Audio-Based Understanding of Audiobook Narration Appeal

Shahar Elisha [1,2], *Mariano Beguerisse-Díaz* [1], *Emmanouil Benetos* [2]

[1]Spotify, [2]Queen Mary University of London

shahar@spotify.com

arXiv:2607.02473v1 [cs.CL] 2 Jul 2026

Abstract

Narration is central to the audiobook listening experience, shaping how listeners engage with and understand the content. This work explores how narration qualities shape an audiobook's appeal, noting that their effects can vary by genre, title, and audience. We extract vocal and acoustic features (e.g., tone, pace, loudness) from LibriVox using pre-trained audio models and analyse their relationship with consumption data (specifically, view-rate) and their interplay with genre and title. Despite limited consumption data, we find that acoustic information alone has a robust association with appeal, even after accounting for title effects. We further validate these findings using more nuanced proprietary engagement metrics. To our knowledge, this is the first systematic computational study linking narration qualities, genre, title, and audiobook consumption, highlighting the potential of data-driven insights to improve audiobook personalisation and narrator casting.

Index Terms: audiobooks, narration style, speech paralinguistics, audio processing

1. Introduction

Narration style and acoustic presentation are important components of audiobooks; they have the power to either elevate or undermine a listener's experience, understanding, and engagement with the story [1]. While the narration alone may not be the determining factor in audiobook selection amongst users, it has a significant impact on whether a user completes listening to the book in full [2]. A narrator offers an interpretation of the story by varying their voice across pitch, volume, timbre, and tempo, portraying characteristics such as gender, age, accent, emotional states and attitudes [3]. This can range from giving each character a distinct voice, to maintaining a consistent tone throughout. Other acoustic qualities such as the recording quality and the use of music and sound effects, can give variety and texture to a narration. Different genres may call for different narration styles; for example, an immersive and expressive style may be better suited to fiction than to non-fiction or academic texts.

Audiobooks are widely consumed, with platforms such as Spotify, Audible, LibriVox, and Libby offering extensive libraries. As catalogues expand, these platforms face unique challenges: distinguishing between multiple recordings of the same title, understanding the narrations that appeal to different users or suit different genres, and informing casting decisions. User-item interaction data is also sparser than in domains such as music - for example, audiobooks are rarely replayed, and users consume far fewer than songs. Improving classification, search, recommendation and personalisation models requires a sophisticated understanding of audiobooks, not just on its content (e.g., text, topic, genre, author), but also its narration style, and their interplay. Even small improvements in modelling appeal can yield substantial gains in large-scale recommendation systems, where marginal ranking improvements can translate into increased user engagement and retention [4, 5].

Here, we investigate the relationship between an audiobook's narration and acoustic features, genres, title, and consumption to understand how narration influences appeal. We analyse vocal and acoustic properties of narrations on a large, publicly available dataset of audiobooks, exploring nuances related to genres, differences within groups of the same texts, and consumption statistics. To our knowledge, this is the first computational study linking acoustic narration features to large-scale, real-world consumption data across genres and multiple recordings of the same title (see Sec. 2 for comparison to prior work). Specifically, our contributions include:

- **Statistical analysis of audiobook consumption** to assess the influence of interpretable acoustic features on a large real-world dataset.
- **Genre-specific modelling** to reveal how and which acoustic features influence appeal across genres.
- **Intra-title comparison framework** for different narrations of the same book (i.e. book-groups) to control for content.
- **Classification and ranking models** to evaluate the predictive power of acoustic features for appeal.

We find a statistically significant relationship between the acoustic features of a narration and an audiobook's appeal. This result is striking for two reasons: the narration style is independent of the book content, and the dataset has substantial limitations, which include variable recording quality of volunteer narrations and coarse, limited consumption data. To assess the robustness of our findings, we conduct follow-up experiments replacing the LibriVox consumption proxy with more granular engagement metrics from a proprietary Spotify dataset. These results are an important step towards understanding audiobook narrations and their appeal, offering insights to improve narrator casting and recommendation systems by matching the right voice to the right book, and the right book to the right listener.

2. Related Works

2.1. Computational Paralinguistics and Voice Perception

Human voices carry paralinguistic information from which a listener perceives qualities about the speaker's identity and intention [6]. Researchers have developed computational models for paralinguistic tasks such as perceived gender and age classification, health predictors, emotions, social standings, vocal qualities, speaking styles, and even vocal attractiveness (e.g., [7–14]). The classic approach is to train machine learning models on predicted acoustic features from the audio (e.g., pitch, loudness, or spectral) [15]. Recent works employ sophis-

ticated techniques, such as transformer models, for improved performance at the cost of interpretability [8, 10, 16].

2.2. Narration Styles

While plenty of audiobook research uses Text to Speech (TTS) and voice conversion (VC) models to deliver expressive speech fit for storytelling (e.g., [17, 18]), studies indicate that listeners prefer human narrations over synthetic ones [19], although some are interested in the ability to customise synthetic narrations [20]. User surveys and qualitative frameworks establish the importance of narrator style to the audiobook experience [1–3], but stop short of identifying which acoustic features drive appeal. We build on computational studies exploring acoustic and vocal features of narration at different levels: utterance, dialogue, and narrator. Prosody and voice quality are salient discriminators between storytelling discourse modes (e.g., narrative, descriptive, and dialogue) [21], a finding that holds cross-lingually [22], though both approaches rely on small datasets. At the dialogue level, female characters are narrated with higher pitch and lower volume relative to males [23]. Other studies cluster narrators by speaking style using glottal source parameters [24] or Convolutional Neural Network (CNN) embeddings [25]. Most relevant to our work, Lange et al. [26] link acoustic features to user-reported absorption and liking, finding higher articulation rates predict both, though feature correlations leave the specific drivers unclear. Critically, none of these studies examine narration appeal at scale across genres and alternative narrations of the same title.

3. Experimental Setup

3.1. LibriVox catalogue

LibriVox [27] is a catalogue of public domain audiobooks, read and recorded by volunteers, with multiple titles and genres. The metadata (e.g., title, author, narrator, genres, text-source) and audio files are available to download freely. The Internet Archive keeps track of the number of page views, favourites, and reviews of each recording [28]. For simplicity, we focus on single-narrator, English audiobooks. Our data contains 8,854 audiobooks read by 1,206 different narrators, across 65 genres (e.g., History, Romance, Comedy). We group audiobooks into *book-groups* if they share the same link to the original text source (i.e., different recordings of the same text). The dataset can be found in the supplementary codebase[1]. We segment recordings (typically a chapter) into 30s intervals; we sample up to 20 segments per recording (the first, last, and 18 random segments in the middle), giving a maximum of 10 minutes of audio per recording (shorter recordings are included in full). This strategy balances representativeness with computational efficiency. In the absence of publicly available information about listener satisfaction, downloads, hours listened, or completion rates, we use the number of views as a proxy for appeal; the number of favourites and reviews are too small to be useful in this analysis. We divide the views by the number of days since publication on LibriVox (i.e. *view-rate*) to account for time-on-platform bias. Note that view-rate is limited as an appeal metric (e.g., no distinction between completed listens, brief listens, or multiple listens by a user), which introduces additional noise to our analysis. Although an initial page view is recorded before any listening occurs, we assume that the view count reflects appeal through repeated visits, as audiobooks are seldom finished in a single sitting. This proxy is therefore biased against shorter recordings, which may require fewer visits to complete and thus accumulate lower view-rates independent of appeal. Nevertheless, it is the only large-scale proxy available, and we use it to ensure transparency and reproducibility. We further extend our analysis using more nuanced engagement metrics from proprietary Spotify data to evaluate the robustness of our findings under a less coarse proxy (Sec. 4.3).

[1] https://github.com/spotify-research/audiobook-narrations-interspeech

Table 1: *Acoustic and vocal feature classes, see SM:Table 1 for all 129 summary statistics.*

Descriptors	Features
Frequency	F0, jitter, F1–3, F1–3 bandwidth
Energy / amplitude	Loudness, shimmer, harmonics-to-noise ratio (HNR), equivalent sound level, loudness peaks per second
Spectral	Alpha ratio, Hammarberg index, spectral slope, F1–3 relative energy, harmonic differences (H1–H2, H1–H3), MFCCs 1–4, spectral flux
Tempo	Word rate, syllable rate, duration
Audio events	Music, sound effects, recording quality, nonverbal vocalisations, speech (e.g. conversation, narration/monologue, screaming, whispering)

3.2. Feature extraction and aggregation

We extract acoustic and prosodic features for every 30s sample using well-known, pre-trained audio models with default parameters (see Table 1), and we concatenate the features for all segments and chapters per audiobook along the time axis and calculate summary statistics:

- ***eGeMAPSv02*** [29]: 25 low-level acoustic descriptors of *frequency*, *energy*, *spectral*, and *tempo* using the openSMILE tool [30]. We replicate openSMILE's aggregation logic to output 84 functional features (see codebase[1]).
- ***YAMNet*** [31] scores: 521 *audio events* classification scores, such as human vocals, musical instruments, animal and environmental sounds. Using the AudioSet ontology [32], we group audio classes into *sound effects*, *music*, *recording quality*, *nonverbal vocalisations*, and *speech*. We keep the 13 individual *speech* subclasses, and aggregate the remaining 4 grouped audio classes by taking the maximum subclass activation, giving a total 17 audio events. We calculate the mean and standard deviation of the scores across time for a final audio events set of 34 features.
- ***whisper-tiny*** [33] transcripts with timestamps. We extract syllables [34] and compute word/syllable counts and rates. We compute the mean, std, min and max of the word and syllable rates, and the total duration, word and syllable counts, giving 11 features.

Finally, we concatenate the features into a 129-dimensional audiobook vector (see Table 1 in Supplementary Materials (SM)[1]).

3.3. Statistical Modelling

We investigate how acoustic and vocal features relate to appeal (view-rate), including genre-specific and intra-title nuances. We report model fit statistics and statistically significant coefficients. We apply the Benjamini-Hochberg (BH) method to correct for false correlations ($p < 0.05$) [35].

Feature pre-processing: To address multicollinearity in our feature set, we perform Variance Inflation Factor (VIF) pruning [36], by iteratively removing the feature with the highest VIF score (i.e. highly correlated with other features) until all features have a factor below 5. We standardise the resulting 70 features prior to model fitting.

Global modelling of consumption: We fit a Generalised Linear Model (GLM) [36] using the pruned and standardised

acoustic features to model audiobook view-rate. The GLM allows us to assess the significance of individual features and identify which acoustic properties are most predictive of audiobook popularity. To account for the long tail in the view-rate distribution, we log-transform the view-rate and assume a Gaussian error distribution.

Genre-specific: To investigate genre-specific effects, we train a GLM per genre (65 genres). Within each genre, we re-standardise the feature set to control for genre-specific distributions. We set the initial parameters to those learned by the global GLM, and drop near-constant features ($std \leq 1e{-}12$). We then compare BH-corrected significant coefficients across genres to examine whether acoustic correlates of appeal vary by genre.

Intra-title: To account for title-specific effects on appeal, we fit a Linear Mixed-Effects (LME) model [36]. We model a random intercept per *book-group*, while the acoustic features serve as the fixed effects. This allows us to disentangle title-driven variability from acoustic influences on view-rate. We transform our view-rate into log scale to align with the GLMs. We compare the model (retrained using maximum likelihood) to the global GLM using Akaike Information Criterion (AIC) [37], a relative measure that helps identify the better-fitting model.

3.4. Predictive Modelling

We also want to understand the predictive power of acoustic features on classifying and ranking appeal. Because the view-rate metric provides limited nuance (see Sec. 3.1), we simplify the problem space by binning view-rates into quartiles, thus converting the task from regression to classification. We use all 129 features as input and standardise during training. We run bootstrapping to estimate the 95% confidence intervals (CIs) and a Kolmogorov-Smirnov test to confirm that the model predictions are better than random. We run 1000 random realisations, and 1000 bootstrapping iterations.

Classification: Given the small and largely interpretable feature set, we train four shallow classifiers: Logistic Regression (LR), Support Vector Machine (SVM), XGBoost, and Multi-Layer Perceptron (MLP). For comparison, we train classification models using genres as input (multi-hot encoding), and a combination of acoustic features and genres. We train and evaluate models with 5-fold cross-validation, grouped by narrator and balanced by view-rate quartile. We evaluate overall accuracy and assess performance across view-rate quartiles.

Ranking: To capture appeal between different narrations of the same title (i.e. *book-groups*), we train ranking models using acoustic features to predict the within-title rankings where an audiobook with a higher view-rate ranks higher within its group. We filter the data to books in groups of size ≥ 2, resulting in 305 groups and 736 audiobooks. Most groups are of size 2, and the maximum group size is 13. We train and evaluate models with 5-fold cross-validation, grouped by book-groups. We train three boosted tree models: XGBRanker [38] with a rank:ndcg objective (based on LambdaMART [39]), with a rank:pairwise objective (uses RankNet loss [40]), and LGBMRanker [41] (based on LambdaRank) [42]. While the default training objectives use Normalised Discounted Cumulative Gain (NDCG), we report performance using average Kendall's rank correlation coefficients (Kendall's τ) between predicted and true book-group rankings, as it directly measures order correlation and avoids inflated NDCG scores due to small, fully relevant groups (the average NDCG on random rankings for this dataset is 0.92). Although differentiable adaptations of Kendall's τ exist [43], we retain the standard training objectives as a starting point since they are well-established, and focus on Kendall's τ purely for evaluation. We compare performance to two pointwise models using the LR model implemented for classification: 1) trained and evaluated on the 736 books used for the ranking models, and 2) trained and evaluated on the full dataset of 8,854 books, with different cross-validation splits.

4. Results

4.1. Statistical Modelling Results

Global modelling of consumption: The GLM attains a pseudo-R^2 of 0.09, indicating that narration-related properties explain a measurable portion of variation in appeal despite the coarse proxy (see Sec. 3.1) and omission of title, genre, and promotional factors. In a large and noisy real-world dataset, explaining nearly 10% of the variability using audio features alone suggests that narration characteristics have a consistent and non-trivial association with listener engagement. We find 31 acoustic features show a significant effect on view-rate (see SM:Fig. 1), though no single feature dominates and effect sizes are mostly small ($|\beta| \leq 0.13$), suggesting that appeal is shaped by a combination of acoustic properties rather than any single dominant factor. For example, variation in vocal shimmer (perceived as a change in vocal qualities, such as breathiness [44]) has a positive effect on appeal, whereas higher spectral flux (correlated to speaker gender [12]) is less favourable. Modelling narrator characteristics (e.g., narrator gender) instead of the acoustic correlates could clarify the driving effects and improve interpretability. We find that the duration sum correlates with higher view-rates; this is likely explained by longer audiobooks requiring more page visits to complete rather than a preference for longer recordings. Articulation rate (syllables_per_min_std, which correlates with syllables_per_min_mean: $\rho = 0.66$) has a positive influence on appeal; this aligns with findings from a user study [26], showing that higher articulation rates increase listener absorption and liking. Further analysis of feature correlations is required to refine interpretation.

Genre-specific: The influence of acoustic features on appeal varies across genres, both in magnitude and direction. For example, variation in vocal shimmer (perceived as breathiness [44]), which shows a positive effect globally ($\beta = 0.09$), is considerably stronger within Romance ($\beta = 0.31$), suggesting that this vocal characteristic is particularly relevant to appeal in this genre. In contrast, vocal shimmer shows no significant effect on appeal within History. Instead, variation in the Hammarberg index, a spectral measure of vocal quality associated with vocal effort and emotion [45], has the strongest effect on appeal within History ($\beta = -0.35$), compared to a weak global effect ($\beta = -0.04$) and no significant effect on Romance. Further analysis of outliers is required before drawing firm conclusions about their influence on appeal (see SM:Fig. 1).

Intra-title: Variation in appeal across narrations of the same title (0.52) is nearly as large as variation across different titles (0.54), indicating that narration contributes on a scale comparable to content-level differences. This underscores the importance of vocal delivery in influencing listener engagement. This holds even if we limit the data to audiobooks with more than one version. Furthermore, the mixed-effects model substantially improves model fit relative to the global GLM ($AIC_{GLM} - AIC_{LME} = 210$). This highlights the importance of modelling title biases, which likely influence selection behaviour and page-view rates. Despite the better fit, most features that were statistically significant from the global GLM, re-

Table 2: *Accuracy [95% CIs] of quartile classification models.*

	Genres	Audio Features	Combined
LR	0.31 [0.31, 0.32]	**0.32 [0.32, 0.33]**	**0.35 [0.34, 0.35]**
MLP	0.31 [0.31, 0.32]	0.30 [0.29, 0.30]	0.32 [0.31, 0.33]
SVM	**0.32 [0.31, 0.33]**	0.31 [0.30, 0.32]	0.33 [0.32, 0.34]
XGBoost	**0.32 [0.31, 0.33]**	0.29 [0.28, 0.30]	0.31 [0.31, 0.32]
Random	0.25 [0.24, 0.26]	0.25 [0.24, 0.26]	0.25 [0.24, 0.26]

Table 3: *Kendall's τ [95% CIs] for ranking models. VR = view-rate, RR = return-rate.*

	Full (VR)	Subset (VR)	Subset (RR)
NDCG	0.08 [-0.03, 0.17]	-0.02 [-0.18, 0.13]	**0.26** [0.11, 0.41]
Pair	0.10 [0.00, 0.20]	0.01 [-0.14, 0.16]	**0.26** [0.11, 0.42]
Lambda	0.13 [0.03, 0.23]	0.02 [-0.13, 0.17]	**0.28** [0.13, 0.42]
LR (group)	0.09 [0.02, 0.15]	0.02 [-0.08, 0.12]	0.08 [-0.01, 0.18]
LR (full)	0.10 [0.03, 0.17]	0.04 [-0.04, 0.13]	0.07 [-0.02, 0.17]
Random	0.00 [-0.15, 0.14]	–	–

main significant after modelling effects driven by title (SM:Fig. 1). While the magnitude of the coefficients has a small amount of variation, the directions remain consistent.

4.2. Predictive Modelling Results

Classification: All models performed above the random baseline of 0.25 (see Table 2). Predicting appeal using acoustic features alone can improve accuracy by up to 0.07, reinforcing the findings from our GLM experiments: in spite of the challenging appeal data, acoustic features have a robust predictive power. Genre-only and audio-only models achieved similar performance, while combining them boosts performance up to 0.35, indicating that acoustic features and genres capture related but not identical information. Class-specific analyses further reveal that acoustic features capture nuances in the middle quartiles, whereas genre features contributed more strongly to predictions at the extremes (see SM:Fig. 2). Simpler models such as LR and SVM generally outperform more complex approaches (MLP, XGBoost).

Ranking: All ranking models show a statistically significant improvement over a random baseline (Table 3), consistent with the other experiments, confirming an influence of acoustic features on appeal. We find that the pointwise models (LR) tend to perform similarly to other ranking models, even when trained on the full dataset (which has 10x the amount of samples to learn from). The ranking models benefit from modelling relative appeal, as it removes any title-specific biases in preferences. We note that the performance of the ranking models are sensitive to data shuffling across the folds; further analysis and larger, more granular data on engagement and appeal will help understand which ranking model is more appropriate for the task.

4.3. Analysis Using Proprietary Engagement Metrics

While LibriVox view-rate is publicly available, which ensures that the results in this work are reproducible by all, it is extremely coarse and has many limitations, as discussed above. Thus, we recompute our analysis on a subset of LibriVox audiobooks (Sec. 3.1) that are also hosted on Spotify, for which we have access to granular consumption data. We replace view-rate with the proportion of distinct users who return to an audiobook within 14 days (i.e. *return-rate*). This metric indicates whether a listener enjoyed an audiobook enough to return and continue listening. Return-rate is also biased against shorter recordings, which may be completed in a single session; developing more nuanced engagement metrics remains an important direction for future work. This data is available for 3,428 audiobooks, a subset of the original dataset (8,854 items).

Global GLM: Refitting the GLM on this subset using view-rate increases pseudo-R^2 from 0.09 to 0.13, reflecting differences in sample composition. Replacing view-rate with the number of returning users within 14 days and setting total users as an exposure offset further improves performance (pseudo-$R^2 = 0.16$) and substantially improves model fit ($\Delta AIC \approx 6000$). These results suggest that audio features are more strongly associated with engagement conditional on exposure than raw popularity.

Ranking: The LibriVox ranking experiment includes 736 audiobooks across 305 book groups (Sec. 3.4:Ranking); this subset has 327 audiobooks in 138 groups. In this reduced setting, using view-rate fails to learn a meaningful relationship with audio features (Kendall's $\tau \approx 0$). In contrast, using return-rate to define relative appeal yields a strong and consistent relationship ($\tau \approx 0.26$–0.28), indicating that return-rate captures a more stable and discriminative intra-title appeal (see Table 3).

5. Conclusion

We examined the relationship between audiobook narration, genres, title, and consumption, and consistently found that acoustic features of narration influence appeal. The robustness of these results, despite coarse consumption data and mixed recording quality, validates our hypothesis that narration styles influence appeal, and point the way to exciting further research.

Modelling relative appeal within book-groups emphasises the substantial contribution of narrations beyond content-level differences. We also find that the influence of different acoustic features varies across genres. Both findings highlight the need to account for title and genre information alongside audio features. While specific acoustic effects should be interpreted cautiously given the dataset limitations and feature correlations, several interesting interpretable patterns emerge. For example, greater variation in articulation rate is positively associated with appeal, while higher spectral flux (linked to vocal characteristics such as perceived gender) tends to be negatively associated. These trends show that measurable vocal and acoustic properties contribute to listener engagement beyond content-level factors. Exploratory analyses using more granular proprietary engagement data further support the conclusion that narration influences appeal, suggesting that this effect may be more pronounced using richer behavioural data. Together, these results provide converging evidence linking paralinguistic and acoustic features of narration to audiobook appeal and establish a foundation for future work using larger, more diverse datasets.

Future work with richer engagement data (e.g., completion behaviour, user journeys across narrations and titles) and broader catalogues including professional narrations will enable a deeper understanding of narration appeal. Testing expanded feature sets, such as perceived narrator characteristics (accent, age, gender), whether the author narrates their own work, and how vocal performance and characterisation suit the narrative tone and context, will support more holistic and interpretable models. Analysing appeal across demographic listener segments will allow us to uncover audience-specific patterns of narration preference, which we hypothesise also shape engagement. Such analyses will allow us to move beyond coarse correlates and towards nuanced models that can be employed to improve personalisation, promotions, casting and other downstream uses.

6. Acknowledgments

We thank R. Dall, R. Jones, D. Korkinof, A. Lima, A. McDowell, S. Reddy, B. Regan, A. Torrisi, L. Vongsathorn, J. Walker, H. Zhang, E. zu Erbach for their useful feedback.

7. Generative AI Use Disclosure

Generative AI tools were used to assist with language editing, formatting, and improving clarity of the manuscript. All experimental design, analysis, and results were conducted and verified by the authors.

8. References

[1] D. Ji, B. Liu, J. Xu, and J. Gong, "Why do we listen to audiobooks? the role of narrator performance, bgm, telepresence, and emotional connectedness," *Sage Open*, vol. 14, no. 2, 2024.

[2] M. Dakic, "Preferences and attitudes of audiobook users in Sweden : Surveying Swedish audiobook groups on Facebook," Master's thesis, University of Borås, Faculty of Librarianship, Information, Education and IT, 2019.

[3] L. Kosch, A. Schwabe, H. Boomgaarden, and G. Stocker, "Experiencing literary audiobooks: A framework for theoretical and empirical investigations of the auditory reception of literature," *Journal of Literary Theory*, vol. 18, no. 1, pp. 67–88, 2024. [Online]. Available: https://doi.org/10.1515/jlt-2024-2005

[4] G. Fazelnia, S. Gupta, C. Keum, M. Koh, T. Heath, G. Carrasco Hernández, S. Xie, N. Singh, I. Anderson, M. Hristakeva, P. Pehrson Skidén, and M. Lalmas, "Generalized user representations for large-scale recommendations and downstream tasks," in *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, ser. RecSys '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 962–966. [Online]. Available: https://doi.org/10.1145/3705328.3748132

[5] C. A. Gomez-Uribe and N. Hunt, "The Netflix recommender system: Algorithms, business value, and innovation," *ACM Trans. Manage. Inf. Syst.*, vol. 6, no. 4, Dec. 2016. [Online]. Available: https://doi.org/10.1145/2843948

[6] B. J. Kröger, "Neurocomputational models of voice and speech perception," in *The Oxford Handbook of Voice Perception*, S. Frühholz and P. Belin, Eds. Oxford University Press, 12 2018. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780198743187.013.34

[7] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, M. Chetouani, and M. Mortillaro, "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge," *Computer Speech & Language*, vol. 53, pp. 156–180, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230816303928

[8] E. Goron, L. Asai, E. Rut, and M. Dinov, "Improving domain generalization in speech emotion recognition with Whisper," in *ICASSP 2024*, 2024, pp. 11 631–11 635.

[9] Y. Obuchi, *Multidimensional Mapping of Voice Attractiveness and Listener's Preference: Optimization and Estimation from Audio Signal*. Singapore: Springer Singapore, 2021, pp. 281–295. [Online]. Available: https://doi.org/10.1007/978-981-15-6627-1_15

[10] S. Elisha, A. McDowell, M. Beguerisse-Díaz, and E. Benetos, "Classification of spontaneous and scripted speech for multilingual audio," in *2024 SLT*, 2024, pp. 489–495.

[11] F. Jalali-najafabadi, C. Gadepalli, D. Jarchi, and B. M. Cheetham, "Acoustic analysis and digital signal processing for the assessment of voice quality," *Biomedical Signal Processing and Control*, vol. 70, p. 103018, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421006157

[12] G. Yasmin, S. Dutta, and A. Ghosal, "Discrimination of male and female voice using occurrence pattern of spectral flux," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, 2017, pp. 576–581.

[13] A. Kathan, S. Amiriparian, L. Christ, S. Eulitz, and B. W. Schuller, "Automatic speech-based charisma recognition and the impact of integrating auxiliary characteristics," in *2024 IEEE Conference on Telepresence*, 2024, pp. 148–153.

[14] S. S. Leal, S. Ntalampiras, and R. Sassi, "Speech-based depression assessment: A comprehensive survey," *IEEE Transactions on Affective Computing*, vol. 16, no. 3, pp. 1318–1333, 2025.

[15] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[16] A. Batliner, M. Neumann, F. Burkhardt, A. Baird, S. Meyer, N. T. Vu, and B. W. Schuller, "Ethical awareness in paralinguistics: A taxonomy of applications," *International Journal of Human–Computer Interaction*, vol. 39, no. 9, pp. 1904–1921, 2023. [Online]. Available: https://doi.org/10.1080/10447318.2022.2140385

[17] B. Manoj, J. Jiji, R. Dileep, and N. Manohar, "Emotionally enhanced audiobook reader with character voice differentiation," in *2025 International Conference on Computing Technologies (ICOCT)*, 2025, pp. 1–6.

[18] A. Sini, D. Lolive, N. Barbot, and P. Alain, "Investigating inter- and intra-speaker voice conversion using audiobooks," in *Proc. of the 13th LREC*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 7305–7313. [Online]. Available: https://aclanthology.org/2022.lrec-1.794/

[19] E. Rodero and I. Lucas, "Synthetic versus human voices in audiobooks: The human emotional intimacy effect," *New Media & Society*, vol. 25, no. 7, pp. 1746–1764, 2023. [Online]. Available: https://doi.org/10.1177/14614448211024142

[20] É. Székely, J. P. Cabral, M. Abou-Zleikha, P. Cahill, and J. Carson-Berndsen, "Evaluating expressive speech synthesis from audiobook corpora for conversational phrases," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 3335–3339. [Online]. Available: https://aclanthology.org/L12-1513/

[21] R. Montaño and F. Alías, "The role of prosody and voice quality in indirect storytelling speech: Annotation methodology and expressive categories," *Speech Communication*, vol. 85, pp. 8–18, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639315300108

[22] ——, "The role of prosody and voice quality in indirect storytelling speech: A cross-narrator perspective in four European languages," *Speech Communication*, vol. 88, pp. 1–16, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639315300418

[23] C. Pethe, B. Pham, F. D. Childress, Y. Yin, and S. Skiena, "Prosody analysis of audiobooks," in *2025 19th International Conference on Semantic Computing (ICSC)*, 2025, pp. 217–221.

[24] É. Székely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameters." *Proc. Interspeech 2011*, pp. 2409–2412, 2011.

[25] N. Embretsén, "Representing voices using convolutional neural network embeddings," Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2019.

[26] E. B. Lange, D. Thiele, and M. M. Kuijpers, "Narrative aesthetic absorption in audiobooks is predicted by blink rate and acoustic features." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 16, no. 1, pp. 110–124, 2022. [Online]. Available: https://doi.org/10.1037/aca0000321

[27] LibriVox, "LibriVox: Free public domain audiobooks," https://librivox.org, 2025.

[28] Internet Archive, "LibriVox audio collection," https://archive.org/details/librivoxaudio, 2025.

[29] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[31] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[32] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[33] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[34] K. Gorman, "syllables: A simple syllable counting package for Python," https://pypi.org/project/syllables/, 2025.

[35] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x

[36] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 6th ed. Wiley, 2021.

[37] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[38] XGBoost Developers, "XGBoost documentation," https://xgboost.readthedocs.io, 2025.

[39] C. J. C. Burges, R. Ragno, and Q. V. Le, "Learning to rank with nonsmooth cost functions," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, ser. NIPS'06. Cambridge, MA, USA: MIT Press, 2006, p. 193–200.

[40] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning*, ser. ICML '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 89–96. [Online]. Available: https://doi.org/10.1145/1102351.1102363

[41] LightGBM Developers, "LightGBM documentation," https://lightgbm.readthedocs.io, 2025.

[42] C. J. Burges, "From RankNet to LambdaRank to LambdaMART: An overview," Tech. Rep. MSR-TR-2010-82, June 2010. [Online]. Available: https://www.microsoft.com/en-us/research/publication/from-ranknet-to-lambdarank-to-lambdamart-an-overview/

[43] K. Zheng, H. Zhang, and W. Huang, "DiffKendall: a novel approach for few-shot learning with differentiable kendall's rank correlation," in *Proc. of the 37th NeurIPS*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[44] J. L. Sofranko and R. A. Prosek, "The effect of levels and types of experience on judgment of synthesized voice quality," *Journal of Voice*, vol. 28, no. 1, pp. 24–35, 2014. [Online]. Available: https://www.jvoice.org/article/S0892-1997(13)00103-3/abstract

[45] M. Ekberg, G. Stavrinos, J. Andin, S. Stenfelt, and Ö. Dahlström, "Acoustic features distinguishing emotions in Swedish speech," *Journal of Voice*, vol. 39, no. 6, pp. 1699.e11–1699.e20, 2025. [Online]. Available: https://doi.org/10.1016/j.jvoice.2023.03.010