QUANTIFYING PITCH DRIFT IN UNACCOMPANIED SOLO SINGING: A COMPUTATIONAL EXAMINATION THROUGH DENSITY-BASED CLUSTERING

Sepideh SHAFIEI[1], **Shapour HAKAM**[2], and **Arvin NICK**[3]

[1,2,3]**Cu Test Inc**, Berkeley, CA USA

ABSTRACT

This research delves into the nuanced phenomenon of pitch drift observed in unaccompanied vocal performances, wherein soloists inadvertently deviate from the initial pitch during protracted presentations. The manifestation of pitch drift depends on different factors, including the skill level of the performer, as well as the duration and intricacy of the musical rendition. In response to this, our study introduces a computational methodology designed to measure pitch drift and changes in intonation throughout unaccompanied vocal performances. Leveraging pitch histogram analysis in conjunction with DBSCAN clustering techniques, this innovative approach not only sheds light on the intricate dynamics of pitch variation but also provides a quantitative and visual framework for the assessment of such deviations in the realm of solo vocal expression. Through this comprehensive analysis, the study contributes to a deeper understanding of the subtle intricacies involved in pitch control and intonation changes during unaccompanied vocal presentations, which can be used as a visual tool for the vocalists for self-assessment. This novel approach also helps significantly in automatic transcription of vocal pieces with pitch drift.

1. INTRODUCTION

Intonation drift, a hallmark of unaccompanied singing, poses a significant challenge in Music Information Retrieval (MIR) tasks like automatic transcription. This phenomenon, characterized by gradual deviations from a reference pitch, usually in a downward direction, throughout a performance, has been explored in various contexts, including both choral and solo singing [1], [2], and [3]. While harmonic progression has been implicated in choral drift [4], and [5], the underlying mechanisms governing solo drift remain poorly understood. This paper presents a novel investigation into the computational measurement and characterization of intonation drift in solo singing. Leveraging a 5':32" vocal piece exhibiting pronounced drift, we develop a robust methodology for quantifying pitch deviations and analyzing their temporal and musical correlations. Our findings shed light on the nature of drift in solo singing and pave the way for improved automatic transcription and analysis of unaccompanied vocal music.

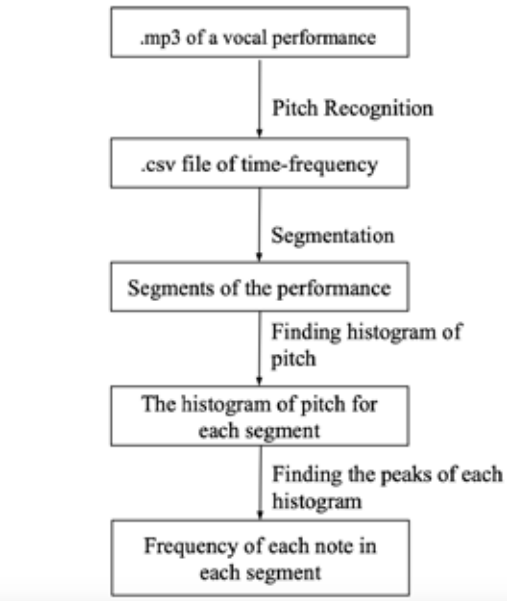

Figure 1. Steps in Data Preprocessing

The methodology is discussed in sections 2 to 4. In order to explain the detail we have used an example of a performance [1] with duration 5':32" (Example 1). It starts with an introduction for about 48 seconds, before the main part starts at the second 00:52". The whole piece has sixteen sentences, which are separated with long silences. The first two sentences are the introduction. A transcription of the whole performance can be found in the corresponding author's dissertation [6]: 69-71.

2. BACKGROUND AND PREPROCESSING

In this section, we present the building blocks of the data preprocessing together with all the needed musical and computational background. Figure 1 gives an overview of the preparation process in this section.

2.1 Pitch Recognition

There are different algorithms for pitch recognition. All these algorithms use the fundamental frequency of the

[1] The mp3 files for the performance in Example 1 is available in the github repository of the software along with some other examples from various musical cultures: https://github.com/SepiSha/PitchDrift

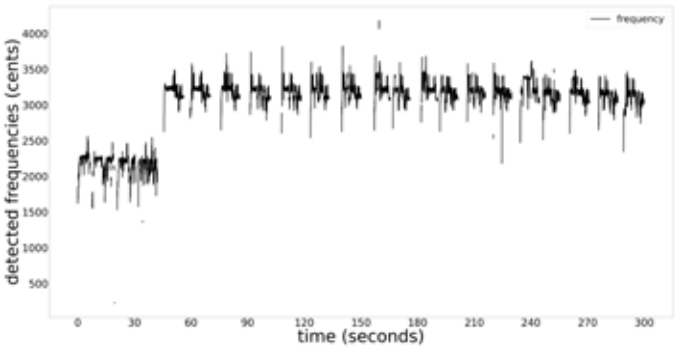

Figure 2. Time-Frequency for Example 1

sound to quantify pitch. Fundamental frequency is the lowest frequency of the sound wave and corresponds to its most dominant perceived pitch. Monophonic voice pitch estimation algorithms have been evaluated and discussed extensively in MIR literature. Gomez et al. have evaluated both algorithms CREPE [7] and pYIN [8] as state-of-the-art pitch recognition methods for monophonic voice [9]. They have evaluated the results of both of these algorithms on iKala dataset and obtained almost similar results for both pYIN and CREPE: 91% of Raw Pitch Accuracy for pYIN for monophonic voice and 90.5% accuracy for CREPE. Among other pitch recognition methods which can be used for monophonic sound one can mention SPICE [10], which is based on a self-supervised learning technique. We chose pYIN for this study, and we have done manual octave-error correction when necessary. Since we use the output of the pitch recognition method (pYIN) as an input for pitch recognition in *.csv* format, with minimum effort, one can replace the current pitch recognition method with any other algorithm for experiment or future improvement.

To make the pitch file we used Sonic Annotator for pitch recognition with the following parameters: step size of 256, block size of 2048, low amplitude suppression of 0.1, onset sensitivity of 0.7, prune threshold of 0.1, and threshold distribution of 2. Figure 2 shows the the time-frequency graph of Example 1. We have changed the frequency of the sound (Hz) to the cents system. Each octave is 1200 cents. The following formula gives the interval between the two notes with the frequencies f_1 and f_2, in cents.

$$cent = 1200 \times \log_2(\frac{f_2}{f_1}) \quad (1)$$

2.2 Identifying Musical Phrases

Phrasing and segmentation are crucial elements in music theory and cognitive musicology, helping to understand how listeners perceive and interpret musical structure. These concepts involve the ways in which music is divided into comprehensible units, influencing both the composition and perception of music. Segmentation is the process of dividing a musical piece into segments or sections based on criteria such as melody, harmony, rhythm, and texture such that it conveys a sense of beginning, continuation, and closure. These concepts are foundational for understanding musical form and expression, serving as the basis for more complex analyses of musical structure. There are various approaches to segmentation of music: The structuralist approach involves analyzing the syntactic rules governing musical structure. Leonard Meyer challenges the structuralist's premise of objective abstraction incorporating elements of musical education and cultural context that are both more subjective and more diverse [11]. He emphasized the importance of expectation in musical experience, suggesting that musical meaning arises from the interplay between learned patterns and their fulfillment or violation. This perspective highlights the importance of cultural and historical contexts in shaping musical understanding [12].

Cognitive approaches focus on how listeners mentally process and organize musical information. David Temperly uses computational approach to fundamental questions about music cognition [13]. Empirical studies often involve experiments where listeners' segmentations are recorded and analyzed. These studies provide insights into how listeners perceive and organize musical elements, revealing patterns that align with theoretical predictions and uncovering new dimensions of musical perception.

Computational models employ machine learning techniques to simulate human segmentation processes, offering a powerful tool for testing theoretical hypotheses and exploring new musical domains. Cross-cultural studies of phrasing and segmentation reveal both universal principles and culture-specific patterns. Genre-specific analyses highlight how different musical styles employ unique segmentation strategies, reflecting their distinct aesthetic and structural conventions.

Integrative theories such as Lerdahl and Jackendoff's Generative Theory of Tonal Music (GTTM) provides a formal framework for analyzing the hierarchical structure of tonal music, combining insights from music theory, cognitive science, and linguistics. The GTTM's main components—Grouping Structure, Metrical Structure, Time-Span Reduction, and Prolongational Reduction which in this framework signifies hierarchical pitch connection—offer a comprehensive model for understanding how listeners parse musical surfaces into structured representations [14].

Since we started this research as part of a larger project on the computational approach to Iranian classical vocal music, we have noted that the phrasing in this genre often follows the structure of classical poetry, with phrases frequently separated by long silences. In this paper, we perform segmentation based on these relatively long silences between the sentences sung by the vocalist. This method helps us identify musical phrases, or sentences. For future reference in this paper, we will call this type of segmentation *Method I*. An alternative approach for computing pitch drift is to consider congruent segments, i.e., every n seconds. We refer to this equal-duration partition of the vocal piece as *Method II*. In this example, we have chosen the length of each segment to be $T/24$, where T is the total duration of the piece in Example 1. In the next section, we consider both methods. In the first case, the segments vary

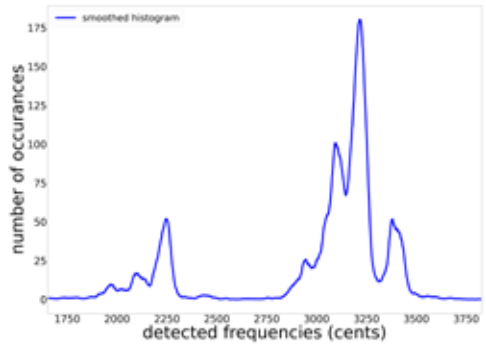

Figure 3. Pitch Histogram for Example 1

in length depending on the duration of the sentences, while in the second method, the segments have equal duration. There is always the potential to employ various types of segmentation depending on the genre of music and differing theoretical approaches, which can be explored in future research.

2.3 Pitch Histogram

The next step in preparation of the data points is to find the histograms of the performed pitches for each segment. In order to find the histogram, we find the total number of occurrences of each fundamental frequency. Figure 3 shows the pitch histogram of Example 1. The vertical axis shows the proportional total duration of each frequency (in cents). Pitch histogram has been used extensively in MIR literature. For example, Bozkurt et al. have used it for analysis of Turkish Makam Music [15]. In the musical traditions where intervals and note frequencies are not absolute, using pitch histograms are useful in identification of the frequency of each note and the intervals between the notes in the performed pieces. The use of pitch histogram in finding intervals of a performed vocal piece in Iranian music has been discussed extensively in [16] and [17]. Koduri et al. have used pitch histograms in analysis of intonation for Carnatic music [18].

2.4 Gaussian Model of the Peaks

We smooth the histogram in order to find the exact peak of each mountain by finding the moving average to smooth out short-term fluctuations and highlight the general trend of the data and we get a semi-Gaussian curve. Choosing reasonable peaks is very crucial. To find the peak of each mountain, which represents the median of the performed pitch for each "note", we need to find the range of the mountain which is itself a challenging task and is discussed in [16]. After finding the range of each mountain, we model each mountain by a tilted Gaussian curve so that we can find a better peak. We fit the following curve to our data to find the parameters $c_1, \ldots, c_5$

$$y = c_1 + c_2 x + c_3 e^{-(x-c_4)^2/c_5} \quad (2)$$

After finding the peak for each mountain in the histogram

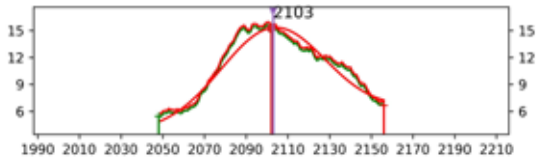

Figure 4. Gaussian peak fit for the first significant peak of the pitch histogram Example 1

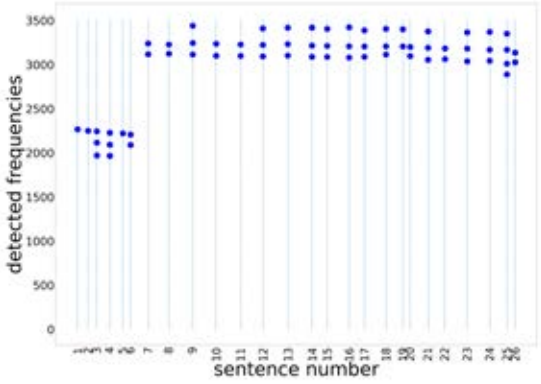

Figure 5. The main detected frequencies for each sentence in Example 1 using *Method I*

of audio, we have the frequency of every note in each sentence in cents. Since we only want reliable estimates, we have picked the mountains that have at least certain heights, so that we have enough data for fitting a Gaussian curve. Figure 4 shows the result of our Gaussian peak fit for the first significant peak of the pitch histogram in Example 1.

3. EXPLANATION OF THE DATA

After performing the preprocessing which was explained in sections 2.1 through 2.4, we have gathered the frequencies associated to the main notes of each segment. Figure 5 shows the data points gathered using *Method I* for segmentation. The horizontal axis shows the segment number and the vertical axis shows the frequency of the main notes in each sentence in cents. Since in *Method I* the duration of the sentences varies, the numbers on the x-axis are not equidistant, the distance of each number from its consecutive number is proportional to its duration. If we use *Method II* and find the frequency of the notes in the same size intervals (every n seconds), we will get Figure 6 for the vocal performance in Example 1.

4. CLUSTERING THE PEAKS USING DBSACN

To see the pattern of drift in various notes throughout the piece, we first sort the frequency of the notes for each segment of the piece. Assume that we have n segments $\{s_1, ..., s_n\}$. For each of these segments $s_i, 1 < i < n$, we will have a number of frequency peaks i_k: $p_{i_1}, ..., p_{i_k}$, the number of the peaks varies from one segment to the other.

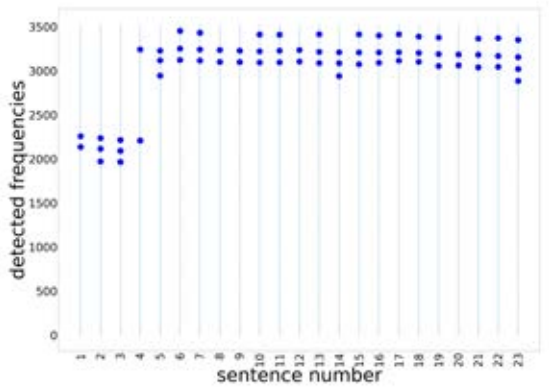

Figure 6. The main detected frequencies for each sentence in Example 1 using *Method II*

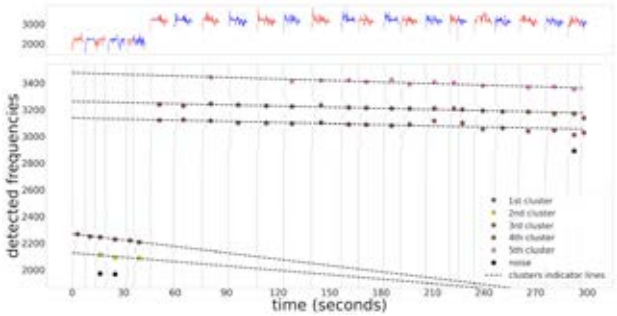

Figure 7. Clustering of the frequencies using DBSCAN, and modeling each cluster with linear regression, using *Method I* for segmentation

For example, in Figure 6, $i_1 = 2$ and $i_2 = 3$, which means that we have two peaks in the first segment, and three peaks in the second segment. We need to cluster these frequency points p_{i_k} for $\forall i, k$, so that we have all the fundamental frequencies associated to a specific note throughout the performance in one class.

In order to cluster the frequency data points we used Density-based spatial clustering (DBSCAN) from *scikit* in Python. We then use linear regression to model each cluster. Figures 7 and 8 show the result of this system for Example 1 using Method I and Method II for segmentation respectively. As can be seen in these Figures, we have detected 5 cluster of points, each corresponds to a note. Each cluster is marked in the graph with a separate color. The first two clusters associated to the lower frequency notes only appear at the beginning of the piece and have fewer samples comparing to the other three clusters. The slope of the lines in the linear regression shows the pattern of change in the tuning during the course of the performance. This method for quantification of pitch drift can be very helpful in automatic transcription of unaccompanied vocal pieces, especially when the pitch drift is significant. As can be seen from the two Figures 7 and 8 the method of segmentation did not effect the result significantly.

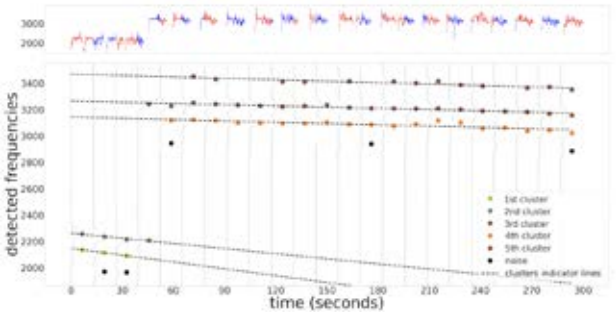

Figure 8. Clustering of the frequencies using DBSCAN, and modeling each cluster with linear regression, using *Method II* for segmentation

Acknowledgments

This work was started as part of the first author's PhD dissertation under the supervision of Prof. Blum in the Music Department of the City University of New York.

5. REFERENCES

[1] M. Mauch, K. Frieler, and S. Dixon, . "Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory." *The Journal of the Acoustical Society of America*, 136(1), pp.401-411. 2014.

[2] J. Dai, M. Mauch, and S Dixon. "Analysis of intonation trajectories in solo singing." In *Proceedings of the 16th ISMIR Conference* (421: 29). October 2015.

[3] J. Dai, and S. Dixon, 2019. "Intonation trajectories within tones in unaccompanied soprano, alto, tenor, bass quartet singing." *The Journal of the Acoustical Society of America*, 146(2), 1005-1014.

[4] H. Terasawa. "Pitch drift in choral music." Music 221A final paper, Center for Computer Research in Music and Acoustics, at Stanford University, CA. Available at https://ccrma. stanford.edu/hiroko/pitchdrift/paper221A. pdf (Last viewed 5 June 2014).

[5] D. Howard.. "Intonation drift in a capella soprano, alto, tenor, bass quartet singing with key modulation." *Journal of Voice*, 21(3), pp.300-315. 2007.

[6] S. Shafiei, *Extracting Theory from Practice: A Computational Analysis of the Persian Radif*. PhD dissertation at the Graduate Center, City University of New York. 2021.

[7] J. Kim, J. Salamon, P. Li, and J. Pablo Bello. "CREPE: A Convolutional Representation for Pitch Estimation." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*: 161-165. IEEE. 2018 .

[8] M. Mauch, and S. Dixon. "pYIN: A fundamental frequency estimator using probabilistic threshold distributions." In *2014 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), pp. 659-663. IEEE, 2014.

[9] E. Gómez, M. Blaauw, J. Bonada, P. Chandna, and H. Cuesta. “Deep Learning for Singing Processing: Achievements, Challenges and Impact on Singers and Listeners." arXiv preprint. arXiv:1807.03046. 2018.

[10] Gfeller, Beat, et al. "SPICE: Self-supervised pitch estimation." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 1118-1128.

[11] Swain, Joseph P. "Leonard Meyer’s New Theory of Style." *Music Analysis* 11 (1992): 335-354.

[12] Meyer, L. B. 1956. and Meaning in Music. University of Chicago Press

[13] Temperley, D., 2004. *The cognition of basic musical structures*. MIT press.

[14] Lerdahl, F., and Jackendoff, R. 1983. *A Generative Theory of Tonal Music*. MIT Press.

[15] B. Bozkurt. “Pitch Histogram Based Analysis of Makam Music in Turkey.” Paper presented at *Les Corpus de l’oralité Colloquium, Strasbourg*. 2011.

[16] S. Shafiei. “An analysis of Iranian Music Intervals based on Pitch Histogram. arXiv preprint arXiv:2108.01283.

[17] S. Shafiei, "Analysis of Vocal Ornamentation in Iranian Classical Music." In *Sound Music Conference Proceedings*. 2019.

[18] Koduri, Gopala Krishna, Joan Serrà Julià, and Xavier Serra. "Characterization of intonation in carnatic music by parametrizing pitch histograms." Gouyon F, Herrera P, Martins LG, Müller M. ISMIR 2012: Proceedings of the 13th International Society for Music Information Retrieval Conference; 2012 Oct 8-12; Porto, Portugal. Porto: FEUP Ediçoes, 2012.. International Society for Music Information Retrieval (ISMIR), 2012.