Learning Phonetic Context-Dependent Viseme for Enhancing Speech-Driven 3D Facial Animation

Anonymous submission to Interspeech 2025

Abstract

Speech-driven 3D facial animation aims to generate realistic facial movements synchronized with audio. Traditional methods primarily minimize reconstruction loss by aligning each frame with ground-truth. However, this frame-wise approach often fails to capture the continuity of facial motion, leading to jittery and unnatural outputs due to coarticulation. To address this, we propose a novel phonetic context-aware loss, which explicitly models the influence of phonetic context on viseme transitions. By incorporating a viseme coarticulation weight, we assigns adaptive importance to facial movements based on their dynamic changes over time, ensuring smoother and perceptually consistent animations. Extensive experiments demonstrate that replacing the conventional reconstruction loss with ours improves both quantitative metrics and visual quality. It highlights the importance of explicitly modeling phonetic context-dependent visemes in synthesizing natural speech-driven 3D facial animation.

Index Terms: Speech-driven 3D Facial Animation, Phonetic Context, Coarticulation

1. Introduction

Speech-driven 3D facial animation aims to predict realistic 3D facial deformation fields, which change a given static facial mesh template, synced with input audio. As this task is often regarded as a key generative AI technology for immersive applications such as VR remote presence, filmmaking, and game character animation [1, 2, 3, 4], it has been actively explored in recent speech and vision research. To produce natural speech-driven 3D facial animation, a generative model is required not only to represent the 3D facial dynamics in depth beyond the vertex or mesh level, but also to understand the visible movements of vocal tract articulators during speech production [5].

We focus on addressing *coarticulation* issue for realistic speech-synchronized 3D facial animation. Coarticulation is a phenomenon where the articulation of a speech segment is influenced by both the preceding segment (backward coarticulation) and the following segment (forward coarticulation), resulting in smoother and more natural speech transitions [6]. As an example, Fig. 1 shows that the lip movements for the word “A” vary according to the surrounding words spoken after it. When pronouncing “A crab”, the lips tend to close after saying “A”. On the other hand, the lips form a widely open shape when saying “A calico”. Each viseme corresponding to a phoneme emerges gradually rather than appearing abruptly, ensuring smooth transitions between visemes. That is, a viseme is influenced not only by the phoneme being uttered but also by its surrounding phonetic context due to the continuous lip motion and inertia of the articulators, i.e., *phonetic context-dependent viseme* [7, 8].

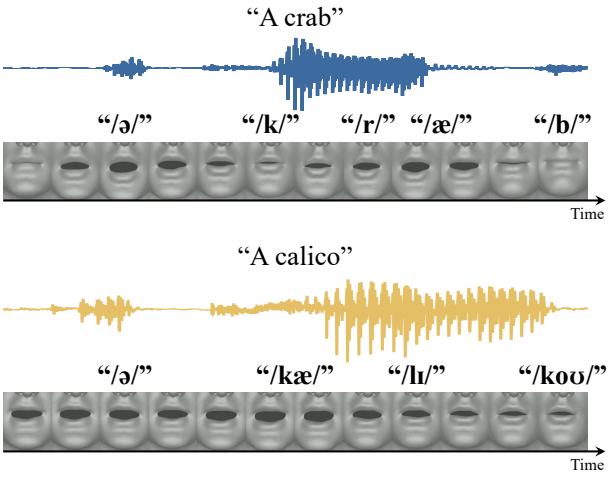

Figure 1: *Visualization of the audio and viseme when pronouncing the word "A crab" and "A calico". Note that the phoneme represents the corresponding dominant viseme.*

Previous works for speech-driven 3D facial animation used temporal-specific networks such as Temporal CNN (T-CNN), LSTM, and Transformer to capture the dynamics of audio [9, 10, 11, 12, 13, 14]. However, similar to motion estimation in 2D video, these methods primarily focus on reconstruction quality at each time step and facial dynamics. As a result, they often produce jittery frames and inaccurate lip synchronization because they do not explicitly account for the various smooth viseme transitions caused by coarticulation.

With the importance of coarticulation, in this paper, we propose a simple yet effective method that explicitly considers phonetic context in speech-driven 3D facial animation. To make the model learn the phonetic context-dependent viseme during training, we design a novel objective function for viseme coarticulation. Unlike the conventional reconstruction loss used in existing methods, we first quantitatively measure how much the facial vertices moved within a temporal window to consider the surrounding lip movements (i.e., phonetic context-dependent viseme). By normalizing it, we can quantify the extent of the relationship with the surrounding phonetic context on each viseme over time. We call it viseme coarticulation weight. Then, by applying our viseme coarticulation weight to the reconstruction loss, the phonetic context-aware loss can be defined. Extensive experiments demonstrate that learning the speech-driven facial animation models with our phonetic context-aware loss can improve their performances quantitatively and qualitatively.

Our contributions are summarized as follows:

- We introduce a simple but effective objective function named as *phonetic context-aware loss* to learn phonetic context-dependent viseme for addressing coarticulation issue and enhancing speech-driven 3D facial animation.
- We demonstrate the effectiveness of the proposed objective function by replace the conventional reconstruction loss in the recent models with the phonetic context-aware loss through extensive experiments on various datasets.

2. Method

In this section, we illustrate the proposed method with the common objective functions used in existing speech-driven 3D facial animator baselines such as FaceFormer [16] and CodeTalker [17]. This section consists of two subsections: 1) common objective functions in speech-driven 3D facial animator baselines as a preliminary and 2) learning phonetic context-dependent viseme to improve the baselines with the proposed phonetic context-aware loss.

2.1. Objective Functions in Speech-Driven 3D Facial Animator Baselines

Most existing speech-driven 3D facial animation models mainly focus on minimizing differences between the ground-truth and the predicted 3D facial vertices equally at each time step. To strictly align with the ground-truth for each time t, they employ the reconstruction loss $\mathcal{L}_{rec}$, which is the Euclidean distance between ground-truth and the generated 3D facial vertices.

$$\mathcal{L}_{rec} = \sum_{t=1}^{T} \left\| v^t - \hat{v}^t \right\|^2, \tag{1}$$

where v^t and $\hat{v}^t$ are ground-truth and the synthesized 3D facial vertices at time t, respectively. T is the length of the sequence. Note that $\hat{v}^t$ is obtained by adding the predicted 3D facial deformation fields $\hat{d}^t$ to a given static 3D facial template as in speech-driven 3D facial animator baselines [16, 17, 18, 19].

Utilizing the reconstruction loss $\mathcal{L}_{rec}$ alone for optimization can induce jittery output because it does not take into account facial dynamics in the time domain. In addition, it uniformly processes the local dynamics at each time t, meaning that it disregards the phonetic context.

To mitigate the problem of jittery output frames caused by relying solely on reconstruction loss $\mathcal{L}_{rec}$ [18], the velocity loss $\mathcal{L}_{vel}$ is generally employed to promote smoother and more natural lip movements over time. The velocity loss can be written as

$$\mathcal{L}_{vel} = \sum_{t=2}^{T} \left\| (v^t - v^{t-1}) - (\hat{v}^t - \hat{v}^{t-1}) \right\|^2. \tag{2}$$

2.2. Learning Phonetic Context-Dependent Viseme

The recent speech-driven 3D facial animators have demonstrated remarkable performances. However, it is insufficient to merely minimize the Euclidean distance between the ground truth and predicted facial vertices, as well as the dynamics of consecutive frames over time for producing clear and intelligible lip movements. To synthesize a realistic and natural 3D facial animation, we introduce a novel phonetic context-aware loss that attends to variations of facial movements in both the preceding speech segment (backward coarticulation and the following speech segment (forward coarticulation). In various audio-visual tasks, such as lip reading [7, 22] and 2D video-based speech recognition [23, 24], explicitly modeling context-dependent visemes has been shown to improve the robustness and accuracy of alignment between visual and auditory streams.

Based on these studies, to learn phonetic context-dependent viseme in a 3D facial animation, we design a novel phonetic context-aware loss $\mathcal{L}_{pc}$ that captures the characteristics of phonetic context-dependent visemes, where the articulation of a phone gradually varies depending on the preceding or following phones. To observe surrounding phones, we first define a temporal window size, which can be written as

$$\Omega_\sigma^t = \{k \mid t - \sigma \leq k \leq t + \sigma\},\ t = 1 + \sigma, \cdots, T - \sigma, \tag{3}$$

where t is the current frame index and T is the total number of frames. σ is the window radius that determines the temporal neighborhood around frame t (i.e., window size can be obtained by $2\sigma + 1$). In this work, the window size is used as 5 ($\sigma = 2$).

Then, we can measure the changes of the articulation of a phone at time t within the temporal window $|\Omega_\sigma^t|$ (i.e., small speech segment).

$$w^t = \frac{1}{|\Omega_\sigma^t|} \sum_{k \in \Omega_\sigma^t} \left|\left| v^k - v^{k-1} \right|\right|^2. \tag{4}$$

The measured w^t represents the extent of change in facial vertices within the vicinity of frame t due to coarticulation. By measuring the local facial and lip movement w^t around time t and normalizing it over time, we can approximately estimate the extent of the phonetic context-dependent viseme.

With the normalization, finally, the proposed viseme coarticulation weight $\tilde{w}^t$ can be defined as

$$\tilde{w}^t = \frac{\exp\left(w^t\right)}{\sum_{i=1}^{T} \exp\left(w^i\right)}. \tag{5}$$

The proposed viseme coarticulation weight $\tilde{w}^t$ focuses more on regions where articulatory movement changes relatively more due to coarticulation. By applying the weight $\tilde{w}^t$ to the conventional reconstruction loss $\mathcal{L}_{rec}$, we can define the proposed phonetic context-aware loss $\mathcal{L}_{pc}$.

$$\mathcal{L}_{pc} = \frac{1}{T} \sum_{t=1}^{T} \tilde{w}^t \cdot \left\| v^t - \hat{v}^t \right\|^2, \tag{6}$$

In training stage, by minimizing $\mathcal{L}_{pc}$, the model can generate more natural 3D facial animations with continuous and smooth transition.

3. Experiments

3.1. Experiments Settings

3.1.1. Datasets

In our experiments, we conduct extensive experiments on four widely-used datasets, which are VOCASET [15], BIWI [20], BIWI_6 [19, 20], and MultiFace [21]. These datasets include pairs of audio and the corresponding 3D facial scans that show the pronunciation of English speech.

VOCASET. VOCASET comprises a total of 480 facial motion sequences from 12 people, recorded at 60 frames per second for roughly 4 seconds each, and 255 distinct words, 5 sentences of which are shared by all speakers. Every 3D face model has 5,023 vertices and is registered to the FLAME [25] topology.

Dataset	Method	Objective Function (Original / Ours)			
		FVE ↓	LVE ↓	LDTW ↓	Lip-max ↓
VOCASET [15]	FaceFormer [16]	0.637 / **0.633**	0.414 / **0.397**	0.482 / **0.465**	0.618 / **0.608**
	CodeTalker [17]	0.690 / **0.642**	0.468 / **0.425**	0.496 / **0.464**	0.631 / **0.615**
	SelfTalk [18]	0.606 / **0.590**	0.395 / **0.368**	0.460 / **0.444**	0.591 / **0.587**
	ScanTalk [19]	0.624 / **0.590**	0.428 / **0.359**	0.507 / **0.460**	0.600 / **0.570**
BIWI [20]	FaceFormer [16]	0.992 / **0.964**	0.212 / **0.209**	**0.140** / **0.140**	0.375 / **0.373**
	CodeTalker [17]	0.979 / **0.929**	0.207 / **0.195**	0.143 / **0.139**	0.377 / **0.371**
	SelfTalk [18]	1.110 / **0.933**	0.240 / **0.201**	0.150 / **0.137**	0.401 / **0.364**
$BIWI_6$ [19, 20]	ScanTalk [19]	0.463 / **0.454**	0.114 / **0.112**	0.112 / **0.110**	0.837 / **0.827**
Multiface [21]	ScanTalk [19]	0.551 / **0.531**	0.100 / **0.092**	0.104 / **0.099**	0.821 / **0.802**

Table 1: *Quantitative evaluations for 3D facial animations before and after replacing the original reconstruction loss with our phonetic context-aware loss on the existing models (Original / Ours) format. Lower values indicate better performance.*

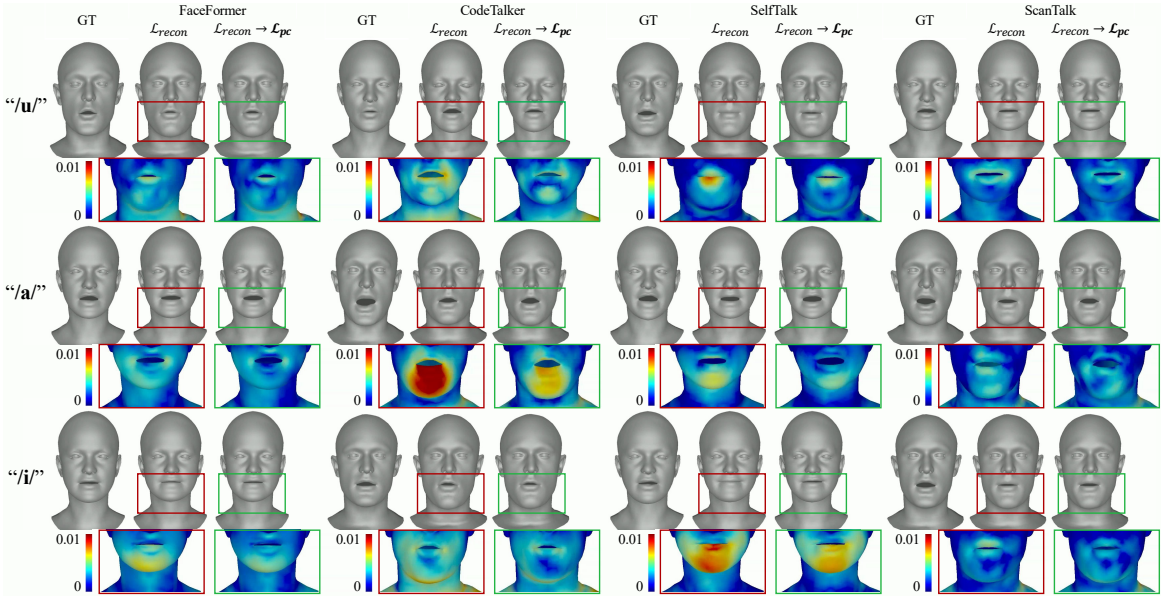

Figure 2: *Visual comparisons of baseline models vs. baseline models trained with our objective function on VOCASET. Note that the second, fourth, and sixth rows represent the visualization of per-vertex errors in lip regions.*

BIWI. BIWI consists of 40 unique sentences shared across all speakers. There are 40 sentences uttered by 14 subjects. A dynamic 3D facial scan at 25 frames per second is captured by repeating each recording twice in either an emotional or neutral setting. There are 23,370 vertices in the registered topology, and the average sequence length is approximately 4.67 seconds.

$BIWI_6$. $BIWI_6$ is a down-sampled version of BIWI with a fixed topology and 3,895 vertices and 7,539 faces used in [19].

Multiface. Multiface includes 13 persons delivering up to 50 utterances, each around 4 seconds in duration, sampled at 30 frames per second and possess a static topology including 5,471 vertices and 10,837 faces.

3.1.2. Models

To verify the effectiveness of the proposed phonetic context-aware loss, we employ four recent speech-driven 3D facial animator as our baseline models. When training the model with our objective function, the conventional reconstruction loss is replaced with our phonetic context-aware loss.

FaceFormer. FaceFormer [16] introduces a transformer-based 3D facial animator to capture the relationship between audio and previous motions to generate 3D facial animations.

CodeTalker. CodeTalker [17] discretely stores 3D motion priors using the VQ-VAE structure, which is known to be effective for image restoration.

SelfTalk. SelfTalk [18] can restore more realistic lip shapes by learning to make the lip-reading result distribution of the generated mesh similar to the ASR result distribution of the audio using the commutative diagram structure.

ScanTalk [19] introduces a diffusion-based approach to enable facial animation synthesis for various topologies with a single model regardless of the topology between various datasets.

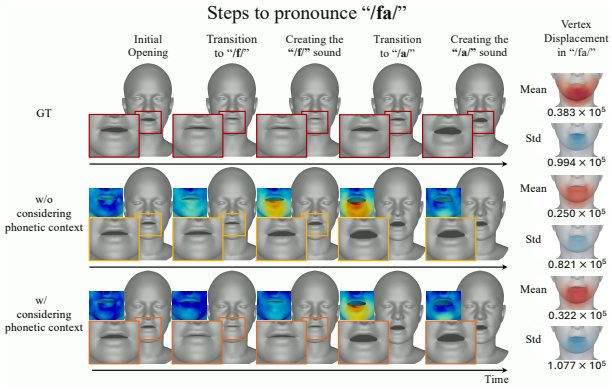

Figure 3: *Qualitative comparison of optimizing the baseline models with their original objective functions vs. with our phonetic context-aware loss instead of reconstruction loss.*

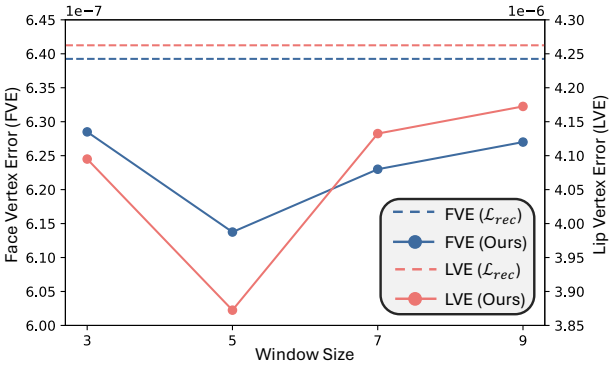

Figure 4: *Performance analysis according to different window sizes of the viseme coarticulation weight in FVE and LVE.*

3.1.3. Metrics

For quantitative evaluation, we employ four metrics, which are FVE, LVE, LDTW, and Lip-max.

Face Vertex Error (FVE). FVE quantifies the geometric discrepancy between a reference and the generated meshes by Euclidean distance for each vertex across the entire facial region.

Lip Vertex Error (LVE). LVE quantifies the Euclidean distance for each vertex specifically within the lip region.

Lip Dynamic Time Warping (LDTW). LDTW is an index that measures the similarity of lip movements over time using dynamic time warping (DTW) [26, 13] for temporal consistency.

Lip-max. Lip-max is a metric used in [27, 13] that calculates the largest vertex error within the lip region.

3.2. Quantitative Results

For performance evaluations of the proposed phonetic context-aware loss, we trained the baseline models with $\mathcal{L}_{pc}$ instead of $\mathcal{L}_{rec}$. Except for the reconstruction loss $\mathcal{L}_{rec}$, we did not change other objective functions used in the official code of each baseline model. To quantitatively evaluate the performance of the proposed method, we calculate FVE, LVE, LDW, and Lip-max as seen in Tab. 1. In our experiments, we follow the official split criteria used in each baseline model for training and inference. As seen in Tab. 1, the models trained with our phonetic context-aware loss achieved better performances, compared to the those of their original versions without any modifications. This means that, instead of treating the error of each facial vertex equally, considering the importance of changes in visemes based on phonetic context enables the creation of more natural speech-driven 3D facial animations.

3.3. Qualitative Results

Fig. 2 shows shows visual comparisons between the ground-truth, FaceFormer [16] w/ $\mathcal{L}_{rec}$ or $\mathcal{L}_{pc}$, CodeTalker [17] w/ $\mathcal{L}_{rec}$ or $\mathcal{L}_{pc}$, SelfTalk [18] w/ $\mathcal{L}_{rec}$ or $\mathcal{L}_{pc}$, and ScanTalk [19] w/ $\mathcal{L}_{rec}$ or $\mathcal{L}_{pc}$. This means that essential to consider phonetic context-dependent viseme and as a results, it enable achieve more highr vertex alignment.

Fig. 3 shows an examples of the pronunciation of "/fa". In Fig. 3, the sixth column visualizes the mean and variance of the movement of vertices at the time of the pronunciation. When considering phonetic context-dependent viseme, more accurate viseme can be generated in the transition region from one phoneme to another than when not considering context. Importantly, learning phonetic context-dependent viseme can produce more perceptually natural and smooth transition, compared to the existing models that did not account for coarticulation.

3.4. Ablation Study

We conduct performance analysis to demonstrate the impact of window size in the proposed viseme coarticulation weight. Specifically, we investigate both FVE and LVE on VOCASET [15] when we train baselines with or without the proposed phonetic context-aware loss according to different window sizes. In Fig. 4, the dotted line represents the average performance of original baseline models trained using their official code. In contrast, the solid line represents the average performance of the same baseline models trained with the our objective function instead of the reconstruction loss. As shown in Fig. 4, when considering the phonetic context in synthesizing facial animation, both FVE and LVE values are much lower than when it is ignored, regardless of window sizes (i.e., the length of small speech segment for phonetic context). In our experiments, we set the window size to 5, leading the lowest FVE and LVE.

4. Conclusion

In this paper, we introduced a phonetic context-aware loss to enhance the naturalness of speech-driven 3D facial animation by explicitly modeling viseme transitions influenced by phonetic context. Unlike existing approaches that treat all facial vertices equally in reconstruction loss, our method leverages viseme coarticulation weight, which assigns greater importance to regions with pronounced articulatory changes. Experimental results across multiple datasets confirm that our approach not only reduces prediction errors but also produces smoother and more realistic facial animations. Additionally, our ablation study shows that the window size for capturing phonetic context plays a crucial role in optimizing animation quality. These findings emphasize the necessity of incorporating phonetic context modeling in speech-driven 3D animation frameworks. Future work will explore extending our approach to more diverse linguistic contexts and integrating additional multimodal cues for further improvements.

5. References

[1] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[2] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM transactions on graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[3] H. K. Kim, S. Lee, and H. G. Kim, "Analyzing visible articulatory movements in speech production for speech-driven 3d facial animation," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 3575–3579.

[4] H. EunGi, O. Hyun-Bin, K. Sung-Bin, C. Nivelet Etcheberry, S. Nam, J. Ju, and T.-H. Oh, "Enhancing speech-driven 3d facial animation with audio-visual guidance from lip reading expert," in *Interspeech 2024*, 2024, pp. 2940–2944.

[5] J. M. De Martino, L. P. Magalhães, and F. Violaro, "Facial animation based on context-dependent visemes," *Computers & Graphics*, vol. 30, no. 6, pp. 971–980, 2006.

[6] S. J. Park, M. Kim, J. Choi, and Y. M. Ro, "Exploring phonetic context-aware lip-sync for talking face generation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4325–4329.

[7] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6447–6456.

[8] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8717–8727, 2022.

[9] K. D. Yang, A. Ranjan, J.-H. R. Chang, R. Vemulapalli, and O. Tuzel, "Probabilistic speech-driven 3d facial motion synthesis: New benchmarks methods and applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 294–27 303.

[10] H. Wu, S. Zhou, J. Jia, J. Xing, Q. Wen, and X. Wen, "Speech-driven 3d face animation with composite and regional facial movements," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 6822–6830. [Online]. Available: https://doi.org/10.1145/3581783.3611775

[11] S. Aneja, J. Thies, A. Dai, and M. Nießner, "Facetalk: Audio-driven motion diffusion for neural parametric head models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 263–21 273.

[12] X. Fan, J. Li, Z. Lin, W. Xiao, and L. Yang, "Unitalker: Scaling up audio-driven 3d facial animation through a unified model," in *European Conference on Computer Vision*. Springer, 2024, pp. 204–221.

[13] B. Thambiraja, I. Habibie, S. Aliakbarian, D. Cosker, C. Theobalt, and J. Thies, "Imitator: Personalized speech-driven 3d facial animation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 20 621–20 631.

[14] Z. Sun, T. Lv, S. Ye, M. Lin, J. Sheng, Y.-H. Wen, M. Yu, and Y.-j. Liu, "Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1–9, 2024.

[15] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. Black, "Capture, learning, and synthesis of 3D speaking styles," *Computer Vision and Pattern Recognition (CVPR)*, pp. 10 101–10 111, 2019. [Online]. Available: http://voca.is.tue.mpg.de/

[16] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong, "Codetalker: Speech-driven 3d facial animation with discrete motion prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 780–12 790.

[18] Z. Peng, Y. Luo, Y. Shi, H. Xu, X. Zhu, H. Liu, J. He, and Z. Fan, "Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, p. 5292–5301.

[19] F. Nocentini, T. Besnier, C. Ferrari, S. Arguillere, S. Berretti, and M. Daoudi, "Scantalk: 3d talking heads from unregistered scans," in *European Conference on Computer Vision*. Springer, 2024, pp. 19–36.

[20] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591–598, 2010.

[21] C.-h. Wuu, N. Zheng, S. Ardisson, R. Bali, D. Belko, E. Brockmeyer, L. Evans, T. Godisart, H. Ha, X. Huang, A. Hypes, T. Koska, S. Krenn, S. Lombardi, X. Luo, K. McPhail, L. Millerschoen, M. Perdoch, M. Pitts, A. Richard, J. Saragih, J. Saragih, T. Shiratori, T. Simon, M. Stewart, A. Trimble, X. Weng, D. Whitewolf, C. Wu, S.-I. Yu, and Y. Sheikh, "Multiface: A dataset for neural face rendering," in *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2207.11243

[22] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6319–6323.

[23] Y. Sun, H. Zhou, K. Wang, Q. Wu, Z. Hong, J. Liu, E. Ding, J. Wang, Z. Liu, and K. Hideki, "Masked lip-sync prediction by audio-visual contextual exploitation in transformers," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–9.

[24] A. Gupta, R. Tripathi, and W. Jang, "Modeformer: Modality-preserving embedding for audio-video synchronization using transformers," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[25] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[26] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[27] A. Richard, M. Zollhöfer, Y. Wen, F. de la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1153–1162.