VoxSim: A perceptual voice similarity dataset

Junseok Ahn[1], *Youkyum Kim*[1], *Yeunju Choi*[2], *Doyeop Kwak*[1], *Ji-Hoon Kim*[1], *Seongkyu Mun*[2], *Joon Son Chung*[1]

[1]Korea Advanced Institute of Science and Technology, South Korea
[2]Samsung Research, South Korea

junseok.ahn@kaist.ac.kr

Abstract

This paper introduces VoxSim, a dataset of perceptual voice similarity ratings. Recent efforts to automate the assessment of speech synthesis technologies have primarily focused on predicting mean opinion score of naturalness, leaving speaker voice similarity relatively unexplored due to a lack of extensive training data. To address this, we generate about 41k utterance pairs from the VoxCeleb dataset, a widely utilised speech dataset for speaker recognition, and collect nearly 70k speaker similarity scores through a listening test. VoxSim offers a valuable resource for the development and benchmarking of speaker similarity prediction models. We provide baseline results of speaker similarity prediction models on the VoxSim test set and further demonstrate that the model trained on our dataset generalises to the out-of-domain VCC2018 dataset.

Index Terms: speaker similarity, neural speech synthesis, mean opinion score, automatic assessment

1. Introduction

In many areas of machine learning, the objective is to train models that emulate human cognitive abilities and aim to match human-level performance [1, 2]. However, there are cases where AI has outperformed human capabilities [3, 4]. Speaker recognition is a notable domain where AI models have shown superiority over human abilities. ECAPA-TDNN [5] represents significant progress in the field of speaker verification, demonstrating an Equal Error Rate (EER) of less than 1% on the VoxCeleb benchmark dataset [6]. Recently, self-supervised learning-based models have shown even superior verification performance [7, 8]. In contrast, human performance in speaker identification falls considerably short. Huh et al. [9] report that Amazon Mechanical Turk crowdworkers achieve an EER of 26.51% on the same dataset, and even expert researchers in speaker recognition demonstrate an EER of 15.77%. This indicates a substantial gap between the speaker characteristics that speaker recognition systems can extract and what humans can discern.

This gap leads to several issues, particularly when evaluating the speaker similarity of synthesised speech using speaker recognition systems. One common goal of speech generative models is to produce a consistent voice that closely matches a reference voice [10, 11, 12, 13]. Therefore, part of the evaluation of these systems is to verify the speaker similarity between the synthesised speech and the reference voice. An objective verification method involves extracting speaker feature embeddings from both voices using a speaker recognition model and measuring their cosine similarity [14, 15, 16, 17]. However, this score often significantly deviates from what humans perceive [18, 19]. As a result, the evaluation of synthesised speech relies heavily on subjective evaluation, a process that requires considerable time and resources.

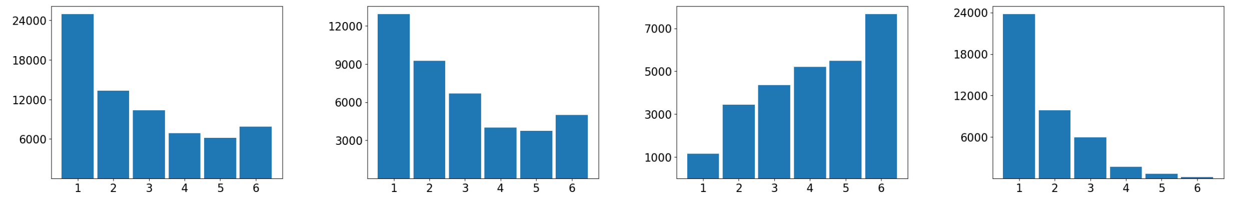
Table 1: *Data statistics for VCC2018, internal dataset from Deja et al. [19], and our VoxSim.* **# spks.**: *Total number of speakers.* **# pairs**: *Total number of utterance pairs.* **ratings**: *Total number of similarity ratings.* **Unseen test spks.**: *Whether the speakers in the test split were unseen during training.* **Public**: *Whether the dataset is public.*

Dataset	# spks.	# pairs	# ratings	Unseen test spks.	Public
VCC2018	12	21,562	30,864	✗	✓
Deja et al. [19]	13	18,493	730,308	✗	✗
VoxSim	1,251	41,578	69,409	✓	✓

To address this problem, techniques to automate speaker similarity assessment for synthetic speech are needed but have not been well-explored. To the best of our knowledge, there have been only two attempts at this automation. SVSNet [18] is the first end-to-end neural network model designed to evaluate speaker similarity between converted and natural speech in a voice conversion task. SVSNet takes raw waveforms as input rather than using handcrafted features to analyse speech more accurately and introduces a co-attention mechanism to resolve length and content mismatches between the two voices. This structure achieves an utterance-level linear correlation coefficient of 0.574 on the VCC2018 [20] speaker similarity evaluation dataset, released by the voice conversion challenge. Deja et al. [19] propose an automated method to evaluate speaker similarity by extending the speaker verification system. They synthesise speech samples using 354 modern text-to-speech systems and collect MUSHRA scores [21] from listening tests to build their own dataset. The authors build a regression model to predict MUSHRA speaker similarity scores from speaker embeddings of two utterances and propose a loss to compensate for data imbalance.

The primary challenge in developing automated speech evaluation models is the scarcity of public data [22, 23]. The VCC2018 dataset, utilised for training SVSNet, comprises 30,864 speaker similarity scores for 21,562 pairs of converted and natural utterances from 36 voice conversion systems. Each pair is evaluated by 1 to 8 subjects through crowdsourced listening tests. In Deja et al.'s study, the model is trained and evaluated on data collected by the authors, which has not been made publicly available. The training data are limited not only in size but also in the diversity of speakers, with both datasets featuring a very small number of speakers. The VCC2018 dataset

The dataset is available from https://mm.kaist.ac.kr/projects/voxsim



(a) individual scores (b) average scores (c) same speaker pairs (d) different speaker pairs

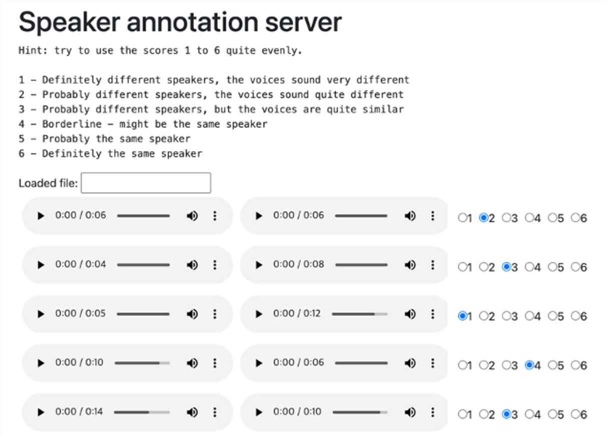
Figure 1: *Distributions of (a) individual scores, (b) average scores per utterance pair, (c) scores for the same speaker pairs, and (d) scores for different speaker pairs. The x-axis represents the score label and the y-axis represents the number of ratings.*

includes 12 speakers and 32 speaker combinations, whereas the other dataset contains only 13 target speakers. This limitation suggests that the trained similarity prediction models may not perform well on speech pairs from unseen speakers and are not suitable for evaluating zero-shot systems [24, 25]. Moreover, both datasets only consist of pairs of natural speech and speech synthesised by a few selected speech synthesis systems, indicating a lack of generalisability across different domains.

In this work, we introduce VoxSim, a large-scale open dataset that evaluates cognitive speaker similarity scores for speech pairs. To the best of our knowledge, this is the first dataset specifically crafted for training models that automate voice similarity assessment. VoxSim consists of approximately 70k similarity ratings from over 1k speakers. Since the utterances are sampled from VoxCeleb1 [6], the evaluation model is exposed to a variety of channel effects and noise during training, enhancing its generalisation performance across speech domains. Table 1 compares VoxSim with the previous two datasets in terms of the number of speakers, pairs, ratings, the unseen status of test speakers, and their availability to the public. We provide baseline results for various speaker similarity prediction models on the VoxSim dataset and demonstrate the generalisability of our data through testing on the VCC2018 dataset.

2. VoxSim Dataset

Data source. Speaker similarity ratings are collected using utterances from VoxCeleb1, which serves as the benchmark dataset for speaker identification and verification tasks. It consists of utterances extracted from videos uploaded to YouTube, thus the speech segments contain various acoustic environments. The speakers encompass a wide range of nationalities and ages. We create 50k random utterance pairs and conduct a listening test to evaluate the voice similarity. Diverse speaker combinations are created by organising speaker pairs to ensure a 1.5 times higher number of different speaker pairs compared to the same speaker pairs.



Speaker annotation server

Hint: try to use the scores 1 to 6 quite evenly.

1 – Definitely different speakers, the voices sound very different
2 – Probably different speakers, the voices sound quite different
3 – Probably different speakers, but the voices are quite similar
4 – Borderline – might be the same speaker
5 – Probably the same speaker
6 – Definitely the same speaker

Loaded file:

0:00 / 0:06 — 0:00 / 0:06 — ○1 ●2 ○3 ○4 ○5 ○6
0:00 / 0:04 — 0:00 / 0:08 — ○1 ○2 ●3 ○4 ○5 ○6
0:00 / 0:05 — 0:00 / 0:12 — ●1 ○2 ○3 ○4 ○5 ○6
0:00 / 0:10 — 0:00 / 0:06 — ○1 ○2 ○3 ●4 ○5 ○6
0:00 / 0:14 — 0:00 / 0:10 — ○1 ○2 ●3 ○4 ○5 ○6

Figure 2: *Speaker similarity annotation page.*

Table 2: *VoxSim Train and Test set statistics.* **# spks.**: *Total number of speakers.* **# spk combs.**: *Total number of speaker combinations.* **# pairs**: *Total number of utterance pairs.* **# ratings**: *Total number of similarity ratings.*

Set	# spks.	# spk combs.	# pairs	# ratings
Train	1,142	24,764	38,802	63,845
Test	109	904	2,776	5,564
Total	1,251	25,668	41,578	69,409

Listening test procedure. Twelve evaluators participate in this listening test and they are asked to evaluate the speaker similarity of each pair on a 6-point Likert scale; *1: Definitely different speakers, 2: Probably different speakers, 3: Possibly different speakers, 4: Possibly the same speaker, 5: Probably the same speaker, and 6: Definitely the same speaker*. Evaluators are requested to rate as evenly as possible from 1 to 6 points, aiming to prevent the dataset's score distribution from being skewed towards either 1 or 6. Given the diverse environmental and linguistic contexts present in VoxCeleb dataset, the evaluators are guided to focus solely on the voice characteristics of the main speaker, regardless of the content of the utterance, language, and the acoustic environment. An example annotation page is shown in Fig. 2.

Quality control. During the listening test, the quality of the collected scores is controlled through periodic reviews of the speaker verification rate. Min-max normalisation is utilised to project the scores between 0 and 1, enabling the calculation of the EER against the actual speaker labels. If an evaluator's EER significantly deviates from the average EER of all evaluators, we provide feedback and ask the evaluator to conduct a re-assessment. At the end of the collection, the average speaker verification rate for all evaluators is an EER of 17.7%. After the listening test, pairs with a difference of more than 3 points between the highest and lowest evaluation scores are considered outliers and excluded.

Data statistics. The total number of speakers in the collected data is 1,251, which matches the total number of speakers in VoxCeleb1. Train and test sets are structured to include distinct speakers, thereby ensuring the trained models' generalisation capabilities for unseen speakers. This division results in 1,142 speakers for training and 109 speakers for testing. After segregating the speakers, refining the collected scores yields 69,409 ratings for 41,578 pairs. In summary, dataset statistics and the train/test split are provided in Table 2, and the distributions of scores are illustrated in Fig. 1.

Table 3: *VoxSim test set results.* pt.: *pre-train.* ft.: *fine-tune.*

Model	LCC ↑	SRCC ↑	R2 ↑	MSE ↓	ACC ↑
ECAPA-TDNN					
pt. speaker recogniser	0.768	0.758	0.521	1.471	0.316
↳ ft. on individual scores	0.827 ±0.002	0.824 ±0.004	0.681 ±0.003	0.981 ±0.008	0.412 ±0.003
↳ ft. on average scores	**0.829** ±0.001	**0.828** ±0.001	**0.685** ±0.001	**0.967** ±0.002	**0.419** ±0.006
WavLM-ECAPA					
pt. speaker recogniser	0.752	0.736	0.505	1.520	0.306
↳ ft. on individual scores	0.833 ±0.001	0.835 ±0.000	0.690 ±0.002	0.951 ±0.005	0.402 ±0.007
↳ ft. on average scores	**0.835** ±0.002	**0.836** ±0.001	**0.693** ±0.003	**0.943** ±0.010	**0.405** ±0.004
SVSNet					
train on individual scores	**0.758** ±0.001	**0.753** ±0.002	**0.549** ±0.003	**1.384** ±0.009	**0.397** ±0.006
train on average scores	0.747 ±0.006	0.742 ±0.006	0.530 ±0.018	1.443 ±0.054	0.378 ±0.005

3. Experimental Setup

3.1. Model architectures

We adopt three model architectures for speaker similarity prediction experiments. ECAPA-TDNN [5] is a state-of-the-art model designed for automatic speaker verification. Given the attempts [26, 27] to apply self-supervised learning (SSL) based models to develop Mean Opinion Scores (MOS) prediction models to enhance generalisation performance across multiple speech datasets, WavLM-ECAPA is adopted. This model uses SSL-based WavLM [8] as a feature encoder. Finally, we also experiment with SVSNet [18], the only publicly available model specifically designed for predicting speaker similarity.

ECAPA-TDNN. ECAPA-TDNN enhances the traditional Time Delay Neural Network (TDNN) architecture by incorporating Squeeze-and-Excitation (SE) blocks to recalibrate channel-wise feature responses dynamically. Additionally, the model employs a multi-layer feature aggregation mechanism that enhances its ability to capture speaker characteristics across various temporal resolutions. This architecture leverages the power of attention mechanisms and convolutional layers to achieve superior performance in speaker verification.

WavLM-ECAPA. In Chen et al. [7], the authors leverage speech representations extracted from SSL-based pre-trained models for automatic speaker verification. The integration of a pre-trained feature encoder and ECAPA-TDNN downstream network demonstrates significant improvement in verification performance. In our experiments, we use WavLM as the feature extraction model, which has shown strong performance in several speech processing tasks. WavLM is trained with masked speech denoising and prediction in the pre-training, which makes it robust in complex acoustic environments and effective in preserving speaker identity.

SVSNet. SVSNet takes raw waveforms as input to fully utilise speech information for prediction and aligns the representations of the two inputs in two directions through a co-attention module. The model can be trained either in a regression manner using an L2 loss or in a classification manner using a cross-entropy loss, with the regression approach being shown to be superior.

3.2. Implementation details

Our implementation is based on the PyTorch [28] framework and is trained on an NVIDIA RTX A6000 with 48GB of memory. During the training of ECAPA-TDNN and WavLM-ECAPA, a 3-second segment from each utterance is randomly sampled to form a batch. Additionally, for ECAPA-TDNN, an 80-dimensional filterbank is extracted as input. These models extract a 256-dimensional speaker embedding for each utterance and predict speaker similarity by computing the cosine similarity between the extracted embeddings. The predicted score is compared to a similarity label projected on a 0 to 1 scale, and the model is trained using MSE loss. To facilitate effective speaker feature extraction at the beginning of training, we use models pre-trained with a speaker identification approach on the VoxCeleb dataset. An Adam [29] optimizer is employed with an initial learning rate of 10^{-5}, which decreases by 5% at each epoch. For SVSNet, a regression-based model is adopted, following the experimental setup of the original paper [18][1]. Every experiment is conducted three times independently to reduce the impact of random initialisation, and the average and standard deviation of these experiments are reported.

3.3. Evaluation metrics

The evaluation is based on the model's predicted similarity score compared to the average similarity score of the utterance pair. The model is evaluated with the following metrics: linear correlation coefficient (LCC), Spearman's rank correlation coefficient (SRCC), coefficient of determination (R2), mean squared error (MSE), and accuracy (ACC). For accuracy, a prediction is considered a true positive if the predicted similarity score is within 0.5 from the ground-truth label.

4. Results

4.1. Results on VoxSim test set

We first compare ECAPA-TDNN, WavLM-ECAPA, and SVSNet on the VoxSim test set, training each model using either individual scores or average scores per utterance pair. ECAPA-TDNN and WavLM-ECAPA are fine-tuned from a pre-trained speaker recognition model, whereas SVSNet is trained from scratch as it is not specifically designed for speaker recognition. The experimental results are summarised in Table 3. Although the pre-trained speaker recogniser ECAPA-TDNN and WavLM-ECAPA exhibit strong performance on VoxCeleb1, with speaker verification EERs of 0.96% and 0.43%, respectively, they achieve LCCs of only 0.768 and 0.733 for speaker similarity prediction, showing much lower correlation than the models fine-tuned with the VoxSim train set. Notably, WavLM-

[1] https://github.com/n1243645679976/SVSNet

Table 4: *Speaker recognition pre-training ablation.*

Model	**LCC ↑**	**MSE ↓**
SVSNet	0.747 ±0.001	1.443 ±0.054
ECAPA-TDNN		
w/o speaker pre-train.	0.761 ±0.001	1.297 ±0.005
w/ speaker pre-train.	**0.829** ±0.001	**0.967** ±0.002
WavLM-ECAPA		
w/o speaker pre-train.	0.806 ±0.002	1.090 ±0.009
w/ speaker pre-train.	**0.835** ±0.002	**0.943** ±0.010

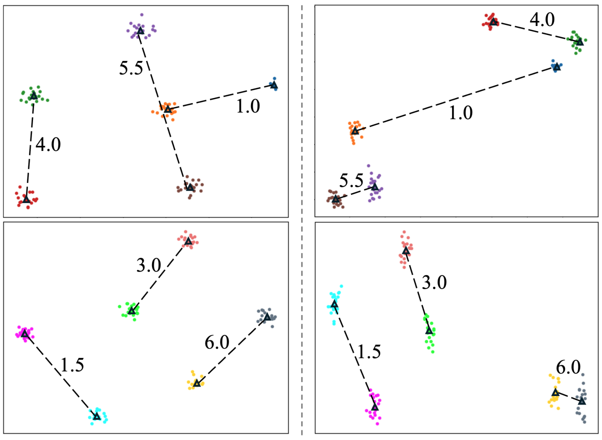


Figure 3: *t-SNE plot for embeddings extracted from (left) speaker recognition model and (right) speaker similarity prediction model. Different colours represent distinct speakers, and the number next to the dashed line represents the human-assessed speaker similarity between two utterances.*

ECAPA, despite its higher speaker verification performance, shows a lower speaker similarity prediction rate compared to ECAPA-TDNN. This suggests that the speaker recogniser may lose features related to speaker similarity while focusing on distinguishing between different speakers and aggregating embeddings of the same speaker. There is no clear superiority between models trained on individual scores and those trained on average scores. WavLM-ECAPA outperforms the others in all metrics except for accuracy, where ECAPA-TDNN achieves the highest accuracy.

Speaker recognition pre-training ablation. In particular, the fine-tuned ECAPA-TDNN and WavLM-ECAPA exhibit significant predictive performance compared to SVSNet, which can be attributed to the speaker recognition pre-training. To establish the effectiveness of speaker recognition pre-training, we train ECAPA-TDNN and WavLM-ECAPA from scratch. As shown in Table 4, both models perform similarly to SVSNet when trained without speaker recognition pre-training. However, this pre-training improves LCC by 8.9% and 3.6%, respectively. This demonstrates that speaker recognition pre-training significantly enhances speaker similarity prediction performance.

Qualitative results from t-SNE plot. To demonstrate that the speaker similarity prediction model fine-tuned on VoxSim captures perceptual similarity, we visualise the speaker embeddings extracted from the model using t-SNE [30] plots. To clearly show the effect of training, plots for embeddings extracted from the pre-trained speaker recognition model are also provided for comparison. As illustrated in Fig. 3, the speaker recognition model separates different speaker embedding clusters far apart regardless of perceived speaker similarity, whereas the similarity prediction model accurately reflects perceived speaker similarity. Furthermore, each speaker cluster in the speaker similarity prediction model exhibits soft boundaries and is relatively spread out. This indicates that the model has learnt that human-perceived speaker characteristics vary depending on the acoustic environment of the utterance.

Table 5: *VCC2018 test set results.*

Model	**LCC ↑**	**MSE ↓**
ECAPA-TDNN		
pt. speaker recogniser	0.512	1.090
↳ ft. on VCC2018	0.576 ±0.001	0.862 ±0.000
pt. on VoxSim	0.562 ±0.004	0.901 ±0.010
↳ ft. on VCC2018	**0.605** ±0.000	**0.806** ±0.001
WavLM-ECAPA		
pt. speaker recogniser	0.439	1.286
↳ ft. on VCC2018	0.594 ±0.000	0.828 ±0.002
pt. on VoxSim	0.566 ±0.003	0.884 ±0.004
↳ ft. on VCC2018	**0.609** ±0.000	**0.800** ±0.000
SVSNet		
SVSNet [18]	0.574	0.844
SVSNet (Ours)	0.575 ±0.001	0.849 ±0.003
pt. on VoxSim	0.509 ±0.004	3.237 ±0.266
↳ ft. on VCC2018	**0.586** ±0.001	**0.839** ±0.006

4.2. Results on VCC2018 test set

To verify the utility of our dataset, we test the generalisability of the models trained on VoxSim using the VCC2018 dataset, which contains natural and synthesised speech from voice conversion systems. Experiments are conducted in both zero-shot and fine-tuning settings, using the same train and test sets as those in the original SVSNet [18] paper. Table 5 shows the results on the VCC2018 test set. The fine-tuning results show that all models pre-trained with VoxSim outperform those trained solely on VCC2018. ECAPA-TDNN and WavLM-ECAPA with VoxSim pre-training demonstrate LCC improvements of 5.0% and 2.5%, respectively, over models trained without pre-training. Interestingly, models trained solely on VoxSim perform similarly to those trained on VCC2018, with scores of 0.576 vs. 0.562 for ECAPA-TDNN and 0.594 vs. 0.566 for WavLM-ECAPA. This demonstrates the excellent generalisability of models trained with VoxSim.

5. Conclusion

We present VoxSim, a speaker similarity evaluation dataset featuring nearly 70k scores for 41k pairs of utterances. This is the first large-scale dataset specifically collected to develop speaker similarity prediction models. Our dataset includes 1,251 speakers, and the utterances span a wide range of acoustic environments and contents. We collect speaker similarity scores from 12 evaluators through listening tests and document a detailed test design for data reliability. We provide baseline results for three speaker similarity prediction model architectures on VoxSim and demonstrate their generalisability through zero-shot and fine-tuning experiments on VCC2018 data.

6. Acknowledgement

This work was supported by Samsung Research.

7. References

[1] M. I. Jordan and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. ACCV*, 2017.

[4] J.-H. Kim, J. Kim, and J. S. Chung, "Let There Be Sound: Reconstructing High Quality Speech from Silent Videos," in *Proc. AAAI*, 2024.

[5] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017.

[7] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-Scale Self-Supervised Speech Representation Learning for Automatic Speaker Verification," in *Proc. ICASSP*, 2022.

[8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[9] J. Kang, J. Huh, H. S. Heo, and J. S. Chung, "Augmentation Adversarial Training for Self-Supervised Speaker Representation Learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1253–1262, 2022.

[10] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in *Proc. ICML*, 2022.

[11] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "Naturalspeech 2: Latent Diffusion Models Are Natural and Zero-Shot Speech and Singing Synthesizers," in *Proc. ICLR*, 2024.

[12] H.-S. Choi, J. Yang, J. Lee, and H. Kim, "NANSY++: Unified Voice Synthesis with Neural Analysis and Synthesis," in *Proc. ICLR*, 2023.

[13] J. Choi, J. Hong, and Y. M. Ro, "DiffV2S: Diffusion-based Video-to-Speech Synthesis With Vision-Guided Speaker Embedding," in *Proc. CVPR*, 2023.

[14] J.-H. Kim, H.-S. Yang, Y.-C. Ju, I.-H. Kim, and B.-Y. Kim, "CrossSpeech: Speaker-Independent Acoustic Representation for Cross-Lingual Speech Synthesis," in *Proc. ICASSP*, 2023.

[15] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale," in *Proc. NeurIPS*, 2024.

[16] J. Kim, K. Lee, S. Chung, and J. Cho, "CLaM-TTS: Improving Neural Codec Language Model for Zero-Shot Text-to-Speech," in *Proc. ICLR*, 2024.

[17] Y. Jang, J.-H. Kim, J. Ahn, D. Kwak, H.-S. Yang, Y.-C. Ju, I.-H. Kim, B.-Y. Kim, and J. S. Chung, "Faces that speak: Jointly synthesising talking face and speech from text," in *Proc. CVPR*, 2024.

[18] C.-H. Hu, Y.-H. Peng, J. Yamagishi, Y. Tsao, and H.-M. Wang, "SVSNet: An End-to-End Speaker Voice Similarity Assessment Model," *Signal Processing Letters*, vol. 29, pp. 767–771, 2022.

[19] K. Deja, A. Sanchez, J. Roth, and M. Cotescu, "Automatic Evaluation of Speaker Similarity," in *Proc. Interspeech*, 2022.

[20] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods," in *Proc. Odyssey*, 2018.

[21] B. Series, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[22] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A Non-intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in *Proc. ICASSP*, 2021.

[23] S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, "Speechlmscore: Evaluating Speech Generation Using Speech Language Model," in *Proc. ICASSP*, 2023.

[24] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer Learning From Speaker Verification to Multispeaker Text-To-Speech Synthesis," in *Proc. NeurIPS*, 2018.

[25] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural Voice Cloning With a Few Samples," in *Proc. NeurIPS*, 2018.

[26] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization Ability of MOS Prediction Networks," in *Proc. ICASSP*, 2022.

[27] G. Maniati, A. Vioni, N. Ellinas, K. Nikitaras, K. Klapsas, J. S. Sung, G. Jho, A. Chalamandaris, and P. Tsiakoulis, "SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis," in *Proc. Interspeech*, 2022.

[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An Imperative Style, High-Performance Deep Learning Library," in *Proc. NeurIPS*, 2019.

[29] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2015.

[30] L. Van der Maaten and G. Hinton, "Visualizing Data Using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.