



Tonality-Based Accompaniment-Guided Automatic Singing Evaluation

Pei-Chin Hsieh¹, Yih-Liang Shen², Ngoc-Son Tran³, Tai-Shih Chi¹

¹Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan

²Institute of Communications Engineering, National Yang Ming Chiao Tung University, Taiwan

³Realtek Semiconductor Corp., Taiwan

wren.ee12@nycu.edu.tw, dennis831209@gmail.com, ngocsontan19398@gmail.com, tschi@nycu.edu.tw

Abstract

Automatic singing evaluation is highly desirable in industrial performance evaluation scenarios, including education and entertainment, given the high cost of human judges. Current singing evaluation systems assess a singer's performance by comparing reference vocals to the singing track. Unfortunately, reference vocals are not always available. Moreover, the similarity measure may not provide a reliable evaluation due to the inherently variable nature of singing performances. This paper proposes a tonality-based method for song-independent automatic singing evaluation. Experimental results show the proposed method outperforms other non-intrusive evaluation algorithms.

Index Terms: Singing evaluation, non-intrusive, pitch extraction, musical key estimation, tonality, accompaniment-guided

1. Introduction

In recent years, most singing evaluation systems have been developed as intrusive systems. They first extract features from vocals of both the user and the reference (or target) singer for further processing. Next, they measure similarity by computing the distance between the two sets of features, which, in some cases, are aligned using dynamic programming techniques like dynamic time warping (DTW) [1, 2, 3, 4, 5, 6] or the NW algorithm [7]. The similarity value then served as an objective score for the singing assessment. However, in practical applications, the reference vocals are not always available. For instance, commercial karaoke systems rely heavily on their pre-established databases to avoid license fees, which only contain accompanying music without reference vocals. In addition, for artistic expression, people sometimes adjust their singing styles for different audiences and environments such that the singing assessment based on reference vocals is inappropriate. Therefore, the singing assessment model without vocal references is in great demand.

Past non-intrusive methods for singing evaluation often focus on pitch accuracy and stability. For instance, Nakano et al. [8] used pitch interval accuracy to measure the pitch stability of the singer. Their approach struggles with vibrato sections, as the pitch estimation algorithm may misjudge these sections as unstable F0 contour segments. Similarly, Gupta et al. [9] constructed pitch histograms across the 12-semitone grid, which reflect intonation accuracy but are biased against specific musical genres, leading to unfair evaluations across different genres.

More recently, data-driven methods have gained popularity [10, 11]. While these approaches have shown promising results on their respective test sets, they face several challenges. Primarily, obtaining manually annotated data is highly time-consuming [12]. Automated annotation processes, alternatively,

rely heavily on large volumes of metadata, including the number of likes and comments for each song from a karaoke application, whose annotation quality may be influenced by factors unrelated to singing performance, such as song popularity and the singer's reputation [13]. These approaches allow models to make inferences by extracting patterns directly from data without understanding music theory. In other words, they are driven by deep learning techniques rather than musical theoretical judgments, potentially resulting in biased inferences due to the lack of grounding in music theory.

This paper presents an accompaniment-guided objective singing quality assessment method based on tonality, addressing the need for evaluation independent of judges' familiarity with specific songs. We postulate that such evaluation should transcend melodic variations across different musical pieces. To our knowledge, this study pioneers the application of tonality theory to singing evaluation tasks. Our algorithm was tested on various singers, songs, and musical segments, generating objective scores that were subsequently validated against subjective ratings from human judges. The results demonstrated a high correlation between our algorithmic assessments and human evaluations, thereby introducing a new, robust paradigm in the field of automated singing performance assessment.

The remainder of this paper is structured as follows. Section II reviews previous approaches and the challenges they faced, summarizing the issues that need to be addressed in the task of singing evaluation without using the reference vocal. Section III describes the proposed algorithm to assess the quality of a given singing clip. Section IV presents the experiment settings and discusses the experimental results. Finally, Section V concludes our work.

2. Related Works

Nakano et al. [8] used pitch interval accuracy, which computed the offset of the singing pitch from the semitone grid throughout the song to measure pitch stability. Using a vibrato section estimator, they achieved a binary classification rate of 0.835. However, their approach fails when the pitch estimation algorithm misjudges vibrato sections as unstable pitch contour segments. Another issue would arise when a singer consistently sings out of tune but in a steady manner throughout the song.

On the other hand, Gupta et al. [9] constructed pitch histograms by mapping the pitch values onto the 12-semitone grid and used Gaussian mixture models (GMMs) to measure the sharpness of peaks of the histograms. Their idea is based on the fact that, compared to amateurs, professional singers' pitch tends to be concentrated around a few specific notes, reflecting the singers' ability to control their intonation accuracy. Similarly, it also suffers from vibrato misjudgments, as the dis-

persion of pitches is regarded as unstable. The most significant drawback of this pitch histogram-based method is that such analysis is reasonable only when singers cover the same song. Under different song arrangements and melody designs, the pitch histograms may exhibit highly complex and variable patterns, leading to unsatisfactory evaluations. For instance, compared to rap songs with fast-changing notes, there is a bias in favor of pop songs rich in steady and long notes. This bias makes the latter more advantageous in terms of the scoring metric.

An alternative approach involves gathering vocal characteristic data, including pitch, rhythm, and timbre, from various cover performers on the platform, leveraging the statistical tendency that high-quality singers exhibit similar features. In contrast, low-quality performers tend to show more dispersed characteristics [6]. However, this method is also limited in that it can only evaluate different singers performing the same song.

Research findings indicate that professionally trained music experts can provide highly consistent singing quality, even for unfamiliar melodies [14] [8], implying that the assessment of singing quality should be independent of the test song, as these evaluations are based on the singing skills. Therefore, the absence of musical theory integration and the limitations of current non-intrusive methods motivate us to develop a music theory-based singing evaluation algorithm independent of specific song melodies and does not require reference vocal sounds.

3. Proposed Method

Perceptual characteristics like pitch accuracy, rhythm consistency, volume dynamics, and singing skills such as appropriate vibrato are often considered when developing a singing assessment method [15, 8]. Given that pitch accuracy exerts the most significant influence on singing evaluation among all factors [2, 16, 5], we proposed a music theory-based singing quality assessment method focusing on pitch accuracy without the vocal reference. This section describes the proposed key-based scoring method with two major components: pitch extraction and musical key classification models.

3.1. Tonality-Based Singing Assessment

In tonal music, such as pop music, each musical key corresponds to a specific scale. For example, the scale for the key of G minor is {G, A, Bb, C, D, Eb, F}. Tones within the tonal scale are perceived as more consonant, stable, and pleasant, while chromatic tones add tension unless contextually justified [17, 18]. Hence, notes falling within the scale are considered in-key and should contribute positively to the pitch score, while those falling outside of the scale often sound dissonant, leading to poor pitch evaluations by judges. Grounded in music theory and psychology, we built a system to calculate the proportion of notes falling within the key’s scale for the singing segment. A higher value indicates that the singer’s pitch aligns well with the accompaniment music, sounding harmonious and pleasant; therefore, the singer should receive a higher score.

To facilitate analysis within the framework of musical tonality, we adopted the MIDI representation for analyzing pitch. The MIDI note number ranges from 0 to 127, with each number corresponding to a specific musical note. The relationship between the frequency f (in Hz) and the corresponding MIDI note number can be outlined using the following formula:

$$\text{MIDI note number} = 69 + 12 \cdot \log_2 \left(\frac{f}{440} \right) \quad (1)$$

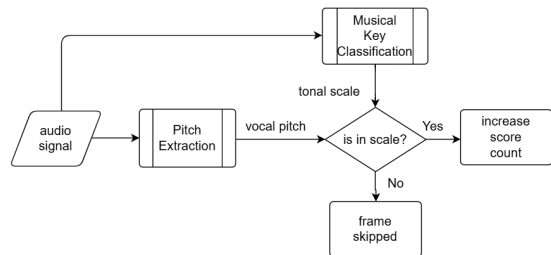


Figure 1: Flow diagram of the proposed scoring method.

Then, the scoring formula is designed as follows:

$$S = \frac{\sum_{i=1}^N \mathbf{I}_A(p_i) \cdot v_i}{\sum_{i=1}^N v_i} \quad (2)$$

where S is the score for the singing segment, N is the total number of frames, p_i is the pitch of the i -th frame, v_i is the flag indicating whether the i -th frame is voiced (1 if voiced, 0 if unvoiced), and $\mathbf{I}_A(p_i)$ is the indicator function defined as:

$$\mathbf{I}_A(p_i) = \begin{cases} 1 & \text{if } p_i \in A \\ 0 & \text{if } p_i \notin A \end{cases} \quad (3)$$

where A is the set of notes in the key’s scale. Fig. 1 illustrates the proposed scoring method.

3.2. Pitch Extraction

There are many well-known fundamental frequency (F0) estimators, such as PRAAT [19], YIN [20], and its variant pYIN [21]. To emphasize the perceptual role of pitch in the context of music theory, we refer to the output of the F0 estimator as “pitch” throughout this paper. These estimators achieve decent accuracy with low computational overhead, yet they are unable to handle situations with more than one melody present simultaneously, such as in duets and harmonic accompaniments [22]. To address the diverse scenarios encountered in real-world applications, we adopted a multi-pitch extraction model from our previous research [22] in our system. This model utilizes a log-scaled short-time Fourier transform (STFT)-based spectrogram and was trained on the MIR-1k dataset [23]. The architectural details are presented in Table 1.

Table 1: Architecture of the pitch extraction model.

Layer	Configuration	Input	Output
Conv Block 1	kernel 8, filters 3, stride 1; max pool (2×2)	1 × 176	8 × 88
Conv Block 2	kernel 8, filters 3, stride 1; max pool (2×2)	8 × 88	8 × 44
Residual Block	Conv1D: kernel 8, filters 1, stride 1	8 × 44	8 × 44
Harmonic Block	Conv1D: kernel 8, filters 1, stride 1	8 × 44	8 × 44
Conv Block 3	kernel 1, filters 6/10/12, stride 1; max pool (2×2)	8 × 44	1 × 22
LSTM	hidden size 256	1 × 22	1 × 256
Linear	hidden size 512	1 × 256	127

Considering the vocal conditions of human singing, we set the frequency of the output pitch to span from 31 to 4857 Hz, with each octave divided into 24 half-semitones, resulting in a 50-cent resolution output. The model was trained by utilizing the binary cross-entropy loss function to calculate the error between the ground truth pitch vector and the predicted pitch vector, and the Adam optimizer [24] with a learning rate of 10^{-3} , which enables faster convergence by dynamically adjusting the learning rate throughout training.

3.3. Musical Key Classification

The user’s singing voice, along with the accompanying music, is used as input to the key classification model, as the musical key is typically associated with the background music’s chord progression. In music theory, there are 24 distinct keys, corresponding to 12 tonics and two modes (major and minor) within the natural major and minor scales. However, pitch accuracy is assessed based on the concept of scales in the proposed method. Therefore, those keys sharing the same scale, known as relative keys, are merged into a single class. This approach is adopted because the scoring system should only be interested in whether two classes have differing tonal scales.

3.3.1. Musical Key Classifier

Our model builds upon the 5-layer CNN proposed in [25], incorporating structural modifications for improved accuracy. To sharpen the feature map, we added a 2D max-pooling layer after each convolutional layer, which was zero-padded for dimension preservation. We also incorporated batch normalization in each convolutional layer to avoid gradient-vanishing during training [26], as shown in Table 2. Following the settings in [25], we trained the classifier based on cross-entropy loss with SGD as the optimizer, an initial learning rate of 10^{-3} , a momentum of 0.9, and a weight decay factor of 10^{-4} . An early-stop strategy was also applied during training, with patience set to 50 epochs.

Table 2: Architecture of the musical key classification model.

Layer	Configuration	Input Shape	Output Shape
Conv Block 1	kernel 8, (5×5), stride (1×1), padding 2; batch norm; max pool (2×2)	1×103×175	8×51×87
Conv Block 2	kernel 8, (5×5), stride (1×1), padding 2; batch norm; max pool (2×2)	8×51×87	8×25×43
Conv Block 3	kernel 8, (5×5), stride (1×1), padding 2; batch norm; max pool (2×2)	8×25×43	8×12×21
Conv Block 4	kernel 8, (5×5), stride (1×1), padding 2; batch norm; max pool (2×2)	8×12×21	8×6×10
Conv Block 5	kernel 8, (5×5), stride (1×1), padding 2; batch norm; max pool (2×2)	8×6×10	8×3×5
Flatten	-	8×3×5	120
Linear 1	Linear (120 → 48), batch norm	120	48
Linear 2	Linear (48 → 12), batch norm	48	12

3.3.2. Data Preparation

Since we did not find accessible public datasets with annotated key labels, we used datasets for other music information retrieval (MIR) tasks, including GTZAN [27], iKala [28], and MedleyDB [29] for training the musical key classification model. It is worth noting that we trained the model using mixed-genre datasets, which include classical, jazz, funk, hip-hop, blues, country, heavy metal, and electronic music. While mixed-genre training may decrease accuracy [25], it aligns with our objective of achieving fair evaluation for music inputs from various genres, i.e., building a general and robust scoring system across diverse songs.

We used the musical key estimations from the commercial software Tunebat [30] as our training targets, assuming Tunebat’s key detection is highly accurate; thus, the performance metrics of our scoring method were evaluated relative to Tunebat’s results. Experimental findings, discussed in Section 4.4, show that certain key classification errors minimally impact the overall singing evaluation.

4. Experimental Results And Discussion

4.1. Subjective Rating

20 participants with varying musical backgrounds were recruited to assess a group of singers on 22 Chinese pop songs. These participants provided subjective ratings ranging from 1 to 3 based on the following criteria: singers with perfect intonation accuracy were rated level 3, those with a few mistakes were rated level 2, and those with poor pitch accuracy were rated level 1. Some clips were chosen from the MIR-1k and iKala datasets to increase the diversity of the data, which helps to capture real-world conditions better. Additionally, we randomly shifted the pitch of some of the clips upwards or downwards by 100 to 200 cents, word-wise, to ensure the singing quality of the test clips was spreading across the three levels. Experimental results of listening tests verified that the quality ratings given by human judges decreased as the extent of the manipulated pitch-shift distortion increased, as expected.

Each test clip was cropped to approximately 15 seconds, considering that such duration was sufficient for assessing singing quality. To alleviate the influence of outliers, we used the median of all scores provided by 20 participants as the final quality score for each test clip.

4.2. Experiment Results

To examine the effect of subsystems on overall performance, we tested various pitch extraction and musical key classification models, calculating Pearson correlation coefficients with listening test scores. Table 3 lists these coefficients, including baseline pitch interval accuracy [8] and GMM-based methods, which use features like *PeakBW*, *PeakConc₁₁₀*, *second moment* and *slope*, chosen for their strong performance in previous studies [9]. In addition, pYIN [21], representing a high-resolution pitch extraction method, was also incorporated into the system to compare the impact of pitch subsystem accuracy on the overall scoring system’s performance. Results in Table 3 indicate that the higher the pitch resolution is, the closer the overall system’s performance aligns with human evaluations.

To have fair comparisons, we restricted baseline methods to only using pitch-related features. In the method proposed in [9], we used the two best-performing features, GMM peak bandwidth *PeakBW* and GMM peak concentration *PeakConc₁₁₀*, according to the original study. In the method proposed in [8], we used the two best-performing features, the *second moment* of long-term averaged semitone stability and the *slope* of the linear regression line from the symmetric smoothing function. Results in Table 3 show the proposed method outperforms the two baseline methods and produces more consistent performance across different resolutions on pitch extraction.

Table 3: Correlation coefficients produced by the baseline and our proposed methods. GMM-based and pitch interval accuracy denote the reference methods proposed in [9][8].

Method	pYIN	Pitch Model [22]
GMM-based <i>PeakBW</i>	0.228	0.034
GMM-based <i>PeakConc₁₁₀</i>	0.148	0.051
GMM-based <i>PeakBW</i> + <i>PeakConc₁₁₀</i>	0.232	0.051
Pitch interval accuracy - second moment	0.364	0.083
Pitch interval accuracy - slope	0.010	0.146
Proposed-Tunebat	0.595	0.530
Proposed-Musical Key Classification Model	0.611	0.524

4.3. Analysis of Discrepancies in Scoring Results

The objective scores predicted by the proposed tonality-based method are plotted against subjective ratings in Fig.2. The horizontal axis represents the 1 to 3 ratings given by the human judges, while the vertical axis represents the predicted scores the proposed method gives. This chart clearly shows that the objective score positively correlates with the subjective rating, validating the concepts of the design of the proposed method. However, we observed some scoring results deviated from expectations. For example, certain musical pieces rated poorly by judges received high scores in the proposed method, as seen in the top left of Fig.2, while some pieces that sounded pleasant to listeners scored lower, as seen in the middle right of the figure.

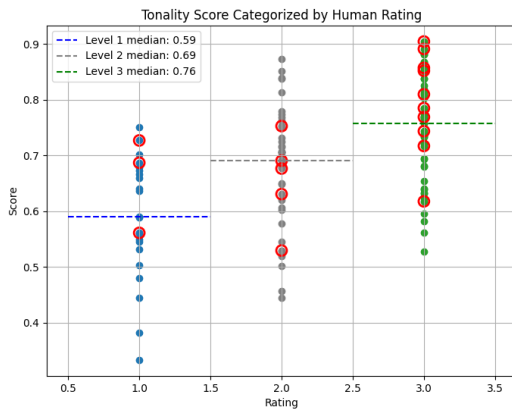


Figure 2: *Tonality-based score categorized by human rating. The red circles are samples with musical keys misclassified as the dominant key.*

Our further analysis revealed several reasons for these cases. The main reason is that although the overall singing melody centers around the correct tonality scale, there are brief instances of severe pitch deviations. These brief deviations might not significantly impact the pitch distribution. Still, such localized errors dominate human judgment for audio quality, causing listeners to focus on these short but jarring mistakes [31], leading to lower ratings.

Second, the resolution of the pitch extraction model introduces quantization errors. When these errors cause the estimated pitch to be classified as a dissonant interval [32], it impacts the correlation with listening tests. For instance, in an Ab minor musical piece, sustained B notes were misclassified as C notes, forming a dissonant minor second. For these cases, replacing the pitch extraction model with a higher-resolution method (pYIN) improved score alignment with human evaluations.

4.4. Effects of Dominant Key Confusion on Evaluation

The confusion matrix (Fig.3) reveals instances of dominant key confusion, marked by red lines. The dominant key, whose tonic is a perfect fifth above the tonic of the original key while sharing the same mode, is challenging to classify due to two main reasons in music theory: the high degree of scale overlap and the frequent use of the same chords in both keys. This difficulty is not unique to the model; even experienced musicians may struggle to distinguish closely related keys, as noted by the

MIREX campaign, which highlights the perceptual closeness of keys related by a perfect fifth, relative major/minor, or parallel major/minor [33].

Although this perceptual proximity is reflected in the data, such errors do not significantly impact the accuracy of the proposed key-based pitch evaluation method. Fig.2 shows misclassified samples as red circles, evenly distributed across all score ranges. Importantly, samples in the top-left and middle-right regions—representing low ratings with high scores or high ratings with low scores—constitute only 4 percent of total clips, further supporting the limited influence of these errors on system performance.

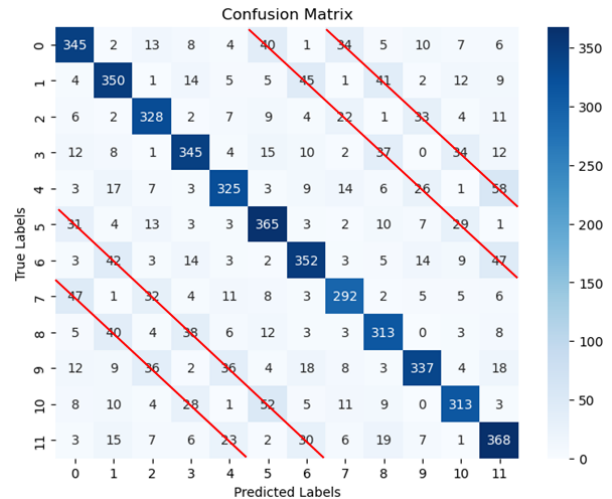


Figure 3: *Confusion matrix of the musical key classification model. Red lines indicate dominant key confusions.*

5. Conclusion

In this study, we propose an automatic singing evaluation system that does not require reference to original vocal tracks. Built on a solid foundation of tonal music theory, the system does not rely on specific melodic features of songs. Instead, it provides objective scores that effectively reflect the singer's pitch control abilities. The proposed system achieves a strong correlation with subjective listening tests in the aspect of pitch control. The data demonstrate that the system exhibits a certain degree of robustness in its key classification subsystem, making it a functional algorithm independent of other vocal feature extractors.

Limitations: The proposed method still relies on the accompaniment information to achieve accurate evaluations. This limitation restricts its application in scenarios where only vocal sounds, such as in a cappella performance, are available. In the future, we aim to develop an automatic singing evaluation system that can function effectively in the absence of accompaniment, thus expanding its usability to more diverse singing contexts.

6. Acknowledgments

This research is supported in part by the National Science and Technology Council, Taiwan under Grant NSTC 113-2221-E-A49-146, and in part by Realtek Semiconductor Corp., Taiwan.

7. References

- [1] W.-H. Tsai and H.-C. Lee, "An automated singing evaluation method for karaoke systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, 2011, pp. 2428–2431.
- [2] E. Molina, I. Barbancho, E. Gómez, A. M. Barbancho, and L. J. Tardón, "Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 744–748.
- [3] W.-H. Tsai and C.-H. Ma, "Singing performance evaluation by reference to cd music recordings," in *Proc. IEEE Int. Conf. Consumer Electron. - Taiwan (ICCE-TW)*, Taipei, Taiwan, 2014, pp. 149–150.
- [4] C.-H. Lin, Y.-S. Lee, M.-Y. Chen, and J.-C. Wang, "Automatic singing evaluating system based on acoustic features and rhythm," in *2014 International Conference on Orange Technologies*, 2014, pp. 165–168.
- [5] C. Gupta, H. Li, and Y. Wang, "Perceptual evaluation of singing quality," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Kuala Lumpur, Malaysia, 2017, pp. 577–586.
- [6] —, "Automatic leaderboard: Evaluation of singing quality without a standard reference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 13–26, 2020.
- [7] W. Yang, X. Wang, B. Tian, W. Xu, and W. Cheng, "A multi-stage automatic evaluation system for sight-singing," *IEEE Transactions on Multimedia*, vol. 25, pp. 3881–3893, 2023.
- [8] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," vol. 4, 09 2006.
- [9] C. Gupta, H. Li, and Y. Wang, "Automatic evaluation of singing quality without a reference," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 990–997.
- [10] T. Maka, "Attributes of audio feature contours for automatic singing evaluation," in *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*, 2013, pp. 517–520.
- [11] N. Zhang, T. Jiang, F. Deng, and Y. Li, "Automatic singing evaluation without reference melody using bi-dense neural network," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 466–470.
- [12] Y. Ju, C. Xu, Y. Guo, J. Li, and S. Lui, "Improving automatic singing skill evaluation with timbral features, attention, and singing voice separation," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 612–617.
- [13] X. Sun, Y. Gao, H. Lin, and H. Liu, "Tg-critic: A timbre-guided model for reference-independent singing evaluation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [14] N. Tomoyasu, G. Masataka, and H. Yuzuru, "Subjective evaluation of common singing skills using the rank ordering method," *Proceedings of the 9th International Conference on Music Perception and Cognition*, 8 2006.
- [15] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1233–1243, 2012.
- [16] C. Cao, M. Li, J. Liu, and Y. Yan, "A study on singing performance evaluation criteria for untrained singers," 11 2008, pp. 1475–1478.
- [17] R. Parncutt and G. Hair, "Consonance and dissonance in music theory and psychology: Disentangling dissonant dichotomies," *Journal of Interdisciplinary Music Studies*, vol. 5, pp. 119–166, 01 2011.
- [18] O. van Dillen, *Outline of Basic Music Theory*. The Hague: Donemus Publishing, 2021.
- [19] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.13)," 2009. [Online]. Available: <http://www.praat.org>
- [20] 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 4, pp. 1917–30, 2002. [Online]. Available: <https://api.semanticscholar.org/CorpusID:1607434>
- [21] 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)*, 2014, pp. 659–663.
- [22] N.-S. Tran, P.-C. Hsieh, Y.-L. Shen, Y.-H. Chu, and T.-S. Chi, "Real-time monophonic dual-pitch extraction model," in *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Macau, China, 2024.
- [23] C. L. Hsu and J. S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 310–319, 2009.
- [24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [25] F. Korzeniewski and G. Widmer, "End-to-end musical key estimation using a convolutional neural network," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 966–970.
- [26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2015, pp. 448–456.
- [27] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [28] T.-S. Chan, T.-C. Yeh, Z.-C. Fan, H.-W. Chen, L. Su, Y.-H. Yang, and R. Jang, "Vocal activity informed singing voice separation with the ikala dataset," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 718–722.
- [29] R. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Mediterranean: A multitrack dataset for annotation-intensive MIR research," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 155–160.
- [30] Tunebat, "Key bpm database and music finder," <https://tunebat.com/>, accessed: 2024-08-21.
- [31] M. Hollier, M. Hawksford, and D. Guard, "Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain," *Vision, Image and Signal Processing, IEE Proceedings -*, vol. 141, pp. 203–208, 07 1994.
- [32] G. M. Bidelman and A. Krishnan, "Neural correlates of consonance, dissonance, and the hierarchy of musical pitch in the human brainstem," *Journal of Neuroscience*, vol. 29, no. 42, pp. 13 165–13 171, 2009. [Online]. Available: <https://www.jneurosci.org/content/29/42/13165>
- [33] Music Information Retrieval Evaluation eXchange (MIREX), "Mirex key detection evaluation," 2010, https://www.music-ir.org/mirex/wiki/2010:Audio_Key_Detection.