ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413372

RELIABILITY ASSESSMENT OF SINGING VOICE F0-ESTIMATES USING MULTIPLE ALGORITHMS

Sebastian Rosenzweig[1], *Frank Scherbaum*[2], *Meinard Müller*[1]

[1]International Audio Laboratories Erlangen, Germany
[2]University of Potsdam, Germany

ABSTRACT

Over the last decades, various conceptually different approaches for fundamental frequency (F0) estimation in monophonic audio recordings have been developed. The algorithms' performances vary depending on the acoustical and musical properties of the input audio signal. A common strategy to assess the reliability (correctness) of an estimated F0-trajectory is to evaluate against an annotated reference. However, such annotations may not be available for a particular audio collection and are typically labor-intensive to generate. In this work, we consider an approach to automatically assess the reliability of F0-trajectories estimated from monophonic singing voice recordings. As main contribution, we propose three reliability indicators that are based on the outputs of multiple algorithms. Besides providing a mathematical description of the indicators, we analyze the indicators' behavior using a set of annotated vocal F0-trajectories. Furthermore, we show the potential of the proposed indicators for exploring unlabeled audio collections.

Index Terms— singing voice, F0, reliability assessment

1. INTRODUCTION

Fundamental frequency (F0) estimates often serve as mid-level representation [1] in music information retrieval (MIR) tasks such as automatic music transcription [2] and performance analysis [3, 4]. There exist a variety of approaches for monophonic F0-estimation, ranging from model-based methods [5–7] to more recent deep-learning-based methods [8, 9]. A monophonic F0-estimation algorithm typically outputs one F0-value per time instance together with a confidence value that indicates the algorithm's certainty whether the sound source is active or not (sometimes referred to as "voicing"). However, high confidence does not necessarily imply high reliability (correctness) of an estimated F0. For example, typical estimation errors are confusions of the F0 with higher or lower harmonics (in particular octaves). The performance of a specific F0-estimation algorithm depends on the audio signal's acoustic properties (e.g., microphone characteristics, recording conditions) and musical properties (e.g., instrumentation, singing/playing styles).

In order to assess the accuracy of F0-estimates, a commonly used strategy is to evaluate an algorithm's output against a manually annotated reference, e.g., using the standard metrics defined in [10, 11] or a recently proposed variant [12]. However, manual F0-annotations are labor-intensive to generate and sometimes not available. This motivates the need for automatic approaches that deliver cues on the reliability of F0-estimates. In prior work [13], the authors have suggested a deep-learning-based approach for reliability assessment of F0-estimates from speech recordings. The approach requires access to the algorithms' internal computations, as well as algorithm-specific adaptation and training.

In this work, we have developed a more generic approach that is independent of the algorithms' working principle and available implementations. Conceptually similar to the studies in [14, 15], our approach makes use of F0- and confidence outputs of multiple algorithms. As one main contribution, we introduce three reliability indicators (denoted as $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$) that measure the reliability of an F0-estimate with respect to three different criteria. $\mathcal{I}_1$ measures the agreement of the algorithms' F0-estimates, $\mathcal{I}_2$ measures the overall confidence of the algorithms, and $\mathcal{I}_3$ measures the stability of the estimated F0-trajectories in a temporal context. The latter criterion is based on the observation that some algorithms tend to output random-like values in parts where no singing voice is active. Furthermore, in parts where F0-estimation is ambiguous or problematic (e.g., for consonants), estimated F0-trajectories often exhibit abrupt jumps. As a test scenario for our indicators, we consider a collection of multitrack field recordings of polyphonic Georgian vocal music (GVM) [16, 17]. The GVM collection comprises 216 performances recorded with multiple close-up microphones, including headset and throat microphones attached to individual singers [18, 19]. Besides being musically relevant, the collection is suitable for testing our indicators due to its diversity of singers, singing skills, and acoustic conditions.

In the following, we provide mathematical definitions of the reliability indicators in Section 2 and evaluate the indicators' performance on a set of manual F0-annotations extracted from selected songs of the GVM collection in Section 3. We indicate the potential of the proposed indicators for exploring unlabeled audio collections in Section 4 and conclude our results in Section 5.

2. RELIABILITY INDICATORS

In Section 2.1, we formalize the notion for our scenario. Then, we summarize the algorithms and annotations used in our investigations in Section 2.2. Subsequently, we introduce our three reliability indicators that measure F0-agreement (Section 2.3), overall confidence (Section 2.4), and F0-trajectory stability (Section 2.5).

2.1. Formalization

In our experiments, we consider several F0-estimation algorithms applied to one audio recording. Let M be the number of algorithms. In order to account for the logarithmic nature of pitch perception, we convert the estimated F0-values (given in Hertz) into the log-

This work was supported by the German Research Foundation (DFG MU 2686/13-1, SCHE 280/20-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS.

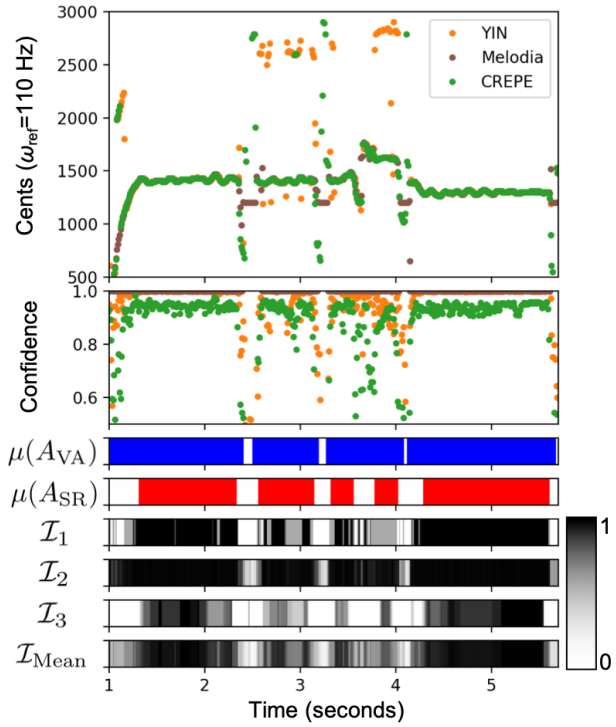

Fig. 1: Estimated F0-trajectories, confidences, annotations, and reliability indicators for the middle voice in the song "Kriste Aghsdga".

frequency domain by defining

$$F_{\text{cents}}(\omega) := 1200 \cdot \log_2\left(\frac{\omega}{\omega_{\text{ref}}}\right). \quad (1)$$

F_{cents} measures the distance (given in cents) between frequency ω and a reference frequency ω_{ref}. In the following, we set $\omega_{\text{ref}} = 110$ Hz. Let us assume, a given F0-estimation algorithm outputs a frequency value as well as a confidence value (a value between 0 and 1) for each discrete time index $n \in [1:N]$. Then, let $T : [1:N] \to \mathbb{R}$ be the resulting frequency trajectory and $C : [1:N] \to [0,1]$ the corresponding confidence trajectory. For our M algorithms, let

$$\mathcal{T} := \{T_1, ..., T_{\text{M}}\} \quad (2)$$

and

$$\mathcal{C} := \{C_1, ..., C_{\text{M}}\} \quad (3)$$

be the corresponding sets of trajectories. Let T_m be the frequency trajectory and C_m the confidence trajectory for the m^{th} algorithm, $m \in [1:M]$.

Furthermore, let $A : [1:N] \to \mathbb{R} \cup \{*\}$ be an F0-annotation, with $A(n) = *$ where the frequency value is unspecified. We denote the set of all time frames where the annotation is active as $\mu(A) := \{n \in [1:N] : A(n) \neq *\}$.

2.2. Algorithms and Annotations

In our investigations, we consider the algorithms YIN [5], Melodia [20], and CREPE [8] ($M = 3$). While YIN and CREPE are designed for monophonic F0-estimation, Melodia was originally developed for the task of predominant melody estimation. Note that the selection of algorithms in this work is exemplary and our measures are not restricted to this specific set of algorithms. For extracting F0- and confidence trajectories, we use the publicly available YIN and Melodia Vamp plugins[1] together with the open-source tool Sonic Annotator [21], as well as the CREPE Python package[2]. All algorithms are applied with default parameter settings. For YIN and CREPE, we use the continuous confidence output of the implementations, whereas for Melodia, we derive binary confidence trajectories from the voice activity decision made by the algorithms.

Additionally, we consider two types of manual annotations. A_{VA} assumes annotated F0-values in cents for parts where the singing voice is active (VA) and the symbol '$*$' elsewhere. Similarly, A_{SR} assumes annotated F0-values in cents for roughly stable regions (SR) of the F0-trajectory and the symbol '$*$' elsewhere. Note that typically, the F0-values in A_{SR} form a subset of the F0-values in A_{VA}. We manually generate A_{SR} using the publicly available tool Tony [22], which is based on the algorithm PYIN [7]. Furthermore, we generate A_{VA} by restricting automatically extracted PYIN trajectories (also obtained using a Vamp plugin) to manually annotated regions where the singing voice is active using Sonic Visualiser [23]. In order to account for different hop sizes of the algorithms and annotations, we resample all F0-trajectories, annotations, and confidences to a time grid with a resolution of 10 ms. Furthermore, we quantize the F0-trajectories and annotations to a frequency resolution of 10 cents.

As a running example in this section, we consider a recording of the three-voice song "Kriste Aghsdga", which is part of a publicly accessible subset[3] of the GVM collection. The three performing singers frequently use pitch slides at the beginning and end of sung notes, which is a characteristic stylistic element in traditional Georgian vocal music. Figure 1 shows a superposition of the resulting F0-trajectories extracted from the throat microphone recording of the middle voice for a short excerpt from our running example. Furthermore, the color-coded activities $\mu(A_{\text{VA}})$ and $\mu(A_{\text{SR}})$ are visualized.

Given the sets $\mathcal{T}$ and $\mathcal{C}$, we now introduce three reliability indicators $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$. The frame-wise arithmetic mean of the three indicators is denoted as $\mathcal{I}_{\text{Mean}}$.

2.3. F0-Agreement

For measuring the agreement of the F0-trajectories, we consider $P = \binom{M}{2}$ trajectory pairs $(T_i, T_j) \in \mathcal{T} \times \mathcal{T}$, with $i < j$. For each pair, we compute the difference between the trajectories by

$$\Delta_p(n) = \begin{cases} 1, & \text{for } |T_i(n) - T_j(n)| \leq \varepsilon_{\mathcal{I}}, \\ 0, & \text{otherwise}, \end{cases} \quad (4)$$

with pair-index $p \in [1:P]$ and $\varepsilon_{\mathcal{I}}$ being a threshold in cents which defines the strictness of the measure. In our experiments, we set $\varepsilon_{\mathcal{I}} = 10$ cents. Compared to a 50 cents tolerance, which is typically used in standard evaluation metrics for evaluating pitch accuracy [11, 12], the chosen threshold is rather strict. Considering the 10 cents quantization of our trajectories, the threshold accounts for possible rounding artifacts caused by quantization. For practical reasons, we work with a fixed $\varepsilon_{\mathcal{I}}$ in our experiments and leave further investigations on the role of $\varepsilon_{\mathcal{I}}$ to future research. Our first reliability indicator is defined as the arithmetic mean of the differences over

[1] https://vamp-plugins.org/
[2] https://github.com/marl/crepe
[3] https://www.audiolabs-erlangen.de/resources/MIR/2018-ISMIR-LBD-ThroatMics

all pairs:

$$\mathcal{I}_1(n) := \frac{\sum_{p=1}^{P} \Delta_p(n)}{P}. \quad (5)$$

Only if the F0-estimates of all algorithm pairs agree, $\mathcal{I}_1(n) = 1$, as shown in our running example in Figure 1. In parts where the F0-estimates strongly deviate (e.g., at 2.5 sec) one obtains $\mathcal{I}_1(n) = 0$. In the part between 2.5–4 sec, there are some octave jumps by YIN and CREPE, which cause the agreement to decrease.

2.4. Overall Confidence

Our second reliability indicator combines the confidence outputs of the algorithms and is defined as the arithmetic mean of the confidences over all algorithms:

$$\mathcal{I}_2(n) := \frac{\sum_{m=1}^{M} C_m(n)}{M}. \quad (6)$$

Note that in order for $\mathcal{I}_2$ to deliver meaningful indications, all trajectories are required to have values in the same value range, ideally making use of the entire $[0, 1]$ interval. If this requirement is not fulfilled, we use suitable normalization techniques or a binarization of the confidence using the algorithm's voice activity decision to balance out the confidence value distributions. In particular, we use binarized confidence trajectories for Melodia. In Figure 1, $\mathcal{I}_2$ indicates high overall confidence in most of the parts where the voice is active, thus showing high agreement with $\mu(A_{\mathrm{VA}})$.

2.5. F0-Trajectory Stability

Our third indicator $\mathcal{I}_3$ measures reliability with respect to the local stability of the estimated F0-trajectories. A trajectory region is considered stable if it exhibits a roughly horizontal structure (up to some tolerance). In order to detect such stable regions in an F0-trajectory, we make use of an automatic approach based on morphological filters proposed in [24]. In a first step, we compute two filtered versions of the trajectory, one by using a min-filter (erosion) and one by using a max-filter (dilation) with filter lengths $L \in \mathbb{N}$. L controls the smoothness of the filtered trajectories and affects the sensitivity of the stable region detection to sudden jumps in the trajectories. For practical reasons, we fix $L = 15$ (150 ms) in our experiments. The value roughly corresponds to the filter length determined in a previous study on Georgian vocal music [24] and might need to be adapted to other application scenarios. We leave further investigations on the role of L to future work. In a second step, we compute the frame-wise absolute difference between the max- and the min-filtered trajectory (also referred to as envelope width). All regions where the envelope width is lower than or equal to a certain threshold τ given in cents are considered stable. The algorithm outputs an indicator $S : [1 : N] \to \{0, 1\}$, where $S(n) = 1$ in stable regions and $S(n) = 0$ in unstable regions. In order to account for trajectory fluctuations of different extent, we consider a set of envelope-width thresholds $\mathcal{W} = \{20, 40, 60, 80, 100\}$, with 20 cents being a very strict threshold allowing for almost no trajectory fluctuations, and 100 cents being a generous threshold allowing for fluctuations of up to a semitone (e.g., vibrato).

Let $S_{m,\tau}$ be the stability indicator for the m^{th} algorithm $m \in [1 : M]$ and threshold $\tau \in \mathcal{W}$. Then, $\mathcal{I}_3$ is defined as the arithmetic mean as follows:

$$\mathcal{I}_3(n) := \frac{\sum_{m=1}^{M} \sum_{\tau \in \mathcal{W}} S_{m,\tau}(n)}{M \cdot |\mathcal{W}|}, \quad (7)$$

for $n \in [1 : N]$. As one can see in Figure 1, $\mathcal{I}_3$ indicates high reliability in regions where all estimated F0-trajectories are roughly stable and therefore strongly coincides with $\mu(A_{\mathrm{SR}})$.

3. EVALUATION USING LABELED DATA

In order to study the behavior of our indicators, we apply different thresholds $\kappa \in [0, 1]$ on our reliability indicators $\mathcal{I} : [1 : N] \to [0, 1]$. The resulting (enduring) subsets of our discrete time axis are given as $\mathcal{E}_\kappa = \{n \in [1 : N] : \mathcal{I}(n) \geq \kappa\}$. The higher κ, the smaller is the obtained subset. For a given subset $\mathcal{E}_\kappa$, we evaluate the agreement with an annotated voice activity $\mu(A)$ using the standard retrieval metrics precision (P), recall (R), and F-measure (F) defined as

$$\mathrm{P} := \frac{|\mathcal{E}_\kappa \cap \mu(A)|}{|\mathcal{E}_\kappa|}, \quad \mathrm{R} := \frac{|\mathcal{E}_\kappa \cap \mu(A)|}{|\mu(A)|}, \quad \mathrm{F} := \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}. \quad (8)$$

Furthermore, we set $\mathrm{P} := 0$ for $|\mathcal{E}_\kappa| = 0$, $\mathrm{R} := 0$ for $|\mu(A)| = 0$, and $\mathrm{F} := 0$ for $\mathrm{P} + \mathrm{R} = 0$. The standard definition of the F-measure equally weights precision and recall. The weighting may have to be adapted depending on the application scenario. As a further analysis step, we evaluate the F0-accuracy of estimated F0-trajectories within the subsets with respect to a reference annotation. Given an F0-trajectory T restricted to the given subset $\mathcal{E}_\kappa$ and an annotation A, we define the F0-accuracy ϕ as

$$\phi := \frac{|\mathcal{E}_\kappa \cap \mu(A) \cap \{n \in [1 : N] : |T(n) - A(n)| \leq \varepsilon_{\mathrm{e}}\}|}{|\mathcal{E}_\kappa \cap \mu(A)|}, \quad (9)$$

with ε_{e} being the evaluation tolerance parameter in cents. In our experiments, we use a strict value of $\varepsilon_{\mathrm{e}} = 10$ cents, to basically allow for quantization errors.

In our evaluation, we expand the scenario described in Section 2.2 to all five songs of the GVM subset. In the following, we consider the three algorithms YIN, CREPE, and Melodia applied on the throat microphone recordings of the middle voices. Furthermore, we manually generated the annotations A_{VA} and A_{SR} for these middle voice tracks (we crosschecked the annotations in spot-checks). The F-measure and F0-accuracy with respect to A_{VA} are denoted as F_{VA} and ϕ_{VA}, whereas the evaluation measures with respect to A_{SR} are denoted as F_{SR} and ϕ_{SR}, respectively.

Figure 2 shows the evaluation metrics for all algorithms averaged over all five recordings for each reliability indicator and algorithm with respect to the threshold κ. For almost all algorithms and reliability indicators, we observe an increasing F0-accuracy along with an increasing κ. The sudden drop in Melodia's ϕ curves for $\mathcal{I}_2$ occurs due to a high number of octave errors in regions with high confidence in one of the five recordings. For CREPE, the F0-accuracy is close to 1 for all values of κ, which indicates that the algorithm performs well on our annotated data. Note that the F-measure curves for a specific reliability indicator are identical for all algorithms, since the F-measure only depends on the chosen indicator $\mathcal{I}$, threshold κ, and annotation A. For high values of κ, only few F0-values remain, which causes the voice activity F-measures to decrease.

In conclusion, our indicators give cues on the reliability of F0-estimates at a given time instance in the audio signal. However, they are less suitable to assess the accuracy of a specific algorithm's estimate, since high reliability does not guarantee correct estimates (e.g., in the case of all algorithms outputting wrong estimates). The choice of a suitable threshold κ depends on the algorithms' individual performances, the chosen reliability indicator, and the target annotation or application.

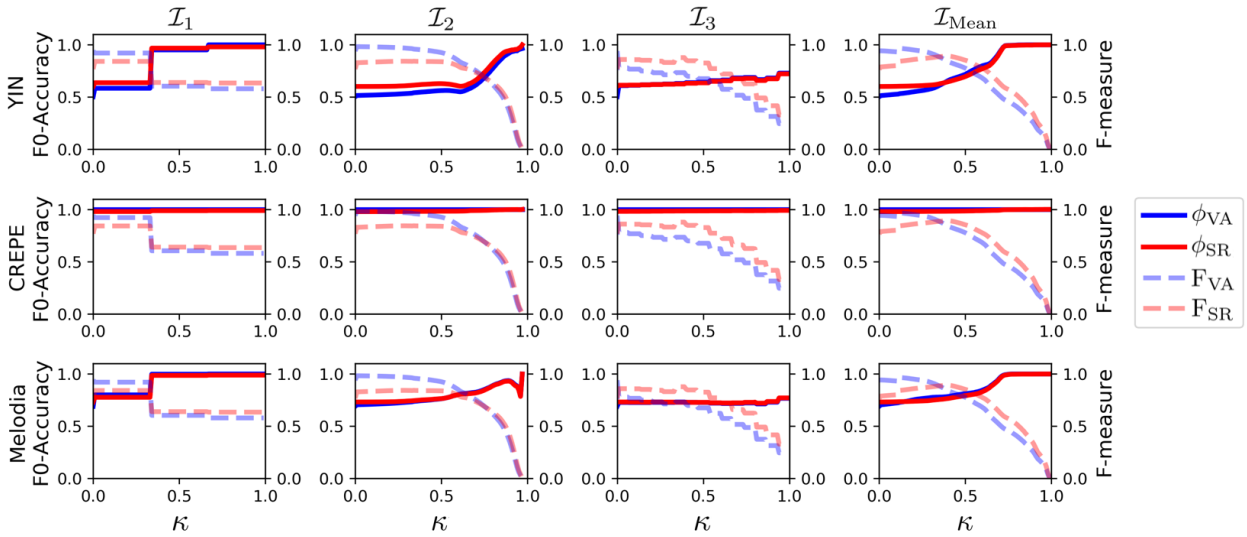

Fig. 2: F-measure and F0-accuracy for all indicators and algorithms with respect to reliability threshold κ averaged over five recordings.

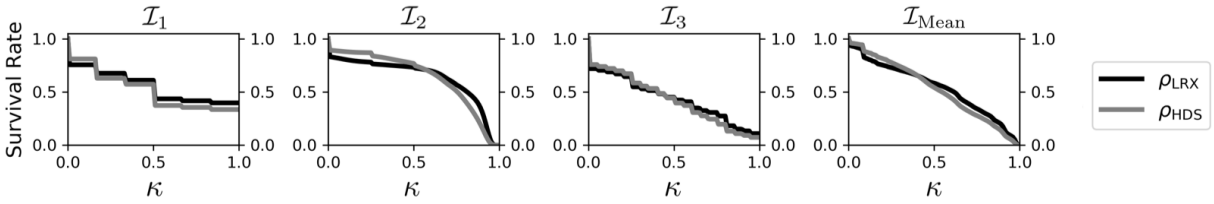

Fig. 3: Survival rates for LRX and HDS microphones with respect to threshold κ averaged over 249 tracks.

4. EXPLORING UNLABELED AUDIO COLLECTIONS

In this section, we want to demonstrate the potential of our reliability indicators for exploring unlabeled datasets. When approaching new audio collections, one may want to have a compact overview on how reliably automatic F0-extraction algorithms perform under the acoustical and musical conditions provided by the data. In the case of the GVM collection, we find two different acoustic conditions due to the different types of close-up microphones used. Throat or larynx microphones (referred to as LRX microphones in the following) capture the vibrations of the throat directly from the skin and are therefore less sensitive to cross-talk of neighboring singers [18, 19]. Headset (HDS) microphones pick up the human voice placed in front of (or next to) the mouth and are prone to bleeding of other voices. The presence of noise or other interfering sounds typically complicates the task of F0-estimation.

In the following, we consider a large subset of the GVM collection consisting of 85 polyphonic songs. More specifically, the subset includes 249 tracks (ca. 9 hours duration) for each microphone type. In order to explore the reliabilities measured by our indicators for the two different microphone types, we introduce a measure referred to as *survival rate* and denoted as ρ. The measure indicates the portion of remaining trajectory values after thresholding $\mathcal{I} : [1 : N] \to [0, 1]$ with $\kappa \in [0, 1]$ and is defined as follows:

$$\rho := \frac{|\mathcal{E}_\kappa|}{N}. \tag{10}$$

The survival rates for LRX and HDS microphone signals are denoted as ρ_{LRX} and ρ_{HDS}, respectively. In this experiment, we expand the setup described in Section 2.2 by adding PYIN to our set of algorithms. Figure 3 shows the two survival rates averaged over all 249 tracks with respect to the threshold κ. The graphs show that for high values of κ, ρ_{LRX} is larger than ρ_{HDS}, whereas for low values of κ, ρ_{HDS} is larger than ρ_{LRX}. This suggests a slightly better discriminability between reliable and unreliable frames for LRX signals.

Rather than advocating a specific indicator or a specific threshold κ, we see the proposed reliability indicators as a toolkit for measuring reliability of automatically extracted F0-trajectories with respect to F0-agreement, overall confidence, and F0-trajectory stability. Depending on the application, one may consider different indicators or suitably weighted combinations of them. Furthermore, one may adapt the selection of F0-extraction algorithms and fine-tune the individual indicators' parameters ($\varepsilon_{\mathcal{I}}$, L, and τ) to account for the specific acoustical and musical properties of the audio material.

5. CONCLUSIONS

In this paper, we presented three indicators for measuring the reliability of F0-trajectories extracted from singing voice recordings. The indicators are based on the outputs of multiple algorithms and measure reliability with respect to F0-agreement, overall confidence, and F0-trajectory stability. As one of our main contributions, we introduced the reliability indicators in a mathematically rigorous way. Furthermore, we evaluated the behavior of the indicators on a set of manually annotated vocal F0-trajectories. While our indicators cannot replace manual F0-annotations, they can be used as an efficient tool to obtain cues on the reliability of automatically extracted F0-trajectories from unlabeled audio collections. Future work will be concerned with using and further exploring our indicators for tonal analysis of Georgian vocal music.

6. REFERENCES

[1] Rachel M. Bittner, Justin Salamon, Juan J. Bosch, and Juan Pablo Bello, "Pitch contours as a mid-level representation for music informatics," in *Proceedings of the AES International Conference on Semantic Audio*, Erlangen, Germany, 2017, pp. 100–107.

[2] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[3] Jiajie Dai and Simon Dixon, "Analysis of interactive intonation in unaccompanied SATB ensembles," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 599–605.

[4] Johanna Devaney, *An Empirical Study of the Influence of Musical Context on Intonation Practices in Solo Singers and SATB Ensembles*, Ph.D. thesis, McGill University, Montreal, Canada, 2011.

[5] Alain de Cheveigné and Hideki Kawahara, "YIN, a fundamental frequency estimator for speech and music.," *Journal of the Acoustical Society of America (JASA)*, vol. 111, no. 4, pp. 1917–1930, 2002.

[6] Arturo Camacho and John G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[7] Matthias Mauch and Simon Dixon, "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 659–663.

[8] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "CREPE: A convolutional representation for pitch estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 161–165.

[9] Beat Gfeller, Christian Frank, Dominik Roblek, Matthew Sharifi, Marco Tagliasacchi, and Mihajlo Velimirovic, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1118–1128, 2020.

[10] Graham E. Poliner, Daniel P.W. Ellis, Andreas F. Ehmann, Emilia Gómez, Sebastian Streich, and Beesuan Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1247–1256, 2007.

[11] 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*, vol. 31, no. 2, pp. 118–134, 2014.

[12] Rachel M. Bittner and Juan J. Bosch, "Generalized metrics for single-f0 estimation evaluation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 738–745.

[13] Boyuan Deng, Denis Jouvet, Yves Laprie, Ingmar Steiner, and Aghilas Sini, "Towards confidence measures on fundamental frequency estimations," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 5605–5609, IEEE.

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

[15] Stefan Balke, Jakob Abeßer, Jonathan Driedger, Christian Dittmar, and Meinard Müller, "Towards evaluating multiple predominant melody annotations in jazz recordings," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, New York City, USA, August 2016, pp. 246–252.

[16] Frank Scherbaum, Nana Mzhavanadze, Sebastian Rosenzweig, and Meinard Müller, "Multi-media recordings of traditional Georgian vocal music for computational analysis," in *Proceedings of the International Workshop on Folk Music Analysis*, Birmingham, UK, 2019, pp. 1–6.

[17] Frank Scherbaum, Nana Mzhavanadze, and Elguja Dadunashvili, "A web-based, long-term archive of audio, video, and larynx-microphone field recordings of traditional Georgian singing, praying and lamenting with special emphasis on Svaneti," *International Symposium on Traditional Polyphony*, 2018.

[18] Frank Scherbaum, "On the benefit of larynx-microphone field recordings for the documentation and analysis of polyphonic vocal music," *Proceedings of the International Workshop Folk Music Analysis*, pp. 80–87, 2016.

[19] Frank Scherbaum, Sebastian Rosenzweig, Meinard Müller, Daniel Vollmer, and Nana Mzhavanadze, "Throat microphones for vocal music analysis," in *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018.

[20] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[21] Chris Cannam, Michael O. Jewell, Christophe Rhodes, Mark Sandler, and Mark d'Inverno, "Linked data and you: Bringing music research software into the semantic web," *Journal of New Music Research*, vol. 39, no. 4, pp. 313–325, 2010.

[22] 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 *Proceedings of the International Conference on Technologies for Music Notation and Representation*, 2015.

[23] Chris Cannam, Christian Landone, and Mark B. Sandler, "Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the International Conference on Multimedia*, Florence, Italy, 2010, pp. 1467–1468.

[24] Sebastian Rosenzweig, Frank Scherbaum, and Meinard Müller, "Detecting stable regions in frequency trajectories for tonal analysis of traditional Georgian vocal music," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 352–359.