Audio Engineering Society

Conference Paper 19

Presented at the AES International Conference on Machine Learning and Artificial Intelligence for Audio
2025 September 8–10, London, UK

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

A Dataset of Amateur Karaoke Singing and Audiobook Narration

Elena Georgieva[1], Pablo Ripollés[1,2, 3], and Brian McFee[1,4]

[1]*New York University Music and Audio Research Lab (MARL)*
[2]*NYU Center for Language Music and Emotion (CLaME)*
[3]*NYU Department of Psychology*
[4]*NYU Center for Data Science*

Correspondence should be addressed to Elena Georgieva (elena@nyu.edu)

ABSTRACT

Singing and narration are fundamental forms of vocal expression. Listener-evaluated data is essential for developing AI systems that better capture the nuances of vocal expression and resonate with listeners perceptually and emotionally. In this paper, we present a dataset with 4300 ratings of 940 recordings of amateur karaoke singing and audiobook narration, annotated by 86 participants. Participants rated multiple aspects of audio quality and vocal performance including skill, likability, passion, sincerity, and intelligibility, among others. We release this dataset alongside baseline analyses comparing ratings of singing and audiobook narration, as well as ratings of clean excerpts versus those with added degradations. We also release a linear regression model that estimates listener ratings from audio. This dataset will serve as a valuable resource for future research in music and speech, with applications in intelligent music production, vocal audio effect recommendation, audiobook editing, content personalization, and automated assessment of singing and speaking voices, to name a few.

1 Introduction

The intelligent music production community could benefit from listener-evaluated datasets to enhance machine learning models, particularly for vocals—one of the most inherently human forms of musical expression [1]. Notably, among the available AI audio datasets, only a few include information from human subjects.[1] At the same time, audiobooks are becoming more popular than ever, changing traditional notions of reading and storytelling, and reshaping the way listeners engage with literary works. The integration of AI is expected to transform audio literature through AI-generated narration, and human annotations provide crucial insights into how listeners perceive vocal performance, expressiveness, and intelligibility in spoken-word audio [2, 3]. Ultimately, more listener-evaluated data is essential for developing AI systems that replicate or analyze the human voice.

Singing is accessible to anyone with a voice, and even those who do not self-identify as "singers" can express themselves through vocal performance. While much research focuses on trained vocal professionals, who develop their skills through voice lessons and choral singing [4, 5]—less is known about unskilled amateur

[1]https://github.com/Yuan-ManX/ai-audio-datasets

performances, which this study explores. We study karaoke recordings because they are performed by amateurs, likely without formal musical training.

In parallel, we examine the same vocal characteristics in audiobook narration performance. A skilled narrator can contribute considerably to the success of a book, and a poor narrator can ruin it [6]. Understanding what makes an effective narration is valuable, as audiobook narration is a performative vocal art form that calls for significantly more vocal expression than typical conversational speech [7]. We choose to study audiobook narration as a proxy for expressive speech more broadly—including podcasts and radio—due to its performative nature and the availability of LibriSpeech, the dataset we use.

Skilled singing performance and audiobook narration both demand vocal expressiveness, precise rhythm, and emotional depth. Singing requires sustained pitch accuracy and musical phrasing, and audiobook narration involves clear enunciation and pacing. Both challenge performers to effectively convey meaning and engage their audience, with performances varying in skill level and technique.

We collected subjective ratings from 86 participants who listened to isolated karaoke recordings and audiobook narration performances of varying audio quality and levels of performer skill. Listeners provided subjective ratings on performer skill, likability, passion, sincerity, emotional effectiveness, power, intelligibility, familiarity, background noise, intelligibility due to recording quality, and whether there were multiple speakers/singers. This dataset not only offers valuable insights into how amateur performances are perceived, but also serves as a resource for machine learning and AI applications related singing and speech.

We release the responses as a dataset, along with baseline analyses.[2] As preliminary analyses, we include results of a linear regression model to estimate the behavioral responses based on a set of extracted acoustic features, including pitch and spectral characteristics. In addition, we fit linear mixed-effects models to compare subjective ratings between karaoke and audiobook narration excerpts, examining how listeners assess recordings in these distinct contexts. We discuss how listeners rate clean versus noisy excerpts differently, leading to implications for audio effects recommendation in intelligent music production and audiobook production. Through our analysis, we confirm the reliability of the listeners' ratings, demonstrating that the data collection was thorough and that participants provided attentive and meaningful evaluations.

2 Related Work

In the field of Intelligent Music Production, listener evaluations exist for mix datasets [8, 9], and audio quality has been listener-evaluated for popular music [10]. However, other studies have not examined listener evaluations of amateur vocal performances. Audiobooks, more popular than ever, are also being shaped by advancements in artificial intelligence, including generated voices and audiobook recommendation [11]. These developments highlight the need for human evaluations to better understand how listeners perceive AI-generated and human-performed vocal expression. The Kaggle Dataset provides metadata on popular audiobooks, including user ratings and reviews, but lacks direct human evaluations of narration quality. [3] Despite AI's growing role in singing and speech production, listener-evaluated datasets for spoken and sung performances remain scarce—highlighting a critical gap and an opportunity for meaningful contribution.

Previous research has explored the automatic prediction of singing skill using acoustic features. Past work has found that the singing power ratio (SPR) is a useful metric [12]. SPR calculates the ratio between peak intensities in the 2–4 kHz and 0–2 kHz frequency bands. Watts et al., (2006) found that there is a notable difference in SPR values between trained and untrained singers, with trained singers having a more resonant sound with vocal "ring". Other research has focused on automatically assessing perceived singing quality using a variety of classifiers and features [13]. Another approach to skill evaluation relies on pitch-based analysis [14, 15, 16]. While pitch is an important measure of vocal skill, it is only one of many factors that shape listener perception of vocal skill.

We also investigated how listeners perceived audio quality. A widely used listener-based metric for evaluating speech quality is the Mean Opinion Score (MOS), where listeners rate audio quality on a five-point Likert

[2] https://github.com/elenatheodora/KaraokeAudiobook

[3] https://www.kaggle.com/datasets/s4lman/top-audiobooks-dataset

scale (Bad, Poor, Fair, Good, Excellent) [17]. Recently, research has focused on blind MOS estimation, using audio signals online. One model estimates MOS as well as reverberation time (T60) and clarity (C50) [18]. Another uses a convolutional neural network to estimate MOS for individual speech utterances [19].

On the singing side, researchers have explored sung lyric intelligibility—defined as a listener's ability to accurately comprehend lyrics [20]. Despite adjacent work on speech quality assessment and singing skill prediction, no prior study has focused on predicting subjective listener evaluations of vocal recordings in different domains (i.e., karaoke, audiobook narration) and across the range of characteristics studied here.

3 Methods

3.1 Dataset

Participants in this study listened to karaoke and audiobook narration recordings. The karaoke singing voice recordings were sourced from the Digital Archive of Mobile Performances (DAMP), recorded and collated by Smule, Inc [21]. The archive includes solo vocal recordings from smartphone users using Smule's karaoke mobile application,[4] and has at least 35,000 vocal performances of contemporary American pop songs like "Love Yourself" by Justin Bieber and "Rockabye" by Clean Bandit. The dataset includes a great diversity of vocal training and performance styles, as it is mostly amateur singers singing karaoke for fun. Recordings come from a wide variety of locations around the world, and have varying recording qualities depending on the device used and level of background noise. The recordings were captured using smartphones and may include compression artifacts, background noise (depending on the recording environment), and device-specific limitations such as poor microphone quality, that could result in an unbalanced frequency response, limited dynamic range, distortion at high input levels, and other artifacts.

The audiobook narration performance recordings were drawn from the LibriSpeech dataset—a corpus of approximately 1000 hours of English audiobook narration [22]. Similarly to the DAMP dataset, LibriSpeech has a variety of narrators with different levels of training, different performance styles, and varying audio quality. LibriSpeech is derived from audiobooks that are part of the LibriVox project [23], a selection of free public domain audiobooks read by volunteers from around the world. We randomly sampled a subset of recordings from both datasets and manually screened them based on specific selection criteria.

For the karaoke recordings, we selected a diverse set of performances while ensuring variety in song choices. The dataset contained multiple performances of the same songs, and we limited the number of times any single song appeared to avoid overrepresentation. We also included a range of performance skill levels, assessed perceptually by co-author E. Georgieva, who has 18 years of experience singing in various vocal ensembles. For audiobook narration, we selected passages from different genres, fiction and non-fiction, while avoiding content with violent or distressing themes.

In both datasets, recordings were chosen to minimize background noise based on objective and subjective listening. We first used spectral contrast to estimate the most clean audio clips. Then, co-author E. Georgieva, who also has training and experience as an audio engineer, reviewed the the files resulting from the spectral contrasts selection and chose the cleanest files. The final study included 940 audio clips: 470 karaoke and 470 audiobook narration. All audio files were monophonic with a 48 kHz sampling rate.

For 840 of these recordings, we used the clean, unedited clip. To control for the quality of the ratings, we degraded 100 recordings (50 karaoke, 50 audiobook) using one of four methods designed to intensify existing recording degradations: low distortion, high distortion, low noise, and high noise. For the distortion files, we used the Pedalboard library to apply nonlinear harmonic distortion via a hyperbolic tangent waveshaping function at two effect levels: low (20 dB drive) and high (a greater drive setting) [24]. This simulates distortion from mic clipping in recordings, but with a greater amount than is naturally present in the data. For the noise conditions, we added pink noise to the audio files at two signal-to-noise ratio (SNR) levels: low noise (0.02 amplitude) and high noise (0.1 amplitude), using SoX's pink noise generator [25]. This process introduced controlled background noise to simulate environmental background noise, alongside the original audio signal. All clips, clean and with added noise, were normalized to the European Broadcasting Union (EBU) R128 loudness standard using FFmpeg [26].

[4] https://apps.apple.com/us/app/smule-sing-record-karaoke/id509993510

3.2 Listener Study Methodology

The vocal characteristics that we are interested in are subjective; therefore, evaluations from human listeners are the best data to describe them. Data were collected online through the SONA platform[5] for recruitment of participants at New York University, as well as via a university email list, and word of mouth. Experiments were hosted on Qualtrics.[6] The Qualtrics survey had audio playback functionality, allowing participants to listen to each 15-second clip before responding. All responses were collected anonymously.

Participants were first presented with the informed consent and then with instructions about the task to be completed. Participants were required to use headphones and underwent a headphone test to determine if they were doing so. This test has been previously validated and is widely used in online experiments using auditory stimuli [27, 28]. Listeners judge which of three pure tones is quietest, with one tone presented 180° out of phase between stereo channels. The task is easy with headphones but difficult with loudspeakers due to phase cancellation. The test is repeated six times, accounting for all possible order combinations of the three pure tones presented.

Participants were then presented with a series of 15-second audio clips of either karaoke singing or audiobook narration and asked to evaluate each recording. The 15-second audio clips included 500 ms of fade-in and 500 ms of fade-out. Listeners could pause/play as they liked and listen to the audio clip as many times as desired. Each participant rated 50 audio recordings (25 karaoke songs; 25 audiobook narrations), where 42 recordings were clean and eight (four karaoke; four audiobook) had the added degradations. Audio files were presented in an interspersed and randomized order, and completed some demographic questions along with the Goldsmith-MSI (measuring musical training) and Barcelona Music Reward Questionnaire (measuring sensitivity to musical reward) [29, 30]. The study lasted around 60 minutes and listeners were compensated for their time either with SONA academic credit or a $15 gift card.

After listening to each audio clip, respondents rated their agreement with seven statements regarding the performance and audio quality of the recording on a five-point Likert scale with options "Completely Disagree(1)", "Disagree(2)", "Neither Agree Nor Disagree(3)", "Agree(4)", and "Completely Agree(5)". Statements assessed perceptions of skill, performance quality, emotional expression, vocal power, intelligibility, and familiarity with the audio content. Participants were required to spend at least 25 seconds on each page before proceeding to the next question. The seven statements were:

- "The performer was highly skilled in delivering the spoken or sung text."
- "I really liked the way the excerpt was performed."
- "The performer conveyed a lot of passion in their performance."
- "The performer's expression was very sincere."
- "The performer conveyed emotions effectively (regardless of whether emotions were positive or negative)."
- "The performer's voice was very powerful."
- "I could clearly discern the words being spoken/sung."

A second section of questions invited participants to evaluate specific attributes of the audio recordings. Respondents answered three trinary-choice questions (options: "no(1)", "maybe(3)", "yes(5)"), addressing intelligibility and sound quality. The number choices one, three, and five were used during analysis to preserve the scale among all questions. The three questions were:

- "Is there more than one singer/speaker in the recording?"
- "Are there sounds other than the speaker/singer in the recording?"
- "The recording had unintelligible words due to recording quality."

For succinctness, we abbreviate the eleven questions to: skill, likability, passion, sincerity, emotional effectiveness, power, intelligibility, familiarity, multiple speakers, background noise, and recording quality.

We used two different strategies to identify and discard low-quality responses where listeners were likely not

[5] https://www.sona-systems.com/
[6] https://qualtrics.com

focused on the task: the headphone test and a set of catch trials. Throughout the study, participants encountered six "catch trials," in which they were presented with an audio file and asked: 'Is the following audio the same or different from the audio you heard on the previous page?' In half of these trials, the correct answer was "yes" and the audio file was identical to the previous one. In the other half, the correct answer was "no" and the audio file was different, though it came from the same dataset (e.g., audiobook or karaoke).

147 participants completed the study. The majority of these participants were recruited via the SONA platform at New York University, and were granted course credits for taking part in the study. Some participants were recruited from a NYU music technology program email list, and finally four participants were recruited by word of mouth. The non-SONA participants received a $15 gift card as compensation.

Data from 52 participants were removed because listeners failed three or more of the headphone test questions (see section 3.2). 25 participants failed two or more of the catch trials (16 of them had already been discarded for failing the headphone test). We kept recruiting participants until each recording had been evaluated by at least four individuals. Ultimately, 86 participants were included in the analysis. 63 were female, 1 non-binary/ third gender. The median age was 20.32 years, with a standard deviation of 2.89. Among the participants, the most common identities were Asian (47), White (23), and Black or African American (9). Eight participants identified with multiple racial/ethnic groups, and eight identified as Hispanic or Latino. When asked about their mother tongue, 37 participants reported English, 18 mentioned English and another language, and 31 listed "Other" as their mother tongue. In the open text field for "Other," 19 participants wrote Chinese (without specifying a dialect), eight mentioned Mandarin, and two cited Cantonese. Additionally, seven participants listed Spanish, four said Russian, three mentioned Korean, and several other languages were mentioned once or twice. We collected a total of 4300 ratings, with each of the 940 recordings being evaluated by at least four participants, averaging 4.6 ratings per audio file with a standard deviation of 1.6. All participants provided written informed consent, as approved by the local institutional review board (New York University's Committee on Activities Involving Human Subjects). All data were anonymized for sharing purposes.

We split questions into two categories: stimulus questions and participant-stimulus questions. Participant-stimulus questions were quite subjective, and responses vary between participants, as indicated by their larger standard deviations. In the stimulus questions: intelligibility, background noise, recording quality, and multiple speakers, the participants evaluated the stimuli in a more objective manner. These have smaller standard deviations across responses (see Table 2).

3.3 Statistical Analyses

To examine the relationship between acoustic features and listener ratings, we applied an Elastic Net linear regression model with 5-fold cross-validation using Python Scikit-learn [31]. In preparing this linear regression, we extracted a range of acoustic features from the audio files to capture various aspects of vocal expression. These features included the mean, standard deviation, minimum, maximum, and 25th and 75th percentiles of the estimated fundamental frequency (f_0) and the root mean square (RMS) energy. Fundamental frequency (f_0) and RMS energy were chosen because they capture key characteristics of vocal expressiveness and acoustic properties. These features provide a representation of the pitch and energy, which are essential for distinguishing differences in vocal performance. The use of f_0 estimation via Pyin and RMS energy analysis aligns with common practices in speech and music research, as they are both reliable indicators of vocal and audio signal characteristics. For f_0 estimation we used Pyin as implemented in Librosa, with default parameters, other than the sampling rate (48 kHz) [32, 33]. We computed an f_0 estimate approximately every 11 milliseconds.

Additionally, we analyze several acoustic features of vocal expressiveness: total variation of pitch (measuring the rate of pitch change [34]), the fraction of time with no voice, and pitch class entropy (which quantifying the degree of unpredictability for the set of vocal pitches). The pitch features and fundamental frequency (f_0) estimation reflects pitch variation in singing and speech, which can provide insight into performer expression and skill [15, 7].

We also computed spectral features, including spectral contrast, centroid, bandwidth, 85% spectral rolloff, and the ratio of spectral energy in the 2–4 kHz frequency range to the total spectral energy of an audio signal. These spectral features offer insights into the

timbral qualities of the voice, which can influence listener perceptions [12]. All features were computed with a standard 2048 frame length and hop size of 512. Finally, we calculate the mean, standard deviation, min, and max of the first 20 MFCCs, widely used speech features, aggregated over time [35]. These features were used as input for the Elastic Net models.

The models aimed to predict listener ratings for each stimulus, using average ratings for each of the 940 stimuli, each rated by four or more listeners. Each model was trained separately for each question, and included both karaoke and audiobook recordings in the same model. We optimized the alpha (regularization strength) and ℓ_1 ratio (balance between ℓ_1 and ℓ_2 regularization) parameters using a grid search on the validation set. The dataset was split into five folds using GroupKFold cross-validation, ensuring that all samples from the same artist appeared in either the training or test set, but not both. We performed the cross-validation five times, each producing distinct training and test sets that were saved for subsequent use.

We evaluated model performance using mean R^2 score and root mean squared error (RMSE). R^2 is the percentage of the variation in the data that the model can explain, and RMSE measures how far off the model's predictions are from the actual values, with lower RMSE indicating better predictions. We repeated this process for each question, allowing us to assess which aspects of vocal expression were most predictable from acoustic features. We evaluate our model twice: once with a full test set, and once on a clean-only test set with the noisy excerpts omitted.

To analyze the ratings, we used R (4.2.2) and RStudio (2024.12.1+563) to implement a linear model with the lm function to quantify the effect of excerpt type (karaoke vs. audiobook; i.e. skill $\sim$ ExcerptType) and noise type (clean vs. noisy; i.e. skill $\sim$ Noise) on each of the questions. We used the lm function to fit the models and used the ggpredict function from the ggeffects package to generate predictions. The models were computed separately for each question. With these linear models, we quantified how the conditions shape listeners' evaluations of the performance.

Finally, we fit a linear mixed-effects model using maximum likelihood estimation to examine the relationship between a stimulus-participant questions and the stimulus questions (e.g., Skill $\sim$ Intelligibility + Multiple_Speakers + Background_Noise + Recording_Quality + (1 | Filename) + (1 | ParticipantID)). The models included random intercepts for filename and participant identification code to account for variability across recordings and participants. To do so we used the lme4 package in R.

4 Results

The results of the user study are illustrated in Figure 1. Among the 11 questions assessed using Elastic Net linear regression models, we obtained the highest R^2 score for "Are there sounds other than the speaker/singer in the recording?" (R^2 = 0.33, RMSE = 0.75), suggesting that background noise was relatively well predicted by the extracted acoustic features. The high RMSE indicates considerable prediction error, possibility due to the variability of background noise and differences in how listeners interpret the question. We observed the second highest R^2 for "The recording had unintelligible words due to recording quality" (R^2 = 0.20, RMSE = 0.56, α = 0.01, ℓ_1 ratio = 0.01), followed by "I could clearly discern the words being spoken/sung" (R^2 = 0.20, RMSE = 0.50, α = 1.0, ℓ_1 ratio = 0.01).

Predictions for more subjective qualities, such as passion (R^2 = 0.07, RMSE = 0.51, α = 0.7, ℓ_1 ratio = 0.01) and skill (R^2 = 0.06, RMSE = 0.53, α = 1.0, ℓ_1 ratio = 0.01), showed lower predictive power, suggesting that these listener judgments were less directly tied to extracted acoustic features. Similarly, ratings for emotional effectiveness (R^2 = 0.047, RMSE = 0.49, α = 0.1, ℓ_1 ratio = 0.01), sincerity (R^2 = 0.041, RMSE = 0.46, α = 0.5, ℓ_1 ratio = 0.01), and power (R^2 = 0.037, RMSE = 0.54, α = 0.7, ℓ_1 ratio = 0.01) were only modestly predictable from acoustic features (see Table 1).

We evaluated the Elastic Net linear regression without clips with added noise in the test set (see Table 1). When comparing the results of the complete test set (including both noisy and clean data) to the clean-only test set, almost all rating questions showed a decrease in R^2 scores after removing the noisy clips, with background noise dropping significantly from 0.33 to 0.16, and recording quality dropping from 0.20 to -0.01.

To quantify the effect of excerpt type (karaoke vs. audiobook) on the average ratings provided for each question, we ran a series of linear regression models to determine the relationship between the variable of interest and excerpt type (i.e. skill $\sim$ ExcerptType). For

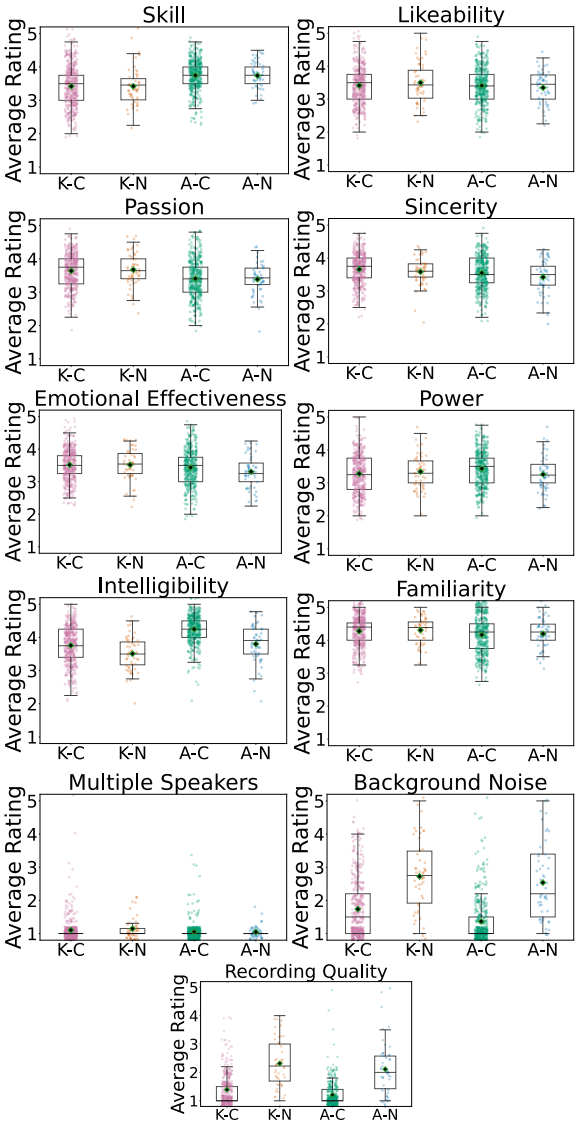

Fig. 1: Distribution of mean ratings for the 11 questions, split up by Excerpt Type (karaoke vs. audiobook) and Noise Type (clean vs. noisy). Each vertical box represents the 25th percentile, median, and 75th percentile, with whiskers extending 1.5x the interquartile range. The mean is represented by a diamond.

skill, excerpt type had a significant negative effect (β = -0.324, t = -9.459, p < 0.001), indicating that karaoke excerpts were rated lower in skill. Likability showed no significant difference between excerpt types (β = 0.019, t = 0.539, p = 0.59). In contrast, passion was

Table 1: Comparison of Elastic Net linear regression results with 5-fold cross-validation, including R² and RMSE for each question. Results with the full test set are in the left column, and results and with only clean clips in the test set (degraded clips omitted) are in the right. The model parameters were optimized via grid search for each listener-rated question.

Question	Full Test		Clean Test	
	R²	RMSE	R²	RMSE
Skill	0.06	0.53	0.06	0.53
Likability	-0.01	0.53	-0.01	0.53
Passion	0.07	0.51	0.07	0.51
Sincerity	0.04	0.46	0.03	0.46
Emotional Effect.	0.05	0.49	0.04	0.50
Power	0.04	0.54	0.03	0.54
Intelligibility	0.20	0.50	0.19	0.49
Familiarity	-0.02	0.52	-0.02	0.53
Multiple Speakers	-0.01	0.31	-0.02	0.31
Background Noise	0.33	0.75	0.16	0.74
Recording Quality	0.20	0.56	-0.01	0.51

rated significantly higher for karaoke excerpts (β = 0.237, t = 7.072, p < 0.001). Sincerity (β = 0.115, t = 3.793, p < 0.001), emotional effectiveness (β = 0.094, t = 2.841, p = 0.005), and familiarity (β = 0.109, t = 3.276, p = 0.001) were also rated higher for karaoke excerpts. However, power (β = -0.129, t = -3.619, p < 0.001) and intelligibility (β = -0.471, t = -14.21, p < 0.001) were rated significantly lower for karaoke performances. These results are expected and validate the reliability of the listener ratings we release in this work.

Background noise ratings were significantly higher for karaoke excerpts than audiobook excerpts (β = 0.352, t = 6.015, p < 0.001). Recording quality was also higher for karaoke excerpts (β = 0.182, t = 4.472, p < 0.001), indicating "the recording had unintelligible words due to recording quality," and reinforcing the perception that the karaoke recordings had poorer audio quality. "Multiple speakers" was also higher for karaoke (β = 0.057, t = 2.881, p = 0.004).

The regression models examining the impact of noise type (clean vs. noisy) on various perceptual ratings showed minimal effects across most attributes. Noise had no significant effect on skill (β = -0.003, t = -0.051, p = .959), likability (β = 0.011, t = 0.187, p = .852), passion (β = 0.008, t = 0.136, p = .892), emotional

effectiveness (β = -0.065, t = -1.209, p = .227), power (β = -0.057, t = -0.986, p = .324), familiarity (β = 0.030, t = 0.560, p = .576), and multiple speakers (β = 0.019, t = 0.601, p = .548), suggesting that those elements remain unchanged in the presence of noise, at least the two types of noise in the levels added in this work. For sincerity, noise significantly reduced ratings (β = -0.102, t = -2.059, p = .040), indicating that voices in noisy conditions are perceived as less sincere. For intelligibility, noise significantly lowered ratings (β = -0.349, t = -6.000, p < .001), indicating that speech in noisy conditions is perceived as less clear. Finally, for background noise and recording quality, noise significantly increased ratings (that is, more perceived noise) (β = 1.079, t = 11.98, p < .001) and (β = 0.913, t = 15.26, p < .001), respectively, confirming that listeners perceived more background noise and lower intelligibility due to recording quality in noisy conditions. Our linear mixed-effects model examined

Table 2: Average standard deviations for each of the 11 questions.

Question	St. Dev.
Skill	0.97
Likability	1.05
Passion	0.92
Sincerity	0.85
Emot. Effectiveness	0.93
Power	0.97
Intelligibility	0.75
Familiarity	1.06
Multiple Speakers	0.14
Background Noise	0.78
Recording Quality	0.58

the relationship between the more subjective stimulus-participant questions and the more objective stimulus questions (intelligibility, background noise, recording quality, and multiple speakers), and included random intercepts for filename and participant ID code to account for variability across recordings and participants. For skill, the conditional R^2 value, which accounts for both fixed and random effects, was 0.23, indicating that 23% of the variance in skill ratings was explained by the full model. The marginal R^2, representing the variance explained by fixed effects alone, was 0.05.

For likability, the conditional R^2 value was 0.21, while the marginal R^2 was 0.02. For passion, the conditional R^2 was 0.20, and the marginal R^2 was 0.01. For sincerity, the model explained 20% of the variance (conditional R^2 = 0.20), with fixed effects contributing minimally (marginal R^2 = 0.02). For emotional effectiveness, the model explained 21% of the variance (conditional R^2 = 0.21), with fixed effects contributing modestly (marginal R^2 = 0.02). For power, the model explained 24% of the variance overall (conditional R^2 = 0.24), with fixed effects contributing about 4% (marginal R^2 = 0.038). Finally, for familiarity, the model explained 57% of the variance overall (conditional R^2 = 0.57), but the fixed effects contributed very little (marginal R^2 = 0.0025), suggesting that most of the variance is due to differences between respondents rather than the predictors tested.

5 Discussion

As may be expected, the results of our Elastic Net linear regression indicate that aspects of intelligibility and recording quality were more predictable from audio than subjective ratings of vocal recordings (see Table 1). Three of our four stimulus questions (intelligibility, background noise, recording quality, and multiple speakers) had the highest R^2 scores. The exception is "multiple speakers," where there was a floor effect. Likability and familiarity, arguably the most subjective questions, had the lowest R^2 scores, as well as the highest average standard deviations. As expected, more subjective questions were harder to predict with a model using acoustic features.

The results of our linear mixed-effects model, which explored how objective stimulus responses predict subjective participant-stimulus responses, can be compared with those from our Elastic Net linear regression, where acoustic features predicted subjective participant-stimulus responses (see Table 1). The linear mixed-effects models explained between 20% and 24% of the variance in subjective ratings (conditional R^2 .20-.24), but the fixed factors explain only 2% to 5% of the variance (marginal R^2 .02-.05) which is in line with the results from the Elastic Net analysis. Both analyses show that it is difficult to predict the subjective participant-stimulus ratings from the selected objective acoustic features and more objective stimulus ratings. Notably, familiarity differed from the other questions, as it was primarily explained by individual differences, with minimal contribution from the fixed factors (marginal R^2 = 0.0025) and a high overall model fit (conditional R^2 =

0.57). This is expected, and suggests that the variability in familiarity ratings is driven more by participant-specific factors rather than the audio files' characteristics, and further validates the methodology used.

Listeners clearly picked up on the degradations we added to some of the audio files, as seen in the higher (more noisy) average ratings for background noise and recording quality, for both excerpt types. Intelligibility also decreased in noisy files for both excerpt types (see Figure 1), though less dramatically than for background noise and recording quality. The success of the linear regression for background noise and recording quality decreases greatly when the test set contained only clean data (with the added-noise excerpts omitted). The drop is likely due to the difference in listener ratings between clean and noisy recordings. Also, the ratings for the clean karaoke and audiobook files are quite low, and there is a floor effect, as illustrated in Figure 1.

For Excerpt Type, we found that karaoke excerpts were rated lower in skill but higher in passion, sincerity, emotional effectiveness, and familiarity compared to audiobook excerpts. In contrast, karaoke performances were rated significantly lower in power and intelligibility. These findings suggest that listeners are controlling for excerpt type in their likability ratings, find singing performances to be more emotionally engaging, but are more harsh when rating singer skill and power. As may be expected, participants found speech in karaoke harder to understand than in narrations. The recording quality was likely lower in the Smule dataset than in the Librispeech dataset, based on the higher ratings of background noise and recording quality for both excerpt types (Figure 1).

This study provides a new dataset and a baseline predictive model. Future work can expand acoustic features to incorporate into the model, and explore alternative models to improve performance. It is also important to acknowledge some limitations of the dataset: the listener ratings primarily come from individuals with a Western cultural background, and the participant pool largely consists of young, college students. These factors may shape perceptual judgments and limit the generalizability of our findings across diverse populations. Additionally, perceptions of traits like sincerity, passion, and likability may vary significantly across cultures, and our current design does not account for potential cultural biases in how these qualities are interpreted. Future work can include more demographically and culturally diverse listeners.

References

[1] Noufi, C., May, L., and Berger, J., "The Role of Vocal Persona in Natural and Synthesized Speech," in *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1–4, 2023.

[2] Walsh, B., Hamilton, M., Newby, G., Wang, X., Ruan, S., Zhao, S., He, L., Zhang, S., Dettinger, E., Freeman, W. T., and Weimer, M., "Large-Scale Automatic Audiobook Creation," in *Interspeech*, ISCA, Dublin, Ireland, 2023.

[3] Speechify Inc., "Speechify: AI-Powered Text-to-Speech App," `http://speechify.com`, 2024.

[4] Scherer, K., "The expression of emotion in the singing voice: Acoustic patterns in vocal performance," *J. Acoust. Soc. Am.*, 142(1805), 2017.

[5] Sundberg, J., Salomão, G. L., and Scherer, K. R., "Emotional expressivity in singing. Assessing physiological and acoustic indicators of two opera singers' voice characteristics," *Journal of the Acoustical Society of America*, 2024.

[6] Komar, S., "Listened To Any Good Books Lately? The Prosodic Analysis of Audio Book Narration," *ELOPE: English Language Overseas Perspectives and Enquiries*, 3, p. 85–98, 2006.

[7] Ji, D., Liu, B., Xu, J., and Gong, J., "Why Do We Listen to Audiobooks? The Role of Narrator Performance, BGM, Telepresence, and Emotional Connectedness," *Sage Open*, 14(2), 2024.

[8] De Man, B., Boerum, M., Leonard, B., and et al., "Perceptual evaluation of music mixing practices," *AES Journal*, 1, pp. 129–136, 2015.

[9] Man, B. D. and Reiss, J. D., "The Mix Evaluation Dataset," in *Digital Audio Effects (DAFx)*, 2017.

[10] Wilson, A. and Fazenda, B., "Perception of Audio Quality in Productions of Popular Music," *AES Journal*, 64, pp. 23–34, 2016.

[11] Palanisamy, B. and V, R., "The listening renaissance: a theoretical exploration of audio-based digital narratives in literature," *Humanities and Social Sciences Communications*, 12, p. 96, 2025.

[12] Watts, C., Barnes-Burroughs, K., Estis, J., and Blanton, D., "The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers," *Journal of Voice*, 20(1), pp. 82–88, 2006.

[13] Böhm, J., Eyben, F., Schmitt, M., and et al., "Seeking the SuperStar: Automatic assessment of perceived singing quality," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1560–1569, 2017.

[14] Nakano, T., Goto, M., and Hiraga, Y., "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. INTERSPEECH*, pp. 1706–1710, ISCA, 2006.

[15] Bozkurt, B., Baysal, O., and Yüret, D., "A dataset and baseline system for singing voice assessment," in *International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2017.

[16] Panteli, M., Bittner, R., Bello, J. P., and Dixon, S., "Towards the Characterization of Singing Styles in World Music," *IEEE ICASSP*, 2017.

[17] International Telecommunication Union, "Subjective evaluation of speech quality with a crowdsourcing approach," ITU-T Recommendation P.808, Geneva, 2018.

[18] Akrami, H. and Gamper, H., "Speech MOS multi-task learning and rater bias correction," in *ICASSP*, pp. 1–5, IEEE, 2023.

[19] Gamper, H., Reddy, C., Cutler, R., Tashev, I., and Gehrke, J., "Intrusive and Non-Intrusive Perceptual Speech Quality Assessment Using a Convolutional Neural Network," in *WASPAA*, 2019.

[20] Ibrahim, K. M., Grunberg, D., Agres, K., Gupta, C., and Wang, Y., "Intelligibility of Sung Lyrics: A Pilot Study," in *ISMIR Conference*, 2017.

[21] Smule, Inc., "DAMP-VPB: Digital Archive of Mobile Performances - Smule Vocal Performances," 2017, doi:10.5281/zenodo.2616690.

[22] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S., "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE ICASSP*, South Brisbane, QLD, Australia, 2015.

[23] LibriVox, "LibriVox | Free Public Domain Audiobooks," 2005, accessed: 2025.

[24] Sobot, P., "Pedalboard," 2021, doi:10.5281/zenodo.7817838.

[25] SoX Development Team, "SoX - Sound eXchange," 1991, accessed: 2025.

[26] FFmpeg Developers, "FFmpeg," 2000, accessed: 2025.

[27] Woods, K. J. P., Siegel, M. H., Traer, J., and et al., "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, 79, pp. 2064–2072, 2017.

[28] Orpella, J., Bowling, D. L., Tomaino, C., and Ripollés, P., "Effects of music advertised to support focus on mood and processing speed," *PLOS ONE*, 20(2), p. e0316047, 2025.

[29] Mas-Herrero, E., Marco-Pallares, J., Lorenzo-Seva, U., and et al., "Individual Differences in Music Reward Experiences," *Music Perception*, 31(2), pp. 118–138, 2013.

[30] Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L., "The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population," *PLOS ONE*, 2014.

[31] Pedregosa, F., Varoquaux, G., Gramfort, A., and et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2011.

[32] Mauch, M. and Dixon, S., "Pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE ICASSP*, 2014.

[33] McFee, B., McVicar, M., Faronbi, D., and et al., "librosa/librosa: 0.10.2.post1," 2024.

[34] Georgieva, E., Ripollés, P., and McFee, B., "The Changing Sound of Music: An Exploratory Corpus Study of Vocal Trends Over Time," in *ISMIR Conference*, 2024.

[35] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357–366, 1980.