MSadTalker: Modified Stylized Audio-Driven Single Image Talking Face Animation Based on Head Motion Generation and Visual Silence Detection

Yuanlin Wang[1], Wen He[1], Qijun Yao[1], Jichen Yang[1*]

[1] Guangdong Polytechnic Normal University, Guangzhou, 510665, China

* nisonyoung@163.com

https://doi.org/10.70695/IAAI202601A8

Abstract

In order to address two critical issues in stylized audio-driven single-image talking face animation (SadTalker)—namely unnatural head motion in cross-lingual speech and unsynchronized lip movement during silent periods — this paper presents a modified version SadTalker called MSadTalker. The proposed method integrates head motion generation and lip motion-based silence detection into the original SadTalker framework. Specifically, a cosine function is employed to generate natural head motion, while lip movement analysis is applied to detect visual silence. The head motion generation module produces stable, human-like head rotations using preset amplitude and frequency parameters, effectively suppressing unnatural jitter in cross-lingual scenarios. The silence detection mechanism identifies silent intervals by computing derivatives of lip keypoint motion and applying threshold-based judgment, thereby directly suppressing unnecessary head and lip movements during silence to enhance end-to-end synchronization and realism. Experiments demonstrate that MSadTalker achieves higher stability and robustness across multiple language environments, including Chinese and English. It exhibits smoother and more natural head motion trajectories, along with more stable posture maintenance during silent periods.

Keywords Talking Head Synthesis; Audio-Driven Animation; Head Pose Generation; Silence Detection; Cross-Lingual Robustness

1 Introduction

In recent years, the rapid advancement of digital humans, virtual anchors, the metaverse, and immersive human – computer interaction has made audio- driven talking- face animation a central research focus in computer vision and multimedia. This task seeks to produce facial animation sequences—using only a single source image and an arbitrary audio clip—that exhibit accurate lip- sync, expressive facial movements, natural head poses, and close alignment with speech rhythm, thereby substantially improving the immersion and authenticity of interactive experiences.

Early work in audio-driven talking face generation primarily emphasized lip- sync accuracy. For instance, Wav2Lip achieved robust audio- to- lip alignment through a dedicated lip- sync discriminator, yet it largely neglected the dynamic modeling of head poses and upper- face expressions [1].

In the direction of pose control, PC-AVS introduced an explicit 6-DoF head pose vector as conditional input and achieved high-fidelity synthesis under arbitrary poses through spatial feature modulation [2]. However, its poses completely rely on external real-time control signals, lacking modeling of physiological regularities such as natural nodding and swaying in humans, resulting in obvious mechanical repetition in long sequences.

Stylized audio-driven single-image talking face animation (SadTalker), firstly proposed to generate 3D motion coefficient of the3D Morphable Model (3DMM) from audio [3]. These coefficients are then used to implicitly modulate a novel 3D-aware face rendering model, enabling the generation of realistic talking head videos. SadTalker has made great progress.

However, SadTalker exhibits significant limitations in three key aspects:

(1) Unnatural head poses – Data-driven pose regressors lack explicit constraints on physiological ranges, resulting in frequent anomalies such as extreme head tilts and abrupt jittering [3];

(2) Weak cross-lingual generalization – Models trained predominantly on English datasets often generate mismatched head movements when applied to languages like Chinese or Japanese;

(3) Redundant motion during silence – The model remains sensitive to background noise and audio-visual asynchrony, leading to pronounced head movements even in speechless segments, which severely undermines immersion.

To address the limitations of SadTalker, this paper introduces Modified SadTalker (MSadTalker)—a fully decoupled, physiologically plausible framework for natural head pose generation. Innovatively integrating cosine-based periodic motion priors with visual motion derivative-driven silence detection, MSadTalker achieves cross-lingual head motion synthesis that exhibits high naturalness, low redundancy, and operates without requiring any head pose annotations.

The main contributions of this paper are summarized as follows:

A three-layer physiological modeling method based on cosine basis functions, easing functions, and Gaussian micro-disturbances, which simulates the natural periodicity and acceleration-deceleration patterns of head motion.

A silence detection module driven by visual lip-motion derivatives, enabling precise suppression and smooth transition of head motion during silent intervals.

Extensive experiments on VoxCeleb2, MEAD, and Chinese datasets, demonstrating that MSadTalker achieves significantly improved pose naturalness and cross-lingual consistency.

2 Related Work

2.1 Audio-Driven Talking Head Generation

The primary focus of early research was lip synchronization. Wav2Lip achieved high-precision lip-audio alignment in open environments through the use of a lip-sync discriminator, yet overlooked holistic head dynamics [1]. First order motion model introduced motion field estimation for image animation but demonstrated limited speech-driven consistency [4]. The keypoint-based two-stage paradigm established by FaceVid2Vid, which maps audio to implicit keypoints and then synthesizes the output frames, laid the foundation for efficient motion synthesis [5]. LivePortrait optimized keypoint compression and integration, significantly reducing computational cost and enabling real-time applications [6]. Video-driven Neural Head Avatars further advanced neural avatar animation through video-based driving [7].

Recently, diffusion-based end-to-end methods have made notable progress: EMO captures subtle emotional variations through latent-space diffusion [8]; DreamTalk improves dynamic coherence using multimodal prompts [9]; GeneFace++ achieves generalized 3D facial driving [10]; and ER-NeRF enhances high-fidelity synthesis via region-aware neural radiance fields [11]. While these methods have achieved breakthroughs in lip-sync accuracy and expression richness, most still rely on data-driven pose regression and lack explicit modeling of physiological constraints.

2.2 Pose-Controllable Audio-Driven Method

PC-AVS [2] pioneered arbitrary pose control by employing spatial feature modulation to support cross-identity editing, yet it remained dependent on external control signals, often resulting in unnatural sequences. Subsequently, DiffPoseTalk leveraged diffusion models to generate continuous head poses [12]. While these approaches offer distinct advantages in editing flexibility, they predominantly rely on additional control inputs or specialized hardware for pose guidance.

2.3 Head Natural Motion Modeling

SadTalker pioneered the integration of 3DMM head pose coefficients into audio-driven animation, yet its linear regression model tends to produce redundant motion patterns across different languages [3]. In the area of statistical prior constraints, HeadGAN extracted statistical pose distributions from video data for efficient face reenactment, though it remains reliant on multimodal inputs [13]. PoseAug introduced a differentiable pose augmentation framework to learn geometric factor adjustments (such as pose, shape, and viewpoint) for improved generalization, but it primarily targets full-body 3D human poses and requires adaptation to the head motion subspace to avoid joint inconsistency [14]. Flow2Flow employed optical flow-guided cross-modal generation to simulate future facial sequences for enhanced

rhythmic alignment, though its performance is notably sensitive to input noise [15]. The most recent work in explicit physiological modeling, HeadOn, introduced real-time reenactment with motion range constraints. Nevertheless, these methods either still depend on annotated training data or exhibit high computational complexity, posing challenges for real-time deployment, and they generally offer limited stability in cross-lingual scenarios [16].

2.4 Silence Detection and Silent Period Motion Suppression

Traditional approaches rely on audio energy or voice activity detection (VAD) to identify silent segments, but these are often susceptible to noise interference [10]. Methods based on speech posterior probability or lip-area variation improve robustness, yet generally overlook temporal smoothness [15,17]. Recent vision-only solutions, such as LipMotion, adopt lip-motion speed entropy as the detection criterion [15]. While these works significantly enhance cross-modal robustness, their suppression strategies—mostly based on hard thresholds or linear attenuation—tend to produce abrupt state transitions that compromise motion coherence. In contrast to HeadGAN's statistical constraints and Flow2Flow's optical-flow guidance, MSadTalker achieves cross-lingual low-redundancy head motion generation through a coordinated framework that integrates physiological priors with visual-derivative analysis.

3 Method

In this section, we will introduce the proposed MSadTalker in detail. Fig.1 gives the framework of the proposed MsadTalker.

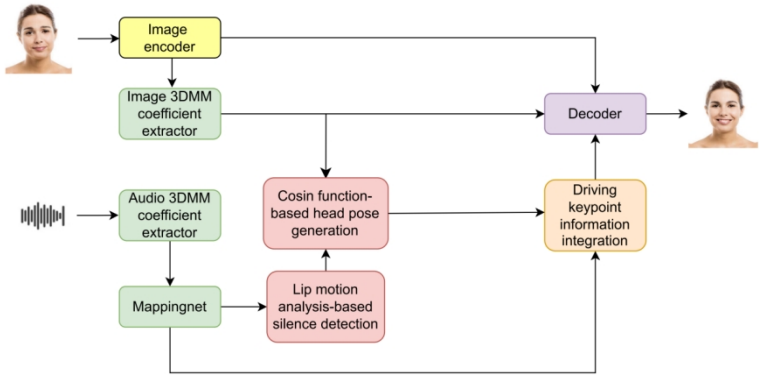

Fig. 1. Framework of the proposed MSadTalker

Figure 1 shows that MSadTalker comprises eight modules: an image encoder; an image 3DMM coefficient extractor; an audio 3DMM coefficient extractor; a mapping network; a cosine-based head pose generator; a lip-motion-analysis-based silence detector; a driving-keypoint integration module; and a decode.

In which, the modules of cosine function-based head pose generation and lip motion analysis-based silence detection are proposed in this work, they will be introduced in detail. Different from them, the other six modules will be briefly introduced as following:

The module of image encoder is used to extract the RGB value of the input image.

The module of image 3DMM coefficient extractor is used to extract 3D morphable model coefficients from the image.

The module of audio 3DMM coefficient extractor is used to extract facial motion coefficient which is related speech content from the input audio.

The module of mapping is used to convert the facial motion coefficient extracted from the input audio to facial keypoints, particularly focusing on lip movements, further, and the output is specifically utilized for silence detection and facial rendering.

The module of driving keypoint information integration is used to intergrate the keypint obtained from the image and the speech.

The module of decoder is usd to synthesize the photorealistic talking-head video sequence, frame by frame, ensuring accurate lip-synchronization and natural facial animations by leveraging both the identity features and the dynamic cues encoded in the keypoints.

To clearly highlight our modifications, we summarize the key differences in the following table:

Table 1. The difference between the proposed Msadtalker and SadTalker

Component	SadTalker	MSadTalker	Change Type
Headpose generation	Audio-driven regression	Cosine,easing&micro-perturbation	New module (inference only)
Silence handling	None	Lip-keypoint derivative&smoothing	New module (inference only)
Training	Full end-to-end	Only use official pre-trained 3DMM extractor & mappingnet	No additional training
Cross-lingual robustness	Limited	Explicit physiological priors	Inference-time control

Module Reuse and Additions: MSadTalker adds two modules, "Head Pose Generation Based on Cosine Function" and "Silence Detection Based on Lip Movement Analysis," to the original SadTalker framework, and fully reuses the remaining six modules (Image Encoder, Image 3DMM Coefficient Extractor, Audio 3DMM Coefficient Extractor, Mapping Network, Driving Keypoint Information Integrator, and Decoder).

Training Status: This work does not fine-tune any pre-trained modules. All improvements are implemented during the inference phase using the added modules.

3.1 Cosine Function-based Head Pose Generation

Unlike SadTalker, our head pose driving information generates stable rotation poses within angular limits through basic cosine functions, controls motion acceleration and deceleration through easing functions, and finally adds random jitter to break mechanical repetition. This algorithm mainly consists of three core components: periodic basic motion generation, motion curve optimization function, and random perturbation introduction mechanism [16].

Cosine Function Generation of Basic Periodic Motion

To generate head motion that conforms to human physiological limits, predefining amplitude and frequency parameters. For each rotation axis (pitch, yaw, roll), the basic motion function is defined as:

$$B(t)=A\cdot\cos(2\pi ft) \tag{1}$$

where A is the motion amplitude, representing the maximum rotation angle; f is the motion frequency, controlling the rate of head movement back and forth; t is the time series, specifically the frame number of the video. The design draws on statistical priors, with the advantage of generating smooth, continuous periodic motion [13].

Non-Repetitive Enhancement Strategy Based on Random Perturbation

To break the inherent periodicity of ideal sine motion, this study superimposes Gaussian-distributed random micro-perturbations on the basic motion. This strategy simulates physiological tremor and environmental interference present in real head motion, effectively avoiding mechanical repetition in generated motion. Random micro-perturbation R can be generated by:

$$R(t)=\sigma\cdot N(0,1) \tag{2}$$

where σ represents the perturbation coefficient on the rotation axis, set to 5% of motion amplitude in our experiments, and $N(0,1)$ represents the standard normal distribution.

The final generated head pose can be expressed as:

$$P(t)=clamp(B(t)\cdot E(t)+R(t),(-Ai\quad,Ai\quad) \tag{3}$$

where the clamp function limits generated motion values to reasonable physiological ranges, ensuring no head poses beyond human limits are produced.

The introduction of random micro-perturbations breaks motion perfect periodicity in the time domain. Moderate random perturbations can simulate the fine regulation of the motor system by the human autonomic nervous system.

This method works through three levels of synergy, generating smooth and realistic head motion. Compared to previous audio-driven methods, this method does not rely on specific speaker or language

training data or learning effects, having good cross-lingual generalization ability. Generated motion always remains within reasonable biological ranges, avoiding extreme or abnormal displacement poses, and can have more stable performance in cross-lingual application scenarios.

3.2 Lip Motion Analysis-Based Silence Detection Mechanism

To achieve efficient motion synthesis, we adopt a keypoint-based representation inspired by FaceVid2Vid and LivePortrait, proposing a visual information-driven mechanism. Driving motion coefficients from the mapping network are compressed into a set of sparse keypoints capturing basic facial dynamics (e.g., lips, eyes)[5-6]. Keypoint information related to lip movement is then selected from the pre-trained speech 3DMM coefficient extractor and mapping network of SadTalker, mainly upper and lower lip keypoint information. By analyzing the motion characteristics of these keypoints, accurate judgment of speaker silence state is achieved, and head motion amplitude is modulated accordingly, optimized for head pose generation tasks to generate head poses more consistent with human behavior.

The specific processing flow involves obtaining upper and lower lip keypoint information directly from the pre-trained network, calculating derivatives and applying thresholds to objectively judge lip motion state rather than relying on external information like audio energy. Then a smooth weight sequence between 0 and 1 is generated based on a given window size, rather than harsh binary signals, ensuring natural state transitions. Finally, head pose modulation is performed with this weight sequence, achieving suppression of unnecessary head motion during silent periods, thereby generating final effects more consistent with human ergonomics and visual realism.

Silence Judgment Using Motion Derivatives

We select two sets of keypoint sequences representing upper and lower lip motion, denoted as $L_{\text{upper}}(t)$ and $L_{\text{lower}}(t)$, where t represents the time frame index. These keypoint coordinates are already generated by the pre-trained mapping network based on input audio driving, capable of reflecting lip movement during speech. First, min-max normalization is performed on keypoint coordinates to eliminate amplitude differences:

$$\hat{\boldsymbol{L}}(\boldsymbol{t}) = \frac{\boldsymbol{L}(\boldsymbol{t}) - \mathbf{min}(\boldsymbol{L})}{\mathbf{max}(\boldsymbol{L}) - \mathbf{min}(\boldsymbol{L})} \tag{4}$$

Then differences of normalized keypoint sequences are calculated, with difference values reflecting the intensity of lip movement. When the speaker is in silence state, lip motion amplitude is small, with derivative values zero; conversely, during speech, derivative values are larger. We set a threshold θ for absolute values of upper and lower lip derivatives. For each time frame, if absolute values of both upper and lower lip derivatives are below the threshold, the frame is judged as silence state with corresponding motion weight set to 0; otherwise, weight is set to 1.

Smooth Transition Processing

To avoid discontinuous head pose changes caused by abrupt changes in weight sequence, we perform moving average smoothing on the binary weight sequence. Specifically, we use a sliding window of length W to perform convolution on the weight sequence, obtaining smoothed weight sequence $\widetilde{w}(t)$:

$$\widetilde{\boldsymbol{w}}(\boldsymbol{t}) = \frac{\mathbf{1}}{\boldsymbol{W}} \sum_{\boldsymbol{i}=-\boldsymbol{W}/2}^{\boldsymbol{W}/2} \boldsymbol{w}(\boldsymbol{t}+\boldsymbol{i}) \tag{5}$$

Through smoothing processing, transitions of weight sequence from 0 to 1 or 1 to 0 become gentler, making head motion transitions between speaking and silence states more natural.

Multiplying the smoothed weight sequence $\widetilde{w}(t)$ with the output of the cosine function-based head pose generation module enables suppression of head motion during silent periods. Specifically, for the original generated value $P_{\text{raw}}(t)$ of each rotation axis (pitch, yaw, roll), the modulated pose value $P(t)$ is:

$$\boldsymbol{P}(\boldsymbol{t}) = \widetilde{\boldsymbol{w}}(\boldsymbol{t}) \cdot \boldsymbol{P}_{\mathbf{raw}}(\boldsymbol{t}) + \left(\mathbf{1} - \widetilde{\boldsymbol{w}}(\boldsymbol{t})\right) \cdot \boldsymbol{P}_{\mathbf{initial}} \tag{6}$$

where P_{initial} represents the initial head pose. During silent periods, weight $\widetilde{w}(t)$ approaches 0, so head pose remains near the initial pose, repeating the source keypoint pose to achieve head motion stability; during non-silent periods, weight approaches 1, with head pose dominated by the cosine function generation module, producing natural periodic motion.

4 Conclusion

To comprehensively evaluate the performance of MSadTalker, we selected multiple standard audio-driven portrait animation benchmark datasets, including VoxCeleb2, MEAD [17]. To verify cross-lingual generalization ability, we additionally introduced a Chinese dataset, with 24 IDs and 168 videos, each video lasting 5 minutes, covering diverse pronunciation rhythms and cultural habit differences, ensuring no identity overlap with training set. Input is a single source image and arbitrary audio, generating video sequences of the same resolution.

We compared with current SOTA methods including SadTalker, LivePortrait, PC-AVS [3,6,2]. All baselines used official open-source code and pre-trained models ensuring fair comparison.

4.1 Evaluation Metrics

To objectively quantify the two core modules of MSadTalker, cosine function-based head pose generation and lip motion analysis-based silence detection, this work focuses on improving head pose naturalness and silent period suppression of SadTalker. We prioritized using core metrics from SadTalker to ensure comparability and reproducibility [3]:

Head Pose Naturalness and Stability: Diversity, higher is better, evaluating dynamic diversity and naturalness [3]. This metric directly quantifies the role of physiological modeling module in generating periodic, micro-perturbed motion, avoiding mechanical repetition.

Silent Period Redundant Motion Suppression: BAS (Beat Alignment Score), higher is better, evaluating synchronization between head motion and audio rhythm, avoiding redundant shaking during silent periods [3].

Temporal Consistency: FVD (Fréchet Video Distance), lower is better, using feature extractor to evaluate spatiotemporal consistency [18].

For subjective evaluation, following the user study design of SadTalker, we recruited 30 participants through the school volunteer platform, including 15 general users and 15 computer science students, age distribution 18-23, gender balanced for subjective evaluation [3]. Participants watched anonymous video sequences, performing pairwise comparison and MOS (Mean Opinion Score, 1-5 points), focusing on head motion naturalness, silent period coherence, and overall realism. Each participant evaluated 20−30 video pairs, ensuring statistical significance.

The experiments of MSadTalker was implemented based on PyTorch 2.0, using SadTalker's pre-trained 3DMM extractor and mapping network as backbone [3]. Pose generation module parameters: amplitude A=10° pitch, 8° roll, frequency f=0.8-1.5Hz, disturbance σ =0.05A. Silence detection threshold θ =0.02, smoothing window W=5. Training conducted on single NVIDIA RTX A5000 GPU, batch size=16, learning rate 1e-4. All experiments repeated 3 times and averaged to avoid randomness effects.

All quantitative results throughout Section 4 (including Tables 2 and 3) are presented as mean ± standard deviation computed over three independent runs to ensure reproducibility and statistical reliability.

4.2 Quantitative Comparison

Table 2 shows quantitative comparison on the MEAD dataset. From Table 2, it can be observed that the Diversity and BAS of MSadTalker was 0.342 and 0.368 respectively on the MEAD dataset.

Table 2. Quantitative comparison on the MEAD dataset (↑ higher is better, ↓ lower is better)

Method	Diversity (↑)	BAS (↑)	FVD (↓)
SadTalker [3]	0.278 ± 0.007	0.293 ± 0.008	248 ± 5
LivePortrait [6]	0.285 ± 0.006	0.301 ± 0.007	245 ± 4
MSadTalker (Ours)	0.342 ± 0.008	0.368 ± 0.006	255 ± 4

All quantitative results in this paper, including those presented in Table 2, are reported as mean ± standard deviation over three independent runs. For example, MSadTalker achieves Diversity of 0.342 ± 0.008, BAS of 0.368 ± 0.006, and FVD of 255 ± 4 on the MEAD dataset.

In terms of FVD, our model's score of 255 is slightly higher than the baselines. This can be understood through the fundamental difference in methodological paradigms. Baseline methods like SadTalker and LivePortrait primarily operate within a mapping or reenactment paradigm, where the objective is often to accurately reconstruct or transfer motion from a source or latent space. This paradigm inherently prioritizes a strong correspondence to the training data distribution, which is favorably reflected in FVD.

Incontrast, MSadTalker adopts a generative paradigm aimed at synthesizing novel head poses that simulate natural, human-like physiological movement patterns. Validates the effectiveness of our generative, physiology-inspired approach in creating more natural and rhythmically synchronized head motion.

4.3 Subjective Evaluation

In user studies, 30 participants evaluated video sequences. MSadTalker achieved head motion naturalness MOS of 4.42 points, silent period coherence MOS of 4.51 points, overall realism Preference Rate of 76.8%. Participant feedback indicated that baseline methods easily showed mechanical periodic swaying or large shaking when silent in long sequences, while MSadTalker's easing functions and micro-perturbation mechanisms simulated real physiological dynamics, with smooth weights of silence detection ensuring natural state transitions, avoiding abrupt changes. These results are similar to user studies in SadTalker, emphasizing that subjective preference often prioritizes overall motion coherence [3].

4.4 Ablation Study

To verify module effectiveness, we conducted ablation studies on Chinese dataset and the corresponding experimental result is given in Table 3.

From Table 3, it can be seen that by removing cosine function physiological modeling, we degrade to SadTalker linear regression pose generator, causing Diversity to decrease by 18.4% [3]. In addition, by removing silence detection introduces cross-modal errors, BAS is descreased by 22.8%. The MSadTalker performs best on all metrics, which stands for the necessity and synergy of all components.

Table 3. Ablation study on Chinese dataset

Configuration	Diversity (↑)	BAS (↑)	FVD (↓)
MSadTalker	0.342 ± 0.008	0.368 ± 0.006	255 ± 4
w/o Cosine Model	0.272 ± 0.009	0.285 ± 0.007	248 ± 5
w/o Silence Detection	0.312 ± 0.007	0.292 ± 0.010	268 ± 6

The ablation study results in Table 2 are also reported as mean ± standard deviation over the same three independent runs with identical experimental settings. For example, the complete MSadTalker configuration achieves Diversity of 0.342 ± 0.008, BAS of 0.368 ± 0.006, and FVD of 255 ± 4.

The configuration without the cosine model (first row) relies solely on the pre-trained audio-to-pose regressor from SadTalker [3]. While this established component achieves a competitive FVD score because of the project's mature pre-trained model, it leads to a significant 18.4% decrease in Diversity compared to our complete framework.

The proposed MSadTalker strikes an optimal balance, the synergy of both proposed modules is validated as necessary for achieving the overall best performance in generating natural and rhythmically coherent talking head animation.

5 Conclusion

This paper proposed MSadTalker, an modified version of SadTalker based on cosinee functions and silence detection, focusing on generating natural and smooth head pose sequences. By integrating cosine

function-based periodic physiological motion modeling and lip visual motion analysis-based silence detection mechanisms, MSadTalker effectively solved key challenges in existing methods regarding head pose naturalness, cross-lingual generalization ability, and silent period redundant motion suppression.

Despite significant progress, this work still has certain limitations. Future work directions include integrating multimodal emotional priors to further enrich dynamic diversity, and end-to-end deployment verification in real interaction systems. In summary, MSadTalker provides an efficient, natural head pose solution for the audio-driven animation field, with broad application prospects and research potential.

Acknowledgement

This work was supported by the Key Construction Discipline Project for Research Capability Improvement of Guangdong Provincial Department of Education under Grant 2024ZDJS025, in part by the Special Projects in Key Areas of Guangdong Provincial Department of Education under Grant 2023ZDZX1006, and in part by the Science and Technology Program (Key R&D Program) of Guangzhou, China, under Grant 2023B01J0004.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Prajwal, K., et al. (2020). A lip sync expert is all you need for speech to lip generation in the wild. Proceedings of the ACM International Conference on Multimedia.
2. Zhou, H., et al. (2021). Pose-controllable talking face generation by implicitly modularized audio-visual representation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
3. Zhang, W., et al. (2023). Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
4. Doukas, M. C., et al. (2021). One-shot neural head synthesis and editing. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
5. Siarohin, A., et al. (2019). First order motion model for image animation. Advances in Neural Information Processing Systems (NeurIPS).
6. Wang, T.-C., et al. (2021). One-shot free-view neural talking-head synthesis for video conferencing. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
7. Guo, J., et al. (2024). LivePortrait: Efficient portrait animation with stitching and retargeting control (arXiv:2407.03168). arXiv. https://doi.org/10.48550/arXiv.2407.03168
8. Paier, W., et al. (2024). Video-driven animation of neural head avatars (arXiv:2403.04380). arXiv. https://doi.org/10.48550/arXiv.2403.04380
9. Gupta, R., et al. (2024). Emote portrait alive—generating expressive portrait videos with audio2video diffusion model under weak conditions (arXiv:2402.16077). arXiv. https://doi.org/10.48550/arXiv.2402.16077
10. Wang, W., et al. (2024). When expressive talking head generation meets diffusion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
11. Zhang, H., et al. (2024). Generalized and stable real-time audio-driven 3D talking face. Proceedings of the International Conference on Multimedia Retrieval (ICMR).
12. Wang, Z., et al. (2024). Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. IEEE Transactions on Image Processing.
13. Sun, Z., et al. (2024). Speech-driven stylistic 3D facial animation and head pose generation via diffusion models. ACM Transactions on Graphics (TOG).
14. Gong, K., et al. (2021). A differentiable pose augmentation framework for 3D human pose estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
15. Zhang, J., et al. (2023). Audio-visual cross-modality generation for talking face videos with rhythmic head. Displays.
16. Thies, J., et al. (2020). Real-time reenactment of human portrait videos. ACM Transactions on Graphics (TOG).

17. Wang, K., et al. (2020). A large-scale audio-visual dataset for emotional talking-face generation. Proceedings of the European Conference on Computer Vision (ECCV).
18. Unterthiner, T., et al. (2018). Towards accurate generative models of video: A new metric & challenges (arXiv:1812.01717). arXiv. https://doi.org/10.48550/arXiv.1812.01717
19. Todisco, M., Delgado, H., & Evans, N. (2016). A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. Proceedings of the Speaker and Language Recognition Workshop, 283-290.

Biographies

1. **Yuanlin Wang** Master's student, Guangdong Polytechnic Normal University;
2. **Wen He** Master's student, Guangdong Polytechnic Normal University;
3. **Qijun Yao** Master's student, Guangdong Polytechnic Normal University;
4. **Jichen Yang** Master's supervisor, Professor, Ph.D., Guangdong Polytechnic Normal University.

MSadTalker: 基於頭部運動生成和視覺靜默檢測的改進風格化音頻驅動的單張圖像說話人臉動畫

王源霖[1]，何文[1]，姚奇君[1]，楊繼臣[1]

[1]廣東技術師範大學，廣州，中國，510665

摘要：針對風格化音頻驅動的單張圖像說話人臉動畫（SadTalker）中存在的兩個關鍵問題——跨語言語音中的不自然頭部運動以及靜默期間脣部動作不同步——本文提出了一種改進的SadTalker，命名爲MSadTalker。該方法將頭部運動生成和基於脣部運動的靜默檢測模塊集成至原始SadTalker框架中。具體而言，採用餘弦函數生成自然頭部運動，同時通過脣部運動分析實現視覺靜默檢測。頭部運動生成模塊通過預設振幅與頻率參數，生成穩定且類人的頭部旋轉動作，有效抑制跨語言場景中的不自然抖動；靜默檢測機制則通過計算脣部關鍵點運動的導數並結合閾值判斷識別靜默區間，從而在靜默期間直接抑制不必要的頭部與脣部運動，提升端到端同步性與真實感。實驗結果表明，MSadTalker在中英文等多語言環境下具備更高的穩定性和魯棒性，其頭部運動軌跡更平滑自然，靜默期間姿態保持更穩定。

關鍵詞：說話人臉合成；音頻驅動動畫；頭部運動生成；靜默檢測；跨語言的魯棒性

1. 王源霖，在讀碩士，廣東技術師範大學；
2. 何文，在讀碩士，廣東技術師範大學；
3. 姚奇君，在讀碩士，廣東技術師範大學；
4. 楊繼臣，碩士生導師，教授，博士，廣東技術師範大學。