ARTEUS: an Algorithmic Rating Tool for Educating Untrained Singers

Arthur Nicholas dos Santos[1], **Bruno Sanches Masiero**[1]

[1]Communication Acoustics Lab – FEEC/Universidade Estadual de Campinas
Av. Albert Einstein, 400 - Cidade Universitária, Campinas - SP

arnics@unicamp.br, masiero@unicamp.br

Abstract. *This study presents the development of a prototype software system for evaluating karaoke performances using a hybrid scoring methodology that integrates objective and non-intrusive subjective metrics. The proposed framework employs scientifically grounded techniques to assess vocal dynamics, timing, pitch, and timbral similarity, complemented by data-driven emotion recognition to capture the perceptual aspects of the performance. A key feature of the system is a composite scoring mechanism with customizable weightings, which enables flexible adaptation to user preferences in pedagogical contexts or entertainment scenarios. By leveraging artificial intelligence models, the system introduces a novel paradigm in performance evaluation, offering a structured yet extensible approach to quantifying aspects of artistic expression that are traditionally considered subjective. Designed with modularity and future scalability in mind, the framework can be expanded to include user-specific evaluation profiles with potential applications in both professional and recreational settings.*

1. Introduction

Karaoke (カラオケ) is a portmanteau of the Japanese terms *kara* (空), meaning "empty", and *ōkesutora* (オーケストラ), meaning "orchestra". It refers to the practice of singing along to pre-recorded instrumental accompaniments, typically with synchronized lyrics displayed on a screen. The absence of a lead vocal track invites the performer to substitute their voice, thereby engaging in a participatory form of musical expression [1].

1.1. Historical Origins

Karaoke is widely regarded as one of Japan's most influential cultural exports. Its invention is typically attributed to the Japanese musician *INOUE Daisuke* (井上 大輔), who frequently performed at singing cafés where patrons would sing along to live instrumental accompaniment, known as *utagoe kissa* (歌声喫茶), which were popular in Japan ca. 1955–1975 [2].

In 1971, in response to the growing demand, Inoue collaborated with an electronics technician, woodworker, and furniture finisher to develop the first karaoke machine prototype, known as the *8 Juke*. This early version lacked visual display components, such as lyric screens, focusing solely on audio playback [2].

Despite Inoue's pioneering contribution, the only officially recognized patent for a karaoke machine was granted to the Filipino inventor Roberto del Rosario (1919–2003). In 1975, he patented an enhanced "sing-along system", which was commercialized in 1978. In 1996, the Supreme Court of the Philippines affirmed del Rosario's status as the legitimate patent holder [2, 3].

1.2. Global Dissemination and Cultural Integration

Following its initial success in Japan, karaoke rapidly spread across East and Southeast Asia, eventually gaining popularity worldwide. Its cultural reception varies significantly: in some regions, it is primarily a leisure activity, while in others, it is approached as a disciplined art form requiring skill and rehearsal [1, 2].

An illustrative example of karaoke's global reach is the establishment of the Karaoke World Championship (KWC) in Finland in 2003. Now comprising participants from over 30 countries, the KWC evaluates performances using the International Karaoke Federation's (IKF) International Scoring System (ISS), which emphasizes four key criteria: vocal quality, technical accuracy, artistic expression, and entertainment value. Each category is scored on a 50-point scale, with a maximum aggregate score of 100 points. Table 1 summarizes the interpretation of scores based on the ISS guidelines. Judges are encouraged to provide qualitative feedback in addition to numerical evaluations, assisting performers in identifying specific areas for improvement [4, 5, 6, 7].

Table 1: Interpretation of performance scores based on the ISS framework.

Score Range	Interpretation
0–20	Low
20–24	Poor
25–29	Average
30–39	Very Good
40–44	Excellent
45–49	Outstanding
50	Perfect

Despite its institutionalization through formal competitions, karaoke remains most prevalent in informal social environments, such as karaoke bars, private rooms, and home systems, where the emphasis lies on enjoyment rather than critical appraisal.

1.3. Technological Evolution

The longevity and adaptability of karaoke are closely tied to its technological evolution and development. From early implementations using cassette tapes, the format expanded to VHS and DVD media. In the digital era, karaoke has migrated to online platforms and video games. Notable

examples include commercial titles such as Karaoke Revolution, Lips (Xbox®), SingStar (PlayStation®), and Rock Band, as well as open-source initiatives such as UltraStar and mobile streaming platforms such as Smule [8, 9].

These digital systems have introduced interactive and gamified features that have broadened karaoke's appeal and accessibility, reinforcing its role in both personal entertainment and communal experiences.

1.4. Research Opportunities and Objectives

The growing use of technology in karaoke settings has given rise to the development of Automatic Singing Assessment (ASA) and Singing Information Processing (SIP) systems. These computational tools aim to provide real-time feedback on performance features such as pitch accuracy and rhythm alignment. The objective is to assist expert human evaluators [10].

Despite these advancements, most karaoke scoring systems described in the patent literature lack rigorous theoretical underpinnings, transparent evaluation protocols, or empirical validation [11]. Most existing systems rely on simplified visual feedback mechanisms that focus only on dynamics and melody.

Given that the human voice is widely considered the most expressive musical instrument, with emotional nuances deeply intertwined with vocal performance, current methodologies fall short of capturing the full expressive scope of singing. A significant research gap remains in the automated analysis, transcription, and classification of expressive singing attributes, including emotional delivery, phrasing, and performance style [12]. This study aims to address these limitations by exploring new computational models and evaluation frameworks for expressive singing assessments in karaoke contexts.

The remainder of this paper is organized as follows: Section 2 outlines the methodology employed in this study; Section 3 presents the experimental results; and an analysis of the research findings is presented in Section 4, summarizing the key points and suggesting directions for future research. Section 5 concludes the paper.

2. Study Objectives and Methodology

This section presents the objectives of the study and the methodology employed to achieve them. An overview of the system is illustrated in Figure 1, which represents the sequential stages of the proposed framework.

The system operates on two primary inputs: (i) the karaoke singer's voice recorded in a noisy environment and (ii) the original song track. The methodological approach is structured into five core objectives.

1. **Source Separation:** The initial step involves isolating the vocal signal from both the karaoke performance and the original track. This ensures the availability of clean *predictor* and *target* signals for the model.
2. **Feature Extraction:** Subsequent to isolation, the signals undergo feature extraction to enable a detailed and meaningful comparison.
3. **Objective Evaluation:** Model-driven, objective metrics are then applied to evaluate the technical fidelity of the karaoke performance.
4. **Subjective Assessment:** Data-driven, non-intrusive metrics are employed to assess perceived performance quality, incorporating subjective evaluation paradigms.
5. **Composite Scoring:** Finally, the system generates a composite performance score, incorporating tunable weights across objective and subjective dimensions.

The following subsections elaborate on the methodological strategies implemented to address each of the five core objectives.

2.1. Objective 1: Source Separation

The first objective is the separation of the *predictor* and *target* signals, referring to the karaoke singer's performance

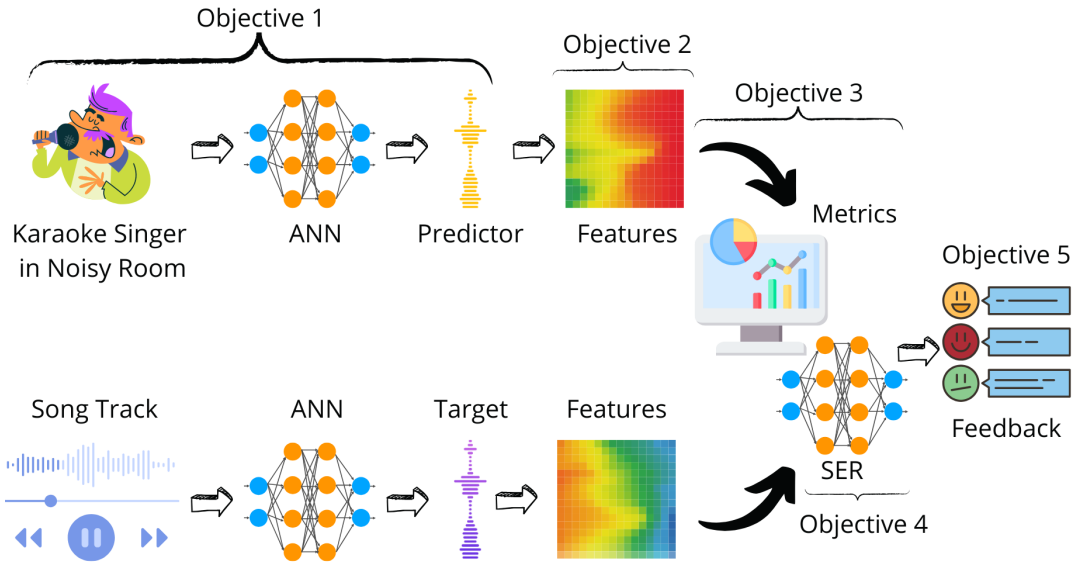

Figure 1: Flowchart illustrating the general objective of this study.

and the isolated vocals from the original studio recording, respectively. This step is foundational because it enables subsequent stages of analysis and evaluation.

The methodology is framed within a realistic *karaoke bar* context, in which the singer performs using a handheld microphone. The performance is accompanied by on-screen lyrics and background music delivered through a sound reinforcement system. This environment introduces multiple sources of noise contamination, which significantly complicate the signal separation process. These noise types can be broadly categorized as follows:

(a) Convolutive Noise: Acoustic reflections from surfaces such as walls, ceilings, and furniture result in reverberation and echo. These effects are modeled as convolutive noise, which alters the temporal and spectral characteristics of the recorded signal [13, 14, 15, 16].

(b) Additive Noise: Background music reproduced via the public address (PA) system may be inadvertently captured by the microphone. Although directional microphones with cardioid polar patterns help suppress ambient interference [17], residual additive components often persist.

(c) Competitive Noise: Audience-generated noise (e.g., cheering, booing, or singing along) constitutes a particularly challenging form of interference. Because this noise is often spectrally similar to the desired vocal signal, it presents a unique challenge for the separation algorithms.

These noise sources compromise the integrity of the predictor signal and necessitate the application of sophisticated speech enhancement and source separation techniques. A similar issue arises for the target signal: although access to multitrack recordings enables straightforward vocal isolation, such resources are rarely available in commercial music. Therefore, blind source separation methods must be employed to extract the target vocal components from mixed stereo tracks.

Traditional model-based approaches, such as Wiener filtering, beamforming, and time-frequency (T-F) masking, have long been employed for source separation tasks. However, the emergence of data-driven methodologies, particularly those based on machine learning (ML) and deep learning (DL), has significantly advanced this field. This evolution has given rise to a subfield often referred to as *Machine Hearing* [18], which is analogous to Computer Vision but focuses on the interpretation and processing of auditory data.

A notable example is presented in the work of researchers at Spotify [19], who implemented a two-dimensional deep U-Net architecture to separate vocals and instruments. Their model processes magnitude spectrograms derived from short-time Fourier transform (STFT) representations of approximately 20,000 song triplets (i.e., original, vocal, and instrumental tracks). The architecture employs a typical encoder-decoder design with a bottleneck layer optimized using an ℓ_1 loss function. This configuration enables the model to function as a Single-Input Multiple-Output (SIMO) system capable of producing spectrogram masks for vocals, drums, bass, and other instrumental components from a single mixed input.

Building upon this foundation, **Objective 1** of our proposed framework is to establish a data-driven preprocessing pipeline utilizing `spleeter` [20], an artificial neural network designed for source separation. This model is based on the architecture introduced in [19], but is trained on a customized dataset to improve the separation quality under real-world conditions. The successful realization of this objective is foundational because it enables and supports the execution of the remaining components within the evaluation framework.

2.2. Objective 2: Feature Extraction

To facilitate a robust comparison between the predictor (karaoke performance) and target (original vocal) signals, it is necessary to move beyond direct waveform analysis. While time-domain representations provide useful information about amplitude and timing, they are insufficient for capturing perceptually and musically relevant differences in sound. Therefore, this stage involves transforming both signals into more informative T-F representations using feature extraction techniques.

Once the vocal signals are cleanly separated, feature extraction must enable a detailed, multidimensional analysis of the audio data. These extracted features should serve as the basis for both model-driven objective metrics and data-driven subjective evaluation. The combination of these two classes of metrics can support a hybrid scoring system capable of delivering a nuanced and perceptually informed assessment of karaoke performance quality.

According to [21], the inherent multidimensionality of musical expression requires a structured representation of audio signals, in which high-level perceptual characteristics are typically inferred from low-level features that are more easily computable. Some examples can be grouped based on the musical dimensions they describe. For instance:

- **Timbre-related features:** These include the mel spectrogram, mel-frequency cepstral coefficients (MFCCs), spectral centroid, spectral roll-off, and zero-crossing rate. These features capture the tonal color and spectral shape of the audio signals.
- **Harmony-related features:** Chief among these is the *chromagram*, which condenses spectral information into the 12 pitch classes of the Western chromatic scale, providing a compact representation of harmonic content over time.

These features are widely supported by standard libraries in Python (e.g., `librosa`) and MATLAB toolboxes, making them readily accessible for music information retrieval applications [22].

While timbre-related features are essential for capturing differences in vocal quality and articulation, chromagrams offer a unique advantage in performance

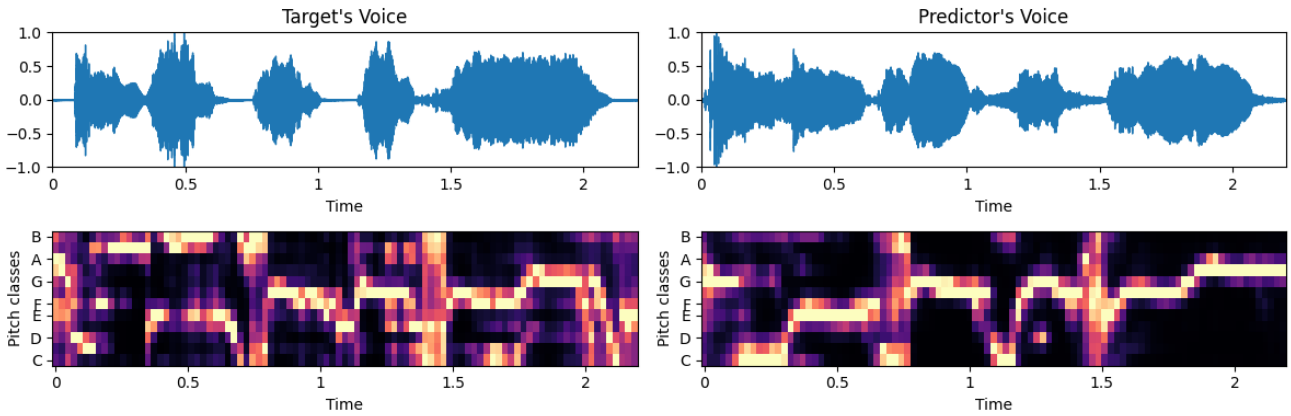

Figure 2: Waveform (top) and chromagram (bottom) of the isolated target (left) predictor (right) voice signals.

evaluation because of their sensitivity to pitch content and harmonic structure. Unlike traditional spectrograms, which display detailed frequency information, chromagrams provide a pitch-class representation that is invariant to octave transposition, making them particularly suitable for analyzing melodic and harmonic consistency [22].

Figure 2 illustrates a comparative chromagram analysis using `librosa`'s `chroma_stft`. On the left, the waveform (top) and chromagram (bottom) of the opening line of Queen's *I Want to Break Free* are shown, using isolated vocals from the original studio multitrack recording. On the right, an untrained singer attempts to reproduce the same phrases. While the waveform highlights timing and dynamic differences, the chromagram captures discrepancies in pitch accuracy and the melodic structure. Such comparative analyses enable both qualitative and quantitative assessments of vocal performance fidelity. Owing to copyright restrictions, no excerpts are provided.

2.3. Objective 3: Objective Evaluation

Following the successful extraction of relevant features from both the predictor and target signals, model-driven approaches are designated to quantitatively evaluate the technical aspects of vocal performance, such as dynamics, timing, pitch, and timbre. The comparison is performed using chromagram matrix representations of the vocal signals, enabling a robust and systematic evaluation framework that distinguishes this system from conventional karaoke-scoring mechanisms.

The proposed evaluation methodology draws on established principles in signal processing and aligns them with vocal performance criteria inspired by the KWC [6]. Each criterion is operationalized using a corresponding signal comparison technique applied to the chromagram representations of the target and predictor signals. This allows for a quantitative assessment while maintaining interpretability in relation to perceptual attributes:

- **Dynamics:** This refers to the variation in vocal intensity throughout the performance. The dynamics are captured using the amplitude of the chromagram bins. To quantify the difference in dynamic expression between the predictor and target, the Euclidean distance (ℓ_2 norm) is computed using `SciPy`'s `euclidean` function. This provides a direct measure of amplitude-level dissimilarity across the chromagram bins, thereby reflecting dynamic consistency.
- **Timing:** Timing encompasses both rhythmic accuracy and expressive deviations from strict tempo. It is evaluated using Dynamic Time Warping (DTW), which aligns the temporal sequences of chromagrams from the predictor and target. The DTW distances are computed with `librosa`'s `dtw` implementation, allowing for an assessment of temporal alignment and detection of timing irregularities or expressive liberties.
- **Pitch:** Pitch accuracy is assessed through cross-correlation (X-Cor), which measures the degree of alignment between the frequency components of the predictor and target chromagrams. This is implemented using `SciPy`'s `correlate` function. Higher cross-correlation values indicate greater pitch congruence, serving as a proxy for intonational accuracy and control.
- **Timbre:** Timbre, or the tonal quality of the voice, is evaluated through cosine similarity (S_C), which measures the angular similarity between harmonic content vectors extracted from the chromagrams. This is implemented via `scikit-learn`'s `cosine_similarity` function. Higher values of S_C indicate a closer resemblance in the tonal character between the two signals.

Table 2 provides an overview of the proposed model-driven objective metrics, detailing their performance criteria, value ranges, and interpretations with respect to the evaluation outcomes.

Table 2: Summary of objective metrics.

Metric	Dimension	Range	Optimal Values
ℓ_2	Dynamics	$[0, \infty)$	Lower
DTW	Timing	$[0, \infty)$	Lower
X-Cor	Pitch	$[-1, 1]$	Higher
S_C	Timbre	$[-1, 1]$	Higher

The integration of these model-driven metrics is

intended to establish a robust and interpretable framework for evaluating karaoke performances. Each metric is designed to capture a distinct technical dimension of singing, thus enabling the analysis of vocal proficiency. Collectively, these metrics facilitate a comprehensive and multidimensional assessment of performance quality. In the following section, this objective framework is further enhanced by incorporating data-driven subjective metrics.

2.4. Objective 4: Subjective Assessment

Contemporary karaoke scoring systems predominantly rely on objective metrics that focus on technical parameters such as pitch accuracy, timing, and dynamics. While such metrics represent a significant advancement, they do not fully capture the expressive or emotional aspects of a musical performance. Integrating subjective metrics, once calibrated against human evaluations, can offer complementary insights through probabilistic and empirical learning methods. These metrics have the potential to replicate human judgment in a non-intrusive and scalable manner.

One promising application lies in Song Emotion Recognition (SER), wherein artificial neural networks classify vocal performances into emotional categories (e.g., happiness, sadness, and neutrality). Musical performances often aim to convey specific emotional content, aligning the singer's expressive intent with the listener's perceived emotion. However, this alignment is nontrivial, as Music Emotion Recognition (MER) involves complex cognitive processes shaped by individual experiences and cultural backgrounds [22].

Although ML provides a powerful framework for SER, it introduces methodological challenges. A comprehensive review by [22] examined common feature representations and model architectures for SER, particularly in the *a cappella* context. Their findings identified chromagram features as highly effective, with a 2D convolutional neural network (CNN) achieving a test accuracy of 0.84 ± 0.02 on the RAVDESS dataset [23]. The emotion classes in that study included neutral, calm, happy, sad, angry, and fearful. Data augmentation techniques were used to improve the generalization of the model.

Within the context of karaoke scoring, emotional expressiveness can be conceptualized as the capacity to "*exude emotion and passion appropriate to the song, conveyed through voice, body, and facial expression*", a criterion that is frequently emphasized in vocal performance competitions. Incorporating emotion recognition into the scoring framework not only deepens the evaluative granularity, but also aligns the system with the inherently artistic and performative nature of singing. In this study, emotional analysis is conducted using the best-performing CNN model identified in [22], which is applied to the chromagrams of the isolated vocal signals obtained via source separation.

2.5. Objective 5: Composite Scoring

The final objective focuses on computing a composite score by integrating model-driven objective metrics with data-driven, non-intrusive subjective evaluations. The overarching aim is to develop a scientifically grounded scoring system that delivers meaningful feedback to karaoke performers, encompassing both technical accuracy and artistry.

The composite score is computed as a weighted average of the normalized metrics derived from both objective and subjective analyses. These weights can be adjusted, enabling the tuning of the evaluation's strictness based on the intended application (e.g., casual entertainment vs. competitive performance). To complement the final score, a detailed breakdown of the individual components is also presented, offering granular insights into specific performance dimensions.

In summary, this composite scoring approach integrates multiple analytical perspectives into a coherent framework, balancing technical precision with perceptual relevance. This establishes a robust foundation for both evaluative feedback and future research extensions in automated performance assessment.

3. Experimental Results

To evaluate the proposed methodology, two primary input signals were utilized: (i) a karaoke performance recorded in a real-world setting and (ii) the original studio recording of the same song performed by the original artist. Figure 3a presents the waveform of the professionally produced original song ("Paranoid" by Black Sabbath). As shown in the figure, aside from a brief introductory segment, the audio signal exhibits consistent dynamic compression, with loudness shaped in an aesthetically controlled manner during mixing and mastering. Owing to copyright restrictions, no excerpts are provided.

In contrast, the karaoke performance was captured in a bar environment, featuring an untrained singer accompanied by loud and enthusiastic audience members. Consequently, the recording was characterized by significant additive noise originating from both the PA system and the crowd. Figure 3b shows the waveform of this recording, clearly illustrating the aforementioned artifacts. Additionally, instances of clipping are evident.

Figures 3c and 3d depict the waveforms following source separation using `spleeter`. In the case of the *clean target* (Figure 3c), the output contains largely muted regions with minor artifacts, corresponding to sections that previously contained instrumental components (e.g., electric guitar solo). Notably, the vocal signal envelope is well preserved.

In contrast, the *clean predictor* (Figure 3d) retains substantial vocal noise components. This discrepancy arises because crowd cheering shares spectral characteristics similar to those of the lead vocal signal. Consequently, the neural network extracts not only the singing voice but also other vocal-like sounds, highlighting a key limitation of this separation approach.

The impact of this limitation is further observ-

able in the chromatograms shown in Figures 3e and 3f. The chromagram of the *clean predictor* exhibits a high degree of noise and disorganization, whereas the *clean target* displays clearer harmonic content. Both chromagrams were computed with the parameters set as follows: $N_{FFT} = 2048$, hop length $= 512$, $N_{chroma} = 12$, and a Hann window. Despite the improved clarity in the *clean target*, some bins still exhibited energy during segments that were expected to be silent, suggesting residual artifacts and incomplete separation.

To assess the similarity over time and pitch classes, the metrics were computed between the chromagrams on a frame-by-frame basis. The resulting time series of these metrics is plotted in Figure 3g, with the y-axis representing the scores and their mean values for interpretability.

A key design consideration in constructing the composite score is to ensure the compatibility and comparability of its constituent parts. Certain objective metrics, such as the ℓ_2 norm and DTW, are unbounded above, which may result in a disproportionate influence on the overall score if they are left unbounded. In contrast, similarity measures, such as X-Cor and S_C, are inherently bounded within $[-1, 1]$, which facilitates a more stable integration.

To address this imbalance, all metrics were rescaled or transformed into a common bounded domain prior to aggregation. This normalization step ensures that no single metric dominates the composite score and that each contribute equitably to the final evaluation. The normalization schemes are defined as follows:

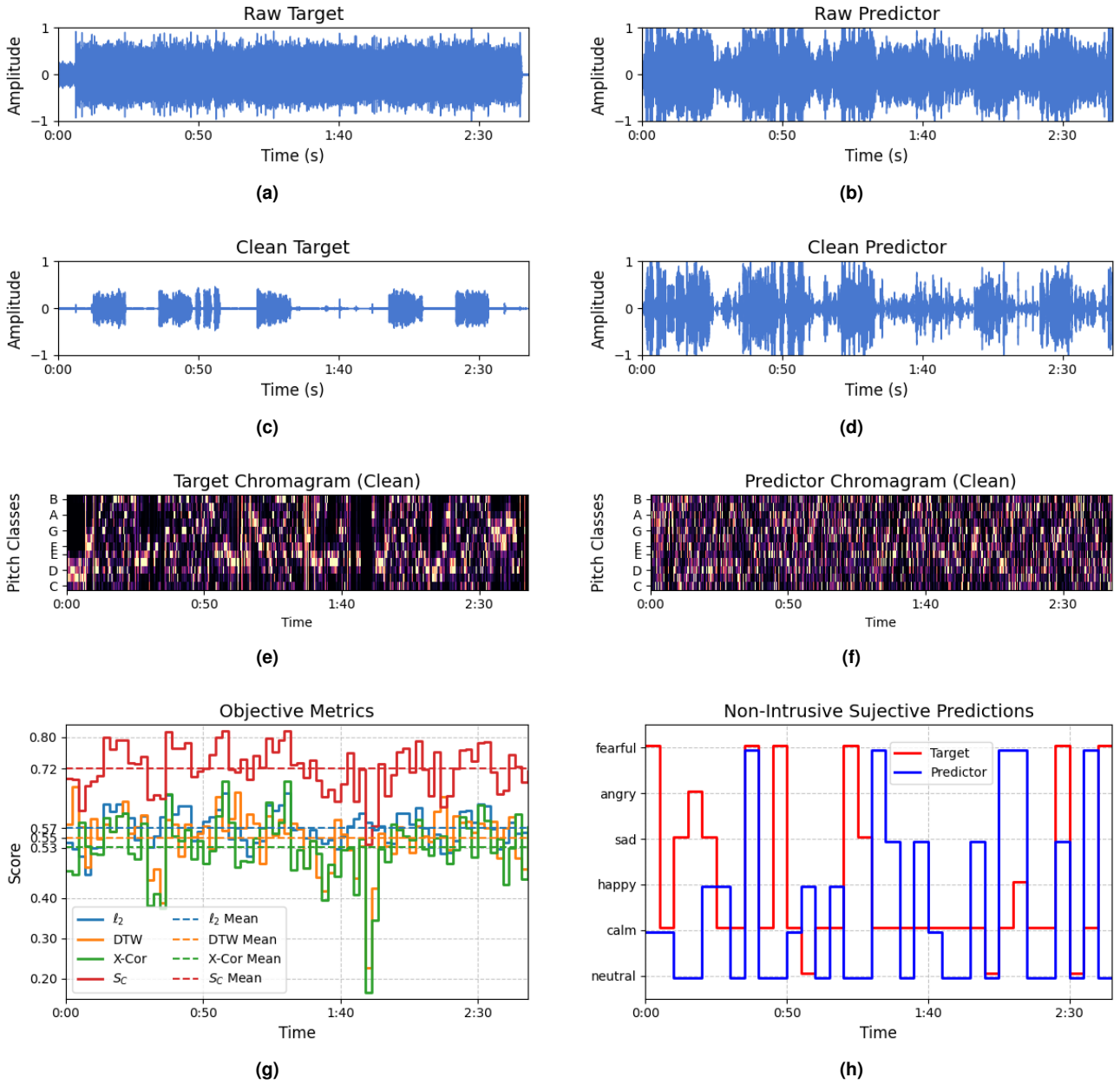

Figure 3: Visualization of various stages: recording, pre-processing, scoring, and analysis.

$$\ell_{2,\text{normalized}} = 1 - \frac{\ell_2}{\ell_{2,\text{max}}} \tag{1}$$

$$\text{DTW}_{\text{normalized}} = 1 - \frac{\text{DTW}_{\text{distance}}}{\text{DTW}_{\text{max}}} \tag{2}$$

$$X_{\text{cor, normalized}} = \max\left(\frac{X_{\text{cor}}}{\|\text{target}\| \cdot \|\text{predictor}\|}\right) \tag{3}$$

$$S_{C,\text{normalized}} = \frac{S_C + 1}{2} \tag{4}$$

in which DTW_{max} was assumed to be 10, while $\ell_{2,\text{max}}$ was defined as $\sqrt{12 \times N}$, in which N denotes the number of time frames in the predictor signal.

Emotion recognition was conducted using input chunks of 211 frames at a time. For the *clean target*, predictions were predominantly associated with the emotions of "sadness", "anger", and "fear", although the model also generated predictions for segments that should have been silent (e.g., instrumental parts removed during source separation). In the case of the *clean predictor*, the model predicted a more diverse emotional profile, including "happiness", "fear", "neutrality", and "sadness".

4. Discussion

Traditional karaoke scoring systems typically assign a single numerical grade to performances. This score is often derived from factors such as signal amplitude thresholds or the alignment of musical note intensity, duration, and pitch accuracy relative to a predefined musical structure, commonly extracted from MIDI files.

In contrast, the methodology proposed in this study evaluates karaoke performances against the reference performance of the original artist. This human-to-human comparison has both advantages and limitations. Conceptually, it is arguably more meaningful to compare one human performance with another rather than with a synthetic representation such as MIDI. However, this approach also raises concerns about subjectivity and fairness, especially when the original performance includes expressive liberties and stylistic nuances that an amateur singer may not replicate. Nevertheless, any evaluative system must anchor itself to a reference, and in this case, professional performance serves that role.

In the experimental setup presented, the karaoke performer achieved average scores of 57/100 for vocal dynamics, 55/100 for rhythmic timing, 53/100 for pitch accuracy, and 72/100 for timbral similarity, the latter being the highest among the objective measures (cf. Figure 3g). Applying equal weights across these categories yielded a composite technical score of 59.25/100, indicating a relatively *average* overall performance, according to Table 1.

However, several contextual factors must be considered. The source separation process was limited in its ability to isolate the singer's voice from overlapping environmental noise. Because the recording environment included loud audience cheering and ongoing conversations, vocal elements unrelated to the singer were retained and consequently evaluated against the target signal. These extraneous voices were often present during instrumental sections when the singer was silent, thereby skewing the assessments. Therefore, future implementations should incorporate voice activity detection (VAD) to distinguish between vocal segments attributable to the performer and background interference.

Additionally, the audio was captured informally by a member of the audience using a smartphone rather than being directly recorded from the singer's handheld microphone. This significantly affected the signal quality and hindered accurate evaluation. Ideally, a direct feed from the vocal microphone should be used to minimize distortion and isolate the performer's voice more effectively.

Regarding emotion recognition, the emotional trajectory of the original artist (predominantly sadness, anger, and fear) is consistent with the expressive nature of the song. In contrast, the karaoke singer's emotional profile included neutrality, happiness, fear, and sadness. These discrepancies are understandable, given the informal and convivial setting of a bar environment where alcohol was present and audience interaction was high.

These findings suggest several directions for future research. First, improvements in the experimental design are essential, particularly regarding high-quality vocal capture. Second, exploring alternative source separation methods, potentially including approaches tailored to "*cocktail party*" environments [24], could enhance the noise mitigation. Third, integrating VAD post-separation would allow the removal of silent segments and reduce error propagation in chromagram computation.

The temporal evolution of the evaluation metrics, as depicted in Figures 3g and 3h, underscores the importance of integrating real-time visual feedback into the karaoke-performance-assessment framework. A solely post hoc summary of performance is inadequate for pedagogical applications, as singers may find it difficult to associate abstract scores with specific error moments. In contrast, real-time feedback facilitates immediate corrective action, thereby promoting a more interactive and effective learning environment, particularly among untrained vocalists. By delivering timely and detailed assessments, real-time feedback has the potential to transform karaoke systems into practical educational tools, enabling singers to iteratively improve their techniques through continuous constructive evaluation.

5. Conclusion

This study introduced a novel methodology for evaluating karaoke performances by integrating objective and subjective metrics within a unified framework. Unlike conven-

tional scoring systems that rely predominantly on simplified representations, such as MIDI files, the proposed approach evaluates performances relative to professional studio recordings by the original artists.

The methodology leveraged advanced source separation techniques, enabling performance evaluation in acoustically challenging environments. By incorporating both technical and perceptual aspects, the system facilitated a holistic assessment of vocal performance.

Despite these contributions, the experimental findings highlight several limitations that warrant further investigations. Source separation remains a critical challenge, particularly in environments with background noise that is spectrally similar to singing voices, such as crowd chatter. Additionally, the reliance on ambient audio captured via consumer devices negatively impacted the signal quality, reinforcing the need for a dedicated microphone input to ensure a reliable analysis. The absence of voice activity detection complicates the accurate segmentation of vocal content, which is essential for meaningful chromagram and metric calculations.

The dynamic nature of the proposed evaluation metrics, as observed over time, further underscores the value of real-time visual feedback. Such a capability would enhance the system's educational potential by allowing performers to make immediate corrections, thereby transforming karaoke into a practical training tool for improving vocals. Additionally, the emotion recognition component warrants further validation to better understand its role in performance assessments.

In conclusion, while technical challenges remain, the findings presented here contribute to a deeper understanding of vocal-performance assessment in real-world settings.

6. Acknowledgments

We extend our sincere gratitude to Mateus Prauchner Enriconi and Arthur Lucas da Silva Nogueira for their insightful contributions, constructive discussions, and valuable suggestions.

References

[1] Laurence Green. *The Serious Business of Song: Karaoke as Discipline and Industry in Japan*, page 231–243. Amsterdam University Press, 2022.

[2] X. Zhou and F. Tarocco. *Karaoke: The Global Phenomenon*. Reaktion Books, 2007.

[3] Supreme Court of the Philippines. Roberto l. del rosario vs. court of appeals and janito corporation. https://lawphil.net/judjuris/juri1996/mar1996/gr_115106_1996.html, March 1996. G.R. No. 115106, March 15, 1996.

[4] Heikki Ruismäki, Antti Juvonen, and Kimmo Lehtonen. Karaoke–the chance to be a star. *The European Journal of Social & Behavioural Sciences*, 7:1222–1233, 2013.

[5] Karaoke World Championships. Kwc – world's biggest global music contest. https://www.kwc.fi/, 2025. Accessed: 2025-05-30.

[6] Karaoke World Championships USA. Kwcusa official judging criteria – solo division. https://www.kwcusa.org/judging-criteria, 2025. Accessed: 2025-05-30.

[7] United States Karaoke Association. Uska scoring system. https://uskakaraoke.com/about/scoring-system/, 2025. Accessed: 2025-05-30.

[8] UltraStar Deluxe Team. Ultrastar deluxe – free open source karaoke game. https://usdx.eu/, 2025. Accessed: 2025-05-30.

[9] Smule Inc. Smule – social singing app. https://www.smule.com/, 2025. Accessed: 2025-05-30.

[10] Huan Zhang, Yi Jiang, Tao Jiang, and Peng Hu. Learn by referencing: Towards deep metric learning for singing assessment. In *International Society for Music Information Retrieval Conference*, 2021.

[11] Wei-Ho Tsai and Hsin-Chieh Lee. Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features. *IEEE transactions on audio, speech, and language processing*, 20(4):1233–1243, 2011.

[12] Oscar Mayor, Jordi Bonada, and Alex Loscos. Performance analysis and scoring of the singing voice. In *Proc. 35th AES Intl. Conf., London, UK*, pages 1–7, 2009.

[13] Frank Filipanits. Design and implementation of an auralization system with a spectrum-based temporal processing optimization. *Master's Research Project, University of Miami May*, 1994.

[14] Ennio Cruz da Costa. *Acústica técnica*. Bluscher, 2003.

[15] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65:943–950, 4 1979.

[16] Alan V. Oppenheim, Alan S. Willsky, and S. Hamid Nawab. *Signals & Systems (2nd Ed.)*. Prentice-Hall, Inc., USA, 1996.

[17] Franz Zotter and Matthias Frank. *Ambisonics*, volume 19. Springer International Publishing, 2019.

[18] Richard F. Lyon. *Human and Machine Hearing*. Cambridge University Press, 5 2017.

[19] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde. Singing voice separation with deep u-net convolutional networks. Unpublished, October 2017.

[20] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: A fast and stateof-the art music source separation tool with pre-trained models. *Late-Breaking/Demo ISMIR*, 2019.

[21] Renato Panda, Ricardo Malheiro, and Rui Pedro Paiva. Audio features for music emotion recognition: A survey. *IEEE Transactions on Affective Computing*, 14(1):68–88, 2023.

[22] Pedro Benevenuto Valadares, Karen Gissell Rosero Jácome, Arthur Nicholas dos Santos, and Bruno Sanches Masiero. Song emotion recognition: a performance comparison between audio features and artificial neural networks. *Revista Eletrônica de Iniciação Científica em Computação*, 20(4), dez. 2022.

[23] Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, 13(5), 2018.

[24] Andrew J. R. Simpson. Probabilistic binary-mask cocktail-party source separation in a convolutional deep neural network, 2015.