Towards Evaluating Multiple Predominant Melody Annotations in Jazz Recordings

Stefan Balke¹, Jonathan Driedger¹, Jakob Abeßer², Christian Dittmar¹, Meinard Müller¹

¹International Audio Laboratories Erlangen, ²Fraunhofer Institute for Digital Media Technology IDMT

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 work, 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.

Jazz Dataset

Weimar Jazz Database (WJD)
- 299 transcribed jazz solos from monophonic instruments.
- Transcriptions specify a musical pitch for each physical time instance.

Dataset for Case Study
- Created subset of 8 solos and annotated the F0-trajectories by 3 human annotators.
- Approx. 15 min of annotations.
- Annotations are publicly available.

Motivation

Typical F0 Estimation Approach:
1. Estimate active time instances when soloist is active.
2. Estimate course of soloist's F0 at active time instances.

Typical Evaluation Approach:
1. Create ground-truth annotations.
2. Compare estimated F0 trajectory against ground-truth annotation using suitable measures.

Problems:
- Human annotators may disagree.
- Is there a single "ground-truth"?
- How to proceed if there are multiple reference annotations?

Case Study:

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>Human 1, F0-Annotation-Tool</td>
</tr>
<tr>
<td>A₂</td>
<td>Human 2, F0-Annotation-Tool</td>
</tr>
<tr>
<td>A₃</td>
<td>Human 3, F0-Annotation-Tool</td>
</tr>
<tr>
<td>A₄</td>
<td>Human 4, WJD, Sonic Visualiser</td>
</tr>
<tr>
<td>A₅</td>
<td>Computed, MELODIA</td>
</tr>
<tr>
<td>A₆</td>
<td>Computed, pYIN</td>
</tr>
<tr>
<td>A₇</td>
<td>Baseline, all time instances active at 1 kHz</td>
</tr>
</tbody>
</table>
Evaluation: Soloist Activity Detection

Typical Evaluation Approach:

- Fix pairwise evaluation measure (e.g., P/R/F-measure).
- Compute annotations in a pairwise fashion.
- Compute suitable statistics (e.g., average, variance).

Kappa Approach:

- Deal with multiple human reference annotations jointly.
- Compensate for chance-based agreement.
- Typical values for Fleiss’ Kappa:

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0</td>
<td>poor</td>
</tr>
<tr>
<td>0 – 0.2</td>
<td>slight</td>
</tr>
<tr>
<td>0.21 – 0.4</td>
<td>fair</td>
</tr>
<tr>
<td>0.41 – 0.6</td>
<td>moderate</td>
</tr>
<tr>
<td>0.61 – 0.8</td>
<td>substantial</td>
</tr>
<tr>
<td>0.81 – 1</td>
<td>almost perfect</td>
</tr>
</tbody>
</table>

- **Kappa ratio ρ**: Quantify agreement of automatically generated annotations and the human annotations in a single value.

<table>
<thead>
<tr>
<th>Annotator Group</th>
<th>(\kappa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\kappa_H)</td>
<td>0.71</td>
</tr>
<tr>
<td>(\kappa_{H,5})</td>
<td>0.60</td>
</tr>
<tr>
<td>(\kappa_{H,6})</td>
<td>0.55</td>
</tr>
</tbody>
</table>

\[
\rho_5 = \frac{\kappa_{H,5}}{\kappa_H} = 0.85
\]

\[
\rho_6 = \frac{\kappa_{H,6}}{\kappa_H} = 0.78
\]

Fleiss’ Kappa [1]

\[
\kappa := \frac{A^o - A^e}{1 - A^e} \in [-1, 1]
\]

\(A^o\) := Mean observed agreement.
\(A^e\) := Mean expected agreement.

Mathematical details and a simple toy example can be found in the paper.

Literature & Acknowledgments


This work has been supported by the German Research Foundation (DFG Mu 2686/6-1 and DFG PF 6697/1-1). 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.

Dataset download: http://bit.ly/2ee2Z1Z