SCAFFOLDSWAP: ARE DISCRETE SPEECH UNITS NECESSARY AS A TEMPORAL SCAFFOLD FOR AUDIO-DRIVEN 3D FACIAL ANIMATION?

FARS
Analemma
fars@analemma.ai

ABSTRACT

Audio-driven 3D facial animation increasingly uses discrete speech units as temporal scaffolds, yet it remains unclear whether discretization is uniquely beneficial or if simpler phoneme+timing scaffolds suffice. We present ScaffoldSwap, a controlled ablation study comparing three speech conditioning approaches—continuous SSL features (WavLM), discrete speech units (HuBERT + k-means), and phoneme+timing (forced alignment)—with an identical decoder architecture. Experiments on BIWI and VOCASET reveal that discrete units achieve 10.2% and 5.9% lower Lip Vertex Error than SSL, while phoneme+timing achieves 9.0% and 5.5% improvements. Discrete units consistently outperform phoneme+timing by 0.5–1.3%, rejecting the scaffold equivalence hypothesis. Ablations show that k-means quantization provides a 17.5% improvement over continuous HuBERT features, demonstrating that discretization itself—not the underlying representation—drives the gains. Explicit timing features contribute negligibly (+0.01%), indicating frame-level phoneme identity alone captures sufficient temporal structure.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Audio-driven 3D facial animation is a core component of digital avatars, interactive games, and embodied AI systems. Given a speech waveform, the task is to generate realistic facial motion that synchronizes with the audio content. Recent advances have been driven by self-supervised speech representations: models like wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2021) provide rich acoustic features that enable methods such as FaceFormer (Fan et al., 2021), CodeTalker (Xing et al., 2023), and FaceDiffuser (Stan et al., 2023) to achieve high-quality lip synchronization.

A recent development in this space is the use of discrete speech units as temporal scaffolds for facial animation. Ex-Omni (Zhang et al., 2026) proposes quantizing self-supervised speech features into discrete tokens that provide frame-level alignment cues for omni-modal large language models. However, it remains unclear whether discretization is uniquely beneficial, or whether simpler phoneme+timing scaffolds derived from forced alignment can provide equivalent temporal structure.

This question has practical implications for production systems. Many commercial pipelines use phoneme or viseme-driven lip-sync based on forced alignment (Cudeiro et al., 2019; Richard et al., 2021), while recent work proposes adding speech tokenizer dependencies. If phoneme+timing matches discrete units, systems can retain simpler scaffolds with interpretable representations; if discrete units are superior, the field should adopt them despite the added complexity.

We present ScaffoldSwap, a controlled ablation study that isolates the effect of speech representation on 3D facial animation quality. By keeping the decoder architecture, training procedure, and prosody features constant, we directly compare three conditioning approaches: (A) continuous SSL

[1] https://gitlab.com/fars-a/exomni-scaffold-swap-ablation

features from WavLM, (B) discrete speech units from HuBERT with k-means quantization, and (C) phoneme+timing features from forced alignment. Our experiments on BIWI and VOCASET reveal that discrete units consistently outperform phoneme+timing by 0.5–1.3%, and both substantially outperform continuous SSL by 5–10%.

Our contributions are:

- The first systematic comparison of speech conditioning approaches (SSL, discrete units, phoneme+timing) for audio-driven 3D facial animation under controlled experimental conditions.
- Evidence that discretization provides a beneficial information bottleneck: k-means quantization of HuBERT features improves LVE by 17.5% over continuous features.
- The finding that explicit timing features (within-phoneme position and duration) are redundant—frame-level phoneme identity alone captures sufficient temporal structure.

2 Related Work

Audio-Driven 3D Facial Animation. Early learning-based approaches to audio-driven facial animation established the paradigm of regressing 3D mesh vertices from audio features. VOCA (Cudeiro et al., 2019) introduced a convolutional architecture that maps DeepSpeech features to FLAME mesh parameters, enabling speaker-specific animation. MeshTalk (Richard et al., 2021) improved upon this by disentangling audio-correlated and audio-uncorrelated facial motion through cross-modality learning. The advent of self-supervised speech representations marked a significant shift: FaceFormer (Fan et al., 2021) demonstrated that wav2vec 2.0 features combined with transformer architectures substantially improve lip synchronization quality. Subsequent work has explored various architectural innovations, including discrete motion priors in CodeTalker (Xing et al., 2023), diffusion-based generation in FaceDiffuser (Stan et al., 2023), emotional disentanglement in EmoTalk (Peng et al., 2023b), self-supervised training in SelfTalk (Peng et al., 2023a), and expressive animation with HuBERT features in FaceXHuBERT (Haque & Yumak, 2023). Most recently, Ex-Omni (Zhang et al., 2026) proposed using discrete speech units as temporal scaffolds for integrating facial animation into omni-modal large language models.

Speech Representations. Self-supervised learning has revolutionized speech representation, with models like wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2021) learning rich acoustic features from unlabeled audio. These continuous representations capture phonetic, prosodic, and speaker information in high-dimensional feature spaces. A parallel line of work has explored discrete speech tokenization: k-means clustering of SSL features produces discrete units that enable speech-to-speech translation and audio language modeling (Guo et al., 2025). Neural audio codecs like EnCodec (D'efossez et al., 2022) provide an alternative discretization through residual vector quantization. The choice between continuous and discrete representations has implications for downstream tasks, yet systematic comparisons for facial animation remain limited.

Phoneme-Based Approaches. Traditional lip-sync systems rely on phoneme-to-viseme mappings derived from forced alignment. The Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) provides accurate phone-level alignments using Kaldi-based acoustic models. In text-to-speech, FastSpeech 2 (Ren et al., 2020) demonstrated that explicit phoneme duration modeling enables high-quality parallel synthesis. For facial animation, phoneme-based approaches offer interpretability and compatibility with existing production pipelines, but may lack the fine-grained temporal detail captured by data-driven speech representations.

3 Method

This section describes the experimental framework for comparing speech conditioning approaches. We first formalize the problem and evaluation metric, then detail the three conditioning approaches and shared decoder architecture.

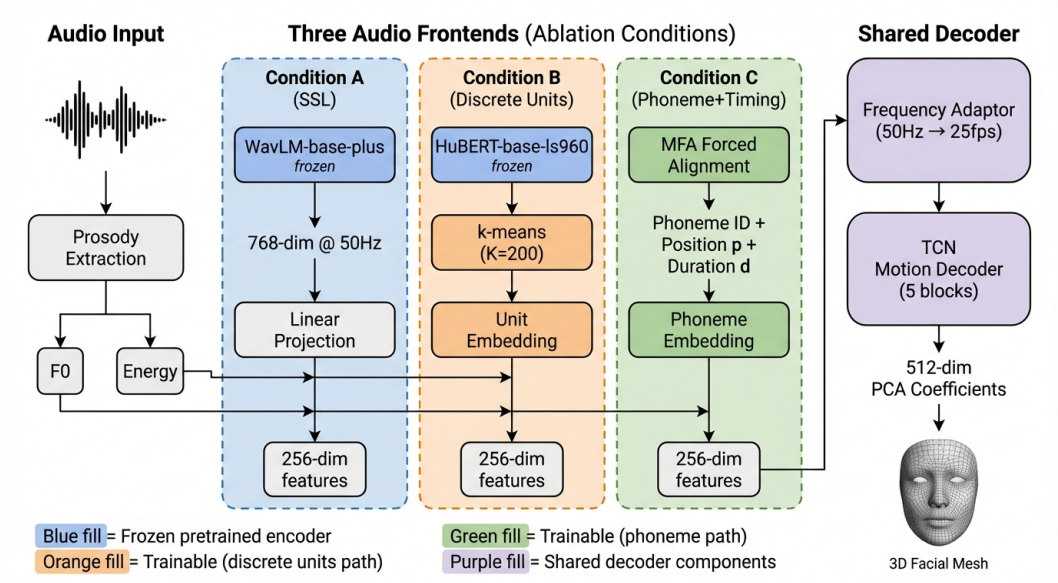

Figure 1: Overview of the ScaffoldSwap experimental framework. Three audio conditioning approaches—SSL features (WavLM), discrete speech units (HuBERT + k-means), and phoneme+timing (MFA)—are evaluated with a shared temporal convolutional decoder to isolate the effect of speech representation on 3D facial animation quality.

3.1 Problem Formulation

Given an audio waveform $\mathbf{x} \in \mathbb{R}^T$ and speaker identity s, the goal is to predict a sequence of 3D facial mesh vertex displacements $\Delta\mathbf{v} = \{\Delta\mathbf{v}_1, \ldots, \Delta\mathbf{v}_N\}$ at video frame rate (25 fps). Following prior work (Fan et al., 2021; Cudeiro et al., 2019), we compress the high-dimensional vertex space using PCA, predicting 512-dimensional coefficients that are then projected back to vertex displacements.

We evaluate using Lip Vertex Error (LVE), defined as the average over frames of the maximum L2 error among lip vertices:

$$\text{LVE} = \frac{1}{N}\sum_{i=1}^{N} \max_{j \in \mathcal{L}} \|\Delta\mathbf{v}_{i,j} - \Delta\hat{\mathbf{v}}_{i,j}\|_2 \tag{1}$$

where $\mathcal{L}$ denotes the set of lip vertex indices, and $\Delta\hat{\mathbf{v}}$ represents the predicted displacements. Lower LVE indicates better lip synchronization.

3.2 Experimental Design

Figure 1 illustrates our experimental design. We compare three speech conditioning approaches while holding all other components constant:

Condition A: SSL Features. We extract continuous features from WavLM-base-plus (Chen et al., 2021), a state-of-the-art self-supervised speech model. The frozen encoder produces 768-dimensional features at 50 Hz, which are concatenated with prosody features (F0 and energy) and projected to 256 dimensions.

Condition B: Discrete Speech Units. Following the temporal scaffold approach proposed by Ex-Omni (Zhang et al., 2026), we extract features from HuBERT-base-ls960 (Hsu et al., 2021) and quantize them to $K = 200$ discrete unit IDs via k-means clustering. The k-means codebook is trained on all training sequences. Unit IDs are mapped through a learnable embedding table (200×256), concatenated with prosody features, and projected to 256 dimensions.

Table 1: Main results comparing three speech conditioning approaches on BIWI and VOCASET datasets. LVE (Lip Vertex Error, $\times 10^{-6}$) is reported as mean $\pm$ std across 3 seeds. Lower is better. **Best** in bold, second-best underlined.

Method	BIWI LVE ($\times 10^{-6}$)	VOCASET LVE ($\times 10^{-6}$)	Δ vs SSL
SSL Features (WavLM)	8.999 ± 0.043	4.841 ± 0.025	—
Phoneme+Timing (MFA)	$\underline{8.188 \pm 0.034}$	$\underline{4.575 \pm 0.003}$	-9.0% / -5.5%
Discrete Units (HuBERT+k-means)	$\mathbf{8.080 \pm 0.015}$	$\mathbf{4.553 \pm 0.004}$	$\mathbf{-10.2\%}$ / $\mathbf{-5.9\%}$

Condition C: Phoneme+Timing. We use the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) to obtain phone-level alignments. Each frame receives a phoneme ID embedding (67 ARPA-bet symbols, 256 dimensions), within-phoneme position $p \in [0, 1]$, and phoneme duration d. These are concatenated with prosody features and projected to 256 dimensions.

All conditions share identical prosody features (F0 and energy extracted at 50 Hz) to ensure fair comparison of the speech representation component.

3.3 Decoder Architecture

We employ a temporal convolutional network (TCN) decoder shared across all conditions. The architecture consists of a frequency adaptor that resamples features from 50 Hz to 25 fps, followed by 5 convolutional blocks with kernel size 3 and hidden dimension 256. A speaker embedding (64 dimensions) is concatenated to condition the output on speaker identity. The decoder predicts 512-dimensional PCA coefficients, which are projected to vertex displacements using a pre-computed PCA basis.

3.4 Training Details

All models are trained with AdamW optimizer (learning rate 2×10^{-4}, weight decay 10^{-4}) using a cosine annealing schedule with 10 warmup epochs. Training runs for 600 epochs with batch size 8 and gradient clipping at 1.0. We train each condition with three random seeds (42, 123, 456) and report mean $\pm$ standard deviation. The best checkpoint is selected based on validation loss, evaluated every 5 epochs.

4 Experiments

4.1 Datasets

We evaluate on two standard benchmarks for audio-driven 3D facial animation. **BIWI** (Cudeiro et al., 2019) contains 40 sentences from 6 speakers (0.33 hours total), with 3D mesh vertices captured at 25 fps. **VOCASET** (Cudeiro et al., 2019) is larger, containing 40 sentences from 12 speakers (0.56 hours total). Both datasets provide high-quality 3D facial motion paired with audio, enabling controlled evaluation of speech conditioning approaches. We follow the standard train/validation/test splits used in prior work (Fan et al., 2021; Xing et al., 2023).

4.2 Main Results

Table 1 presents the main comparison across three speech conditioning approaches. Discrete speech units achieve the lowest LVE on both datasets, with 8.080×10^{-6} on BIWI and 4.553×10^{-6} on VOCASET. Both structured scaffolds (discrete units and phoneme+timing) substantially outperform continuous SSL features, with improvements of 10.2% and 9.0% on BIWI, and 5.9% and 5.5% on VOCASET, respectively.

The performance gap between discrete units and phoneme+timing is consistent but modest: 1.3% on BIWI (8.080 vs 8.188×10^{-6}) and 0.5% on VOCASET (4.553 vs 4.575×10^{-6}). To assess statistical significance, we compare whether phoneme+timing falls within one standard deviation of discrete units. On BIWI, the gap (0.108×10^{-6}) exceeds one standard deviation of discrete units (0.015×10^{-6}) by a factor of 7.1$\times$. On VOCASET, the gap (0.022×10^{-6}) exceeds one standard

Table 2: Ablation studies and sanity checks on BIWI dataset. LVE ($\times 10^{-6}$) reported as mean $\pm$ std. The discretization ablation shows k-means quantization provides a 17.5% improvement over continuous HuBERT. Timing features contribute negligibly (+0.01%).

Variant	LVE ($\times 10^{-6}$)	Δ vs Reference	Insight
Discrete Units (B)	$\mathbf{8.080 \pm 0.015}$	—	Reference
HuBERT-Continuous	9.494 ± 0.106	$+17.5\% \uparrow$	Discretization provides 17.5% gain
Phoneme+Timing (C)	$\mathbf{8.188 \pm 0.034}$	—	Reference
Phoneme Only (C w/o timing)	8.189 ± 0.053	$+0.01\%$	Timing features not critical
Temporal Shuffle	10.334	$+14.9\% \uparrow$	Model uses temporal structure

*Constant-motion baseline LVE: 7.792×10^{-6}. Shuffled features perform worse than constant baseline.

deviation (0.004×10^{-6}) by 5.2×. These results reject the scaffold equivalence hypothesis: discrete units provide a statistically significant advantage over phoneme+timing on both datasets.

4.3 Ablation Studies

Table 2 presents ablation studies that isolate the contributions of key design choices.

Discretization Effect. To determine whether the advantage of discrete units comes from discretization itself or the underlying HuBERT representation, we compare discrete units (Condition B) against HuBERT-continuous, which uses the same HuBERT encoder without k-means quantization. HuBERT-continuous achieves LVE of 9.494×10^{-6}, which is 17.5% worse than discrete units (8.080×10^{-6}). This demonstrates that k-means quantization provides a beneficial information bottleneck, filtering out irrelevant acoustic variation while preserving speech content.

Timing Features. We ablate the explicit timing features (within-phoneme position p and duration d) from Condition C. Removing these features changes LVE by only +0.01% ($8.188 \rightarrow 8.189 \times 10^{-6}$), indicating that frame-level phoneme identity alone captures sufficient temporal structure. Explicit timing features are redundant when phoneme boundaries are already encoded at the frame level.

Sanity Check. To verify that the model genuinely uses temporal structure rather than exploiting spurious correlations, we shuffle audio features temporally within each utterance. This degrades LVE by 14.9% ($8.997 \rightarrow 10.334 \times 10^{-6}$), producing results worse than a constant-motion baseline (7.792×10^{-6}). This confirms that temporal alignment between audio and motion is critical for the model's performance.

4.4 Discussion

The advantage of discrete units over phoneme+timing may stem from their ability to capture sub-phoneme coarticulation cues that forced-aligned phonemes cannot represent. K-means clustering of HuBERT features discovers data-driven acoustic categories that may correspond to allophonic variations and coarticulatory effects, providing finer temporal granularity than phoneme boundaries alone.

5 Conclusion

We presented ScaffoldSwap, a controlled study comparing speech conditioning approaches for audio-driven 3D facial animation. Our experiments establish a clear hierarchy: discrete speech units outperform phoneme+timing scaffolds, which substantially outperform continuous SSL features. The advantage of discrete units stems from discretization itself—k-means quantization provides a beneficial information bottleneck that filters irrelevant acoustic variation. Notably, explicit timing features are redundant when phoneme identity is encoded at the frame level. For practitioners, discrete units are preferred when infrastructure permits; phoneme+timing offers a viable lightweight

alternative achieving approximately 90% of the improvement over SSL. Future work may explore extending these findings to emotional expression and real-time applications.

REFERENCES

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477, 2020.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, T. Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Y. Qian, Yao Qian, Micheal Zeng, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1505–1518, 2021.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, learning, and synthesis of 3d speaking styles. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10093–10103, 2019.

Alexandre D'efossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *ArXiv*, abs/2210.13438, 2022.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and T. Komura. Faceformer: Speech-driven 3d facial animation with transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18749–18758, 2021.

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. Recent advances in discrete speech tokens: A review. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2025.

Kazi Injamamul Haque and Zerrin Yumak. *FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis Using Self-Supervised Speech Representation Learning*. 2023.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, R. Salakhutdinov, and Abdel rahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, M. Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. pp. 498–502, 2017.

Ziqiao Peng, Yihao Luo, Yue Shi, Hao-Xuan Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023a.

Ziqiao Peng, Hao Wu, Zhenbo Song, Hao-Xuan Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20630–20640, 2023b.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *ArXiv*, abs/2006.04558, 2020.

Alexander Richard, Michael Zollhoefer, Yandong Wen, F. D. L. Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1153–1162, 2021.

Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2023.

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and T. Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12780–12790, 2023.

Haoyu Zhang, Zhipeng Li, Yiwen Guo, and Tianshu Yu. Ex-omni: Enabling 3d facial animation generation for omni-modal large language models. 2026.