominant melody estimation—a task that can be subdivided into the subtask of soloist activity detection and F0 estimation. The evaluation of this task is not straightforward since the existence of a single "ground-truth" reference annotation is questionable. After having reviewed standard evaluation measures used in the field, one of our main contributions was to adapt Fleiss' Kappa—a measure which accounts for multiple reference annotations. We then explicitly defined and discussed Fleiss' Kappa for the task of the soloist activity detection.

The core motivation for using Fleiss' Kappa as an evaluation measure was to consider an automatic approach to work correctly, if its results were just about as variable as the human annotations. We therefore extended this the kappa measure by normalizing it by the variability of the human annotations. The resulting ρ-values allow for quantifying the agreement of an automatically generated annotation and the human annotations in a single value.

For the task of F0 estimation, we showed that the standard evaluation measures are biased by the activity detection task. This is problematic, since mixing both subtasks can obfuscate insights into advantages and drawbacks of a tested predominant melody estimation procedure. We therefore proposed an alternative formulation for RPA which decoupled the two tasks.

6. ACKNOWLEDGMENT

This work has been supported by the German Research Foundation (DFG MU 2686/6-1 and DFG PF 669/7-1). We would like to thank all members of the Jazzomat research project led by Martin Pfleiderer.

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer-Institut für Integrierte Schaltungen IIS.

7. REFERENCES

[1] Accompanying website. `http://www.audiolabs-erlangen.de/resources/MIR/2016-ISMIR-Multiple-Annotations/`.

[2] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José R. Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In *Proc. of the Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 493–498, Curitiba, Brazil, 2013.

[3] Juan J. Bosch and Emilia Gómez. Melody extraction in symphonic classical music: a comparative study of mutual agreement between humans and algorithms. In *Proc. of the Conference on Interdisciplinary Musicology (CIM)*, December 2014.

[4] Chris Cannam, Christian Landone, and Mark B. Sandler. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proc. of the Int. Conference on Multimedia*, pages 1467–1468, Florence, Italy, 2010.

[5] J. Stephen Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[6] Jonathan Driedger and Meinard Müller. Verfahren zur Schätzung der Grundfrequenzverläufe von Melodiestimmen in mehrstimmigen Musikaufnahmen. In Wolfgang Auhagen, Claudia Bullerjahn, and Richard von Georgi, editors, *Musikpsychologie – Anwendungsorientierte Forschung*, volume 25 of *Jahrbuch Musikpsychologie*, pages 55–71. Hogrefe-Verlag, 2015.

[7] Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical Methods for Rates and Proportions*. John Wiley Sons, Inc., 2003.

[8] Arthur Flexer. On inter-rater agreement in audio music similarity. In *Proc. of the Int. Conference on Music Information Retrieval (ISMIR)*, pages 245–250, Taipei, Taiwan, 2014.

[9] Klaus Frieler, Wolf-Georg Zaddach, Jakob Abeßer, and Martin Pfleiderer. Introducing the jazzomat project and the melospy library. In *Third Int. Workshop on Folk Music Analysis*, 2013.

[10] M. Cameron Jones, J. Stephen Downie, and Andreas F. Ehmann. Human similarity judgments: Implications for the design of formal evaluations. In *Proc. of the Int. Conference on Music Information Retrieval (ISMIR)*, pages 539–542, Vienna, Austria, 2007.

[11] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[12] Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justing Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. In *Proc. of the Int. Conference on Technologies for Music Notation and Representation*, May 2015.

[13] Matthias Mauch and Simon Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7480–7484, 2014.

[14] MIREX. Audio melody extraction task. Website `http://www.music-ir.org/mirex/wiki/2015:Audio_Melody_Extraction`, last accessed 01/19/2016, 2015.

[15] Meinard Müller. *Fundamentals of Music Processing*. Springer Verlag, 2015.

[16] Oriol Nieto, Morwaread Farbood, Tristan Jehan, and Juan Pablo Bello. Perceptual analysis of the F-measure to evaluate section boundaries in music. In *Proc. of the Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 265–270, Taipei, Taiwan, 2014.

[17] Jouni Paulus and Anssi P. Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1159–1170, 2009.

[18] The Jazzomat Research Project. Database download, last accessed: 2016/02/17. `http://jazzomat.hfm-weimar.de`.

[19] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. MIR_EVAL: A transparent implementation of common MIR metrics. In *Proc. of the Int. Conference on Music Information Retrieval (ISMIR)*, pages 367–372, Taipei, Taiwan, 2014.

[20] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[21] Justin Salamon, Emilia Gómez, Daniel P. W. Ellis, and Gaël Richard. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, 31(2):118–134, 2014.

[22] Justin Salamon and Julián Urbano. Current challenges in the evaluation of predominant melody extraction algorithms. In *Proc. of the Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 289–294, Porto, Portugal, October 2012.

[23] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proc. of the Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 555–560, Miami, Florida, USA, 2011.