INTERNATIONAL AUDIO LABORATORIES ERLANGEN
A joint institution of Fraunhofer IIS and Universität Erlangen-Nürnberg

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, [2]Fraunhofer 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.

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

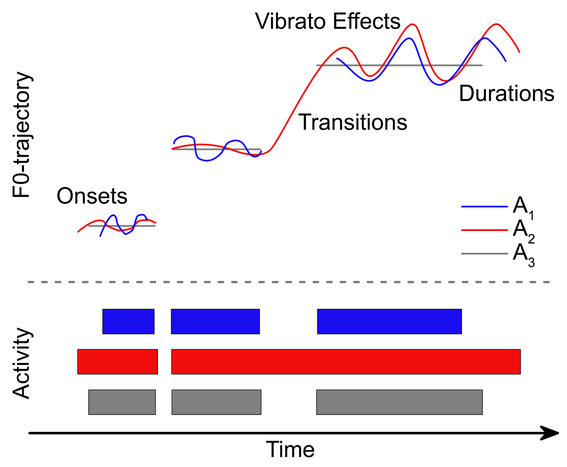
Motivation

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.

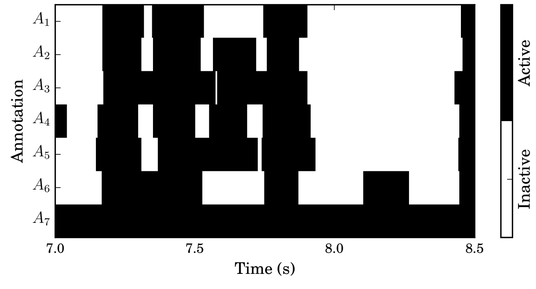


Problems:

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

Case Study:

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
A_6	Computed, pYIN
A_7	Baseline, all time instances active at 1 kHz



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).

Voicing Detection (Recall): $\mathrm{VD} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN}}$

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

Kappa Approach:

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

< 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

Annotator Group	κ
κ_H based on $H \in \{1,2,3,4\}$	0.71
$\kappa_{H,5}$ based on $H \cup \{5\}$	0.60
$\kappa_{H,6}$ based on $H \cup \{6\}$	0.55

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

$$\rho_5 = {}^{\kappa_{H,5}}/_{\kappa_H} = 0.85$$

$$\rho_6 = {}^{\kappa_{H,6}}/_{\kappa_H} = 0.78$$

Evaluation: F0 Estimation

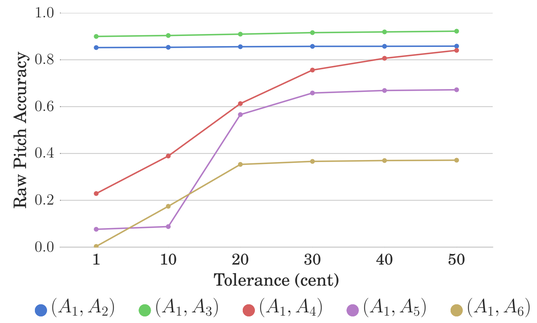


Figure: Raw Pitch Accuracy (RPA) evaluated on all active time instances according to the reference annotation.

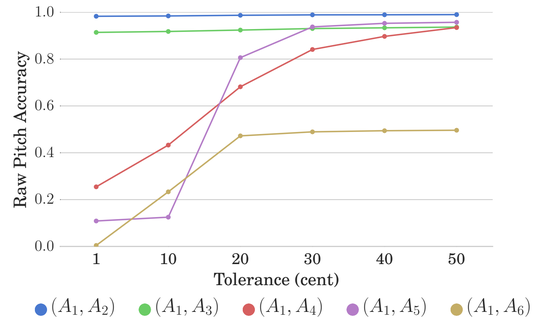


Figure: RPA evaluated on all active time instances according to the union of reference and estimate annotation.

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

[1] J. L. Fleiss, B. Levin, and M. Cho Paik. Statistical Methods for Rates and Proportions. J. Wiley Sons, 2003.

[2] A. Flexer. On inter-rater agreement in audio music similarity. ISMIR, 2014.

[3] J. J. Bosch and E. Gómez. Melody extraction in symphonic classical music: A comparative study of mutual agreement between humans and algorithms. CIM, 2014.

[4] J. Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Transactions on Audio, Speech, and Language Processing, 2012.

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.

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Dataset download:
http://bit.ly/2ae2ZlZ

Fraunhofer IDMT

Fraunhofer IIS