Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

AI-Based Speech-Driven 3D Avatar Animation System

Integrated Gender, Emotion, and Viseme Analysis for Browser-Rendered Avatars

Dr. Anumol Mathai
Department of Computer Science and Engineering
G Madegowda Institute of Technology
Bharathinagara, Karnataka, India

Ankitha G
Department of Computer Science and Engineering
G Madegowda Institute of Technology
Bharathinagara, Karnataka, India

Apoorva
Department of Computer Science and Engineering
G Madegowda Institute of Technology
Bharathinagara, Karnataka,

Inchara K R
Department of Computer Science and Engineering
G Madegowda Institute of Technology Bharathinagara, Karnataka, India

Keerthana M S
Department of Computer Science and Engineering
G Madegowda Institute of Technology Bharathinagara, Karnataka, India

***Abstract* - Speech-driven three-dimensional avatars provide a practical interface between human speech and interactive digital characters. This paper presents an integrated, AI-based pipeline that converts ordinary speech into a synchronized, browser-rendered avatar without specialized capture hardware. A frozen pretrained Wav2Vec2 encoder extracts frame-level speech features, which three lightweight modules interpret to infer speaker gender, emotional state and viseme category. Gender classification fuses a trained GenderCNN with pitch evidence through calibrated weighting, so that avatar selection does not rest on the neural classifier alone. A thirteen-class Transformer predicts visemes, and a second trained convolutional network smooths them before rendering. A FastAPI backend performs inference, and a Three.js frontend renders the avatar with procedural breathing, blinking and audio-reactive motion. Recorded measurements include a mean Wav2Vec2 real-time factor of about 0.18, a GenderCNN accuracy of 0.7130 before augmentation and 0.7207 on the augmented domain, and an eight-class emotion accuracy of about 0.37 for a Random Forest baseline. The BiLSTM emotion model used at inference has no logged accuracy yet, and this is documented as an open item. The system is framed as an explainable, deployment-ready engineering pipeline rather than a direct competitor to continuous speech-to-mesh research.**

Keywords - speech-driven animation; 3D avatar; Wav2Vec2; viseme; lip synchronization; emotion recognition; gender classification; pitch fusion; blendshape animation; CPU-oriented inference

I. INTRODUCTION

Three-dimensional digital humans and talking avatars are used increasingly in virtual assistants, games, education, telepresence and training. This growing range of uses raises the bar for what a convincing avatar must do. A useful speech-driven avatar cannot merely move a mouth: facial behavior must remain temporally coherent, speech-related motion must be visually synchronized, and the character must appear sufficiently expressive to avoid a rigid presentation. Procedural

visemes, learned audio-to-face mappings, pretrained speech representations and personalized 3D facial animation have each been investigated toward this end [1]–[8].

Imitator, the reference work used to frame the organization of this project, formulates personalized speech-driven facial animation through audio encoding, generalized viseme generation and identity-specific motion decoding [1]. Learned mappings from speech to continuous 3D facial motion are likewise demonstrated by VOCA, FaceFormer, MeshTalk and CodeTalker [2]–[5]. These systems motivate the problem, but they generally require personalized, subject-specific capture, which the present work deliberately avoids.

In the proposed system, speech understanding is separated into three interpretable signals, namely gender, emotion and viseme information. These signals are temporally smoothed and mapped onto an existing rigged avatar, instead of a dense facial mesh being predicted directly. Several practical constraints motivate this design: ordinary microphone or file audio is accepted as input, multi-camera capture is avoided, a pretrained speech encoder is used instead of training one from scratch, the final 3D presentation is performed by the browser, and CPU-oriented development hardware is targeted throughout.

System integration is therefore the principal contribution of this work. Multiple control signals are produced from a single speech input, and visible avatar behavior is produced from them through facial blendshapes and skeletal motion. Claims of superiority over research systems whose datasets, tasks and metrics differ from those used here are avoided; comparisons are framed around deployability, interpretability and accessibility instead (Section VI).

II. RELATED WORK

A. Speech-Driven 3D Facial Animation

Early speech-animation systems drove facial motion from phonetic or viseme representations, and coarticulation was later handled through context-dependent visemes [6]. JALI introduced an animator-centric procedural formulation in which jaw and lip actions are controlled directly by acoustic information [7], while VisemeNet learned motion curves for production facial rigs directly from audio [8].

Learned 3D approaches pursued continuous facial motion. VOCA learns speaking styles from synchronized 4D face data [2]; FaceFormer predicts 3D facial motion sequences with a Transformer built on pretrained speech representations [3]; MeshTalk addresses speech-to-face generation through cross-modality disentanglement [4]; and CodeTalker models realistic facial dynamics with a learned discrete motion prior [5]. Personalized speaking-style adaptation is studied by Imitator [1], and audio-guided personalization and photorealistic 3D Gaussian avatars are explored by MemoryTalker and GaussianSpeech [9], [10]. SadTalker represents an alternative audio-driven talking-face approach [11]. Many of these systems are evaluated on continuous facial geometry or perceptual quality, whereas this project evaluates discrete downstream classifiers and an integrated controller, so direct numerical

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

comparison is not attempted.

B. Speech Representation Learning

Wav2Vec 2.0 introduced self-supervised learning of contextual speech representations from raw audio [12]. Such pretrained features are useful for downstream tasks without training a new acoustic model. In this project the pretrained facebook/wav2vec2-base-960h model is used strictly in frozen evaluation mode, and its frame-level outputs are stored for downstream processing rather than fine-tuned.

C. Emotion and Expressive Animation

Speech has been combined with pose and emotion information in emotion-conditioned facial animation [13], and learned models have preserved lip synchronization while emotional behavior is varied [14]. This project adopts a simpler, interpretable formulation: emotion is an additional discrete control signal that modifies avatar expression, and no claim of complete human emotional understanding is made.

III. PROPOSED SYSTEM

A. System Architecture

The complete pipeline is shown in Fig. 1. An audio waveform, supplied as an uploaded file or a live browser-microphone recording, is standardized and passed through a fixed pretrained speech encoder. The resulting embedding is interpreted independently by three lightweight downstream modules, and the combined predictions are smoothed and rendered onto a rigged avatar. Every stage after feature extraction (gender, emotion, viseme interpretation, temporal smoothing and avatar control) was implemented and trained by the authors, whereas the pretrained encoder and the avatar rig are used as fixed, non-fine-tuned building blocks.

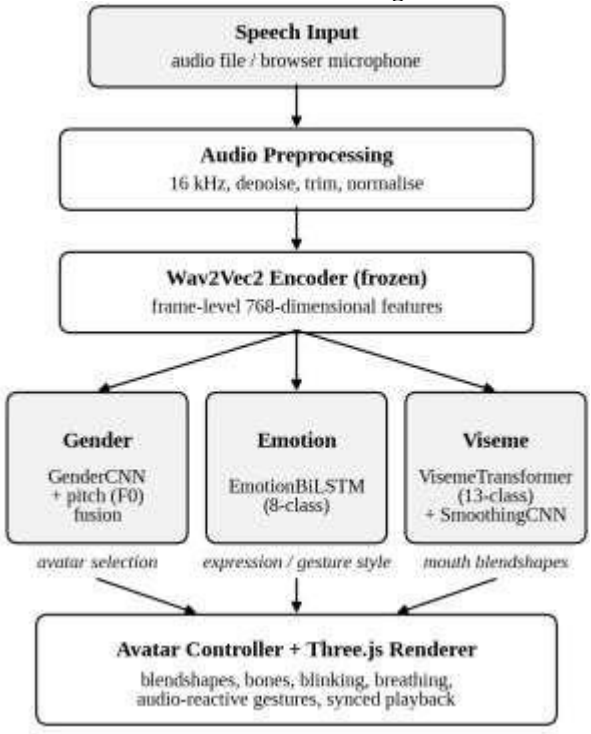


Fig. 1. Overall pipeline of the proposed speech-driven 3D avatar animation system.

B. Audio Preprocessing

Input audio is converted to a consistent 16 kHz representation, after which denoising, trimming and amplitude normalization are applied. WebM/Opus recordings from the browser microphone are first converted to a backend-compatible format. This step reduces variation caused by silence, inconsistent recording level and background noise, and makes downstream feature extraction more stable.

C. Speech Feature Extraction

Let the preprocessed waveform be denoted $x(t)$. The pretrained Wav2Vec2 encoder produces a sequence of contextual feature vectors,

$$A = \text{Wav2Vec2}(x) = \{a_1, a_2, \ldots, a_T\} \quad (1)$$

where each frame a_t is approximately 768-dimensional. Because the model is operated strictly in frozen mode, no gradient updates reach its weights while the downstream modules are trained, and the frame sequences are cached as NumPy arrays so that the encoder is not re-invoked for every experiment.

D. Multi-Task Speech Interpretation

Three interpretable streams are derived in parallel from the speech representation: a male/female probability used for automatic avatar selection, an eight-class emotional category, and a thirteen-class viseme representation describing visually similar mouth configurations. Each stream can be evaluated, retrained or replaced independently, so the overall pipeline need not be retrained.

Gender. A Random Forest and a lightweight convolutional network (GenderCNN) were evaluated on standardized, voiced-frame-pooled Wav2Vec2 embeddings. GenderCNN achieved better pre-augmentation accuracy and F1 (0.7130 / 0.7520) than the Random Forest baseline (0.7083 / 0.7225) and was selected. At inference its softmax female probability is combined with an independent acoustic signal: a median fundamental frequency (F0) is estimated from the raw waveform by short-frame autocorrelation (70–350 Hz search range, outlier-trimmed median over voiced frames) and mapped to a soft pitch-based female probability through a logistic transition centered near 165 Hz. A calibrated weighted fusion then combines the two signals, nominally

$$p_{\text{fused}} = w \cdot p_{\text{CNN}} + (1 - w) \cdot p_{\text{pitch}} \quad (2)$$

with $w = 0.60$. The weighting shifts toward pitch when the CNN output is itself uncertain (female probability between 0.35 and 0.65), and two further consistency adjustments prevent a very strong pitch signal from being completely overridden by a less-confident CNN prediction in either direction. Pitch is used only as supporting evidence and never as an absolute male/female rule. The fused probability is compared against a calibrated threshold of 0.4735 to give the final decision used for avatar selection. This fusion, together with voiced-frame pooling and augmentation, was introduced to improve robustness to the domain shift between controlled training audio and unconstrained microphone recordings.

Emotion. An eight-class Random Forest was trained on 1,440 RAVDESS samples (96 neutral, 192 for each remaining class), and an accuracy of about 0.37 was recorded. A bidirectional LSTM (EmotionBiLSTM) was also trained on the same Wav2Vec2 feature sequences and is the model actually executed by the deployed inference pipeline. A validated accuracy for this BiLSTM is not yet available, so it is not

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

reported as a performance claim, and the gap between the deployed and the benchmarked model is stated explicitly as a limitation (Section VII).

Viseme. Each of the thirteen viseme classes groups phonemes that share a similar visible mouth configuration, and phoneme-level alignment before class assignment was performed with the Montreal Forced Aligner. Unlike a fixed phoneme-to-viseme lookup table, the mapping from speech features to the viseme sequence is produced by a trained

Transformer-based sequence model (VisemeTransformer) operating on the frozen Wav2Vec2 embeddings in fixed-length chunks, which bounds memory use on CPU hardware. The raw per-frame probabilities are passed to the smoothing stage before a mouth state is finalized for each frame.

E. Temporal Smoothing

Frame-level viseme predictions fluctuate rapidly because acoustic evidence is distributed unevenly over time. Smoothing is therefore performed by a second trained model, SmoothingCNN, which takes the full sequence of per-frame viseme probabilities from VisemeTransformer and outputs a smoothed sequence from which the final class is taken by argmax. Because it is learned from labeled sequences rather than fixed by a hand-chosen window and weighting, it can suppress isolated spikes while remaining sensitive to genuine phoneme transitions.

F. Avatar Selection and Control

The male or female GLB/GLTF asset is selected by the predicted gender, mouth-related blendshape weights are determined by the smoothed viseme state, and the predicted emotion supplies an additional expression signal. Deterministic controller logic performs the final mapping to animation values, so machine learning interprets the speech while the rig is driven by non-learned logic. The current avatar root is tracked and cleared whenever a new gender prediction changes the selected identity, which prevents duplicate or merged models in the scene. Because the male and female assets differ in height, pose and framing, automatic normalization and camera framing preserve a consistent presentation scale.

G. Procedural Body Motion

A procedural layer added by the frontend sits on top of the learned facial controls. A small sinusoidal vertical sway and periodic eye blinking, driven by morph-target interpolation, reduce a static appearance between speech events. Visible shoulder, arm, hand and finger variation comes from an audio-reactive gesture system: a Web Audio AnalyserNode computes the real-time RMS amplitude of the playing clip, and discrete gesture events (single-arm, two-handed emphasis, open-hand and head-nod) are triggered probabilistically from it, subject to a cooldown. Each event is shaped by a smoothstep ease-in/hold/ease-out envelope, and gesture intensity, speed and minimum pause are modulated by the predicted emotion; angry speech, for example, produces faster and larger gestures with shorter pauses than sad speech. Arm and forearm bones are driven toward camera-relative, chest-height targets by a per-frame swing-rotation (aim-at-target) computation, so raised-arm gestures converge on a consistent, plausible position regardless of avatar proportions. These behaviors are procedural and audio-reactive, not a learned co-speech gesture model, since no gesture-specific training data was used.

H. Software Architecture

A FastAPI/Uvicorn backend provides the inference service, and the pipeline is instantiated once through an application-lifespan mechanism so that models are not reloaded for every request. The frontend is built with HTML, CSS and JavaScript, with Three.js providing browser-based 3D rendering. File-based audio and live microphone recording are both supported; microphone audio is captured through the MediaRecorder API and encoded client-side as WebM/Opus or OGG/Opus, chosen by runtime feature detection, because most browsers cannot record WAV natively. Windows launcher scripts, Docker packaging and deployment documentation accompany the implementation.

IV. DATASETS AND ALGORITHMS USED

RAVDESS [15] is the principal dataset for the controlled gender and emotion experiments reported here: an emotion experiment of 1,440 samples and an augmented gender dataset of 2,880 samples, balanced between 1,440 male and 1,440 female samples, are drawn from it. Common Voice was considered during development, but no completed training result on that corpus is established, so it is not reported as an experimental source. Table I summarizes the algorithms and technologies used at each pipeline stage.

TABLE I. ALGORITHMS AND TECHNOLOGIES BY PIPELINE STAGE

Stage	Algorithm / Technology	Role
Feature extraction	Wav2Vec2 (facebook/wav2vec2-base-960h), PyTorch, torchaudio	Frozen, pretrained
Gender	GenderCNN (Random Forest baseline) fused with pitch (F0) evidence	Trained + rule-based fusion
Emotion	Bidirectional LSTM, 8-class (deployed); Random Forest (benchmarked)	Trained from scratch
Viseme	VisemeTransformer, 13-class, MFA-aligned targets	Trained from scratch
Smoothing	SmoothingCNN over per-frame viseme probabilities	Trained from scratch
Avatar rig	GLB/GLTF, morph targets and skeletal bones	Pretrained asset
Backend	FastAPI, Uvicorn (Python)	Infrastructure
Frontend	Three.js, WebGL, Web Audio API, MediaRecorder	Infrastructure
ML tooling	scikit-learn, PyTorch (CNN, BiLSTM, Transformer)	Infrastructure

Measurements in the next section are restricted to values explicitly documented in the project record. Published results of Imitator, VOCA, FaceFormer, MeshTalk, CodeTalker, GaussianSpeech and SadTalker are not inserted into the same table, because tasks, datasets and metrics differ and a shared table would misrepresent comparability.

V. RESULTS AND DISCUSSION

The mean Wav2Vec2 real-time factor of about 0.18 means that feature extraction takes roughly 18% of the audio duration, or about 5.5 times real time. End-to-end browser latency additionally involves audio conversion, service overhead, model loading and rendering, and is not captured by this figure alone.

GenderCNN changed from 0.7130 accuracy / 0.7520 F1 before augmentation to 0.7207 accuracy / 0.7358 F1 on the

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

augmented-domain test, with confusion matrix [[215, 100], [81, 252]]. The matrix shows both correct and incorrect predictions, consistent with the documented observation that gender misclassification can occur intermittently for some microphone recordings. This is the pre-fusion CNN result, and the pitch-fusion step of Section III-D is the system's mitigation for exactly this failure mode.

Emotion recognition remains a challenging component: the Random Forest baseline reaches only about 0.37 on eight classes. Because the deployed BiLSTM has not yet had its own accuracy logged, this figure should be read as a reference point for the difficulty of the task, not as a measurement of the model serving predictions. Emotion is therefore used as an additional expression-control signal and not as a claim of reliable human-level emotional understanding.

TABLE II. RECORDED PERFORMANCE OF THE PROPOSED SYSTEM

Component	Metric / Setting	Recorded Result
Wav2Vec2	Mean real-time factor	≈0.18 (≈5.5× real time)
Gender RF	Pre-aug. accuracy / F1	0.7083 / 0.7225
Gender CNN	Pre-aug. accuracy / F1	0.7130 / 0.7520
Gender CNN	Params / inference time	9,250 / ≈1.364 ms
Augmented gender (CNN, pre-fusion)	Accuracy / F1	0.7207 / 0.7358
Emotion RF (baseline)	8-class accuracy	≈0.37
Emotion BiLSTM (deployed)	8-class accuracy	Not yet logged (pending)
Gender calibration	Female-prob. threshold	≥ 0.4735

These figures are carried forward from the project's experiment log and have not been independently re-verified in this revision. The one entry for the deployed BiLSTM is marked pending instead of being estimated, in keeping with the practice of not inserting unmeasured values into a results table.

VI. COMPARISON WITH EXISTING SYSTEMS

Table III positions the proposed system against representative prior work along dimensions of practical deployability, since raw regression accuracy is not comparable across different tasks, datasets and mesh-versus-classifier protocols (Section II-A). The proposed system does not claim an advantage in facial-motion accuracy. Its advantage lies in operating end-to-end on commodity CPU hardware, in requiring no personalized capture of any target speaker, and in exposing gender, emotion and viseme decisions as separately inspectable signals.

TABLE III. COMPARISON WITH PRIOR SPEECH-DRIVEN FACIAL ANIMATION SYSTEMS

Dimension	VOCA / FaceFormer / MeshTalk / CodeTalker / Imitator	Proposed System
Capture	Personalized 4D face scans or subject-specific capture	Public audio only; no custom capture
Hardware	GPU-trained, GPU-inference deep networks	CPU-oriented development and inference
Output	Continuous, dense facial mesh vertices	Discrete gender, emotion and viseme signals mapped to a rig
Interpretability	Largely end-to-end; hard to inspect per decision	Modules independently trainable, testable, replaceable
Avatar identity	Fixed to the training subject's mesh	Automatically selected between distinct avatars
Deployment	Mostly research prototypes / offline rendering	REST API with live browser rendering and microphone input
Accessibility	Needs specialized capture pipelines for new subjects	Extensible with public multi-speaker corpora

Table III is a statement about system scope and accessibility, not a claim that the proposed classifiers outperform continuous mesh-regression networks on shared benchmarks; no such shared benchmark exists. Within its stated scope (automatic, CPU-deployable, gender-and-emotion-aware avatar control without personalized capture) the system is more directly deployable and more transparent to audit than the compared systems, each of which requires subject-specific data, GPU-class hardware, or both.

VII. LIMITATIONS AND FUTURE WORK

Modularity is the main design choice of this work. Continuous speech-to-3D facial animation research often learns facial motion end to end [1]–[5], [9]–[11]; this system instead decomposes the task into interpretable stages that are easier to inspect, replace and connect to existing avatar assets. The cost is that discrete viseme and emotion controls do not model the full continuous space of human facial motion.

Several limitations are documented explicitly. Gender classification can be intermittent when microphone recordings differ from the training domain, and pitch fusion is only a partial mitigation. The emotion module has an unresolved gap between the benchmarked Random Forest (about 0.37 accuracy) and the deployed BiLSTM (accuracy not yet logged); benchmarking the BiLSTM on the same evaluation split is the near-term next step. Microphone operation is record-then-process rather than continuous streaming, body gestures are procedural rather than learned, no completed multilingual experiment exists, and avatar quality depends on the GLB rig, its morph targets, bone names and export settings.

Future directions include chunked streaming inference, a logged accuracy for the deployed BiLSTM, additional microphone-domain data, improved gender calibration, stronger pretrained representations for emotion, more diverse emotional speech, learned speech-to-gesture generation, multilingual speech representations with language-aware phoneme-to-viseme alignment, more customizable avatars, and formal human-subject evaluation of naturalness, lip synchronization, emotion agreement and usability.

VIII. CONCLUSION

This paper presented an AI-based speech-driven 3D avatar animation system that converts ordinary speech into a visible, synchronized 3D character response. A pretrained Wav2Vec2 representation is combined with modular gender, emotion and viseme analysis: gender is fused with pitch evidence, and viseme predictions from a trained Transformer are stabilized by a trained smoothing network. Deterministic avatar-control logic maps the resulting control sequence onto facial blendshapes and skeletal motion, and the browser frontend adds procedural breathing, blinking, audio-reactive body movement, camera

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

normalization and reliable avatar replacement.

The documented measurements show a mean Wav2Vec2 real-time factor of about 0.18, gender classification at about 0.72 augmented-domain accuracy before pitch fusion, and about 0.37 accuracy on the eight-class emotion baseline, with the deployed emotion model's own accuracy identified as an open measurement gap. These values should be read within the project's specific tasks and datasets, not as comparisons with continuous speech-to-mesh research. The principal contribution is an integrated, explainable, CPU-oriented and deployment-ready pipeline connecting speech understanding to a working 3D avatar.

ACKNOWLEDGMENT

The authors thank Dr. Anumol Mathai, Department of Computer Science and Engineering, G Madegowda Institute of Technology, for her guidance and support throughout this project.

REFERENCES

[1] B. Thambiraja, I. Habibie, S. Aliakbarian, D. Cosker, C. Theobalt, and J. Thies, "Imitator: Personalized speech-driven 3D facial animation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 20621–20631.

[2] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 10101–10111.

[3] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "FaceFormer: Speech-driven 3D facial animation with transformers," in *Proc. IEEE/CVF CVPR*, 2022, pp. 18770–18780.

[4] A. Richard, M. Zollhofer, Y. Wen, F. de la Torre, and Y. Sheikh, "MeshTalk: 3D face animation from speech using cross-modality disentanglement," in *Proc. IEEE/CVF ICCV*, 2021, pp. 1153–1162.

[5] 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 *Proc. IEEE/CVF CVPR*, 2023, pp. 12780–12790.

[6] 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.

[7] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "JALI: An animator-centric viseme model for expressive lip synchronization," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 127:1–127:11, 2016.

[8] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "VisemeNet: Audio-driven animator-centric speech animation," in *ACM SIGGRAPH*, 2018.

[9] H. K. Kim, S. Lee, and H. G. Kim, "MemoryTalker: Personalized speech-driven 3D facial animation via audio-guided stylization," in *Proc. IEEE/CVF ICCV*, 2025, pp. 11241–11251.

[10] S. Aneja, A. Sevastopolsky, T. Kirschstein, J. Thies, A. Dai, and M. Nießner, "GaussianSpeech: Audio-driven personalized 3D Gaussian avatars," in *Proc. IEEE/CVF ICCV*, 2025, pp. 13065–13075.

[11] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in *Proc. IEEE/CVF CVPR*, 2023, pp. 8652–8661.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 12449–12460.

[13] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 94:1–94:12, 2017.

[14] S. L. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. K. Hodgins, and I. A. Matthews, "A deep learning approach for generalized speech animation," *ACM Trans. Graph.*, vol. 36, no. 4, p. 93, 2017.

[15] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, p. e0196391, 2018.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)