How Pleasant is Your Voice? On Automatic Prediction of Voice Attractiveness

Anja Bulajić and Stavros Ntalampiras
Department of Computer Science
University of Milan, Milan, Italy
anja.bulajic@studenti.unimi.it

***Abstract*—This work describes an approach to automatically compute the attractiveness of a voice based only on speech audio input, while linking such perceptual criteria to the identified spectro-temporal regions. Voice attractiveness plays a crucial role in the effectiveness of human–computer interaction systems, including speech synthesis technologies and virtual assistants. To this end, both traditional machine learning and deep learning methods were employed. More specifically, a Support Vector Regression model and a Convolutional Neural Network combined with bidirectional Long Short-Term Memory (CNN–BiLSTM) layers, were created and compared. Importantly, we adopted a standardized speaker-independent experimental protocol along with a publicly-available corpus facilitating the reproduction of this study. Interestingly, the CNN–BiLSTM model provided lower prediction errors and a positive coefficient of determination R^2, outperforming the SVR model. The comparison between actual and predicted Mean Opinion Score values indicated that the largest differences were found at the extremities of the perceptual scale. Moreover, we produced Grad-CAM visualizations characterizing the model operation when processing speech coming from different genders and attractiveness levels. There, we observe that the network is focused on high-frequency regions that are related to brightness and harmonic stability. At the same time, non-attractive voices show more scattered activation patterns. These results contribute to the understanding of the relationship between subjective perceptual impressions and objective acoustic modelling mediated by explainable deep learning.**

I. Introduction

Voice attractiveness is one of the key factors influencing the success of human–computer interaction systems of the present day such as speech synthesis, virtual assistants, audiobook narration, and advertising applications [1]–[4]. A voice that sounds pleasant may increase user engagement, loyalty, and emotional connection, which makes vocal attractiveness prediction a relevant direction for affective computing and multimodal user experience design [5], [6]. Biologically, vocal attractiveness emerges from a combination of acoustic properties and perceptual mechanisms shaped by human cognition. It not only includes the measurable characteristics of the voice but also the subjective evaluations of the listeners [7], [8].

As shown by Suda et al. [9], attractiveness judgments depend not only on static spectral cues but also on dynamic prosodic and temporal variation. The question of what makes a voice "attractive" is becoming more and more relevant for the design of expressive and human-like speech synthesis with the fast development of AI-generated voices and text-to-speech technologies [3], [6]. Conventional acoustic models based on handcrafted features—such as Mel Frequency Cepstral Coefficients (MFCCs) and spectral centroid measures [10], [11] are likely to overlook complex nonlinear dependencies that exist between the spectral and temporal dimensions. On the other hand, deep learning algorithms are able to perform automatic hierarchical auditory feature extraction directly from spectro-temporal data, thus, enabling major advances in emotion recognition, speaker identification, and perceptual quality estimation [12], [13].

Motivated by these advancements, this research goes one step further and examines the prediction of vocal attractiveness using both conventional and deep learning models. As a preliminary step, a Support Vector Regression model was fit on low-level acoustic descriptors. Subsequently, a CNN–Bidirectional LSTM network was created directly on Mel-spectrograms. The CNN–BiLSTM model outperformed the SVR both on the validation and test sets. The convolutional layers capture localized spectral structures, while the BiLSTM layers learn long-range temporal dependencies across speech frames. Importantly, all experiments were conducted on a suitable and publicly-available corpus following a speaker-independent protocol.

Afterwards, to thoroughly understand its predictions, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to show the areas of time–frequency that most influenced model predictions. Such explanations showed that the network mainly attends to high-frequency regions related to the brightness and the harmonic clarity of the attractive voices, while in the non-attractive samples the attention is more disjointed and less consistent. Furthermore, the visualizations reveal minor yet persistent differences across gender–attractiveness combinations, following the network's focus on gender-specific spectral patterns.

To the best of our knowledge, this work is among the first to investigate automatic voice attractiveness prediction as a continuous regression task using a deep spectro-temporal model combined with interpretability analysis. In addition to quantitative performance evaluation, we provide qualitative insight through Grad-CAM visualizations showing how acoustic patterns may influence the model's attractiveness predictions. The implementation of the complete pipeline will be made available upon acceptance.

The main contributions of this work are threefold:

- We formulate voice attractiveness prediction as a continuous MOS regression task on a public corpus under a speaker-independent protocol.
- We compare a handcrafted-feature SVR baseline with a CNN–BiLSTM model operating on Mel-spectrograms and show that the latter yields lower prediction error.
- We apply Grad-CAM to analyze spectro-temporal regions associated with attractiveness predictions, providing an interpretable perspective on the learned acoustic correlates.

II. Related Work

Research on vocal attractiveness and pleasantness is becoming increasingly sophisticated and intricate over the last few decades. This domain has attracted attention from various scientific fields like phonetics, psychology, and speech technology. In their pioneering research, Varošanec-Skarić et al. [14] delved into the areas of acoustics that correspond to the production of a pleasant voice, by studying the distribution of spectral energy via long-term average spectra. What emerged from the results was the idea that a pleasant male voice has more relative energy below 300 Hz, while a pleasant female voice showed less differentiation, pointing to the existence of gender-specific perceptual mechanisms in voice evaluation.

Zuta et al. [15] extended the research by investigating the ideas that listeners form of the physical attributes of a person from the vocal attractiveness. They concluded that non-linguistic cues in speech could evoke in the listener's mind a visual image of the speaker's physical appearance. These results had the effect of stressing that vocal pleasantness is perpetually multidimensional not only from an acoustical perspective but also a perceptual one.

Several perceptual experiments have been conducted with the aim of finding a connection between vocal parameters and the biological and aesthetic features that the listener attributes to the source of voice. Collins et al. [16] were able to prove that the voices of women with higher fundamental frequency and lesser formant spacing are perceived as more attractive; at the same time, they associate these features with feminine physical characteristics. Along the same line, Bruckert et al. [17] discovered that composite or "average" voices—obtained through the blending of multiple speech samples—are consistently judged as more attractive, which unfolds that vocal attractiveness follows statistical regularities similar to those of facial attractiveness. These studies set a foundation for empirical evidence that the human preference for specific voice features is not a matter of chance but has a basis in acoustic patterns that can be reproduced.

With the advancements on speech technology, researchers shifted their attention to the objective modelling of voice pleasantness. One of the earliest automatic systems for classifying pleasantness and estimating its intensity constructed by Pinto-Coelho et al. [3] incorporated acoustic, prosodic, and signal-based features in a machine-learning framework. Their system accomplished an error rate of 9.1% in classification and 15.7% in intensity estimation, thus showing that the subjective impressions of voice quality could be convincingly simulated by using quantitative descriptors. In a similar objective, Gocsál and Urbanics [5] took a musical angle on vocal pleasantness, setting forth the theory that pitch intervals and melodic patterns might be the factors that explain the perceived pleasantness. As no direct linkage between musical intervals and pleasantness was established, their outcome pointed to the idea that the appreciation of vocal beauty is primarily related to prosodic structure.

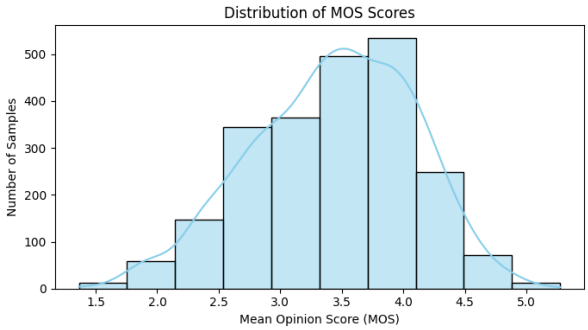

Fig. 1. Distribution of Mean Opinion Score values in the CocoNut-Humoresque corpus.

Recently, Suda et al. [9], [18] developed the CocoNut-Humoresque corpus, which is the first extensive and "in-the-wild" dataset for the study of vocal likability comprising 1,800 speech segments rated by 885 listeners. Significantly, the dataset is publicly available for research purposes, while their work revealed systematic gender and age biases in likability judgments. Overall, prior research has progressed from qualitative acoustic and perceptual analyses toward more data-driven modeling approaches. In contrast to earlier studies that primarily focused on perceptual analysis or traditional machine-learning methods, our work investigates continuous MOS prediction using a spectrogram-based DNN and complements it with Grad-CAM-based interpretability analysis.

III. Dataset Description

Our experiments were conducted on the CocoNut-Humoresque speech corpus [18], a large-scale Japanese speech likability dataset specifically designed for the study of perceptual voice evaluation. The corpus includes 1,800 speech segments of ten different types of speakers and rated by 885 human listeners of different genders and ages. Each listener rated the vocal attractiveness on a 1–6 scale, with higher values indicating greater pleasantness. Fig. 1 shows the distribution of MOS ratings in the CocoNut-Humoresque corpus. Most notably, raters were asked to pay attention only to the sound quality and to ignore the linguistic content of the utterances.

The speech recordings are supplemented with metadata that detailed the speaker (e.g., sex, age, recording conditions) and the listener (e.g., demographic and perceptual profiles) from whom the speech samples were obtained. The detailed information contained in the metadata allows the analysis of sociophonetic biases in listeners' perception, for example by

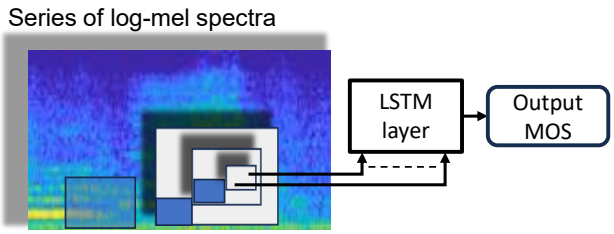

Fig. 2. The architecture of the developed Convolutional Neural Network combined with bidirectional Long Short-Term Memory.

showing how gender, age, or familiarity influence attractiveness judgments.

Such analyses of the CocoNut-Humoresque corpus [9] revealed two consistent perceptual trends: male listeners generally rated female voices as more attractive, and younger listeners tended to assign higher MOS scores overall. In addition to these perceptual patterns, the authors reported several acoustic correlates of likability, showing that fundamental frequency (F0), formant positioning, and x-vector embeddings, i.e., fixed-dimensional speaker representations extracted from deep speaker recognition models, are strongly associated with the preference ratings assigned by human listeners.

CocoNut-Humoresque was selected as a suitable benchmark for vocal attractiveness prediction due to its public availability, scale, and perceptual richness. These properties make it a relevant resource for training and evaluating models that aim to capture the subjective and multidimensional nature of vocal attractiveness. Besides the MOS scores, gender information was used in the Grad-CAM analysis to study attractiveness across speech coming from different genders.

IV. Proposed Methodology

The implemented pipeline for vocal attractiveness prediction was designed as a seven-stage framework including: a) data preprocessing and MOS assignment, b) feature extraction, c) Support Vector Regression, d) CNN–BiLSTM learning, e) model evaluation, f) error analysis, and g) Grad-CAM based interpretation.

A. Feature Extraction and SVR Baseline

Firstly, feature extraction involved calculating a set of handcrafted acoustic descriptors that would effectively represent the core timbral and prosodic aspects of the sound sample. These descriptors included MFCCs, the mean and variance of the fundamental frequency (F0) and the spectral centroid [10], [19]. Collectively, these descriptors characterize vocal color, the harmonic structure, and the pitch distribution, which are low-level acoustic correlates potentially related to attractiveness. The extracted features were normalized over the dataset using the z-score scheme. After preliminary experiments, the SVR model employed a Radial Basis Function kernel. This classical approach was selected because of its ability to model nonlinear relationships and capture perceptual information [20], [21].

B. CNN–BiLSTM Model

Moving beyond handcrafted features, a CNN–BiLSTM model was designed to learn spectro-temporal dependencies directly from Mel-spectrograms. Each audio segment was resampled to 16 kHz, transformed into a Mel-spectrogram with 64 Mel bins and 128 time frames, and normalized through per-sample z-score normalization in order to facilitate training stability. The final model, determined through grid-search-based hyperparameter tuning, consists of two convolutional layers followed by a bidirectional LSTM layer (see Fig. 2). The convolutional layers extract localized spectral patterns related to harmonic energy and formant distribution, whereas the bidirectional LSTM layer captures temporal dynamics in both forward and backward directions.

The CNN–BiLSTM network was optimized using the AdamW algorithm and a mean squared error loss function. The model was trained for up to 20 epochs with a batch size of 8. Training was regularized using two callbacks: EarlyStopping, which stopped training when the validation loss plateaued, and ReduceLROnPlateau, which gradually decreased the learning rate to support fine-tuning of the optimization process. The initial learning rate was set to 5×10^{-3} and was reduced as needed to as low as 10^{-5}.

C. Evaluation Protocol

Assessment on both the validation and held-out test sets was performed using a standardized speaker-independent protocol in order to evaluate generalization to unseen speakers. Both models were evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). These metrics were selected to reflect both the magnitude of prediction errors and the proportion of variance explained in the MOS annotations.

D. Grad-CAM-based Interpretation

Model predictions were further examined using Grad-CAM [22], [23], which computes the gradient of the model output with respect to the final convolutional layer and produces a heatmap indicating the time–frequency regions that most influenced the prediction. Grad-CAM visualizations were generated for four contrasting gender–attractiveness combinations (female/male $\times$ high/low MOS), allowing us to inspect differences not only between attractiveness levels but also across gender-related spectro-temporal patterns.

V. Experimental Results and Analysis

The obtained results are shown in Fig. 3. The SVR model achieved relatively consistent performance on both validation and test sets, suggesting that low-level acoustic features such as MFCCs, F0, and spectral centroid carry information relevant to vocal attractiveness perception, in agreement with previous studies (see Section II).

The CNN–BiLSTM model was able to generalize more accurately as it takes into account nonlinear dependencies between spectral and temporal structures in the speech signal [24], [25]. The bidirectional recurrent layers gave the model

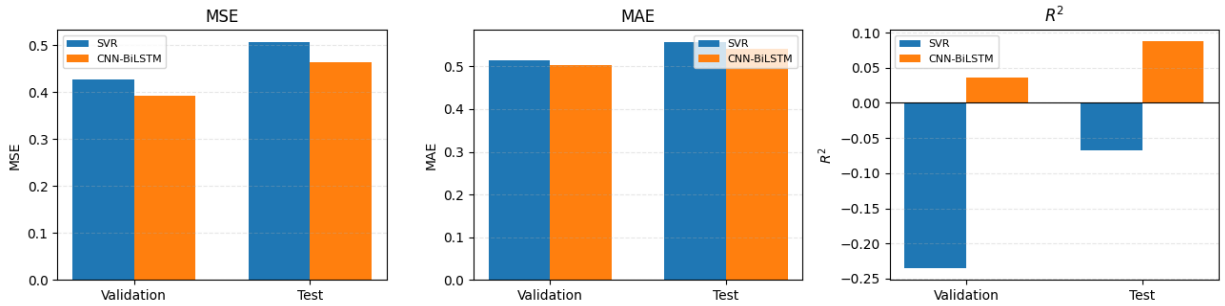

Fig. 3. Comparison of SVR and CNN–BiLSTM on the validation and test sets in terms of MSE, MAE, and R^2. Lower MSE/MAE and higher R^2 indicate better regression performance.

the ability to consider context not only from the past but also from the future, hence, the model became more sensitive to prosodic flow and temporal coherence.

The CNN–BiLSTM model achieved an MSE of 0.4643 and an MAE of 0.5402 on the test set, corresponding to an average prediction error of approximately ± 0.54 MOS points. Moreover, it reached a positive R^2 value of 0.0875, indicating that it explains a measurable portion of the variance in human attractiveness ratings while consistently outperforming the SVR baseline.

Although the absolute performance remains moderate, these results should be interpreted in the context of the difficulty of the task. Voice attractiveness is inherently subjective, and the target MOS values reflect aggregated judgments from multiple listeners rather than a fully objective ground truth. Therefore, the reported improvement is better viewed as a reproducible research baseline and a first step toward practical modeling of vocal attractiveness, rather than as a deployment-ready solution.

To get a better understanding of the model's performance, a continuous residual curve (Fig. 4) was drawn by sorting samples in accordance with their ground-truth MOS. This figure shows that the highest errors are committed mainly in extreme MOS values, whereas for the mid-range scores, the prediction patterns are more stable. The trend is a mirror of the problem of modelling extreme perceptual judgments, which are less represented in the data distribution. Interestingly, this analysis is in line with the global metrics presented in [9].

The results suggest that the CNN–BiLSTM model is able to

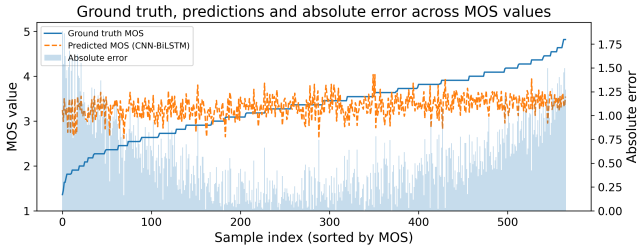

Fig. 4. Continuous prediction residuals of the CNN–BiLSTM model sorted by increasing ground-truth MOS values. The horizontal axis represents test samples ordered by true MOS, while the vertical axis shows the signed prediction residual (predicted MOS minus ground-truth MOS).

capture patterns in the acoustic space that are associated with perceived attractiveness. This improvement can be attributed to the depth of the CNN encoder, bidirectional temporal modelling and adaptive optimization through early stopping and learning-rate scheduling, all of which contributed to stabilizing convergence and lowering overfitting.

Nevertheless, the relatively moderate absolute performance is a manifestation of the inherent difficulty of the task despite the progress made. Vocal attractiveness is a subjective and multidimensional phenomenon whose main factors are emotion, prosody, and speaker identity, which cannot be entirely deduced from acoustic cues. The inter-rater variability within the dataset acts as a limiting factor for the upper bound of achievable accuracy. Actually, Suda et al. [9] conducted an explicit analysis of variability across listener groups. They found that there are statistically significant differences in the variance of MOS between demographic groups and especially pointed out that female listeners showed greater score variance than male listeners and that younger participants also tended to rate voices with higher variability. These results serve as a quantitative confirmation that the perceptual evaluation of vocal attractiveness is inconsistent among listeners, which accounts for the moderate prediction performance encountered in this research. Hence, even relatively small improvements in model accuracy can have a significant impact, as they take place in a context characterized by large subjective variability in human ratings.

This study also has several limitations. First, experiments were conducted on a single publicly available corpus in Japanese, which may limit generalization across languages, cultures, and speaking styles. Second, we did not perform repeated runs across multiple random seeds or statistical significance testing, and therefore the observed performance differences should be interpreted with appropriate caution. Third, the comparison was limited to an SVR baseline and did not include stronger neural or self-supervised speech baselines, which remain important directions for future work.

VI. Explainability and Discussion

A. Grad-CAM Implementation

Even though quantitative metrics characterize predictive accuracy, they do not explain why the model associates specific

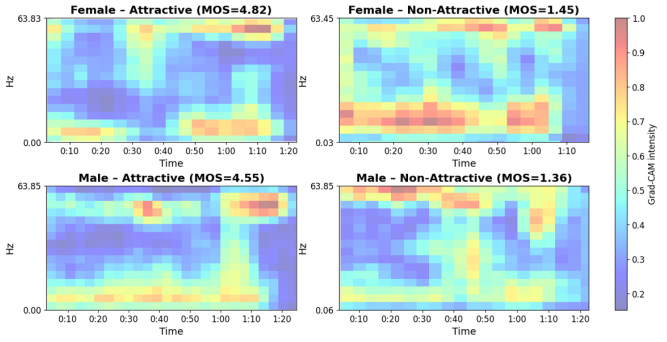

Fig. 5. Grad-CAM visualizations across gender and attractiveness levels. Each panel shows averaged relevance over the test set, highlighting the spectro-temporal regions that most strongly influenced the CNN–BiLSTM model's attractiveness predictions.

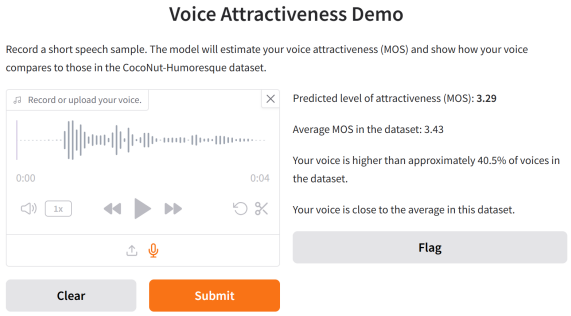

Fig. 6. User interface of the system demonstration. The user can record or upload a speech sample, after which the model predicts the attractiveness score and compares it to the reference distribution of the CocoNut-Humoresque dataset.

acoustic patterns with higher or lower attractiveness. To address this issue, Grad-CAM was applied to the CNN–BiLSTM network in order to identify the time–frequency regions that most strongly influenced the model's predictions. Four gender–attractiveness combinations (female/male $\times$ high/low MOS) were examined, allowing us to analyze whether the model's focus changes not only across attractiveness levels but also across speaker gender.

The explainability analysis relied on Grad-CAM, a visualization technique that highlights spectro-temporal regions with high relevance for the network's output [22], [23]. More specifically, Grad-CAM was applied to the last convolutional layer to generate relevance heatmaps indicating the frequency bands and temporal structures most strongly associated with attractiveness predictions. Fig. 5 visualizes these heatmaps averaged over the test set.

B. Spectro-temporal Patterns Across Attractiveness Levels

Overall, the Grad-CAM heatmaps suggest that the CNN–BiLSTM model primarily focuses on spectro-temporal patterns in the low and high-frequency ranges. In highly attractive voices (both female and male), the model's attention is more strongly concentrated around stable harmonic structures and upper-formant energy, which appear to be associated with higher attractiveness predictions. On the other hand, unattractive voices are characterized by more diffuse, fragmented, and inconsistently distributed activation patterns.

The observed high-frequency emphasis is consistent with perceptual studies suggesting that listeners rely on upper-formant structure and spectral clarity when judging vocal appeal. These regions are related to timbral features associated with vocal tract resonance, the breathiness, and the articulatory precision, all of which are known to be contributors to the perceived attractiveness (see Section II).

C. Relation to Perceptual Research

Moreover, the weak and spatially more spread-out activations in low-MOS samples indicate that the model may be less confident in representing these samples, which often correspond to segments with weak periodicity or unstable harmonic structure. Such a behavior is in line with the larger prediction errors at low MOS values as discussed in Section V. These findings are in line with Suda et al. [9], [18] who have come to similar conclusions focusing on the role of formant-related cues, harmonic stability, and spectral clarity in influencing voice likability. The patterns of the Grad-CAM shown here substantiate the argument that perceived attractiveness cannot be fully explained by simple prosodic or pitch-based cues but higher-level spectral organization plays a crucial role.

Overall, the Grad-CAM analysis provides a perceptually grounded and interpretable view of the model's behavior. Attractive voices tend to elicit stronger and more coherent activation patterns, whereas low-MOS voices are associated with weaker and more scattered relevance maps. These visualizations help connect computational modeling with human perceptual evaluation by illustrating how acoustic cues may influence model predictions across gender and attractiveness levels.

VII. System Demonstration

To illustrate the practical usability of the proposed framework, we implemented an interactive demonstration system. The demo allows a user to upload a speech sample or record a voice segment through a microphone, after which the trained CNN–BiLSTM model predicts the corresponding attractiveness score. An overview of the interface is shown in Fig. 6.

The input audio is first converted into a normalized Mel-spectrogram and then forwarded to the trained neural network, which outputs a continuous MOS estimate. In addition, the system compares the predicted score with the MOS distribution of the CocoNut-Humoresque corpus, providing percentile-based feedback. The demo was implemented using the Gradio framework (https://www.gradio.app/) and is intended primarily as an illustrative research prototype rather than a production-ready application.

VIII. Conclusion

This work presented an approach for speech attractiveness prediction together with an interpretable analysis framework.

The results showed that the proposed CNN–BiLSTM achieved better generalization than the SVR baseline, reaching a validation MSE of 0.3920 and a test MSE of 0.4643, together with a test MAE of 0.5402 and a positive coefficient of determination ($R^2 = 0.0875$). These results suggest that the CNN–BiLSTM is able to identify acoustically meaningful patterns related to vocal attractiveness. The error analysis, conducted through the continuous residual curve, highlighted that prediction difficulty increases at perceptual extremes. This is due to limited data availability and more variable ratings in those ranges. The proposed approach provides a useful basis for further research in speech synthesis, affective voice design, and human–AI auditory interaction [3], [26], [27].

Via an explainable AI technique, we produced visualizations for all four gender–attractiveness combinations and observed that vocal attractiveness is mainly reflected in high-frequency bands characteristic of brightness, harmonic richness, and timbral clarity, whereas the low-frequency energy was not as significant. At the same time, non-attractive voices are associated with more diffused activation patterns.

The present study supports the idea that vocal attractiveness is not solely determined by fixed tonal qualities of a voice, but rather by spectral and temporal variations, which are able to transmit clarity, warmth, and emotional liveliness [5], [14], [16], [17]. The obtained results further support the potential of deep neural architectures to represent complex perceptual characteristics of speech and motivate future work on more perceptually informed and interpretable auditory AI systems.

At the same time, the moderate absolute performance and the use of a single corpus indicate that further validation is required before such systems can be considered for broader real-world deployment. Based on the presented findings, future work will focus on approaches for handling imbalanced data environments, such as sample weighting and selection [28], as well as audio augmentation methods [29], [30]. In addition, stronger baseline comparisons, repeated experiments across multiple random seeds, and the inclusion of self-supervised speech representations will be considered in order to provide a more comprehensive evaluation of the task.

REFERENCES

[1] E. Akhmetshin, G. Meshkova, M. Mikhailova, R. Shichiyakh, G. P. Joshi, and W. Cho, "Enhancing human computer interaction with coot optimization and deep learning for multi language identification," *Scientific Reports*, vol. 14, no. 1, Oct. 2024.

[2] Z. Ding, Y. Ji, Y. Gan, Y. Wang, and Y. Xia, "Current status and trends of technology, methods, and applications of human–computer intelligent interaction (hcii): A bibliometric research," *Multimedia Tools and Applications*, vol. 83, no. 27, p. 69111–69144, Jan. 2024.

[3] L. Pinto-Coelho, D. Braga, M. Sales-Dias, and C. Garcia-Mateo, "On the development of an automatic voice pleasantness classification and intensity estimation system," *Computer Speech & Language*, vol. 27, no. 1, p. 75–88, Jan. 2013.

[4] S. Ntalampiras and I. Potamitis, "On predicting the unpleasantness level of a sound event," in *Interspeech 2014*. ISCA, Sep. 2014, p. 1782–1785.

[5] A. Gocsál and T. Urbanics, "Vocal pleasantness ratings explained by a musical approach," in *The Second International Conference on Laboratory Phonetics & Phonology*. ICLPP, Oct. 2022, p. 59–66.

[6] H. Wang, J. Zhao, Y. Yang, S. Liu, J. Chen, Y. Zhang, S. Zhao, J. Li, J. Zhou, H. Sun, Y. Lu, and Y. Qin, "Speechllm-as-judges: Towards general and interpretable speech quality evaluation," 2025.

[7] J. Shang and Y. Zhang, "Influence of male's facial attractiveness, vocal attractiveness and social interest on female's decisions of fairness," *Scientific Reports*, vol. 14, no. 1, Jul. 2024.

[8] J. Ostrega, V. Shiramizu, A. J. Lee, B. C. Jones, and D. R. Feinberg, "No evidence that averaging voices influences attractiveness," *Scientific Reports*, vol. 14, no. 1, May 2024.

[9] H. Suda, A. Watanabe, and S. Takamichi, "Who Finds This Voice Attractive? A Large-Scale Experiment Using In-the-Wild Data," in *Interspeech 2024*, 2024, pp. 3165–3169.

[10] 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. ACM, Oct. 2010, p. 1459–1462.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The kaldi speech recognition toolkit," in *Proceedings of ASRU 2011*. Hilton Waikoloa Village Resort, Hawaii: IEEE Signal Processing Society, 2011, pp. 1–4.

[12] S. Ntalampiras, "Emotional quantification of soundscapes by learning between samples," *Multimedia Tools and Applications*, vol. 79, no. 41–42, p. 30387–30395, Aug. 2020.

[13] ——, "Species-independent analysis and identification of emotional animal vocalizations," *Scientific Reports*, vol. 15, no. 1, Aug. 2025.

[14] G. Varošanec-Skarić, "Relation between voice pleasantness and distribution of the spectral energy," in *Proc. 14th Int. Congr. Phonetic Sciences (ICPhS)*. San Francisco, CA: ICLPP, Oct. 1999, p. 1013–1016.

[15] V. Zuta, *Voice Pleasantness of Female Voices and the Assessment of Physical Characteristics*. Springer Berlin Heidelberg, 2009, p. 116–125.

[16] S. A. Collins and C. Missing, "Vocal and visual attractiveness are related in women," *Animal Behaviour*, vol. 65, no. 5, p. 997–1004, May 2003. [Online]. Available: http://dx.doi.org/10.1006/anbe.2003.2123

[17] L. Bruckert, P. Bestelmeyer, M. Latinus, J. Rouger, I. Charest, G. A. Rousselet, H. Kawahara, and P. Belin, "Vocal attractiveness increases by averaging," *Current Biology*, vol. 20, no. 2, p. 116–120, Jan. 2010.

[18] A. Watanabe, S. Takamichi, Y. Saito, W. Nakata, D. Xin, and H. Saruwatari, "Coco-nut: Corpus of japanese utterance and voice characteristics description for prompt-based control," in *2023 IEEE ASRU Workshop*. IEEE, Dec. 2023, p. 1–8.

[19] S. Ntalampiras, "Interpretable probabilistic identification of depression in speech," *Sensors*, vol. 25, no. 4, p. 1270, Feb. 2025. [Online]. Available: http://dx.doi.org/10.3390/s25041270

[20] ——, "Model ensemble for predicting heart and respiration rate from speech," *IEEE Internet Computing*, vol. 27, no. 3, pp. 15–20, 2023.

[21] G. H. Mohmad Dar and R. Delhibabu, "Speech databases, speech features, and classifiers in speech emotion recognition: A review," *IEEE Access*, vol. 12, pp. 151 122–151 152, 2024.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE ICCV*, 2017, pp. 618–626.

[23] A. Akman and B. W. Schuller, "Audio explainable artificial intelligence: A review," *Intelligent Computing*, vol. 3, Jan. 2024.

[24] D. Yilihamu, M. Ablimit, and A. Hamdulla, "Speech language identification using cnn-bilstm with attention mechanism," in *2022 PRML*, 2022, pp. 308–314.

[25] A. Saitta and S. Ntalampiras, "Language-agnostic speech anger identification," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, 2021, pp. 249–253.

[26] S. Ntalampiras, "Toward language-agnostic speech emotion recognition," *Journal of the Audio Engineering Society*, vol. 68, no. 1/2, p. 7–13, Feb. 2020.

[27] M. Li, J. Zhang, F. Guo, Y. Liao, X. Hu, and J. Ham, "Audiovisual affective design of humanoid robot appearance and voice based on kansei engineering," *International Journal of Social Robotics*, vol. 17, no. 1, p. 15–37, Jan. 2025.

[28] C. Santiago, C. Barata, M. Sasdelli, G. Carneiro, and J. C. Nascimento, "Low: Training deep neural networks by learning optimal sample weights," *Pattern Recognition*, vol. 110, p. 107585, Feb. 2021.

[29] U. Avci, "A comprehensive analysis of data augmentation methods for speech emotion recognition," *IEEE Access*, vol. 13, p. 111647–111669, 2025.

[30] S. Ntalampiras and G. Pesando Gamacchio, "Explainable classification of goat vocalizations using convolutional neural networks," *PLOS ONE*, vol. 20, no. 4, p. e0318543, Apr. 2025.