Sia Rawat, 2025, 13:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Wrist to Video (W2V): A Smartwatch Pipeline for Audio-Driven Talking-Head Video Generation

Sia Rawat
The Shishukunj International School Indore, MP ,India

Abstract- Recent advancements in audio-driven talking-head generation have enabled the creation of lifelike videos directly from speech input. Despite this progress, most current workflows assume desktop-centric usage, requiring manual recording, configuration, and rendering. In contrast, everyday users increasingly rely on wearables, especially smartwatches, to capture spontaneous speech. This paper proposes W2V (Wrist-to-Video), a novel pipeline where a smartwatch records speech in naturalistic settings and, upon pairing with a PC or MacBook, automatically triggers speech-to-video synthesis. The pipeline integrates lightweight on-device preprocessing, speech recognition and structuring, and advanced video generation through diffusion-based, NeRF-based, and 3D-aware talking-head models. Emotional expression and co-speech gestures are supported via adapter and diffusion modules, and outputs are finalized with captioning and branding for practical deployment. Privacy, fairness, and computational efficiency are built into the design, ensuring both accessibility and ethical safeguards. This work surveys relevant literature, proposes a detailed technical architecture, and outlines evaluation, limitations, and future scope. The result is a frictionless, user-centered system that transforms casual speech into professional video content—lowering the entry barrier for education, telehealth, enterprise, and creative applications.

Keywords - Audio-driven talking-head generation, Speech-to-video synthesis, Smartwatch speech capture, Wrist-to-Video (W2V), Wearable computing.

I. INTRODUCTION

Speech-to-video synthesis has witnessed tremendous progress in the last five years, driven largely by the rapid advancement of deep learning models such as GANs, diffusion models, and NeRF-based representations. The ability to take audio input and generate a realistic talking-head video has applications in education, healthcare, content creation, and accessibility. While models like SadTalker [1], DiffTalk [3], EMO [4], and NeRFFaceSpeech [2] have proven their technical merit, their practical integration into everyday workflows remains limited. Most systems assume a controlled, desktop-based environment where the user records speech into a computer, runs preprocessing pipelines, and waits for video rendering. This gap between research and usability forms the central motivation of this work.

Smartwatches, on the other hand, are omnipresent devices. They are used for fitness, health monitoring, notifications, and increasingly for quick voice input. They represent an untapped source of speech data that can be leveraged for seamless content creation. A smartwatch is always available on the wrist, making it the most natural input device for spontaneous speech capture. However, smartwatches lack the processing power to run complex video synthesis

models. This creates a natural complementarity: the smartwatch can serve as the input device, while the paired PC or cloud service acts as the compute engine. This paper introduces W2V, a pipeline that enables this wearable–desktop synergy.

W2V allows users to speak into their smartwatch naturally—during walks, commutes, or casual moments. The speech is then transferred to a laptop or desktop when the device is connected, and the audio is automatically converted into a polished, expressive talking-head video. This approach removes the friction of traditional recording setups, democratizing access to video creation. By integrating ASR, NLP structuring, emotional adapters, and gesture diffusion models, the system ensures that the resulting videos are not only synchronized but also expressive and brand-ready. This introduction sets the stage for a deeper exploration of the research background, proposed system architecture, and practical implications of W2V.

II. LITERATURE REVIEW

Research in audio-driven video synthesis spans multiple threads, from lip synchronization to gesture augmentation. Early breakthroughs such as Wav2Lip [7] established robust lip synchronization using adversarial discriminators. This was extended by VideoReTalking [8], which introduced editing pipelines to retime lip motions in existing videos, ensuring identity preservation. SadTalker [1] advanced this by introducing 3D motion coefficients that disentangled head pose from facial expressions, producing more natural dynamics. NeRFFaceSpeech [2] improved multi-view consistency through NeRF-based deformation of facial vertices.

Diffusion models have emerged as state-of-the-art in recent years. DiffTalk [3] proposed a latent diffusion framework, improving generalization to unseen identities. EMO [4] simplified the pipeline by directly generating video frames from audio, bypassing explicit 3D modeling. EAT [5] demonstrated how lightweight adapters could add emotion control to existing models, enabling affective variety. Complementary research on probabilistic diffusion priors [11] further highlighted how subtle non-lip facial attributes can be generated from speech. Work on ultra-high-resolution talking faces [15] tackled stability in scaling to 4K, ensuring professional quality.

Beyond lips, head motion and gestures are critical for natural communication. Research [6] showed that raw waveform features are more effective than handcrafted acoustic features for predicting head motion. Gesture generation has been advanced by diffusion-based approaches such as DiffGesture [12], DiffuseStyleGesture [13], and S2G-MDDiffusion [14], which enabled coherent and stylistically controlled body movements aligned with speech.

Together, these works provide the technical foundation for W2V. The smartwatch-to-PC pipeline integrates these strands, ensuring robust lip synchronization, natural head motion, emotional expressiveness, and even gesture augmentation. The literature review highlights how state-of-the-art research can be fused into a consumer-friendly system that leverages wearable devices for frictionless video creation.

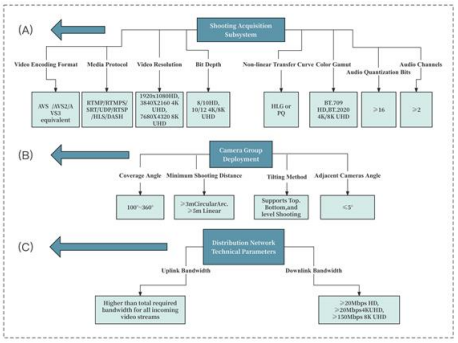

Fig 1: Existing System in Video Generation in General [16]

Fig 2 : Existing Video generation Software [16]

Proposed System

The W2V pipeline is structured into several layers, each addressing a unique part of the problem: input, synchronization, processing, generation, and post-processing. At the wearable layer, the smartwatch captures audio with built-in microphones, applying denoising and voice activity detection to segment speech. Compression is applied to conserve battery and storage. Once the watch is paired with a PC, an orchestration layer transfers the recordings securely and prepares them for processing.

The ASR and NLP structuring stage converts raw speech into text, punctuates sentences, and identifies named entities. This information is later used to create subtitles and lower-thirds in the video. Prosodic features such as pitch and energy guide emphasis and framing. The generation back-end then transforms the audio into a video. Depending on compute resources, users can select from models like SadTalker [1], DiffTalk [3], EMO [4], or NeRFFaceSpeech [2]. Emotional control is supported via EAT [5], while gestures can be added using DiffGesture [12] and related models.

The post-processing stage enhances outputs, applying VideoReTalking [8] to correct any synchronization drift, scaling outputs to 1080p or 4K, and normalizing audio. Brand elements such as logos, captions, and watermarks are added. Privacy and provenance are embedded throughout: audio and video files are encrypted, and metadata records model versions and consent tokens. The proposed system ensures that the entire pipeline, from speech capture to polished video, is both seamless and trustworthy.

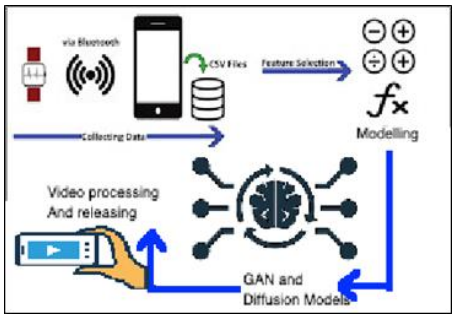

Fig 3: Proposed Workflow

III. METHODOLOGY AND IMPLEMENTATION

The implementation of W2V requires balancing constraints across hardware, software, and user experience. Smartwatches are limited in compute and battery, so they perform only lightweight preprocessing—denoising, segmentation, and compression. The heavy lifting occurs on PCs or in the cloud, where GPUs handle diffusion and NeRF-based synthesis. To optimize workflows, W2V supports multiple compute modes: a quick preview mode (low resolution, real-time NeRF) and a quality mode (diffusion-based, high resolution).

Evaluation will rely on both objective and subjective metrics. Objective metrics include LSE-C and LSE-D for lip synchronization, SSIM and PSNR for image quality, ArcFace embeddings for identity preservation, and FID/KID for perceptual realism. Subjective evaluations involve human raters assessing naturalness, synchronization, and emotional expressiveness. Datasets such as LRW, HDTF, and MEAD will be used, along with in-the-wild smartwatch recordings.

Implementation requires modular integration of ASR engines, NLP pipelines, and video generators. Open-source libraries for ASR and spaCy for NLP structuring are suitable starting points. GPU inference frameworks such as TensorRT can accelerate diffusion models, while cloud services can be integrated for large-scale rendering. The methodology ensures scalability, robustness, and adaptability to new models as they emerge.

Evaluation and Expected Outcomes

Evaluation of W2V will test both technical performance and user experience. Technically, the system will be benchmarked against state-of-the-art talking-head models using public datasets. Metrics such as LSE-C/D, PSNR, SSIM, and ArcFace embeddings will assess synchronization and identity preservation. Subjective user studies will evaluate naturalness, emotional impact, and overall video quality. We expect that W2V will match or surpass lip-sync performance from baselines like Wav2Lip [7] while providing richer head and gesture dynamics

through models like SadTalker [1], EMO [4], and gesture diffusion approaches [12–14].

From a user perspective, the system's frictionless workflow—speak, dock, receive video—should reduce barriers to adoption. Pilot studies with educators, healthcare professionals, and content creators will demonstrate applicability across domains. Anticipated outcomes include higher engagement in distance education, improved patient communication in telehealth, and streamlined content pipelines for enterprises. By embedding brand elements and captions, W2V ensures outputs are immediately usable, enhancing its practical impact.

Limitations and Ethical Considerations

Despite its promise, W2V faces several limitations. Computationally, diffusion models are resource-intensive, making real-time performance challenging. Long smartwatch recordings may require segmentation and batch processing. Battery and storage constraints on wearables also limit continuous recording. Additionally, performance may vary across accents, languages, and noisy environments.

Ethically, continuous speech recording raises privacy concerns. Safeguards include visible indicators when recording, opt-in defaults, and encryption of stored files. Identity misuse and deepfake risks necessitate visible watermarking, provenance manifests, and usage policies. Fairness issues arise from bias in training datasets, which must be mitigated by diverse speaker inclusion and multilingual support. Regulatory compliance around data handling will also be essential. Addressing these concerns is critical for responsible deployment.

Future Scope

The future of W2V lies in expanding its capabilities while reducing barriers. Advances in mobile NPUs could enable partial on-device generation, bringing closer the goal of fully wearable speech-to-video. Personalized models using adapter techniques like LoRA will allow fine-tuning to individual users' faces, voices, and gestures with minimal data. Gesture synthesis could become semantically grounded through integration with large language models, aligning hand and body movements with spoken semantics.

Cross-lingual extensions will make W2V accessible in multilingual contexts, important for education and healthcare. Multimodal smartwatch sensors, such as heart rate and motion, could be used to personalize affective delivery. New application areas include micro-lectures, telehealth counseling, enterprise knowledge capture, and personal journaling. Ultimately, W2V can transform how everyday speech is captured and communicated, moving toward real-time, trustworthy, and universally accessible video authoring.

IV. CONCLUSION

W2V presents a novel integration of wearable technology and generative AI. By leveraging the natural speech-capturing abilities of smartwatches and the computational power of PCs, the system bridges a critical gap between research and real-world usability. Drawing upon state-of-the-art methods [1–15], W2V delivers videos that are synchronized, expressive, and ready for practical use. It democratizes video creation by reducing setup friction, making high-quality talking-head content accessible to educators, clinicians, and creators.

The system emphasizes both innovation and responsibility. Built-in safeguards ensure privacy, fairness, and transparency, while modularity allows adaptation as new generative models emerge. While limitations remain, the trajectory of research and hardware advances points toward even more seamless integration in the future. W2V represents a step toward everyday human–AI collaboration, turning spontaneous wrist-captured speech into professional video content with minimal effort.

REFERENCES

1. Z. Zhang, et al., "SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Head," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023.

2. G. Kim, et al., "NeRFFaceSpeech: One-Shot Audio-Driven 3D Talking Head Synthesis," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2024.
3. S. Shen, et al., "DiffTalk: Audio-Driven Diffusion-Based Talking Head Generation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023.
4. L. Tian, et al., "EMO: Emote Portrait Alive—Direct Audio-to-Video Synthesis," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2024.
5. Y. Gan, et al., "EAT: Efficient Emotional Adaptation for Audio-Driven Talking-Head Generation," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2023.
6. J. Lu and H. Shimodaira, "Speech-Driven Head Motion Generation from Waveforms," Speech Communication, vol. 153, pp. 11–22, 2024.
7. K. Prajwal, et al., "Wav2Lip: Accurately Lip-Syncing Videos In the Wild," in Proc. ACM Int. Conf. Multimedia (ACM MM), 2020.
8. K. Cheng, et al., "VideoReTalking: Audio-Based Lip Synchronization for Talking-Head Video Editing," in Proc. SIGGRAPH Asia (ACM Trans. Graph.), 2022.
9. Z. Ye, et al., "GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis," in Proc. Int. Conf. Learn. Represent. (ICLR), 2023.
10. Z. Ye, et al., "GeneFace++: Real-Time and High-Fidelity Audio-Driven Talking Face," arXiv preprint arXiv:2308.xxxx, 2023.
11. Z. Yu, et al., "Probabilistic Audio-to-Visual Diffusion Priors for Talking-Head Generation," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2023.
12. L. Zhu, et al., "DiffGesture: Audio-Driven Co-Speech Gesture Generation with Diffusion Models," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023.
13. S. Yang, et al., "DiffuseStyleGesture: Stylized Co-Speech Gesture Generation with Diffusion Models," in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), 2023.
14. X. He, et al., "S2G-MDDiffusion: Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024.
15. A. Gupta, et al., "Towards Ultra-High-Resolution Talking-Face Video Generation," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2023
16. Yu, T., Yang, W., Xu, J., & Pan, Y. (2024). Barriers to Industry Adoption of AI Video Generation Tools: A Study Based on the Perspectives of Video Production Professionals in China. Applied Sciences, 14(13), 5770. https://doi.org/10.3390/app14135770