Voice Conversion for Likability Control via Automated Rating of Speech Synthesis Corpora

Hitoshi Suda[1], *Shinnosuke Takamichi*[2,3,1], *Satoru Fukayama*[1]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2]Keio University, Japan
[3]The University of Tokyo, Japan

suda.h@aist.go.jp, shinnosuke_takamichi@keio.jp, s.fukayama@aist.go.jp

Abstract

Perceived voice likability plays a crucial role in various social interactions, such as partner selection and advertising. A system that provides reference likable voice samples tailored to target audiences would enable users to adjust their speaking style and voice quality, facilitating smoother communication. To this end, we propose a voice conversion method that controls the likability of input speech while preserving both speaker identity and linguistic content. To improve training data scalability, we train a likability predictor on an existing voice likability dataset and employ it to automatically annotate a large speech synthesis corpus with likability ratings. Experimental evaluations reveal a significant correlation between the predictor's outputs and human-provided likability ratings. Subjective and objective evaluations further demonstrate that the proposed approach effectively controls voice likability while preserving both speaker identity and linguistic content.

Index Terms: speech synthesis, voice conversion, voice likability, paralinguistic voice control

1. Introduction

Speech conveys three types of information: (1) linguistic information, which is represented by sequences of discrete symbols; (2) paralinguistic information, such as speaking styles, which can be intentionally controlled by speakers; and (3) non-linguistic information, such as speaker identity, which is typically beyond their control [1]. Numerous studies have explored methods for controlling paralinguistic and non-linguistic features in synthesized speech. For instance, in text-to-speech (TTS) systems, various studies have proposed methods to control the emotional tone and speaking style [2, 3]. Voice conversion techniques, such as speaker conversion and style transfer, have also been studied to control specific paralinguistic or non-linguistic features [4, 2]. Typically, these techniques rely on reference audio signals or emotion labels to control the desired speech characteristics, whereas others employ natural language prompts [5]. However, practical applications, such as designing voices for advertisements targeting distinct customer segments, require controlling speech characteristics based on more subjective criteria, such as target demographics. The direct manipulation of paralinguistic and non-linguistic elements for subjective purposes would enable more adaptable and targeted voice designs. Furthermore, this capability could facilitate the development of voice training applications that analyze a user's voice and generate a synthesized reference voice optimized for the speaker and tailored to specific audiences.

Voice likability is a fundamental and inherently subjective aspect of both paralinguistic and non-linguistic features of speech [6]. Several studies have demonstrated that voice likability significantly influences social outcomes, such as partner selection and leadership quality [7, 8]. Some previous studies have revealed a relationship between voice likability and fundamental frequency (f_o) [9], while others have indicated contributions from factors such as phoneme durations and speech rate [10]. Analyses of a large corpus have confirmed that f_o significantly impacts voice likability and indicated that additional acoustic features also contribute [11]. Since perceived voice likability varies among listeners [6], it is a multifaceted phenomenon that cannot be explained by a single acoustic feature. Therefore, models that integrate multiple acoustic parameters are necessary to predict and control likability effectively.

This paper presents a voice conversion method that controls voice likability while preserving both speaker identity and linguistic content. This method will enable users to enhance their vocal performance by generating exemplary voice samples. However, since its development relies on subjective evaluations from multiple listeners for each training sample, the method suffers from scalability issues. To overcome this limitation, we utilize an existing voice likability dataset [11] to develop an automatic likability predictor. The predictor enables the development of a likability control model without additional manual ratings, as it automatically assigns likability ratings to a speech synthesis corpus. This paper details our approach to automatic likability prediction and voice conversion for likability control. The paper also presents experimental results demonstrating the effectiveness of the proposed method.

2. Automatic prediction of voice likability

This section describes our approach to automatically predicting voice likability. Figure 1 shows the architecture of the proposed likability predictor. The model accepts a log-Mel spectrogram as input and employs time-delay neural networks (TDNNs) along with a statistics pooling layer, similar to the x-vector architecture [12]. The model then outputs a single time-invariant likability rating for each of four listener groups defined by gender and age. We utilize the CocoNut-Humoresque corpus, which provides subjective ratings for 1800 voice samples along with gender and age information of the listeners [11].

Likability prediction is similar to mean opinion score (MOS) prediction for synthesized speech [13, 14]. Unlike MOS prediction, the ratings for a given speech sample can vary among listeners. To leverage this characteristic, we partition the listeners into four groups based on gender and age: males under 40 ($n = 202$), males 40 or older ($n = 323$), females under 40 ($n = 143$), and females 40 or older ($n = 210$), where n denotes the number of listeners in the corpus. The model uses a shared network to predict the mean rating for each group.

The rating distribution in the corpus is concentrated around

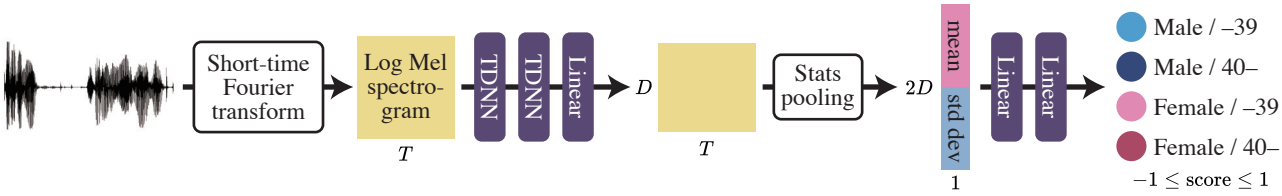

Figure 1: *Architecture of the proposed voice likability predictor. The predictor outputs likability ratings ranging from −1 to 1 for four listener groups defined by age and gender: males under 40, males 40 or older, females under 40, and females 40 or older. T denotes the number of time frames, and D denotes the number of dimensions of the intermediate features.*

the center (i.e., near 0 when normalized to the range $[-1, 1]$). To mitigate the effects of overfitting, we apply a post-filtering process that linearly transforms the initial predictions. Specifically, on the validation set, we first compute the mean μ and variance σ^2 of the human-provided ratings, along with the corresponding mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ of the predictions. We then compute the adjusted ratings $\hat{y}'$ from the predictions $\hat{y}$ by

$$\hat{y}' = \frac{\sigma}{\hat{\sigma}}(\hat{y} - \hat{\mu}) + \mu. \quad (1)$$

Since this transformation is linear, correlation-based evaluation metrics remain unchanged.

Several previous studies have also investigated automatic voice likability prediction [15, 16]. In contrast to these methods, which rely on multiple openSMILE-based acoustic features, such as f_o, energy, and Mel-frequency cepstral coefficients (MFCCs), our single-network approach exhibits improved noise robustness. Furthermore, the differentiable nature of our network facilitates its seamless integration into larger frameworks, such as TTS and voice conversion systems.

3. Voice conversion for likability control

This section describes the proposed voice conversion method that controls voice likability while preserving both speaker identity and linguistic information. Figure 2 depicts the architecture of the method. This method is based on a voice conversion approach that utilizes sequences of discrete speech units extracted from a self-supervised learning (SSL) model [17, 18]. First, the method extracts SSL features from the training data, then applies k-means clustering to derive k cluster centroids. A TTS model is then trained to synthesize speech signals conditioned on three types of inputs: (1) cluster indices derived from the SSL feature sequences as discrete units; (2) speaker embeddings extracted using ECAPA-TDNN [19]; and (3) target likability ratings. We employ hidden-unit BERT (HuBERT) [20] for the SSL component and FastSpeech 2 [21] for the TTS model. During synthesis, sequences of consecutive identical discrete tokens are compressed into a single token. For example, if the extracted discrete token sequence is [13, 7, 7, 21, 21, 5], the model input becomes [13, 7, 21, 5]. This strategy facilitates adjustments of speech rate based on speaker embeddings and target likability ratings.

Training this model requires a multi-speaker corpus annotated with likability ratings. Since the CocoNut-Humoresque corpus is derived from the Coco-Nut [22], which comprises speech samples collected from YouTube and processed via source separation, high-quality synthesis cannot be achieved using this corpus alone. In contrast, manually rating a multi-speaker corpus is costly and is not scalable. To overcome this limitation, we utilize likability ratings that are automatically predicted by the likability predictor described in Section 2.

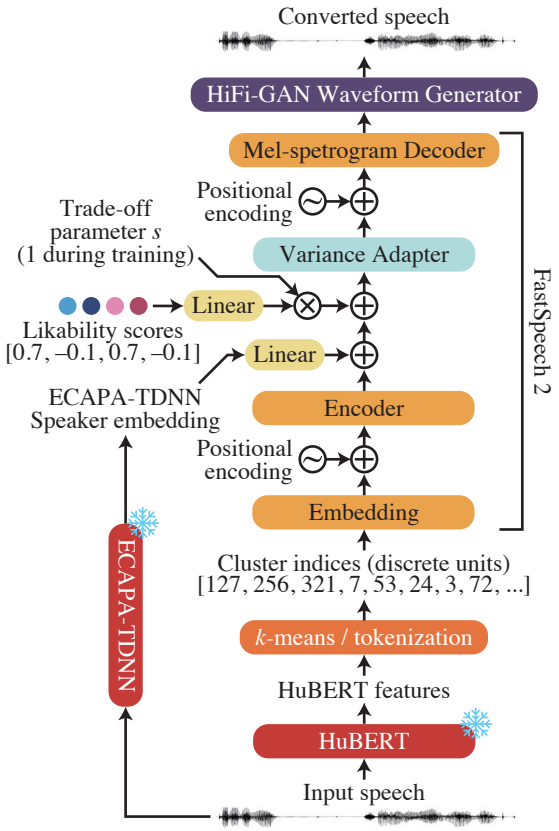

Figure 2: *Architecture of the proposed voice conversion model for likability control. The model is based on FastSpeech 2 [21] and utilizes HuBERT-based discrete units, speaker embeddings extracted with ECAPA-TDNN [19], and target likability ratings.*

This approach entails a trade-off between likability control and speaker identity preservation: excessive control may compromise speaker identity, and vice versa. To improve usability in application scenarios, our approach introduces a scalar multiplier s, which is applied to the embeddings of the target likability ratings. The scalar multiplier is fixed at 1 during training but can be adjusted during inference to modulate the balance between likability control and speaker identity preservation.

4. Experiment 1: Voice likability prediction

4.1. Experimental setup

We utilized the CocoNut-Humoresque corpus [11] and normalized the ratings to the range $[-1, 1]$. The corpus is divided

Table 1: *Time contexts and the number of dimensions for each layer in the likability prediction architecture. T_s denotes the total number of signal samples, T denotes the total number of frames, and t denotes the time-frame indices.*

Layer	Layer context	Input×output
spectrogram	$[0, T_s)$	$T_s \times 80T$
frame1	$\{t-2, t, t+2\}$	240×32
frame2	$\{t-6, t-3, t, t+3, t+6\}$	160×32
frame3	$[t]$	32×32
stats pooling	$[0, T)$	$32T \times 64$
segment4	$[0]$	64×32
segment5	$[0]$	32×4

Table 2: *Performance of likability prediction. MSE and Std represent the mean squared error and the standard deviation of the test set, respectively. LCC, SRCC, and KTAU denote the linear, Spearman rank, and Kendall rank correlation coefficients, respectively. GT Std denotes the standard deviation of the human-provided (ground truth) ratings. LCC values marked with ** are statistically significant ($p < 0.01$).*

Listeners	MSE↓	Std	LCC↑	SRCC↑	KTAU↑	GT Std
M –39	0.17	0.40	0.36**	0.39	0.26	0.32
M 40–	0.12	0.34	0.41**	0.43	0.28	0.28
F –39	0.19	0.44	0.40**	0.41	0.28	0.35
F 40–	0.17	0.41	0.38**	0.39	0.26	0.31
All	0.08	0.29	0.46**	0.49	0.33	0.26

into training, validation, and test sets comprising 1500, 300, and 300 audio samples, respectively. Each audio sample is approximately 4 seconds long, recorded in stereo at a sampling rate of 44 100 Hz. We generated monaural signals by averaging the stereo channels and downsampling them to 22 050 Hz.

Table 1 presents the dimensions and time contexts of each layer. We extracted 80-bin log-Mel spectrograms up to 8000 Hz with a hop length of 256 samples. The loss function was the mean squared error between the predicted and human-provided ratings, and the optimal model was selected based on the Spearman rank correlation coefficient computed on the validation set.

For data augmentation, we randomly appended silence to the beginning and the end of the signals, applied reverberation, added white noise, and reversed the signals in time. For reverberation, we used the impulse response dataset from the MIT Acoustical Reverberation Scene Statistics Survey [23].

4.2. Results

Table 2 shows the performance of likability prediction. A significant correlation ($r = 0.46, p < 3\times10^{-17}$) was observed between the predicted and human-provided ratings; therefore, the results indicate that the proposed approach effectively predicts likability ratings. Moreover, post-filtering based on a linear transformation adjusted the standard deviation of the predicted ratings to follow that of the human-provided ratings while preserving the correlation coefficients. Furthermore, regarding the classification of the audio samples as "liked" or "disliked" based on the mean rating, the proposed method achieved an accuracy of 74% ($F_1 = 0.70$), demonstrating its effectiveness in predicting whether the voices were likable or not.

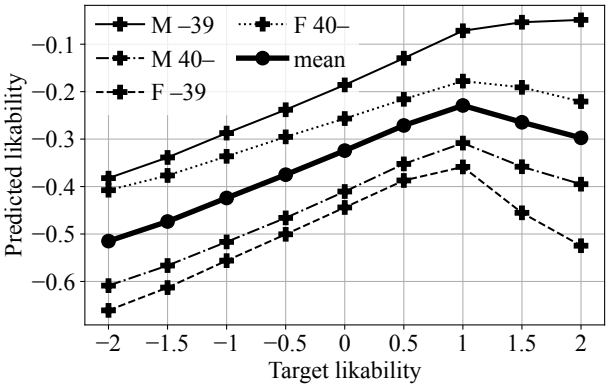

Figure 3: *Predicted likability ratings of utterances with controlled likability. Thin lines represent the ratings for each gender–age listener group, while the thick line represents the average rating across all four listener groups.*

5. Experiment 2: Voice likability control

5.1. Experimental setup

We utilized two speech corpora: the JVS corpus, comprising 14997 utterances from 100 speakers [24], and the JTES corpus, comprising 20000 utterances from 100 speakers in four emotional styles (neutral, angry, joyful, and sad) [25]. We reserved ten neutral sentences (sentences 41–50) uttered by two female speakers (f49 and f50) and two male speakers (m49 and m50) from the JTES corpus for evaluation. The training dataset, comprising the entire JVS corpus and 18240 utterances from the JTES corpus, was constructed to ensure that sentences or speakers did not overlap with the evaluation set. Speaker embeddings and likability ratings were extracted for each utterance individually without averaging across speakers.

We adopted the mini-batch k-means algorithm [26] with $k = 1000$ to cluster the SSL features. We used an open-source FastSpeech 2 implementation[1] and applied its default configurations for the LJSpeech dataset. For waveform generation, we used a pre-trained universal HiFi-GAN model[2] and fine-tuned it using Mel-spectrograms generated via teacher forcing. We employed the HuBERT Large model as an SSL model, which was trained on approximately 60000 hours of Japanese television broadcast audio [27]. For speaker embedding extraction, we used an open-source ECAPA-TDNN model `speechbrain/spkrec-ecapa-voxceleb`[3] [28]. The trade-off parameter s, described in Section 3 and Figure 2, was set to 1 during training and to 2.5 during evaluation.

In the evaluation, we set the target likability range from -2 to 2. Note that since the likability ratings in the training data were normalized to $[-1, 1]$, the synthesis process extrapolates likability values when they fall outside this range.

5.2. Objective evaluation

First, we evaluated the likability of the synthesized speech using the likability predictor described in Section 4. Figure 3 shows the results. For all listener groups, the synthesized speech's likability was successfully controlled to follow the target ratings.

[1] `https://github.com/ming024/FastSpeech2`
[2] `https://github.com/jik876/hifi-gan`
[3] `https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb`

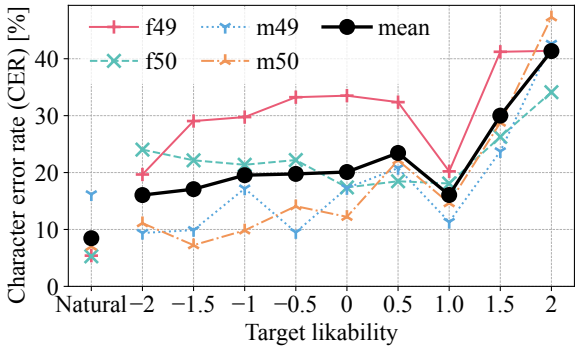

Figure 4: *CER of speech recognition on the synthesized speech. The leftmost column shows the results for natural speech.*

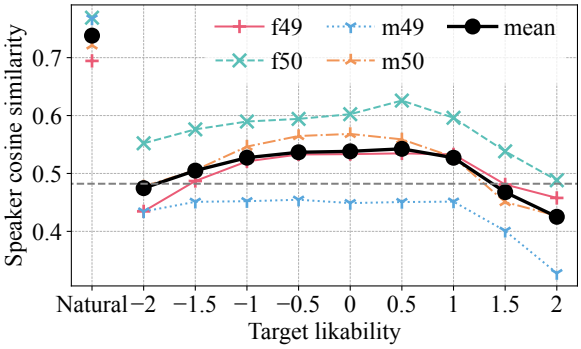

Figure 5: *Cosine similarity between speaker embeddings extracted from reference and synthesized speech. The leftmost column shows the results for natural speech. The gray dashed line indicates the similarity threshold corresponding to the equal error rate (EER) computed on the JTES corpus.*

As shown in Figure 3, the predicted ratings range only from -0.51 to -0.23, which is much narrower than the target range of -2 to 2. Since the proposed method consistently preserves speaker identity, the model appears to adjust voice likability only within each speaker's inherent range.

Next, we evaluated the preservation of linguistic content by performing speech recognition on the synthesized speech and computing the character error rate (CER). For speech recognition, we employed a Conformer model based on HuBERT features [29, 27] trained on the LaboroTVSpeech dataset, a large-scale Japanese speech corpus [30]. Figure 4 shows the CER results. The results indicate that the linguistic content remained consistent under likability control when the target likability was set within the range of -2 to 1. However, Figure 4 shows that the conversion process resulted in a higher CER, especially for female speakers. This degradation indicates limitations in the voice conversion model and suggests that further improvements in clustering and model architecture are needed to achieve higher-quality conversion.

Additionally, we evaluated speaker identity preservation by computing the cosine similarity between the embeddings of the reference and synthesized speech. We used 40 neutral sentences (sentences 1–40) from the JTES corpus as a reference set and extracted embeddings with the ECAPA-TDNN model described in Section 5.1. The open-source implementation[4] uses a same-

[4] https://github.com/speechbrain/speechbrain

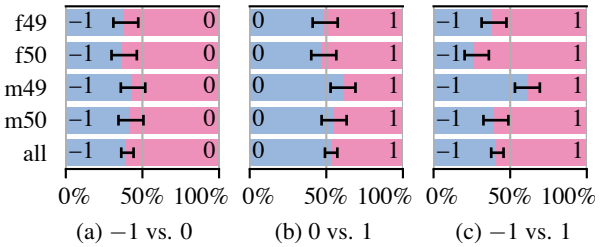

Figure 6: *Likability preference rates for synthesized utterances with different target likability ratings. Each row corresponds to a different speaker, while the "all" row represents the overall ratings. Error bars indicate the 95% confidence intervals.*

speaker threshold of 0.25, whereas the equal error rate (EER) condition on the JTES corpus indicates a threshold of 0.48 for speaker verification. The results shown in Figure 5 indicate that speaker identity was preserved within the target ratings of -1 to 1, and the likability control process successfully maintained speaker identity even as target likability values varied.

5.3. Subjective evaluation

We conducted listening tests with human participants to subjectively evaluate the effectiveness of the likability control. We applied the likability control to 10 utterances per speaker using three target likability values: -1, 0, and 1. Each listener then evaluated three randomly selected sentences per speaker for each pairing condition (-1 vs. 0, 0 vs. 1, and -1 vs. 1), resulting in 36 pairwise comparisons in total. In the evaluation, 100 participants took part and were each paid 120 Japanese yen.

Figure 6 presents the results of the subjective evaluation. The results indicate that, for three of the four speakers, the synthesized speech's likability was successfully controlled to follow the target values, with the exception of speaker m49. Specifically, significant differences in perceived likability were observed between target values of -1 and 0, and an overall significant difference was found between target values of -1 and 1. On the other hand, little difference in likability was observed between target values of 0 and 1, possibly due to a degradation in naturalness when the target value was 1. Among the four speakers, speaker m49 exhibited an unexpected pattern: rather than increasing, perceived likability decreased as the target value increased from 0 to 1, and a significant degradation in likability was observed between target values of -1 and 1. A possible explanation is that, for speaker m49, the synthesized speech failed to effectively preserve speaker identity, as indicated in Figure 5, resulting in an unnatural sound.

6. Conclusion

This paper presents a method for controlling voice likability for any speaker by extending a voice conversion approach that uses discrete speech units. To improve the scalability of training data, we constructed a likability predictor based on a voice likability corpus and used it to automatically annotate speech synthesis corpora with likability ratings. The results of both subjective and objective evaluation indicate the effectiveness of likability control as well as the preservation of speaker identity and linguistic content. Future work includes extending our approach to a voice control application that enables users to specify desired voice design characteristics through natural language or parameters beyond likability.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 21H04900, 23K20017, and 23K24895, and JST FOREST Program, Grant Number JPMJFR226V. This research was partially supported by the AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain." The authors would like to acknowledge Mr. Takizawa (AIST) for his support in the speech recognition evaluation.

8. References

[1] H. Fujisaki, "Prosody, models, and spontaneous speech," in *Computing Prosody*. Springer US, 1997, pp. 27–42.

[2] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80, 2018, pp. 5180–5189.

[3] H. Barakat, O. Turk, and C. Demiroglu, "Deep learning-based expressive speech synthesis: A systematic review of approaches, challenges, and resources," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, pp. 1–34, 2024.

[4] Y. Stylianou, "Voice transformation: A survey," in *Proc. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 3585–3588.

[5] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "PromptTTS: Controllable text-to-speech with text descriptions," in *Proc. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] B. Weiss, J. Trouvain, M. Barkat-Defradas, and J. J. Ohala, Eds., *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*. Springer Nature Singapore, 2020.

[7] S. A. Collins and C. Missing, "Vocal and visual attractiveness are related in women," *Animal behaviour*, vol. 65, no. 5, pp. 997–1004, 2003.

[8] I. Pavela Banai, B. Banai, and K. Bovan, "Vocal characteristics of presidential candidates can predict the outcome of actual elections," *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, vol. 38, no. 3, pp. 309–314, 2017.

[9] I. Skrinda, T. Krama, S. Kecko, F. R. Moore, A. Kaasik, L. Meija, V. Lietuvietis, M. J. Rantala, and I. Krams, "Body height, immunity, facial and vocal attractiveness in young men," *Die Naturwissenschaften*, vol. 101, no. 12, pp. 1017–1025, 2014.

[10] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, ""Would you buy a car from me?" – On the likability of telephone voices," in *Proc. Interspeech 2011*, 2011, pp. 1557–1560.

[11] H. Suda, A. Watanabe, and S. Takamichi, "Who finds this voice attractive? A large-scale experiment using in-the-wild data," in *Proc. Interspeech 2024*, 2024, pp. 3165–3169.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[13] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," in *Proc. NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop*, 2016.

[14] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4521–4525.

[15] D. Lu and F. Sha, "Predicting likability of speakers with Gaussian processes," in *Proc. Interspeech 2012*, 2012, pp. 286–289.

[16] F. Eyben, F. Weninger, E. Marchi, and B. Schuller, "Likability of human voices: A feature analysis and a neural network regression approach to automatic likability estimation," in *Proc. 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013, pp. 1–4.

[17] W.-C. Huang, Y.-C. Wu, and T. Hayashi, "Any-to-one sequence-to-sequence voice conversion using self-supervised discrete speech representations," in *Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5944–5948.

[18] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6562–6566.

[19] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[20] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[21] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. International Conference on Learning Representations*, 2021.

[22] 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 *Proc. 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023.

[23] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proc. National Academy of Sciences of the United States of America*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[24] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," *arXiv [cs.SD] 1908.06248*, 2019.

[25] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," in *Proc. 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, 2016, pp. 16–21.

[26] D. Sculley, "Web-scale k-means clustering," in *Proc. 19th International Conference on World Wide Web*, 2010.

[27] D. Takizawa, T. Nakamura, H. Suda, and S. Fukayama, "Automatic speech recognition of Japanese dialects using large-scale self-supervised learning models (in Japanese)," in *Proc. 2025 Spring Meeting of the Acoustical Society of Japan*, 2025.

[28] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," *arXiv [eess.AS] 2106.04624*, 2021.

[29] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[30] S. Ando and H. Fujihara, "Construction of a large-scale Japanese ASR corpus on TV recordings," in *Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6948–6952.