

Fréchet Oral Distance: A Coordinate-Free Metric and Benchmark for Speech-Driven Oral Animation

Jaehong Myung*

mjhsin@sogang.ac.kr

Department of Artificial Intelligence

Sogang University

Seoul, South Korea

Hyunjung Chung

phone1263@sogang.ac.kr

Department of Computer Science

Sogang University

Seoul, South Korea

Seungho Eum

seunghoeum@sogang.ac.kr

Department of Computer Science

Sogang University

Seoul, South Korea

Unsang Park[†]

unsangpark@sogang.ac.kr

Department of Computer Science

Department of Artificial Intelligence

Sogang University

Seoul, South Korea

ABSTRACT

Recently, speech-driven facial and oral animation technologies have been used in various fields such as education, speech therapy, and healthcare. But methods to objectively evaluate the quality of generated results are still limited. Accordingly, we propose the Fréchet Oral Distance (FOD), which uses embeddings from an Oral Motion Encoder (OME) and a TIMIT-Vis reference distribution. In experiments, FOD reproduced model rankings similar to the coordinate-based metric, Lip Vertex Error (LVE), and showed trends consistent with user preference. Our framework provides objective and reproducible benchmarking without relying on expensive equipment or subjective evaluation, and it can be extended to compare a variety of 2D/3D models.

1 INTRODUCTION

Speech-driven facial animation generation technologies have recently been used in fields such as education and healthcare [12, 17]. When used for language therapy such as pronunciation correction, it is important to represent oral motion naturally and accurately.

While methods for modeling facial animation continue to advance, quantitative evaluation of generated videos remains insufficient. Current evaluations fall into coordinate-based, subjective, and distribution-based approaches. Coordinate-based metrics (e.g., LVE [5, 16, 23]) require precise equipment or manual landmarks, which are hard to obtain. Subjective evaluation (e.g., MOS [19]) is time- and cost-intensive and has low reproducibility. As a result, there is a lack of methods to measure the quality of oral animation accurately and reproducibly.

To address these issues, this study proposes the following contributions:

- **Fréchet Oral Distance (FOD):** a metric that evaluates quality without coordinate GT by converting generated oral animation into latent vectors and comparing distances between distributions.

- **TIMIT-Vis:** a large-scale articulatory simulation dataset that visualizes TIMIT with the Vocal Tract Lab (VTL) [1, 7], providing a standardized evaluation environment with diverse speakers and dialects. We use *TIMIT-Vis (test)* as the reference distribution because it offers stable, audio–video-synchronized articulatory coverage across diverse speakers and dialects.

2 RELATED WORK

Among evaluation methods for generated video, the widely used coordinate-based metric LVE [23] requires GT coordinates or equipment-based tracking, making data collection difficult. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [22] evaluate frame-wise distortion and are useful indicators of restoration quality, but they have limitations in assessing temporal information or the naturalness of motion. Subjective evaluation such as MOS (Mean Opinion Score) [19] has the advantage of reflecting human perception, yet it is costly and has low reproducibility.

In distribution-based evaluation of generated video, Fréchet-based evaluation methods are widely used. Fréchet Inception Distance (FID) [9] approximates the embedding distributions extracted from Inception-V3 [20] as multivariate Gaussians and computes the Fréchet distance. Fréchet Video Distance (FVD) [21] extends FID to video by computing the distance between two distributions using I3D embeddings [2]. Domain-specific Fréchet-based metrics have also been proposed. For example, in audio, Fréchet Audio Distance (FAD) [10] uses VGGish embeddings [8]; for human motion sequences, Fréchet Motion Distance (FMD) [11] applies Fréchet computation to motion-latent embeddings; and for 3D point clouds, Fréchet Pointcloud Distance (FPD) [18] was proposed in the context of TreeGAN. In this way, encoders suitable for each domain are used to evaluate at the distribution level rather than via frame-wise distortion.

Latent-space L1/L2 measure average deviations between embeddings. By contrast, Fréchet-based metrics (e.g., FID [9], FVD [21]) compare distributions by matching both means and covariances, capturing channel correlations, scale differences, and variability. Building on this perspective, our Fréchet Oral Distance (FOD) is

*Lead author.

[†]Corresponding author.

© 2025 Copyright held by the owner/author(s).

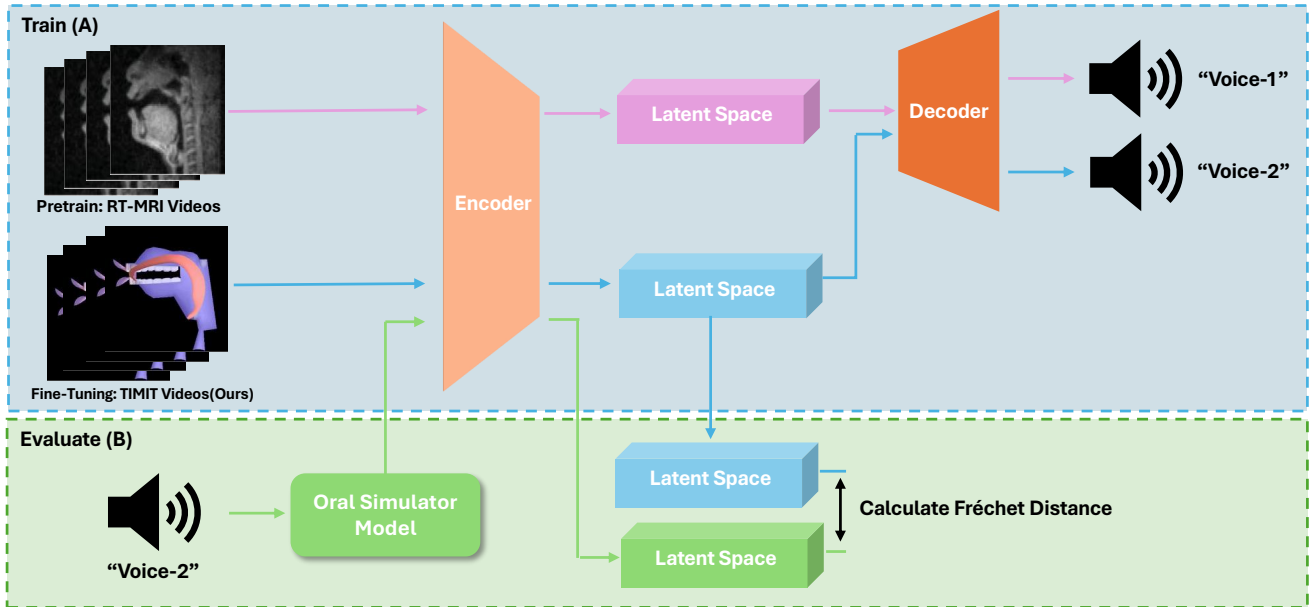


Figure 1: (A) Training pipeline: The Oral Motion Encoder is trained using synthetic articulatory videos generated from *TIMIT-Vis* [1, 7] (VTL-rendered TIMIT). **(B) Evaluation pipeline:** Latent vectors extracted from simulation videos are compared to reference distributions using the Fréchet distance [9]. These two stages collectively define the framework for both learning and assessing articulatory dynamics.

a distribution-based approach configured for the domain of oral animation.

To apply and benchmark such distribution-based metrics in our domain, a suitable articulatory dataset is necessary. RT-MRI [17] allows direct observation of internal articulators (e.g., tongue, velum) with temporal coupling to speech, yet publicly available datasets remain limited in scale and scope. VOCA [3] and BIWI [6] are 4D face datasets paired with audio and offer precise topological synchronization, but they lack internal oral information. This motivates the use of *TIMIT-Vis*, which visualizes TIMIT with the Vocal Tract Lab (VTL) [1, 7], providing audio–video-synchronized articulatory coverage across diverse speakers and dialects. These limitations restrict such datasets utility as training data for encoders that require internal oral-motion information.

3 METHOD

We propose a Fréchet-based metric that compares the latent distributions of generated facial and oral animations. Generated videos are embedded with an Oral Motion Encoder (OME), and detailed evaluation is defined in §3.2.

3.1 OME \times TIMIT-Vis

OME encodes oral motion from generated animations into a latent space using Lip2Wav’s spatiotemporal encoder [15]. It is pretrained on RT-MRI [17] (resampled 83→30 fps) and fine-tuned on *TIMIT-Vis* [1, 7] with the same configuration. Each input covers about 1.5 s (≈ 45 frames) in a sliding window. The overall training process is illustrated in Figure 1A.

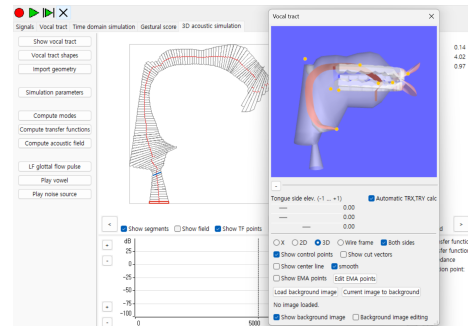


Figure 2: VocalTractLab (VTL) [1] interface. The GUI visualizes articulatory structures and computes acoustic responses; VTL was used to render videos for *TIMIT-Vis* [1, 7].

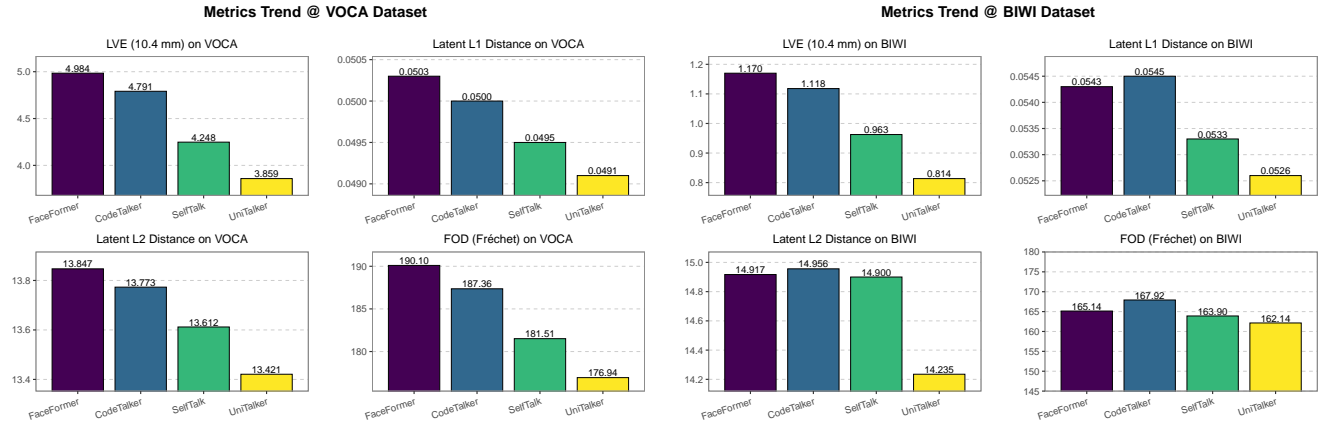
TIMIT-Vis visualizes TIMIT [7] speech via VTL [1], rendering articulator motion at 30 fps and 256×256. It includes 6,200/100 train/test videos from 630 speakers (8 dialects; 192 F / 438 M) (Table 1). The dataset is used for OME training and to build the reference latent distribution. VTL was modified to cover English/German IPA, and its rendering interface is shown in Figure 2.

Table 1: Summary of the TIMIT-Vis benchmark dataset [1, 7]. 6,200/100 videos (train/test) at 30 FPS, 256×256; 630 speakers, 8 dialect regions (192 female, 438 male).

Name	TIMIT-Vis	FPS	30
Resolution	256×256	Regions	8
Samples (Train/Test)	6,200/100	F/M	192/438

Table 2: Absolute scores on VOCA/BIWI: LVE (when available) and latent-space L1/L2/FOD. Lower indicates better alignment.

Metric		FaceFormer [5]		CodeTalker [23]		SelfTalk [14]		UniTalker [13]	
		VOCA [3]	BIWI [6]	VOCA [3]	BIWI [6]	VOCA [3]	BIWI [6]	VOCA [3]	BIWI [6]
LVE		4.98	1.17	4.79	1.12	4.25	0.96	3.86	0.81
Ours	L1	0.0502	0.0542	0.0501	0.0545	0.0494	0.0532	0.0491	0.0526
	L2	13.85	14.92	13.77	14.96	13.61	14.90	13.42	14.23
	Fréchet-D	190.10	165.14	187.36	167.92	181.51	163.90	176.94	162.14

**Figure 3: Trend visualization across datasets: LVE, latent L1/L2 distances, and FOD on VOCA and BIWI [3, 6]. The figure summarizes ranking and separability trends (lower is better), complementing the absolute values in Table 2.**

3.2 Fréchet Oral Distance (FOD)

FOD is the Fréchet distance [9] between two distributions: OME latents from generated videos and the TIMIT-Vis (test) [1, 7] reference. We estimate means and covariances for both and then compute the distance. Lower values indicate better alignment with the reference articulatory distribution.

Evaluation procedure.

- Step 1:** Build the reference once using *TIMIT-Vis (test)*: extract OME representations and record their summary statistics (mean and covariance).
- Step 2:** For each method, generate evaluation videos, extract OME embeddings from its outputs, and compute the method’s summary statistics (mean and covariance).
- Step 3:** Compare the method’s summary to the fixed reference with the Fréchet Oral Distance and report the score (lower is better).

Quantitative results are reported in a leaderboard format. These steps correspond to the evaluation pipeline depicted in Fig. 1B.

4 EXPERIMENTS

4.1 Experimental Setup

Data & Reference. All methods generate rendered videos from the same 100 test audios. We embed the videos with OME and compare the latents to the fixed *TIMIT-Vis (test)* [1, 7] reference. Baselines use publicly released VOCA [3]/BIWI [6]-pretrained weights and are evaluated in batch with the same pipeline.

Encoder Training. The OME is pretrained on RT-MRI (from 83 fps to 30 fps) and then fine-tuned on *TIMIT-Vis-train* with the same input configuration (about 1.5 seconds; ≈ 45 frames) [1, 7, 17]. To

reduce accumulated errors on long sequences, the input length is limited to about 1.5 seconds, and audio–video pairs are randomly shuffled.

System Configuration. The system environment used for training and evaluation is summarized in Table 3.

Table 3: Experimental system configuration for training and evaluation.

OS	Ubuntu 22.04	Python	3.10.12
CPU	Xeon Gold 6248R	Torch	2.0.1
GPU	RTX-A6000	CUDA	12.1

Baselines. We include FaceFormer [5], CodeTalker [23], SelfTalk [14], and UniTalker [4] using public pretrained models. All methods are tested on 100 unused audios from TIMIT-Vis, and evaluation is conducted with the same pipeline.

Metrics. We report FOD as the primary metric and additionally present *latent L1/L2 distances* and LVE. In this paper, latent L1/L2 distances are used as simple baselines that measure the average deviation between embeddings, whereas FOD is a distributional distance that compares the mean and covariance of OME representations against the fixed TIMIT-Vis (test) [1, 7] reference distribution. When the distribution is relatively simple and visibility is good, L1/L2 can also be effective; when channel correlations, scale, or variance structure are present, FOD tends to provide a more robust separation.

4.2 Results

Main Results. Evaluating VOCA [3] and BIWI [6] together, we find that on VOCA the FOD—computed as the Fréchet distance

between each model and the fixed TIMIT-Vis (test) [1, 7] reference statistics—reproduces nearly the same ranking as LVE. The auxiliary baselines (latent L1/L2 distances) show similar trends. In contrast, on BIWI the agreement between FOD and LVE is weaker and the separability is smaller. We attribute this to BIWI’s side-view rendering, which lowers the visibility of lip contours and the lips, making it difficult for the OME to stably extract speech-related lip motion. As a result, on BIWI, FOD does not align with LVE to the same extent, with several mid-tier models swapping adjacent order (see Fig. 3 for interpretation and Table 2 for the numbers).

Qualitative Analysis. Some OME latent channels respond to macroscopic motions (e.g., lip opening/closing and anterior–posterior lip protrusion), whereas others are sensitive to microscopic timing. Several channels directly reflect oral contours or static shape cues, indicating that the encoder captures not only articulatory motion but also structural characteristics. As shown in the activation maps in Fig. 4, these patterns are clearly observed.

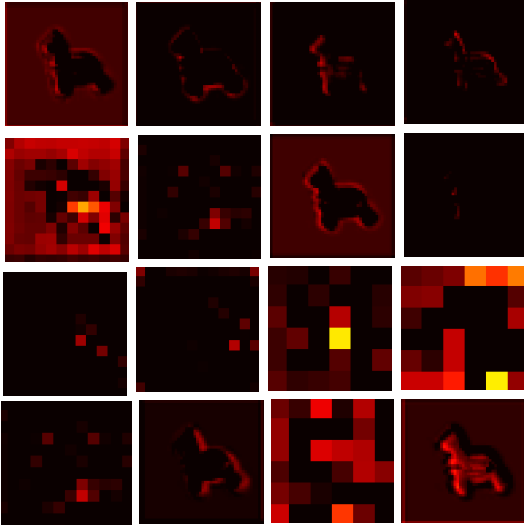


Figure 4: Activation maps of 16 latent channels (reshaped to $H \times W$). Responses highlight both articulatory motion (e.g., lips, tongue) and static oral structure, indicating joint encoding of dynamics and shape.

Table 4: User study (from CodeTalker [23]) on BIWI-Test-B and VOCA-Test using A/B testing. Values are percentages.

Competitors	BIWI-Test-B		VOCA-Test	
	Lip Sync	Realism	Lip Sync	Realism
CodeTalker vs. VOCA	92.47	89.25	86.02	84.95
CodeTalker vs. MeshTalk	80.65	82.80	95.70	92.47
CodeTalker vs. FaceFormer	53.76	56.99	70.97	69.89
CodeTalker vs. GT	43.01	49.46	43.01	43.01

Human Evaluation Results. Participants generally preferred models with lower LVE. The same ordering appears with FOD, indicating qualitative agreement with coordinate-based and perceptual judgments [23]. *The detailed A/B outcomes are summarized in Table 4.* While FOD tracks LVE on VOCA, we stop short of making perceptual claims; direct user studies tailored to FOD are deferred to future work.

5 DISCUSSION

5.1 Limitations

Our study has limits. Ablations of the OME are missing, so impacts of convolution depth, self-attention, and latent size on FOD are unknown. FOD assumes Gaussian latents; sensitivity to higher-order moments or KDE is unverified. Latents also encode static shape cues, implying possible shape-similarity effects on FOD. Evaluation is confined to English TIMIT-based data, leaving tonal languages and special styles (rate changes, whispering, singing) open. The fixed reference from TIMIT-Vis (test), while disjoint from OME training, makes FOD corpus-specific.

5.2 Future Directions

Next steps include architecture/training ablations to quantify FOD stability and discriminability; exploring higher-order statistics, kernel bandwidths, and alternative two-sample distances (e.g., MMD) in Fréchet [9] computations; and enforcing disentanglement (synthetic controls, channel masking, orthogonality) to separate motion from shape. We aim to extend *TIMIT-Vis* [1, 7] to multilingual, multi-speaker and clinical use, and to run user studies [19]. View-invariant or multi-view processing could reduce BIWI side-view sensitivity.

6 CONCLUSION

We presented a framework for evaluating speech-driven facial animation. It combines a distribution-based metric (Fréchet Oral Distance, FOD) with TIMIT-Vis [1, 7], an articulatory simulation dataset with guaranteed audio–video synchronization. Together, they mitigate reliance on coordinate GT, reduce the instability of subjective evaluation, and address data-diversity gaps. In our experiments, FOD reproduced rankings aligned with LVE and showed higher discriminability than latent L1/L2 distances, and it also qualitatively agreed with user preference. In addition, channel activation visualizations support that OME encodes both motion and structural characteristics. TIMIT-Vis provides a stable reference distribution with diverse speakers and dialects and is meaningful as a standard public resource for future evaluation research. The proposed framework functions as an objective and reproducible benchmarking tool and offers extensibility to compare various 2D/3D generative models under a common standard.

ACKNOWLEDGMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00621, RS-2022- I220621, Development of artificial intelligence technology that provides dialog-based multi-modal explainability)

REFERENCES

- [1] Peter Birkholz. 2013. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS One* 8, 4 (2013), e60603.
- [2] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the Kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [3] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. 2019. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10101–10111.
- [4] Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. 2024. UniTalker: Scaling up audio-driven 3d facial animation through a unified model. In *European Conference on Computer Vision*. Springer, 204–221.
- [5] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. FaceFormer: Speech-driven 3D facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18770–18780.
- [6] Gabriele Fanelli, Thibaut Weise, Juergen Gall, and Luc Van Gool. 2011. Real time head pose estimation from consumer depth cameras. In *Joint Pattern Recognition Symposium*. Springer, 101–110.
- [7] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N 93* (1993), 27403.
- [8] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.
- [9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems* 30 (2017).
- [10] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharif. 2018. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466* (2018).
- [11] Antoine Maiorca, Youngwoo Yoon, and Thierry Dutoit. 2022. Evaluating the quality of a synthesized motion with the Fréchet motion distance. In *ACM SIGGRAPH 2022 Posters*. 1–2.
- [12] Simon Moxon and N. S. Thammarat. 2023. Pronunciation Coach 3D. *Comput Assist Lang Learn Electron J (CALL-EJ)* 24, 1 (2023), 205–221.
- [13] Le Thien Phuc Nguyen, Zhuoran Yu, Khoa Quang Nhat Cao, Yuwei Guo, Tu Ho Manh Pham, Tuan Tai Nguyen, Toan Ngo Duc Vo, Lucas Poon, Soochahn Lee, and Yong Jae Lee. 2025. UniTalk: Towards Universal Active Speaker Detection in Real World Scenarios. *arXiv preprint arXiv:2505.21954* (2025).
- [14] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. 2023. SelfTalk: A self-supervised commutative training diagram to comprehend 3D talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5292–5301.
- [15] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Nambodiri, and CV Jawahar. 2020. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13796–13805.
- [16] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. 2021. MeshTalk: 3D face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1173–1182.
- [17] Matthieu Ruthven, Agnieszka M Peplinski, David M Adams, Andrew P King, and Marc Eric Miquel. 2023. Real-time speech MRI datasets with corresponding articulator ground-truth segmentations. *Scientific Data* 10, 1 (2023), 860.
- [18] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 2019. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3859–3868.
- [19] Robert C Streijl, Stefan Winkler, and David S Hands. 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems* 22, 2 (2016), 213–227.
- [20] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.
- [21] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).
- [22] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
- [23] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. CodeTalker: Speech-driven 3D facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12780–12790.