메타버스융합대학원
중앙대학교

MemoryTalker: Personalized Speech-Driven 3D Facial Animation via Audio-Guided Stylization

Hyung Kyu Kim, Sangmin Lee, Hak Gu Kim

중앙대학교 메타버스융합학과
Department of Metaverse Convergence, Chung-Ang University

ICCV OCT 19-23, 2025 HONOLULU HAWAII

INTRODUCTION

Speech-driven 3D Facial Animation

- Animating 3D facial motion from speech using a template mesh.

Personalization Issue in Speech-driven 3D Facial Animation

- ***Personalization*** is a challenging in this task [1].
- Common articulatory patterns exist across speakers. (e.g., /a/ mouth opening)
- Same phoneme can be pronounced with ***distinct styles***. (e.g., Large / Small)
- Realistic animation requires to model ***shared phonetic cues*** and ***speaker-specific styles*** from audio with rich information [2].

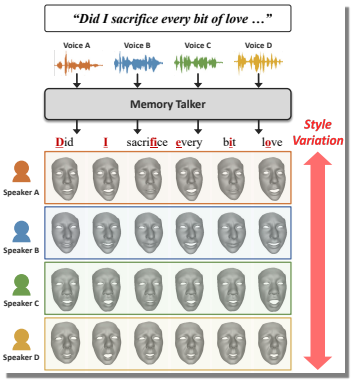

Figure 1. The intuition of MemoryTalker for personalized speech-driven 3D facial animation

Novel Personalization Method

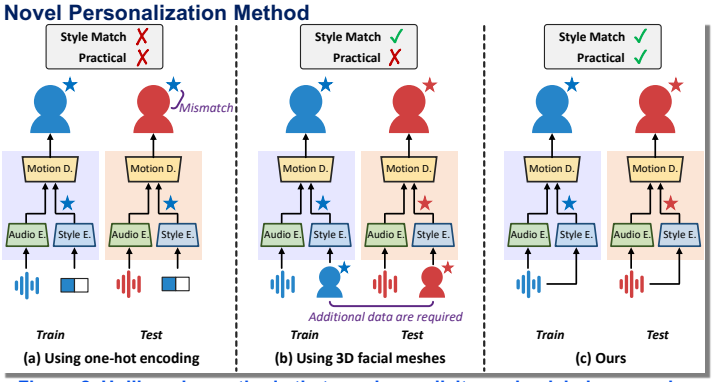

Figure 2. Unlike prior methods that require explicit speaker labels or mesh sequences, MemoryTalker models speaking style implicitly from speech only.

- One-hot encoding methods cannot generalize to unseen speakers.
- Mesh-based methods are resource-intensive.
- Our method generates personalized 3D facial motion using only audio.

EXPERIMENTS

Quantitative Evaluation

Method	FVE	LVE	FID	LDTW	Lip-max
FaceFormer [CVPR`22]	0.639	0.413	3.583	0.507	0.452
CodeTalker [CVPR`23]	0.721	0.498	3.713	0.554	0.484
SelfTalk [ACM MM`23]	0.593	0.382	3.279	0.475	0.416
Imitator [ICCV`23]	0.686	0.456	3.918	0.554	0.472
ScanTalk [ECCV`24]	0.609	0.375	3.623	0.457	0.420
UniTalker [ECCV`24]	0.570	0.382	3.256	0.507	0.407
MemoryTalker	**0.506**	**0.293**	**3.045**	**0.418**	**0.331**

Table 1. Performance Comparisons on VOCASET.

Competitors	Lip Sync (%)	Realism (%)	Speaking Style (%)
vs. FaceFormer	83.9	85.5	80.6
vs. CodeTalker	85.5	83.9	71.8
vs. Imitator	87.1	87.1	78.2
vs. ScanTalk	94.4	91.1	92.8
vs. UniTalker	79.8	80.6	86.3

Table 2. User study: our method vs. competitor on VOCASET.

PROPOSED METHOD

Illustration of the proposed our MemoryTalker

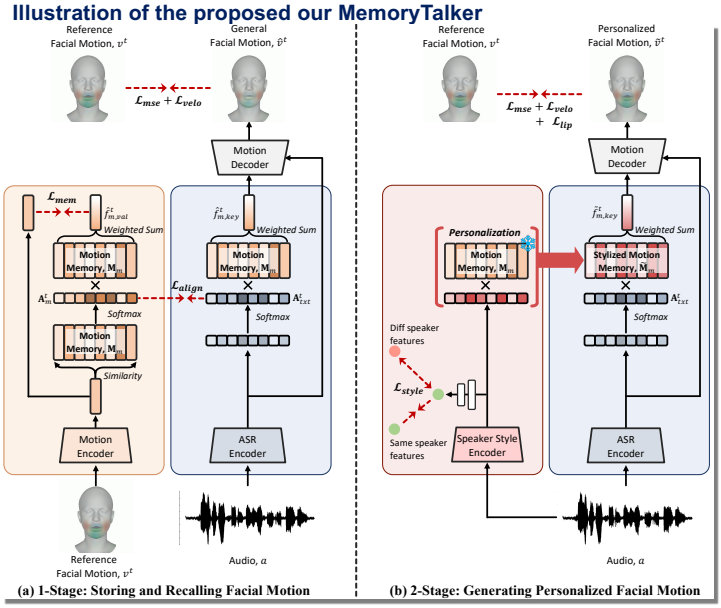

Figure 3. MemoryTalker's architecture: a two-stage framework that (a) memorizes general motion and (b) stylizes general motion

- Stage1: ***Generalized*** facial motion features are ***stored*** in memory and ***retrieved*** using text representations to model ***common articulatory patterns*** across speakers.
- Stage2: Motion memory is ***refined*** per speaker by modulating slot-wise features using audio, enabling the model ***to capture speaker-specific speaking styles***.
- Together, our ***two-stage training strategy*** allows ***disentangling*** shared and personalized dynamics from speech for realistic facial animation.

Qualitative Evaluation

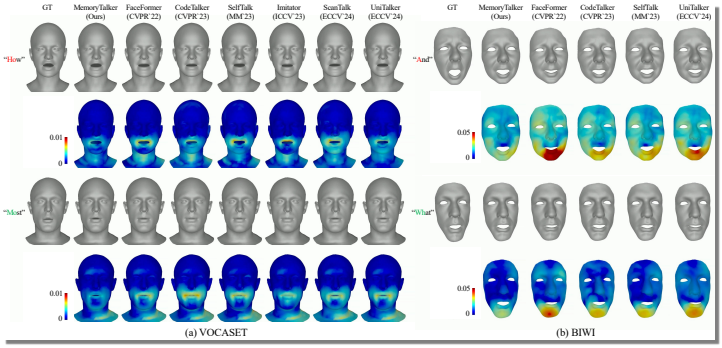

Figure 4. Visual Comparisons with state-of-the-art methods.

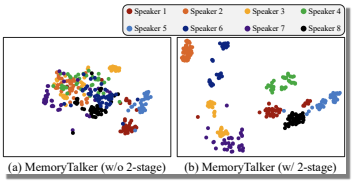

Figure 5. The t-SNE visualization of the recalled motion features.

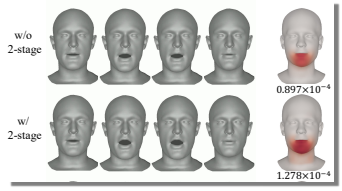

Figure 6. Comparison "w/o 2-stage" and "w/ 2-stage"

IV. REFERENCE

[1] A linear model of acoustic-to-facial mapping: Model parameters, data set size, and generalization across speakers, JASA, 2008

[2] Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron, ICML, 2018.