pare the performance of the six estimation algorithms on the dataset and that support the use of median absolute deviation as a measure of confidence in the estimates. We also introduce two types visualization that shed some light on musical content of the collection. Finally we present an illustration of the web interface of the dataset. [4]

2. PROCESS AND WORKFLOW

2.1 Estimates and alignment

The first step in our process is to apply six algorithms for monophonic f_0 estimation to each recording. The outputs of these algorithms are shifted in time, typically by a few hundredths of a second. We align the six estimates in time by fixing the estimates of one arbitrarily-chosen algorithm (Boersma) and shifting each of the others by multiples of $0.01s$. We choose the shift that minimizes $d(\delta t)$, the time-lagged ℓ_2-distance between the two time series at lag δt. We limit our search to the range $[-0.1, 0.1]$ because the vibrato of the human voice is $5-7Hz$, which is the Nyquist frequency of a sample rate of $10 - 14Hz$ [27].

Figure 1 shows a plot of $d(\delta t)$ for one song in the collection. In most cases δt_0 is very close to 0.

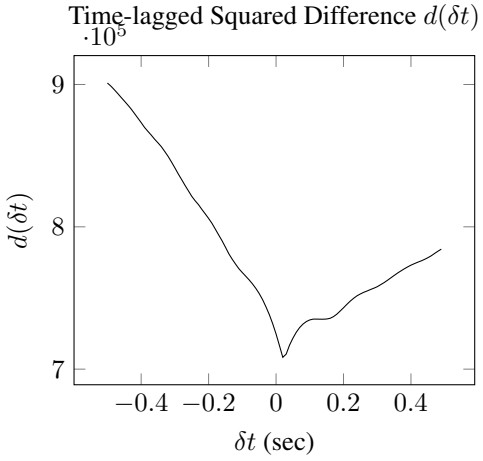

Figure 1. The squared difference between Boersma at time t and another f_0 estimator at time $t + \delta t$ for the middle part of *Ak'a Si Rekisho* from the Megrelian collection.

2.2 Note estimation

The next step in our process is a heuristic method for identifying notes in the target estimate. We chose a method that is simple to code and produces note values that serve well as a visual aid for the analyst, as opposed to a state-of-the-art method such as [28]. For purposes of post-processing non-note sections longer than $0.2s$ are considered unvoiced and the rest are considered voiced consonants.

The idea behind the heuristic is that a note transition is typically a monotone change in pitches between a local peak and a local trough in the singer's vibrato. The heuristic traverses the time series forward and identifies a note transitions as a change of 7% from the most recent local peak or trough. The pitch value of the note estimate is the average of the target estimate pitches over the duration of the note. The heuristic sets a minimum note length of $0.07s$. The heuristic also identifies as *transitions* monotone sequences of pitches between notes and increasing sequences of pitches before notes. Anything else is a non-note. We tuned the parameters during development. The 7% note transition threshold typically errs on the side of false negative note transitions.

[4] The web interface for the dataset makes available the f_0 estimates of the three vocal parts of each song [26].

2.3 Voiced-unvoiced detection

In parallel to the two steps just described, a heuristic "voiced/unvoiced" method removes non-salient voice parts that are present when the salient voice part is silent. This heuristic is a simple threshold based on the histogram of amplitude and the technique of weighted zero crossings. The recording is divided into overlapping frames of length $0.015s$ spaced $0.01s$ apart. For each frame the root mean squared of amplitude is divided by the zero crossing rate. A histogram of the ratio is created and smoothed using a Hanning window. The smoothed histogram typically has a local maximum near 0 that is due to sections of silence in the recording. There are typically one or two other local maxima near 0 that correspond to silence in the salient part. The "voiced" threshold is characterized by a high local maximum preceded by a low local minimum.

We tuned the parameters to result typically in some false positive sections of the recording that require manual correction but only isolated false negatives that can be corrected automatically. Figure 2 shows the smoothed histogram for one part of one song.

In Georgian music there is often a *decrescendo* in each voice at the end of a phrase. The threshold produced by the algorithm tends to correctly classify the ends of phrases but also tends to generate false *voiced* classifications during the subsequent silence. To combat this problem we added an automatic post-processing step which reclassifies short *voiced* sections within *unvoiced* sections.

We hand-tuned each of the heuristics mentioned above on a sample of varied musical sections, with the goal of minimizing the use of the interactive tool.

2.4 Interactive tool

The analyst uses a graphical interface to construct a target f_0 estimate for each part by stitching together the best estimates for different sections of the part. In sections of a part where no estimate is correct the analyst can specify a pitch range. The interface presents a window containing plots of the time series of f_0 estimates for one part of one song. The analyst can switch between parts and can turn the visibility of each time series on and off. The time series include the f_0 estimates of the six algorithms before and after application of the voiced-unvoiced algorithm. The window also includes the *target estimate*, which is initially equal to the CREPE time series, and note estimates for the target estimate. Selection tools and buttons enable the main functionalities of the tool, which are as follows:

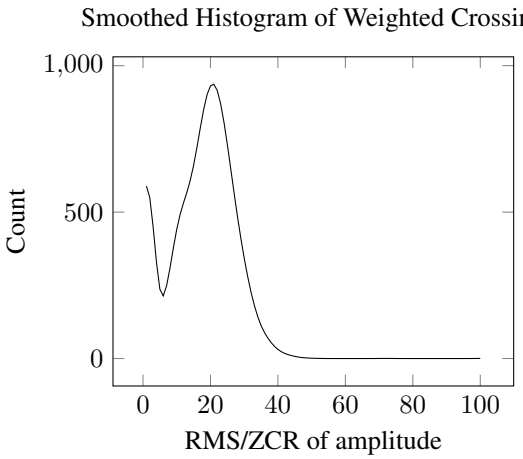

Figure 2. Smoothed histogram of weighted crossing for the top part of *Ak'a Si Rekisho* from the Megrelian collection

- to delete a section of the target estimate, i.e., set $f_0 = 0$ at selected times
- to set a section of the target estimate equal to the estimate of a selected algorithm
- to override a section of the target estimate with a pitch range

After using the tool on one part of a song, the analyst post-processes and saves the target estimate.

Figure 3 shows a screenshot of the tool in operation.

Figure 3. The interactive tool in operation at $84s$ in the middle part of *Adila-Alipasha* from the Gurian collection.

The last step of the process is to rerun the f_0 algorithms Noll, Hermes, Boersma, and Yin, constrained to be within a fixed percentage of the target estimate or within the pitch range. For CREPE and Maddox we use implementations that do not allow pitch constraints as input. If an algorithm is unable to find an estimate within the pitch range, its estimate is dropped.

2.5 Automatic Post-processing

The post-processing algorithm makes the following changes to the analyst's final choice of target estimate:

1. Revert isolated $(0.01 - 0.02s)$ unvoiced pitches to the estimates that were made before the application of the voiced-unvoiced algorithm.
2. Set unvoiced non-note sections (those shorter than $0.2s$) of the target estimate to 0.
3. Set voiced non-note sections of the target estimate to -10 to indicate *undecided*.

2.6 Manual Post-processing

The process described above generally results in a median absolute deviation of the estimates that is less than 1% of the median of estimates. (See Figure 6.) There are exceptions, sections of harmonic content where the estimates disagree. In some cases, typically sustained, well-tuned chords, it is clearly audible that two or more estimates are incorrect and skew the result. In these cases we have manually set a narrow pitch range in the interest of accuracy in the annotations.

3. RESULTS

In this section we discuss qualitatively the effectiveness of our method of combining multiple f_0 estimates. We provide descriptive statistics about the f_0 estimates that we include in the dataset of Gurian and Megrelian songs. First, we assess the estimates produced by each of the six algorithms on this dataset by comparing the initial estimate of each algorithm with the final estimate. Second, we assess the variability of the estimates as a measure of confidence in the final estimate. Third, we provide visualizations of the pitches and intervals of one song, *Gepshvat Ghvini*. Finally, we describe the web interface for the dataset and illustrate it with an example.

3.1 Discussion of the method

In most situations our process of combining multiple f_0 estimates "works" in the sense of producing near agreement among estimates where the fundamental frequency is unambiguous to the human ear. There is one noticeable failure mode. On non-monophonic input Hermes and Yin may fail to produce an estimate within given upper and lower frequency limits, and Boersma may label the sample *unvoiced*. In particular, Hermes, Yin, and Boersma tended to produce no estimate or incorrect estimates on prolonged, well-tuned chords. This is not a noticeable problem for Noll. We do not have a theoretical explanation, but we have found cases where the normalized difference function of Yin either does not have a minimum on the given interval or the location of minimum is unstable. On a small fraction of samples our process produced only two estimates or widely varying estimates.

3.2 Assessing each f_0 estimate

In this subsection we measure the accuracy of each monophonic f_0-estimation algorithm relative to the final f_0 estimate, separately on each collection. We consider the f_0 estimate of each algorithm, after applying alignment and the voiced-unvoiced algorithm, limited to samples that the note estimation algorithm labels as notes. Figures 4 and 5 show one smoothed histogram for each algorithm; it is a histogram of the ratio of the algorithm's f_0 estimate to the final f_0 estimate. By this measure Boersma is the most accurate when it is in the right octave (near 1.0), but Crepe and Noll are in the right octave more often.

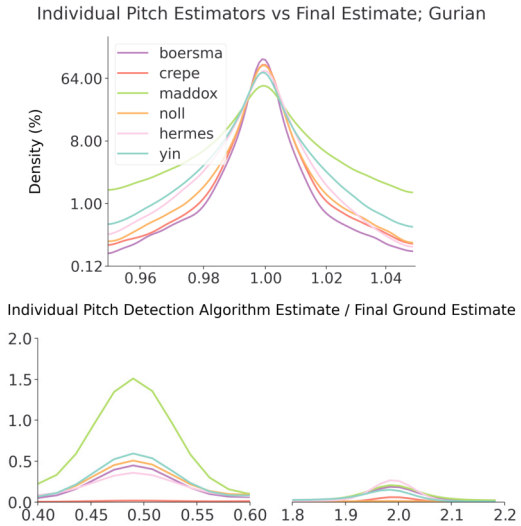

Figure 4. Smoothed histogram of ratio of f_0 estimates, algorithm/final, Gurian collection. Percentage of samples shown: boersma: 85.7%, crepe: 94.7%, maddox: 56.9%, noll: 85.9%, hermes: 86.0%, yin: 81.7%.

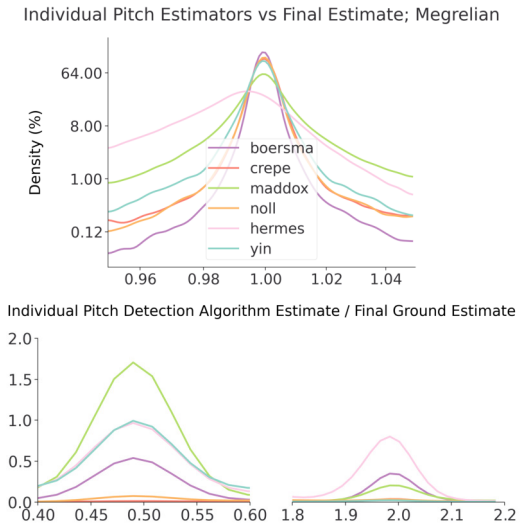

Figure 5. Smoothed histogram of ratio of f_0 estimates, algorithm/final, Megrelian collection. Percentage of samples shown: boersma: 84.3%, crepe: 96.3%, maddox: 63.1%, noll: 91.8%, hermes: 63.0%, yin: 79.4%.

3.3 Variability of the f_0 estimates

At each time step the final f_0 estimate is the median of the estimates of several algorithms. In this subsection we measure the variability of the estimates around the median. This measurement lends credibility to our method by showing that when there is a note (according to the note estimation algorithm) the variability is small (the estimates agree) and when there is not a note the variability is large (the estimates disagree). For our measure of variability we choose the median absolute deviation (MAD) around the median. This measure is robust to outliers, which means that when there is agreement among at least three estimates the MAD will be small. This choice corresponds to our intuition that during a vowel there will be general agreement among estimates and during a consonant there will be general disagreement. Figure 6 shows two graphs of smoothed histograms of MAD (in Hz), one for *note* samples and one for *non-note* samples. The graphs show 97.7% of Gurian note samples and 96.1% percent of Megrelian note samples. They show 70.9% percent of non-note Gurian samples and 74.7% of non-note Megrelian samples.

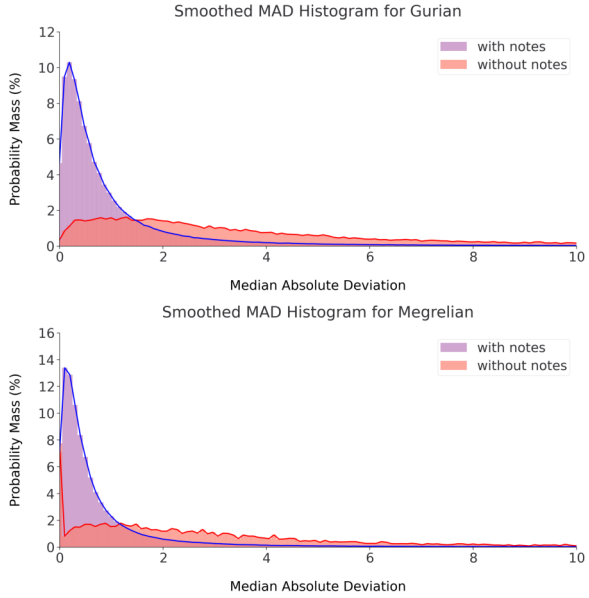

Figure 6. Smoothed histograms of MAD of each estimate, in Hz, *note* vs. *non-note* samples. Top: Gurian collection. Bottom: Megrelian collection.

3.4 Musical Observations

In this subsection we provide visualizations of the pitches and intervals in one Megrelian song, *Gepshvat Ghvini*. The purpose of these visualizations is to illustrate the potential for visualizations of digital data to enable and inform research in ethmomusicology.

We return to the question mentioned in the introduction of what intervals make up the Georgian scale, and to what extent singers deviate from the scale to produce desired harmonies. Figure 7 shows a histogram of the pitches of the three voices for *Gepshvat Ghvini*. The variation in

the peaks of the histograms for the three voices between 210Hz and 220Hz possibly indicates deviation from a fixed scale. Further investigation is needed.

Figure 8 shows a two-dimensional histogram of the chords of *Gepshvat Ghvini*. The x-axis is the ratio of pitches between the middle and bass parts. The y-axis is the ratio of pitches between the top and middle parts. The histogram is shown as a heatmap. The dark patch roughly at coordinates $(1.22, 1.22)$ shows that triads with a perfect fifth are common, and that these triads cluster around those with a "neutral" third between the minor and major third.

To exclude pitches in transition between notes we have limited the data in both figures to pitches at least three time steps away from the first and last pitch of each note, as designated by the note estimation algorithm. In the two-dimensional histogram we have limited the data to samples where we have estimates for all three parts.

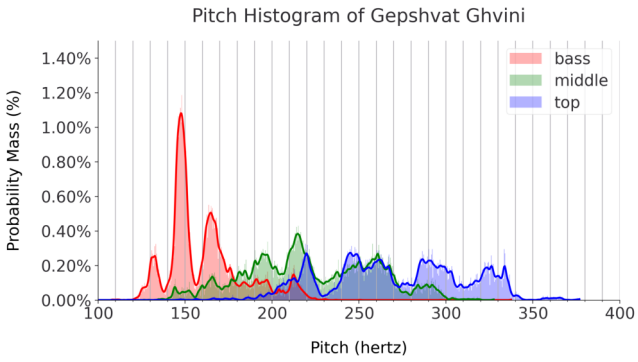

Figure 7. Histogram of pitches in *Gepshvat Ghvini*.

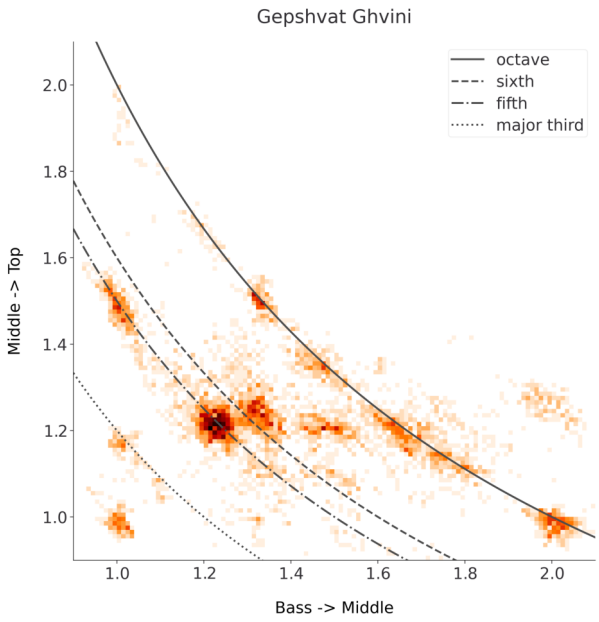

Figure 8. Histogram of chords in *Gepshvat Ghvini*. The x-axis is the ratio of middle to bass pitch. The y-axis is the ratio of top to middle pitch.

3.5 Web interface

In this subsection we describe the web interface for our dataset and illustrate the interface with an example. The web interface serves as a visual tool for the analyst to assess the accuracy of the results, as well as for ethnomusicologists to analyze of songs and for singers to learn the songs. The interface shows a scrolling graph of pitches that is synchronized to an audio player. Figure 9 shows a snapshot of the interface for the Megrelian song *Gepshvat Ghvini*. The graph optionally displays any of the median estimates of the three parts and the median absolute deviations (MAD) of the estimates.

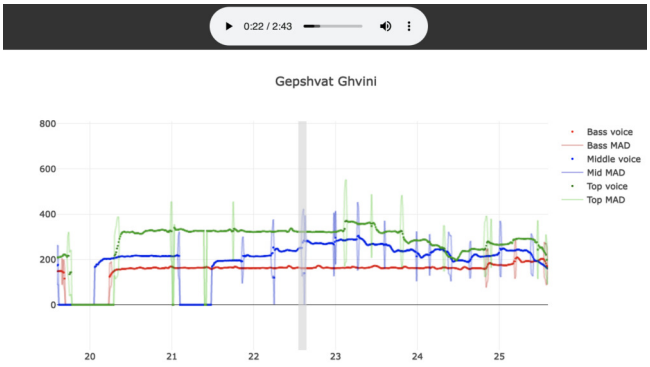

Figure 9. Web interface: *Gepshvat Ghvini*. The shaded line shows the current time. Each solid line shows the median frequency estimate for one part; the faint line of the same color shows the median absolute deviation.

4. CONCLUSION

We have presented a process and workflow for multi-f_0 annotation of a dataset of songs from the Republic of Georgia in which each song is represented by four recordings from different microphones. The annotations in our dataset represent the average of f_0 from different algorithms and are accompanied by the median absolute deviation of the estimates, which we have shown is a reasonable measure of confidence in the estimate. We hope our dataset and web interface are of interest to ethnomusicologists as audiovisual aids for analysis.

We plan to apply our process to collections of recordings from two other regions of Georgia which have been produced recently [29, 30]. We plan to use the resulting labeled datasets to develop algorithms for multi-f_0 estimation using the recordings of the mixed voices from the room microphones in our collections. We plan to develop the web interface further as a learning aid for singers by adding new features, including lyrics as subtitles.

5. ACKNOWLEDGEMENTS

The authors would like to thank the International Centre for Georgian Folk Song and the International Research Center for Traditional Polyphony, for permission to share the audio data with researchers.

6. REFERENCES

[1] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, "MedleyDb: A Multitrack Dataset for Annotation-Intensive MIR Research," in *Proceedings of the 15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, Oct. 2014.

[2] S. Rosenzweig, H. Cuesta, C. Weiß, F. Scherbaum, E. Gómez, and M. Müller, "Dagstuhl ChoirSet: A Multitrack Dataset for MIR Research on Choral Singing," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 98–110, Jul. 2020, number: 1 Publisher: Ubiquity Press. [Online]. Available: http://transactions.ismir.net/articles/10.5334/tismir.48/

[3] S. Rosenzweig, F. Scherbaum, D. Shugliashvili, V. Arifi-Müller, and M. Müller, "Erkomaishvili Dataset: A Curated Corpus of Traditional Georgian Vocal Music for Computational Musicology," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 31–41, Apr. 2020, number: 1 Publisher: Ubiquity Press. [Online]. Available: http://transactions.ismir.net/articles/10.5334/tismir.44/

[4] F. Scherbaum, N. Mzhavanadze, S. Rosenzweig, and M. Müller, "Multi-Media Recordings Of Traditional Georgian Vocal Music For Computational Analysis," in *Proceedings of the 9th International Workshop on Folk Music Analysis*, Birmingham, Jul. 2019.

[5] H. Cuesta, B. McFee, and E. Gómez, "Multiple F0 Estimation in Vocal Ensembles using Convolutional Neural Networks," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*, Montréal, Canada, 2020, arXiv: 2009.04172. [Online]. Available: http://arxiv.org/abs/2009.04172

[6] M. Mauch, C. Cannam, R. M. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon, "Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency," in *Conference Proceedings of the First International Conference on Technologies for Music Notation and Representation*, 2015.

[7] ——, "Tony: a tool for melody transcription, https://code.soundsoftware.ac.uk/projects/tony," 2015. [Online]. Available: https://code.soundsoftware.ac.uk/projects/tony

[8] M. Müller, S. Rosenzweig, J. Driedger, and F. Scherbaum, "Interactive Fundamental Frequency Estimation with Applications to Ethnomusicological Research," in *Proceedings of the Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, Jun. 2017. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=18777

[9] F. Scherbaum, N. Mzhavanadze, S. Arom, S. Rosenzweig, and M. Müller, "Tonal Organization of the Erkomaishvili Dataset: Pitches, Scales, Melodies and Harmonies," *Computational Analysis Of Traditional Georgian Vocal Music*, 2020, publisher: Universitätsverlag Potsdam. [Online]. Available: https://publishup.uni-potsdam.de/frontdoor/index/index/docId/47614

[10] A. Erkomaishvili, *Georgian Folk Music. Guria. From Artem Erkomaishvili's Collection.* Tbilisi, Georgia: International Centre for Georgian Folk Song, 2005.

[11] S. Gelzer, "Testing a Scale Theory for Georgian Folk Music," in *The First International Symposium on Traditional Polyphony, Proceedings*. International Research Center for Traditional Polyphony, Sep. 2002, pp. 194–200. [Online]. Available: https://drive.google.com/file/d/1t1wNS_d2mBnPzeZnqSOe8QUymowiriwU/view?usp=drive_open&usp=embed_facebook

[12] M. Erkvanidze, "On Georgian Scale System," in *The First International Symposium on Traditional Polyphony, Proceedings*. International Research Center for Traditional Polyphony, Sep. 2002, pp. 178–185. [Online]. Available: https://drive.google.com/file/d/1MT2fdZwhrkMitpDjDYZBwTtxS_jl1DT0/view?usp=drive_open&usp=embed_facebook

[13] Z. Tsereteli and L. Veshapidze, "On the Georgian Traditional Scale," in *The Seventh International Symposium on Traditional Polyphony, Proceedings*. Tbilisi, Georgia: International Research Center for Traditional Polyphony, Sep. 2014, pp. 288–295. [Online]. Available: https://drive.google.com/file/d/1iVzCX1jAXnMnv_rJWNYEDdtOPTARqUzj/view?usp=drive_open&usp=embed_facebook

[14] D. Shugliashvili, *Georgian Church Hymns, Shemokmedi School.* Georgian Chanting Foundation, 2014.

[15] N. Razmadze, "Private communication," Mar. 2022.

[16] A. M. Noll, "Short-Time Spectrum and "Cepstrum" Techniques for Vocal-Pitch Detection," *The Journal of the Acoustical Society of America*, vol. 36, no. 296, 1964.

[17] ——, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate," in *Proceedings of the Symposium on Computer Processing in Communication*, vol. 19. New York, NY, USA: Microwave Institute, University of Brooklyn, 1970, pp. 779–797.

[18] D. Hermes, "Measurement of pitch by subharmonic summation," *The Journal of the Acoustical Society of America*, vol. 83, no. 257, 1988.

[19] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *The Journal of the Acoustical Society of America*, vol. 71, pp. 679–688, 1982.

[20] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings 17*, 1993, pp. 97–110.

[21] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[22] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002, publisher: Acoustical Society of America. [Online]. Available: https://asa.scitation.org/doi/10.1121/1.1458024

[23] E. Larson and R. K. Maddox, "Real-Time Time-Domain Pitch Tracking Using Wavelets," in *Proceedings of the University of Illinois at Urbana Champaign Research Experience for Undergraduates Program*, 2005.

[24] J. W. Kim, J. Salamon, P. Li, and J. Bello, "Crepe: A Convolutional Representation for Pitch Estimation," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[25] D. Gillman, U. Goyat, and A. Kutlay, "Teach Yourself Georgian Folk Songs Dataset: Code," 2021. [Online]. Available: https://github.com/New-College-of-Florida/Voice

[26] ——, "Teach Yourself Georgian Folk Songs Dataset: Website," 2021. [Online]. Available: http://131.247.152.67/georgian

[27] M. Vetterli, J. Kovačević, and V. K. Goyal, *Foundations of Signal Processing*. Cambridge University Press, Sep. 2014. [Online]. Available: https://www.cambridge.org/highereducation/books/foundations-of-signal-processing/DCC08E20D354F34E084FC11862E18F18

[28] R. Nishikimi, E. Nakamura, K. Itoyama, and K. Yoshii, "MUSICAL NOTE ESTIMATION FOR F0 TRAJECTORIES OF SINGING VOICES BASED ON A BAYESIAN SEMI-BEAT-SYNCHRONOUS HMM," *New York City*, p. 7, 2016.

[29] M. Khardziani and N. Razmadze, "Acharan Folk Songs – Collection Of Sheet Music With CD For Self-Study – International Research Center for Traditional Polyphony," 2020. [Online]. Available: http://polyphony.ge/en/acharan-folk-songs-collection-of-sheet-music-with-cd-for-self-study/

[30] R. Tsurtsumia and N. Razmadze, "Svan Folk Songs," May 2020, section: Books. [Online]. Available: https://chanting.ge/en/svan-folk-songs/

ADAPTING METER TRACKING MODELS TO LATIN AMERICAN MUSIC

Lucas S. Maia[1] **Martín Rocamora**[2] **Luiz W. P. Biscainho**[1] **Magdalena Fuentes**[3]
[1] PEE/COPPE, Federal University of Rio de Janeiro, Brazil
[2] FING, Universidad de la Republica, Uruguay
[3] MARL-IDM, New York University, United States
lucas.maia@smt.ufrj.br

ABSTRACT

Beat and downbeat tracking models have improved significantly in recent years with the introduction of deep learning methods. However, despite these improvements, several challenges remain. Particularly, the adaptation of available models to underrepresented music traditions in MIR is usually synonymous with collecting and annotating large amounts of data, which is impractical and time-consuming. Transfer learning, data augmentation, and fine-tuning techniques have been used quite successfully in related tasks and are known to alleviate this bottleneck. Furthermore, when studying these music traditions, models are not required to generalize to multiple mainstream music genres but to perform well in more constrained, homogeneous conditions. In this work, we investigate simple yet effective strategies to adapt beat and downbeat tracking models to two different Latin American music traditions and analyze the feasibility of these adaptations in real-world applications concerning the data and computational requirements. Contrary to common belief, our findings show it is possible to achieve good performance by spending just a few minutes annotating a portion of the data and training a model in a standard CPU machine, with the precise amount of resources needed depending on the task and the complexity of the dataset.

1. INTRODUCTION

Meter tracking means following the pulsating temporal structure of music from audio signals, which implies identifying at least beats and downbeats [1]. It is a long-standing area of research in music information retrieval (MIR) with applications ranging from automatic DJ mixing [2] to musicological studies [3]. Meter tracking has gone through a big transformation in the last decade due to the introduction of deep learning (DL) techniques [4–7], which brought an improvement in performance as well as a change in the design paradigm of related methods [8]. Nowadays, beat and downbeat tracking models rely mostly on supervised DL [8], and thus become data-driven and requiring large amounts of annotated data to generalize to different songs, genres or datasets.

This dependence on annotated data poses many challenges to the widespread use and adoption of such models, especially for culturally specific music traditions [9–11], which often lack annotated data as producing annotations requires culturally-aware expertise. For this reason, off-the-shelf general-purpose models typically underperform in these music genres since they are underrepresented in the datasets used for training. Nevertheless, previous work on some Latin American music traditions shows that if annotations are available, training statistical models can produce good performance results [12].

Recent works have started to look at beat tracking from a different perspective. Instead of developing "universal" models capable of performing equally well across various music genres (requiring large quantities of labeled data), recent efforts have shifted towards adapting preexisting models to succeed on a subset of interest [13], which can be as restricted as a single musical piece [14, 15]. This paradigm aligns well with real-world applications, where it is reasonable for a user to spend a short time producing a few seconds of annotations to get a good performance.

We apply this idea to the refinement of a meter tracking model so that it works well in a particular music genre. We argue that if the genre presents enough homogeneity in terms of its instrumentation and metric structure, as is the case with many Latin American music traditions, it is possible to adapt meter tracking models to perform notably well with just a few annotated data points. We explore this adaptability in terms of data, performance and computational cost. While focusing on two Latin American music genres, *samba* and *candombe*, we study the adaptation of a deep learning state-of-the-art model [16] and compare it with a simpler statistical model [17]. Our contributions are: 1) We perform a detailed analysis on how much annotated data and computation time in CPU are needed to achieve close to "full-dataset" performance in *samba* and in *candombe*, including models trained from scratch and fine-tuned, and compare them to off-the-shelf models trained with Western music; 2) We propose initial experiments to understand the homogeneity conditions under which this adaptation will be successful; 3) We open-source our experiments and provide pre-trained models.

1.1 Other adaptive methods

Existing adaptive methods typically feature some form of transfer learning or fine-tuning, use deep learning models, and are concerned with either beat tracking [13–15] or onset detection [18]. Fiocchi et al. [13] adapt a beat tracking model via transfer learning from Western music to a dataset of Greek music. The authors explore both recurrent neural networks (RNNs) and long-short term memory networks (LSTMs), plus a dynamic Bayesian network (DBN) for inference. The model is fine-tuned using a big training set, though with limited success which the authors attribute to the challenges of the dataset. Even though this is an interesting approach for exploring the idea of adapting to a particular music genre, RNNs and LSTMs are known to be computationally expensive, so in the context of real-world model adaptation they are a concerning choice.

On the other hand, Pinto et al. [14] and Yamamoto [15] explore temporal convolutional networks (TCNs), and focus on the adaptation of models to a particular piece of interest of the user. These authors showed that is possible to adapt TCN models using very small quantities of data (in the order of seconds) to work well, in particular with musically challenging pieces. Furthermore, the TCN is a light-weighted model, computationally more efficient.

Finally, Fonseca et al. [18] apply similar ideas to the adaptation of an onset detection model (also featuring TCNs) to a Latin American music tradition: *maracatu de baque solto*. The authors fine-tune the last layers of the TCN with just a few seconds of manual annotations, and show the advantage of instrument-specific models for the automatic annotation of onsets for musicological studies.

1.2 Latin American Music Traditions

Candombe drumming is a musical tradition from Uruguay that constitutes an essential part of its popular culture and African heritage. Its rhythm is structured in 4/4 meter, and it is played while marching in the streets using three types of drums of different sizes and pitches: *chico*, *repique*, and *piano*. Each of these drums has a distinctive rhythmic pattern and musical role. An additional time–line pattern, called *clave* or *madera*, is shared by the three drums. The *chico* drum is the timekeeper; it repeats a one-beat pattern that establishes the pulse throughout the performance. The *repique* drum is the improviser; it alternates *clave* patterns and characteristically syncopated phrases. The *piano* drum delineates the timeline with distinctive one-cycle patterns and occasionally interposes ornamented *repique*-like figurations. The rhythm shares many traits with other musical traditions of the Afro-Atlantic world. Notably, some of its rhythmic patterns have strong phenomenological accents displaced with respect to the metric structure and divide the rhythmic cycle irregularly with few strokes on the beat.

In parallel, *samba* is a Brazilian musical genre deeply rooted in Brazilian culture, and also has African origins. The word "*samba*" actually describes a family of different subgenres, the most famous arguably being *samba de enredo*, *partido alto*, *bossa nova*, and *pagode*. Similarly to *candombe*, *samba* can be played while parading, which is most common during the festivities of *Carnaval*. It can also be performed in more informal settings, in *rodas* and bars, or even as a chamber-music-like style. Its rhythm is commonly perceived in 2/4 meter, and is conveyed by several types of percussion instruments —– *tamborim*, *pandeiro*, *surdo*, *cuíca*, *agogô*, among others. Each instrument has a handful of distinct patterns [19], and more than one instrument may act as the timekeeper. Because of this combination of timbres and pitches, the texture of a performance can become very complex. *Samba* has unmistakable characteristics as the strong accent on the second beat and the development of contrametric structures.

2. METHOD

Following our intuition about the high homogeneity of *candombe* and *samba* as discussed in Section 1.2, our objective is to understand if it is possible to train meter tracking models with small quantities of data from these music traditions, and if so, how much is needed. To that end, we train the models with increasing amounts of annotated data, ranging from less than a minute up to nearly 40 min, and compare the performance and computational cost of each configuration against the others. We contrast three different training strategies: 1) training the model from scratch with either *candombe* or *samba* snippets; 2) fine-tuning a model trained with 38 h of data from diverse datasets of Western music to work on either *candombe* or *samba*; and 3) same as the previous two, but training the models with data augmentation to artificially increase the "small data" input. We use a state-of-the-art temporal convolutional network model [16] for our experiments, as it presents a good compromise between performance and computational cost. We contrast this model against off-the-shelf models trained in Western music. To understand the adaptability and computational cost of deep learning based methods, we compare the TCN against another simple yet effective baseline, a Bayesian model (BayesBeat) [17]. In the following, we explain our methodology in detail.

2.1 Datasets

We have selected datasets of two different Afro-rooted Latin American music traditions for our experiments. First, the *Candombe* dataset [12, 20], which consists of 35 recordings of *candombe* drumming, for a total of nearly 2.5 h. Each track contains an ensemble recording of three to five drummers using different configurations of drums. Tempo varies greatly and often increases along the performance. To represent the *samba* genre, we use the "acoustic mixtures" data from the BRID dataset [21]. These correspond to 93 short tracks (about 30 s each) of musicians playing together rhythm patterns found in *samba* and two of its subgenres (*samba de enredo* and *partido alto*). Ten different instrument classes are represented, and two to four musicians take part in each track. The dataset contains a variety of tempi; however, tempo remains fairly constant within each track. In order to consistently train our models with about the same amount of data from both *candombe*