DETECTING STABLE REGIONS IN FREQUENCY TRAJECTORIES FOR TONAL ANALYSIS OF TRADITIONAL GEORGIAN VOCAL MUSIC

Conference Paper · June 2020

CITATIONS
13

READS
221

3 authors:

Sebastian Rosenzweig
Audoo Ltd
27 PUBLICATIONS **154** CITATIONS
SEE PROFILE

Frank Scherbaum
Universität Potsdam
375 PUBLICATIONS **12,737** CITATIONS
SEE PROFILE

Meinard Müller
Friedrich-Alexander-Universität of Erlangen-Nürnberg
388 PUBLICATIONS **12,503** CITATIONS
SEE PROFILE

DETECTING STABLE REGIONS IN FREQUENCY TRAJECTORIES FOR TONAL ANALYSIS OF TRADITIONAL GEORGIAN VOCAL MUSIC

Sebastian Rosenzweig[1] **Frank Scherbaum**[2] **Meinard Müller**[1]
[1] International Audio Laboratories Erlangen, Germany
[2] University of Potsdam, Potsdam, Germany
{sebastian.rosenzweig, meinard.mueller}@audiolabs-erlangen.de, frank.scherbaum@uni-potsdam.de

ABSTRACT

While Georgia has a long history of orally transmitted polyphonic singing, there is still an ongoing controversial discussion among ethnomusicologists on the tuning system underlying this type of music. First attempts have been made to analyze tonal properties (e. g., harmonic and melodic intervals) based on fundamental frequency (F0) trajectories. One major challenge in F0-based tonal analysis is introduced by unstable regions in the trajectories due to pitch slides and other frequency fluctuations. In this paper, we describe two approaches for detecting stable regions in frequency trajectories: the first algorithm uses morphological operations inspired by image processing, and the second one is based on suitably defined binary time–frequency masks. To avoid undesired distortions in subsequent analysis steps, both approaches keep the original F0-values unmodified, while only removing F0-values in unstable trajectory regions. We evaluate both approaches against manually annotated stable regions and discuss their potential in the context of interval analysis for traditional three-part Georgian singing.

1. INTRODUCTION

Polyphonic singing plays a vital role in many musical cultures. One of the oldest forms of polyphonic singing can be found in Georgia, a country located in the Caucasus region of Eurasia. The traditional three-part songs, which are typically passed down orally from one generation to the next, are acknowledged as Intangible Cultural Heritage by the UNESCO. Although being a long-studied subject, the non-tempered nature of traditional Georgian vocal music is still discussed controversially among musicologists [7,33]. So far, musicological studies on traditional Georgian music have mostly been conducted on the basis of manually transcribed field recordings. Such approaches are problematic, since important tonal cues (as well as many other performance aspects) are likely to get lost in the transcription process. The importance of field recordings in research on Georgian vocal music has raised the demand for computer-based methods to assist ethnomusicologists in analyzing the audio material.

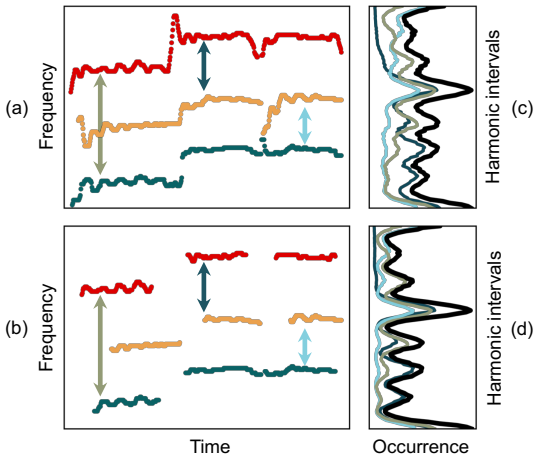

Figure 1. Detection of stable regions in F0-trajectories for a three-part singing recording. **(a)** Original F0-trajectories. **(b)** F0-trajectories restricted to stable regions. **(c)** Harmonic interval histogram based on (a). **(d)** Sharpened harmonic interval histogram based on (b). The histograms in (c) and (d) were computed considering the entire Erkomaishvili dataset.

One source of central importance for ethnomusicological research is a collection of audio recordings of the former master chanter Artem Erkomaishvili (1887–1967). The collection, which was created at the Tbilisi State Conservatory in 1966, comprises 101 three-part songs. Each chant was recorded in a three-stage "dubbing" process using tape recorders, where Erkomaishvili successively sung the individual voices with previously recorded voices being played back. In the study [20], a semi-automatic salience-based approach was applied to determine fundamental frequency (F0) trajectories of all three voices. The extracted F0-annotations are publicly available. [1] In a follow-up study [31], the authors determined from these trajectories harmonic (vertical) as well as melodic (horizontal) intervals, which give cues on the tonal organization [21, 22] of Georgian vocal music.

[1] https://www.audiolabs-erlangen.de/resources/MIR/2017-GeorgianMusic-Erkomaishvili

In general, studies on tonal analysis (see, e. g., [14, 15, 17, 31]) have shown that the usage of previously extracted F0-trajectories leads to various challenges. For example, as a stylistic element of traditional Georgian music, sung notes often start, end, or are continuously connected using pitch slides, see Figure 1a. Furthermore, automated F0-estimation procedures typically introduce inaccuracies such as extraction errors, outliers, or smoothing artifacts. Consequently, tonal analysis of Georgian vocal music based on highly fluctuating and error-prone F0-trajectories is problematic. For example, when computing harmonic interval statistics (as illustrated by Figure 1c), such artifacts may lead to an increased noise level and a less salient peak structure in the computed histograms. When analyzing melodic intervals, the presence of frequency variations (such as pitch slides) have a strong negative impact on subsequent analysis results. To alleviate such issues, contributions such as [17, 31] apply (semi-automatic) post-processing procedures to remove unstable regions in the trajectories and derive note-like events with a stable pitch. Note that for other scenarios (e. g. the tonal analysis of Hindustani Raga [29]), non-stable regions may contain musically important information.

Motivated by such tonal analysis applications, we present in this paper two automatic approaches that aim at identifying stable regions in frequency trajectories. Technically speaking, such regions correspond to horizontal structures (up to some tolerance) of trajectories. In acoustical and musical terms, such regions relate to pitched sounds where a singer has tuned into a harmonically stable pitch synchronized to other singers. In this context, our goal is to remove all frequency values in unstable regions, while keeping the original frequency values unmodified in the stable ones (see Figure 1b). For accomplishing this task, we introduce two conceptually different approaches—one based on morphological operations and the other one based on binary masking. Furthermore, we evaluate both approaches against manually annotated stable regions and indicate their potential for interval analysis using the Erkomaishvili recordings as example.

The remainder of this paper is organized as follows. We discuss related work in Section 2, then give a technical description of our approaches in Section 3, and summarize our experiments in Section 4.

2. RELATED WORK

In the following, we give an overview on work that is related to detecting stable regions in F0-trajectories. First, we want to note that stable region detection is not equivalent to F0-based transcription. In general, automated music transcription (AMT) aims at converting a music recording into some form of music notation [1, 2, 13]. In this process, many AMT systems apply temporal and spectral quantization of previously extracted F0-trajectories to derive pitches, onsets, and offsets of note events [3, 4, 6, 10, 15, 16, 18, 23, 30]. Rather than using quantized or modified F0-trajectories for our analysis, we aim at using trajectories restricted to stable regions (that may or may not correspond to note events) while leaving the original F0-values unmodified.

Detecting stable, transitional, and fluctuating patterns in F0-trajectories plays an important role for various tasks such as vibrato detection [5, 26, 37], singing style classification [24, 27], and motif detection [12, 25]. For example, in [35–37], the authors address the problem of detecting portamento (note transition) regions in Chinese string music. In [15], the authors identify stable regions as an important step towards transcribing recordings of Flamenco singing. In [19], the authors propose a vocal trajectory segmentation algorithm based on hysteresis defined on pitch–time curves. However, the underlying octave equivalence assumption may not be fulfilled in traditional Georgian vocal music. For a recent overview article of singing voice analysis, we refer to [11].

Furthermore, there are various studies on Indian Raga music, which are related to our work. In [9], a global pitch histogram ("pitch inventory") of the whole recording is computed. Then, informed by the histogram's peaks, stable regions are derived using empirically chosen thresholds for duration and fluctuation tolerance. In [14], the authors compute the local slope of the F0-trajectory and obtain stable regions by thresholding and quantization. However, due to the underlying scale assumptions, such approaches can not be directly applied to analyzing traditional Georgian singing, where pitch drifts may occur over the course of the song.

3. STABLE REGION DETECTION

In this section, we formalize the notion of a frequency trajectory as used in this work (Section 3.1). Then, to motivate the subsequent procedures, we introduce a simple median-based filtering approach (Section 3.2). As our main technical contributions, we introduce two conceptually different approaches for determining stable regions in frequency trajectories—one based on morphological operations (Section 3.3) and the other one based on binary masking (Section 3.4).

3.1 Frequency Trajectories

To account for the logarithmic nature of human pitch perception, we convert frequency values into the log-frequency domain. To this end, we fix a reference frequency ω_{ref} given in Hertz (Hz). In the following, we set $\omega_{\mathrm{ref}} = 55$ Hz. Then, an arbitrary frequency value ω is converted into the logarithmic domain by defining

$$F_{\mathrm{cents}}(\omega) := 1200 \cdot \log_2\left(\frac{\omega}{\omega_{\mathrm{ref}}}\right), \tag{1}$$

which measures the distance between ω and ω_{ref} in cents. In this paper, we model a frequency trajectory as a function

$$\gamma : \mathbb{Z} \to \mathbb{R} \cup \{*\}, \tag{2}$$

which assigns to a given time index $n \in \mathbb{Z}$ either a real-valued frequency value $\gamma(n) \in \mathbb{R}$ (given in cents) or the symbol $\gamma(n) = *$ (when the frequency value is left to be

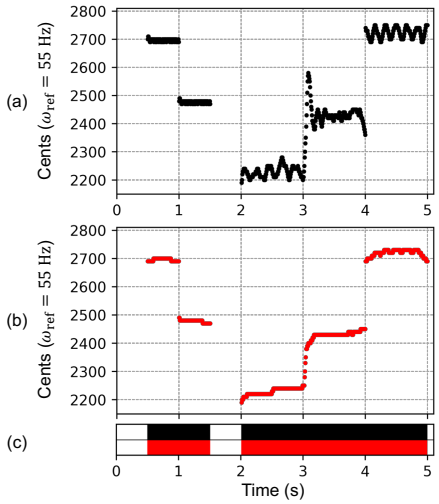

Figure 2. Effect of median filtering. **(a)** Original trajectory γ. **(b)** Median-filtered trajectory γ^{Median}. **(c)** Activation regions of γ (black) and γ^{Median} (red).

unspecified). In our implementation, we use a time resolution of 5.8 ms per time index and a frequency resolution of 10 cents. Figure 2a shows a frequency trajectory, which will serve as our running example in the remainder of this section. In the first two seconds, two notes are played on a piano without interruption. Subsequently, in the next two seconds, there are two sung notes smoothly connected by a pitch slide. Finally, the recording contains a note sung with vibrato.

3.2 Median Filtering

For tonal analysis based on frequency trajectories, one often applies some kind of filtering to remove outliers and other undesired pitch fluctuations [17, 32]. For example, by applying a median filter of odd length $L \in \mathbb{N}$, one obtains a smoothed trajectory γ^{Median} defined by

$$\gamma^{\text{Median}}(n) := \text{median}\big\{\gamma(n - \tfrac{L-1}{2} : n + \tfrac{L-1}{2})\big\} \quad (3)$$

for $n \in \mathbb{Z}$. In this definition, the symbol $*$ is handled as $-\infty$. Figure 2b shows γ^{Median} of our running example using $L = 69$ (corresponding to 0.4 sec). This example shows how median filtering introduces smoothing while removing outliers (such as the peak around the third second). However, the non-stable transition between the two sung notes remains after filtering. This is not what we aim at. First, we do not want to change frequency values in stable regions (with the goal not to introduce smoothing effects in subsequent tonal analysis steps). Second, we aim at explicitly detecting unstable regions, which can then be removed from the frequency trajectory. In the following, we present two conceptually different approaches that fulfill these requirements.

3.3 Morphological Approach

The first approach, which is inspired by work of Vávra et al. [34], uses morphological operations as known in image

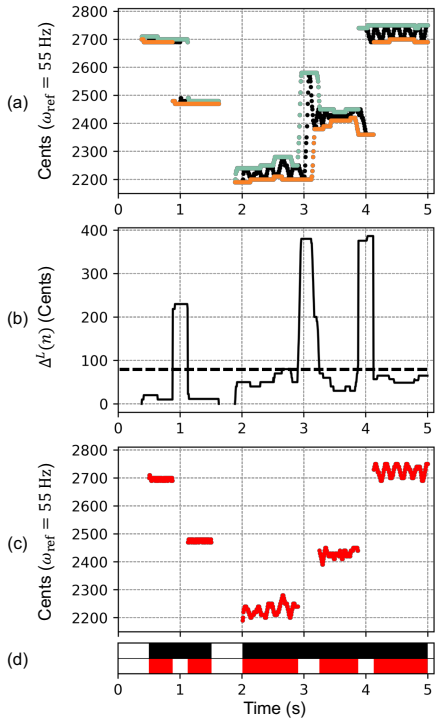

Figure 3. Morphological approach for detecting stable regions. **(a)** Frequency trajectories γ (black), $\gamma^L_{\max}$ (green), and $\gamma^L_{\min}$ (orange). **(b)** Morphological gradient Δ^L with threshold $\tau = 90$. **(c)** Trajectory γ^{Morph} restricted to stable regions. **(d)** Activation regions for γ (black) and γ^{Morph} (red).

processing. Applying these operators to frequency trajectories, dilation corresponds to max filtering, and erosion to min filtering. Given a trajectory γ, this results in a dilated trajectory $\gamma^L_{\max}$ and an eroded trajectory $\gamma^L_{\min}$ defined by

$$\gamma^L_{\max}(n) := \max\big\{\gamma(n - \tfrac{L-1}{2} : n + \tfrac{L-1}{2})\big\}, \quad (4a)$$
$$\gamma^L_{\min}(n) := \min\big\{\gamma(n - \tfrac{L-1}{2} : n + \tfrac{L-1}{2})\big\}, \quad (4b)$$

for $n \in \mathbb{Z}$, where $L \in \mathbb{N}$ is assumed to be an odd integer. In max filtering, the symbol $*$ is handled as $-\infty$, whereas in min filtering it is handled as $+\infty$. Figure 3a shows the resulting trajectories $\gamma^L_{\max}$ and $\gamma^L_{\min}$ for our running example using $L = 43$ (corresponding to 0.25 sec). In a next step, we define the difference Δ^L between the dilated and eroded trajectories, also termed morphological gradient [28]:

$$\Delta^L(n) := \gamma^L_{\max}(n) - \gamma^L_{\min}(n) \quad (5)$$

for $n \in \mathbb{Z}$, where we set $\Delta^L(n) = *$ whenever $\gamma^L_{\max}(n)$ or $\gamma^L_{\min}(n)$ are not defined. As shown in Figure 3b, the difference Δ^L is large in non-stable parts (e. g., around the third second), whereas it is small in stable parts (e. g., within each of the piano notes). Fixing a suitable threshold $\tau > 0$ (given in cents), we define the trajectory γ^{Morph} by setting

$$\gamma^{\text{Morph}}(n) := \begin{cases} \gamma(n), & \text{for } |\Delta^L(n)| \leq \tau, \\ *, & \text{otherwise.} \end{cases} \quad (6)$$

The threshold τ can be seen as a tolerance parameter that specifies the maximally allowed fluctuation under which a

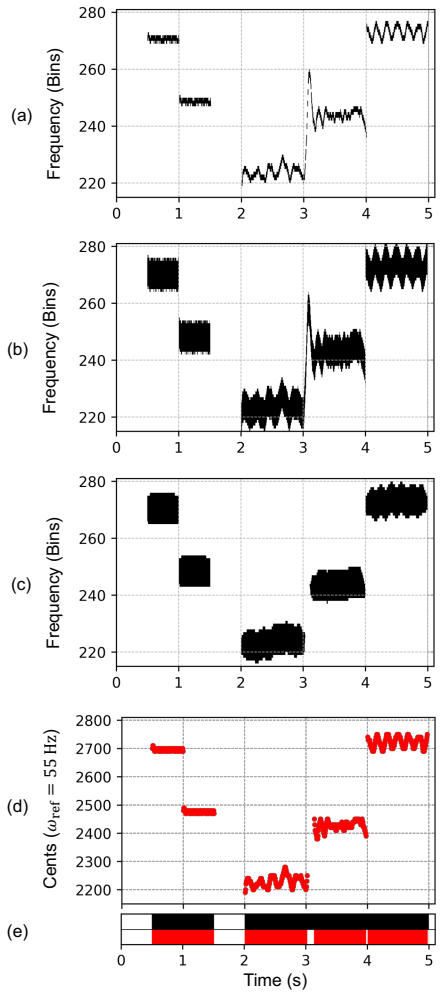

Figure 4. Masking approach for detecting stable regions. **(a)** Binary representation Γ_R. **(b)** Max-filtered representation Γ_R^β. **(c)** Median-filtered binary mask $\Gamma_R^{\beta,L}$. **(d)** Trajectory γ^{Mask} restricted to stable regions. **(e)** Activation regions for γ (black) and γ^{Mask} (red).

trajectory is still considered to be stable. The resulting trajectory γ^{Morph} for our running example is depicted in Figure 3c using a threshold of $\tau = 90$ cents. As shown in Figure 3d, the morphological approach succeeds in identifying stable regions. However, it also introduces a truncation at both sides of sudden jumps (e. g., around the first and fourth second) by half the filter length $(L-1)/2$. In the next section, we show how this truncation effect can be reduced by applying a 2D-masking approach involving some median filtering. Finally, we want to note that considering the morphological gradient is conceptionally similar to the approach based on Gaussian derivate filtering as described in [15]. In our approach, the threshold parameter τ can be adjusted dynamically to account for characteristics of individual trajectories, e. g. by considering the p-quantile of the morphological gradient Δ^L.

3.4 Masking Approach

We now introduce an alternative approach for detecting stable trajectory regions, which works in the 2D-domain. In a first step, we encode a trajectory γ as a binary 2D-representation $\Gamma_R : \mathbb{Z} \times \mathbb{Z} \to \{0, 1\}$. Given a frequency resolution of $R \in \mathbb{R}$ (given in cents), Γ_R is defined by

$$\Gamma_R(n, b) := \begin{cases} 1, & \text{for } \left\lfloor \frac{\gamma(n)}{R} + 0.5 \right\rfloor = b, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

with time index $n \in \mathbb{Z}$ and frequency bin index $b \in \mathbb{Z}$ (corresponding to a logarithmic frequency axis). Figure 4a shows the binary representation Γ_R using $R = 10$ cents for our running example. In the second step, we introduce some tolerance in frequency direction by vertically applying a max-filtering using a filter length parameter $\beta \in \mathbb{N}_0$ (specified in bins). This results in the representation Γ_R^β defined by

$$\Gamma_R^\beta(n, b) := \max\{\Gamma_R(n, b - \beta : b + \beta)\}. \tag{8}$$

This operation is illustrated by Figure 4b using $\beta = 5$ (leading to a frequency width of $2\beta + 1 = 11$ bins corresponding to 110 cents). In a third step, inspired by an algorithm for Harmonic–Percussive Source Separation [8], a median filter of odd length $L \in \mathbb{N}$ is applied in horizontal direction yielding a representation $\Gamma_R^{\beta,L}$:

$$\Gamma_R^{\beta,L}(n, b) := \text{median}\big\{\Gamma_R^\beta(n - \tfrac{L-1}{2} : n + \tfrac{L-1}{2}, b)\big\}. \tag{9}$$

Applying horizontal median filtering suppresses vertical structures (e. g., pitch slides), while enhancing horizontal structures (corresponding to stable regions), see Figure 4c for an illustration when using $L = 43$ (corresponding to 0.25 sec). In the fourth step, the output trajectory γ^{Mask} is obtained by setting

$$\gamma^{\text{Mask}}(n) := \begin{cases} \gamma(n), & \text{if } \Gamma_R^{\beta,L}(n, b) = 1, \\ *, & \text{otherwise,} \end{cases} \tag{10}$$

with $b = \left\lfloor \frac{\gamma(n)}{R} + 0.5 \right\rfloor$. This last step can be thought of as "masking" the input trajectory γ using the binary mask $\Gamma_R^{\beta,L}$. Figure 4d shows the resulting trajectory γ^{Mask} for our running example. Note that, even though the masking procedure involves some quantization parameter R, the final trajectory γ^{Mask} coincides with the original trajectory γ in stable regions. Similar to the parameter τ for computing γ^{Morph}, the parameter β controls the frequency tolerance within stable regions for γ^{Mask}. As also indicated by our running example, the truncation effects at sudden jumps introduced by the morphological approach have been eliminated by our masking approach (compare γ^{Morph} and γ^{Mask} around the first and fourth second). While the 2D-masking approach is computationally more expensive than the 1D-morphological approach, it allows for processing multiple (non-overlapping) trajectories at the same time. Furthermore, one may account for weighted trajectories (e. g., trajectories with assigned amplitude or confidence values) by using real-valued instead of binary masks. Note that both algorithms do not enforce continuity of output trajectories. In particular, strict parameter settings (e. g. small τ and small β) may result in fluctuating sound events (e. g. a note sung with strong vibrato) being split up into several disconnected regions.

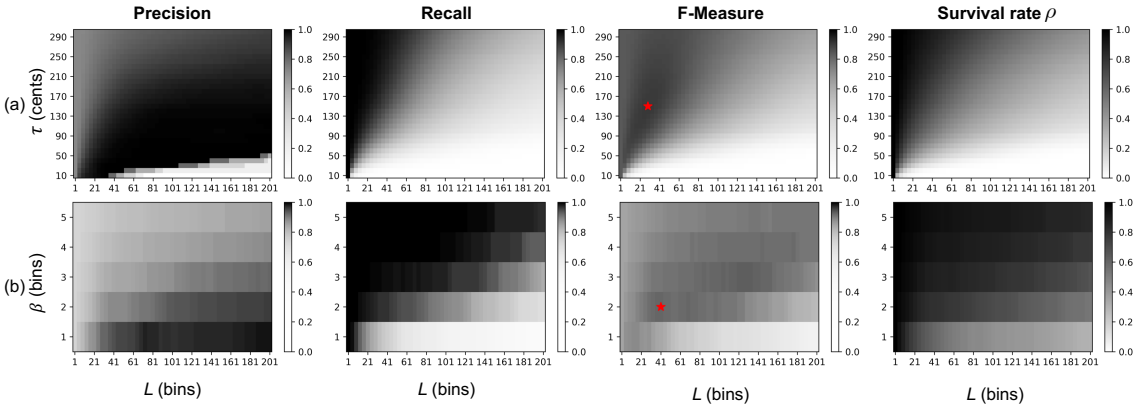

Figure 5. Precision, recall, F-Measure, and survival rate ρ of parameter sweeps averaged over five recordings (see Table 1). The parameter settings chosen for subsequent experiments are marked with red stars. **(a)** Morphological approach. **(b)** Masking approach.

ID	γ^{Anno}	γ^{Morph}				γ^{Mask}			
	ρ	P	R	F	ρ	P	R	F	ρ
001	61%	0.82	0.94	0.88	70%	0.82	0.94	0.88	71%
002	79%	0.94	0.85	0.89	72%	0.93	0.87	0.90	74%
010	68%	0.87	0.92	0.89	72%	0.84	0.95	0.89	77%
087	78%	0.88	0.98	0.93	87%	0.87	0.98	0.92	88%
110	74%	0.90	0.96	0.93	79%	0.88	0.97	0.92	80%

Table 1. Precision (P), recall (R), F-Measure (F), and survival rate (ρ) evaluated on the basis of manually annotated F0-trajectories for five Erkomaishvili recordings.

4. EVALUATION

In this section, we report on experiments that indicate the role of the parameters and the behavior of the morphological and the masking approach. In Section 4.1, we numerically compare both approaches using a set of manually annotated stable regions in F0-trajectories from the publicly available Erkomaishvili dataset [20]. Using suitable parameter settings, we then apply both algorithms to the trajectories of all 101 recordings in the dataset (see Section 4.2). It turns out that a consistent detection of stable regions using the two conceptually different approaches is a good indicator that the results are musically meaningful. Finally, in Section 4.3, we demonstrate the potential of our approaches for enhancing harmonic interval distributions.

4.1 Evaluation Measures and Parameters

In order to compare the algorithms' performance, we annotated stable regions of F0-trajectories extracted from five representative Erkomaishvili recordings. To this end, we used an interactive interface described in [20] to manually remove all unstable trajectory regions that correspond to note transitions and other artifacts. As evaluation metrics, we use standard precision (P), recall (R) and F-measure (F) computed frame-wise on the basis of the trajectories' activations. First, all frames with no specified frequency value in the original trajectory ($\gamma(n) = *$) are left unconsidered. Frames classified as stable by our approaches are counted as *true positives* (TP) if they agree with frames annotated as stable, otherwise they are counted as *false positives* (FP). Furthermore, frames annotated as unstable are counted as *false negatives* (FN), if they are classified as unstable. Then,

$$\mathrm{P} := \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \mathrm{R} := \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \mathrm{F} := \frac{2\cdot\mathrm{P}\cdot\mathrm{R}}{\mathrm{P}+\mathrm{R}}. \quad (11)$$

Note that $\mathrm{P} := 0$ for $\mathrm{TP}+\mathrm{FP} = 0$, $\mathrm{R} := 0$ for $\mathrm{TP}+\mathrm{FN} = 0$, and $\mathrm{F} := 0$ for $\mathrm{P}+\mathrm{R} = 0$. Furthermore, we introduce an evaluation measure referred to as *survival rate* and denoted as ρ. This measure, which indicates the percentage of remaining trajectory values after filtering, is defined as follows:

$$\rho := \frac{|\{n : \gamma^{\text{Stable}}(n) \neq *\}|}{|\{n : \gamma(n) \neq *\}|} \cdot 100, \quad (12)$$

with $\gamma^{\text{Stable}} = \gamma^{\text{Morph}}$ for the morphological approach, $\gamma^{\text{Stable}} = \gamma^{\text{Mask}}$ for the masking approach and $\gamma^{\text{Stable}} = \gamma^{\text{Anno}}$ for an annotated trajectory γ^{Anno}.

In order to analyze the algorithms' behavior for different parameter settings, we conduct parameter sweeps over L, τ, and β, using a fixed frequency resolution of $R = 10$ cents. For each evaluation metric, we construct a matrix with each entry corresponding to a metric's value for a specific parameter setting averaged over the five annotated recordings. The resulting matrices for precision, recall, F-measure, and survival rate are depicted in Figure 5a for the morphological approach and in Figure 5b for the masking approach. The visualizations show that τ and β play a similar role: high values of τ and β make the approaches more tolerant to local frequency fluctuations in the trajectories, thus increasing the survival rates. In contrast, when decreasing τ and β, less values remain in the filtered trajectories, leading to lower survival rates. Furthermore, note that increasing the filter length L leads to an increase in precision and a decrease in recall for both approaches. In the case of the morphological approach, very large filter lengths lead to a survival rate of $\rho = 0$ (nothing is remaining), which also leads to a precision of zero.

For our further experiments, we use fixed parameter settings for both approaches that correspond to maxima in the F-measure matrices (see red stars in Figure 5). The

	P	R	F	$\rho(\gamma^{\text{Morph}})$	$\rho(\gamma^{\text{Mask}})$
μ	0.89	0.94	0.92	73%	77%
σ	0.02	0.01	0.02	5%	5%

Table 2. Evaluation of the masking approach against the morphological approach considering the trajectories of all 101 recordings of the Erkomaishvili dataset (with fixed parameter settings from Section 4.1). The mean μ and standard deviation σ refer to statistics taken over the dataset.

morphological approach reaches a maximum F-measure of 0.90 for $\tau = 150$ cents and $L = 29$ bins, whereas the masking approach reaches a maximum F-measure of 0.90 for $\beta = 2$ bins and $L = 41$ bins. Using these parameter settings, the evaluation results for our five annotated examples (IDs correspond to songs on the publicly available website [2]) are given in Table 1. From the table, we can see that both approaches are able to detect stable regions in all five examples. We want to note that the optimal parameter settings vary from song to song, depending on the occurring note durations, characteristics of pitch slides, and other performance aspects. As an alternative to a fixed setting, one may chose the parameters in a song-dependent way, e. g., by fixing the survival rate. In summary, our experiments on the Erkomashvili dataset showed that the specific choice of parameters is not crucial within a certain range (see also the F-measure matrices of Figure 5).

4.2 Consistency

The two approaches for detecting stable regions in trajectories are conceptually different. Nevertheless, in the case of the five annotated recordings, both approaches worked successfully and performed in a similar fashion. Based on the hypothesis that a consistent performance of both approaches is a necessary condition for obtaining meaningful results, we applied both approaches independently to all 101 recordings of the Erkomaishvili dataset. We then compared the results by evaluating the trajectories obtained by the masking approach against the trajectories obtained by the morphological approach using the evaluation metrics defined in Section 4.1. The mean μ and standard deviation σ (taken over the dataset) of the evaluation results are shown in Table 2. The numbers indicate that both approaches deliver similar results on average with a small standard deviation. Furthermore, both approaches roughly exhibit the same average survival rate for the chosen parameter settings. Beyond these overall measures, we also looked at recordings where the two approaches delivered less consistent results. A manual inspection revealed that these recordings often contain speech-like passages (rather than singing) and extremely short notes such as in the songs with ID 022 and ID 074. Results for all 101 recordings are publicly available through audio–visual interfaces. [3]

[2] https://www.audiolabs-erlangen.de/resources/MIR/2017-GeorgianMusic-Erkomaishvili

[3] https://www.audiolabs-erlangen.de/resources/MIR/2019-ISMIR-StableF0

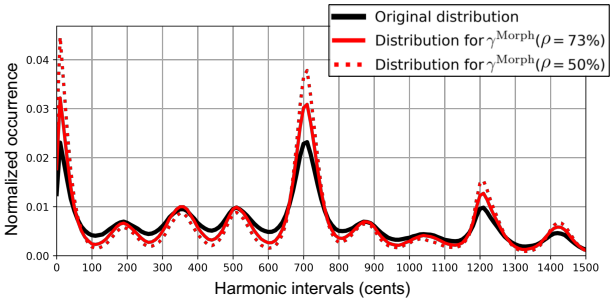

Figure 6. Harmonic interval distributions obtained from the entire Erkomaishvili dataset.

4.3 Harmonic Interval Analysis

In the following, we want to demonstrate the potential of the presented approaches for interval analysis of Georgian vocal music by computing harmonic interval size distributions from the filtered trajectories. To this end, similar to [20, 31], we superimpose the filtered trajectories of lead, middle and bass voice and determine the frame-wise intervals for each voice pair (as indicated in Figure 1). Then, by accumulating the occurrences of the different intervals over time, we obtain interval histograms. These histograms are normalized (using the ℓ^1-norm) to obtain distributions. Figure 6 shows three such distributions obtained by considering all 101 recordings of the Erkomaishvili dataset. The first distribution (black solid line) is based on the original F0-trajectories. The second distribution (solid red line) is obtained by considering only stable regions after morphological filtering. (Here, we use the parameter setting discussed in Section 4.1. Filtering with the masking approach leads to similar distributions.) Note that the filtering leads to a sharper interval distribution emphasizing the peaks at the harmonically relevant intervals while not changing the respective peak locations. Using stricter parameter settings leads to a further sharpening (see red doted line in Figure 6). However, overdoing the filtering may drastically reduce the survival rate. This, in turn, may lead to a distortion or even a loss of peak structures corresponding to relevant harmonic intervals.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented two conceptually different approaches for detecting stable regions in frequency trajectories, which perform equally well with respect to a set of manually annotated trajectories. Rather than advocating a specific parameter setting, our goal was to introduce these concepts in a mathematical rigorous way, while highlighting their potential using the Erkomaishvili dataset as example scenario. Going beyond harmonic interval analysis, future work will be concerned with applying these filtering techniques for the analysis of melodic intervals, singer interaction, and intonation drifts—aspects of foremost importance in ethnomusicological research on traditional Georgian vocal music.

Acknowledgements: 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.

6. REFERENCES

[1] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, 36(1):20–30, 2019.

[2] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434, 2013.

[3] Paul Brossier, Juan Pablo Bello, and Mark D. Plumbley. Fast labelling of notes in music signals. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Barcelona, Spain, 2004.

[4] Paul Brossier, Juan Pablo Bello, and Mark D. Plumbley. Real-time temporal segmentation of note objects in music signals. In *Proceedings of the International Computer Music Conference (ICMC)*, Miami, Florida, USA, 2004.

[5] Georgios Chrysochoidis, Georgios Kouroupetroglou, and Sergios Theodoridis. Vibrato detection in byzantine chant music. In *International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pages 636–639, Athens, Greece, 2014.

[6] José Miguel Díaz-Báñez and Juan-Carlos Rizo. An efficient DTW-based approach for melodic similarity in flamenco singing. In *International Conference on Similarity Search and Applications*, pages 289–300, 2014.

[7] Malkhaz Erkvanidze. The Georgian musical system. In *Proceedings of the International Workshop on Folk Music Analysis*, pages 74–79, Dublin, Ireland, 2016.

[8] Derry FitzGerald. Harmonic/percussive separation using median filtering. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 246–253, Graz, Austria, September 2010.

[9] Kaustuv Kanti Ganguli and Preeti Rao. On the distributional representation of ragas: experiments with allied raga pairs. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 1(1):79–95, 2018.

[10] Emilia Gómez and Jordi Bonada. Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to A cappella singing. *Computer Music Journal*, 37(2):73–90, 2013.

[11] Eric J. Humphrey, Sravana Reddy, Prem Seetharaman, Aparna Kumar, Rachel M. Bittner, Andrew Demetriou, Sankalp Gulati, Andreas Jansson, Tristan Jehan, Bernhard Lehner, Anna Krupse, and Luwei Yang. An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music. *IEEE Signal Processing Magazine*, 36(1):82–94, 2019.

[12] Vignesh Ishwar, Shrey Dutta, Ashwin Bellur, and Hema A. Murthy. Motif spotting in an alapana in carnatic music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 499–504, Curitiba, Brazil, 2013.

[13] Anssi P. Klapuri and Manuel Davy, editors. *Signal Processing Methods for Music Transcription*. Springer, New York, 2006.

[14] Gopala Krishna Koduri, Sankalp Gulati, Preeti Rao, and Xavier Serra. Rāga recognition based on pitch distribution methods. *Journal of New Music Research*, 41(4):337–350, 2012.

[15] Nadine Kroher and Emilia Gómez. Automatic transcription of Flamenco singing from polyphonic music recordings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):901–913, 2016.

[16] 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*, May 2015.

[17] Matthias Mauch, Klaus Frieler, and Simon Dixon. Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory. *Journal of the Acoustical Society of America (JASA)*, 136(1):401–411, 2014.

[18] Andrew McLeod, Rodrigo Schramm, Mark Steedman, and Emmanouil Benetos. Automatic transcription of polyphonic vocal music. *Applied Sciences*, 7(12), 2017.

[19] Emilio Molina, Lorenzo J. Tardón, Ana M. Barbancho, and Isabel Barbancho. Sipth: Singing transcription based on hysteresis defined on the pitch-time curve. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 23(2):252–263, 2015.

[20] Meinard Müller, Sebastian Rosenzweig, Jonathan Driedger, and Frank Scherbaum. Interactive fundamental frequency estimation with applications to ethnomusicological research. In *Proceedings of the AES International Conference on Semantic Audio*, pages 186–193, Erlangen, Germany, 2017.

[21] Aleksey Nikolsky. Evolution of tonal organization in music mirrors symbolic representation of perceptual reality. part-1: Prehistoric. *Frontiers in Psychology*, 6:1405, 2015.

[22] Aleksey Nikolsky. Evolution of tonal organization in music optimizes neural mechanisms in symbolic encoding of perceptual reality. part-2: Ancient to seventeenth century. *Frontiers in Psychology*, 7:211, 2016.

[23] Ryo Nishikimi, Eita Nakamura, Masataka Goto, Katsutoshi Itoyama, and Kazuyoshi Yoshii. Scale- and rhythm-aware musical note estimation for vocal F0 trajectories based on a semi-tatum-synchronous hierarchical hidden semi-markov model. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 376–382, Suzhou, China, 2017.

[24] Maria Panteli, Rachel M. Bittner, Juan Pablo Bello, and Simon Dixon. Towards the characterization of singing styles in world music. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 636–640, New Orleans, LA, USA, 2017.

[25] Preeti Rao, Joe Cheri Ross, Kaustuv Kanti Ganguli, Vedhas Pandit, Vignesh Ishwar, Ashwin Bellur, and Hema A. Murthy. Classification of melodic motifs in raga music with time-series matching. *Journal of New Music Research*, 43(1):115–131, 2014.

[26] Lise Regnier and Geoffroy Peeters. Singing voice detection in music tracks using direct voice vibrato detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1685–1688, Taipei, Taiwan, 2009.

[27] Rafael Caro Repetto, Rong Gong, Nadine Kroher, and Xavier Serra. Comparison of the singing style of two jingju schools. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 507–513, 2015.

[28] Jean-François Rivest, Pierre Soille, and Serge Beucher. Morphological gradients. *Journal of Electronic Imaging*, 2(4):326–336, 1993.

[29] Joe Cheri Ross, Vinutha T. P., and Preeti Rao. Detecting melodic motifs from audio for hindustani classical music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 193–198, Porto, Portugal, 2012.

[30] Matti Ryynänen and Anssi P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008.

[31] Frank Scherbaum, Meinard Müller, and Sebastian Rosenzweig. Analysis of the Tbilisi State Conservatory recordings of Artem Erkomaishvili in 1966. In *Proceedings of the International Workshop on Folk Music Analysis*, pages 29–36, Málaga, Spain, 2017.

[32] Johan Sundberg. Perceptual aspects of singing. *Journal of voice*, 8(2):106–122, 1994.

[33] Zaal Tsereteli and Levan Veshapidze. On the Georgian traditional scale. pages 288–295, Tbilisi, Georgia, 2014.

[34] František Vávra, Pavel Nový, Hana Mašková, Michala Kotlíková, and Arnoštka Netrvalová. Morphological filtration for time series. In *Conference on Applied Mathematics (APLIMAT)*, pages 983–990, Bratislava, Slovakia, 2004.

[35] Luwei Yang, Elaine Chew, and Khalid Z. Rajab. Logistic modeling of note transitions. In *International Conference on Mathematics and Computation in Music (MCM)*, pages 161–172, London, UK, 2015.

[36] Luwei Yang, Khalid Z. Rajab, and Elaine Chew. AVA: an interactive system for visual and quantitative analyses of vibrato and portamento performance styles. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 108–114, New York City, USA, 2016.

[37] Luwei Yang, Khalid Z. Rajab, and Elaine Chew. The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation. *Journal of Mathematics and Music*, 11(1):42–60, 2017.