Journal of Multidisciplinary Engineering Science Studies (JMESS)
ISSN: 2458-925X
Vol. 11 Issue 10, October - 2025

Singing Head Animation Using FFmpeg Expression Overlay

Hung-Che Shen
Department of Emerging Media Design
I-Shou University
Kaohsiung, Taiwan
e-mail: shungch@isu.edu.tw

***Abst[ract*—This paper presents an efficient and fully reproducible framework for 2D singing head animation utilizing an FFmpeg-based expression overlay technique. Unlike computationally demanding approaches such as 3D morphing or deep-learning-driven synthesis, this method emphasizes execution speed and accessibility. By combining a static base head image with a minimal set of predesigned facial expression layers—eyebrows, eyes, and mouth—the system enables expressive singing animation through simple command-line compositing. Each expression is sequentially layered using FFmpeg filter operations, producing perceptually convincing motion without specialized software or training data. Experimental results demonstrate that this lightweight method achieves rapid generation of expressive singing heads suitable for low-resource environments, educational settings, and rapid prototyping, and contributes a reproducible FFmpeg-based baseline for future research in lightweight singing animation synthesis.**

Keywords—Singing head; Animation; FFmpeg; Facial expression; Overlay

I. INTRODUCTION

Recent advances in singing animation have primarily relied on 3D modeling, facial rigging, and deep-learning-based audio-to-video generation. Although these methods achieve high realism and temporal synchronization, they typically demand extensive datasets, GPU resources, and expert knowledge. Such requirements create significant barriers for independent developers or small research teams aiming to explore expressive animation in lightweight and reproducible environments.

Meanwhile, the growing accessibility of AI-based image generation tools has made it remarkably simple to produce multiple, stylistically consistent facial expressions within minutes. This contrast—between the complexity of modern animation pipelines and the ease of expression image generation—motivates the present study.

The goal of this research is to examine whether an efficient and easily reproducible workflow can be realized using only FFmpeg expression overlay and a minimal set of pre-generated facial expression images. By compositing static head and expression layers—such as eyebrows, eyes, and mouth—through FFmpeg commands, the method achieves synchronized facial motion with near-instantaneous rendering and minimal computational cost, eliminating the need for 3D modeling or neural inference.

This study serves as a technical demonstration of how open-source media tools and AI-generated visual assets can jointly lower the entry barrier to singing animation research. It provides a foundation for evaluating the expressive potential of image-based overlay techniques before integrating audio-driven or emotion-based control systems. The overall compositing pipeline for layering facial components, including eyes, mouth, and expression masks, is illustrated in Fig. 1, which shows how the base head image is sequentially merged with dynamic overlay layers.

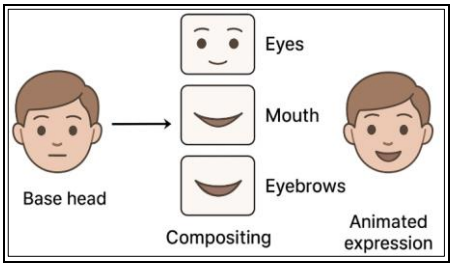

Fig. 1. *Diagram of the overlay pipeline (base head + expression layers).*

All experimental resources—including example head images, expression sets, and FFmpeg command scripts—are publicly available on GitHub, with a demonstration video hosted on the project blog. The remainder of this paper is organized as follows: Section II reviews related works on facial animation and video compositing; Section III describes the FFmpeg-based implementation; Section IV presents results; and Section V concludes with discussion and future work.

II. Related Work

A. Singing and Audio-Driven Facial Animation

Research on singing face animation has largely evolved from the broader domain of talking face synthesis. Early approaches relied on parametric facial models and viseme-to-phoneme mapping, such as those in classic 3D animation and lip-sync systems. While effective for speech, these techniques often struggle with singing, where prolonged vowels, vibrato, and emotional dynamics deviate from conversational prosody.

More recent works employ audio-driven neural models, which learn to generate facial motion directly from speech or song waveforms. Systems such as Wav2Lip [1], MakeItTalk [2], and SadTalker [3] can produce synchronized lip movements by training deep convolutional or transformer-based networks. However, these approaches require large paired datasets, GPU acceleration, and extensive preprocessing steps such as landmark extraction and facial alignment. For singing tasks, additional challenges arise due to pitch variation and longer phoneme duration, often resulting in unstable or exaggerated expressions.

In practical applications, 2D animation software such as Adobe Character Animator [4] and Live2D Cubism [5] provide semi-automatic lip-sync workflows, but they depend on commercial environments and manual keyframe adjustment. As a result, many singing animation pipelines remain complex, resource-intensive, and difficult to reproduce in lightweight experimental settings.

B. Expression-Overlay using FFmpeg Command

In contrast to morph- or learning-based systems, image overlay compositing provides a simpler alternative for generating expressive facial motion. The idea is to sequentially layer pre-rendered facial components—such as eyes, eyebrows, and mouth—onto a base head image, achieving perceptual animation effects through timed transitions.

The FFmpeg multimedia framework [6] offers a powerful yet underexplored platform for implementing such compositing workflows. Its filter_complex command supports layered image blending, transparency control, and time-based switching—all achievable through shell scripting without dedicated animation software. Recent creative coding studies have also highlighted FFmpeg's potential for compositing and automation in multimedia art [7].

This study positions itself at the intersection of these two domains: it combines AI-generated facial expression images with FFmpeg-based overlay scripting to demonstrate a minimal yet reproducible approach for singing head animation. Rather than competing with audio-driven systems, it aims to provide a transparent baseline for studying expressive timing and visual synthesis using open-source tools.

To date, little academic research has focused on FFmpeg-based expression overlay for lip-sync or singing animation. This is primarily because such methods lack an automatic mapping between audio and expression states, offer limited temporal smoothness, and are typically categorized as post-production techniques rather than algorithmic innovations. Nevertheless, their simplicity, reproducibility, and accessibility make them valuable for experimental demonstrations and creative prototyping in low-resource environments.

III. Method

A. Resources and Setup

To ensure full reproducibility, all experiments in this study are conducted using freely available resources and open-source tools. The required materials include (1) FFmpeg installation and (2) an example base image with facial expression layers for eyebrows, eyes, and mouth. The setup can be completed on any standard desktop computer without specialized software. All required files, including example expression images FFmpeg command scripts and resulted videos are provided on the accompanying GitHub repository at https://github.com/shungch-code/Singing-Head-Animation-Using-FFmpeg-Expression-Overlay.

1) Installing FFmpeg: FFmpeg is a cross-platform command-line toolkit for audio and video processing. It is used here as the core animation engine for overlaying and sequencing expression images.

- Official website: https://ffmpeg.org/download.html
- Windows users: download the precompiled "release full build" from https://www.gyan.dev/ffmpeg/builds/ and extract the folder (e.g., C:\ffmpeg).
- Add the FFmpeg binary folder (e.g., C:\ffmpeg\bin) to the system PATH so that the command "ffmpeg -version" can run from the terminal.

2) Base Image with Facial Expression Images: The animation relies on two types of input images: a fixed base layer and a set of dynamic expression layers.

a) Base Image: To create a simple 2D animation, a static PNG frame (e.g., head_base.png) with a transparent background is required, showing the front-facing neutral expression. This image serves as the fixed layer upon which facial expressions are composited. Users can render this from the provided image files (download: https://bit.ly/3QuBdUd) or use a hand-drawn/AI-generated portrait. Consistency in face orientation and lighting is crucial.

b) Expression Images: The overlay-based animation depends on pre-generated expression layers divided into three functional groups:

- Eyebrows: neutral, angry, raised, sad
- Eyes: neutral, blinking, squinting, wide open
- Mouth: neutral, angry, happy, sad

Each expression image must be a separate transparent PNG file precisely aligned to the same base head position and exported with consistent resolution (e.g., 512 x 512 pixels). Example filenames follow the convention: part_expression.png (e.g., eyebrow_neutral.png, eyes_blink.png, mouth_happy.png). All example expression PNGs are bundled in the Expression Images.zip file available in the GitHub repository as shown in Fig. 2.

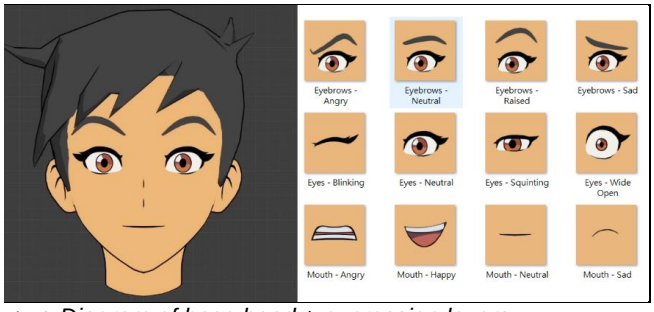

Fig. 2. *Diagram of base head + expression layers*

B. Single-Part Expression Overlay (Mouth Demo)

To facilitate reproducibility and highlight the modular design of the system, Table I summarizes the key FFmpeg commands used to achieve core compositing tasks. These include static image overlay, sequential frame blending, alpha-channel handling, and video assembly.

TABLE I. THE KEY FFMPEG COMMANDS

FFmpeg Command	Description
-i	Input file
-filter-complex	Specifies filtergraph
overlay	Places one image over another
-t	Specifies output duration
-c.v	Specifies video codec

1) Preparing Mouth Images: Select 2–3 mouth expressions for testing, e.g., mouth_neutral.png, mouth_happy.png, mouth_sad.png. Ensure they are aligned to the base head image (head_base.png) with transparent background.

2) FFmpeg Mouth Overlay Animation Command: A minimal FFmpeg command to cycle through mouth expressions for a short duration (e.g., 6 seconds total) is as follows:

```
ffmpeg \
 -loop 1 -i head_base.png \
 -loop 1 -i mouth_neutral.png \
 -loop 1 -i mouth_happy.png \
 -loop 1 -i mouth_angry.png \
 -loop 1 -i mouth_sad.png \
 -filter_complex "
   [1:v]scale=120:120[m1];
   [2:v]scale=120:120[m2];
   [3:v]scale=120:120[m3];
   [4:v]scale=120:120[m4];
   [0:v][m1]overlay=x=240:y=360:enable='between(t,0,1)'[v1];
   [v1][m2]overlay=x=240:y=360:enable='between(t,1,2)'[v2];
   [v2][m3]overlay=x=240:y=360:enable='between(t,2,3)'[v3];
   [v3][m4]overlay=x=240:y=360:enable='between(t,3,4)'[v]
" \
 -map "[v]" \
 -t 4 \
 -pix_fmt yuv420p \
 result.mp4
```

Explanation:

- [0:v] is the base head image, serving as the static background layer.
- [1:v] to [3:v] represent the mouth expression images composited sequentially.
- overlay=x=240:y=360 positions each mouth image precisely on the facial region of the base head
- enable='between(t,start,end)' defines the time interval during which each expression is visible.
- -t 6 sets the total duration of the rendered animation to 6 seconds.
- -pix_fmt yuv420p ensures the output video is compatible with standard media players.

C. Multi-Part Expression Overlay

Building on the single-part test, we demonstrate the integration of multiple facial parts: eyebrows, eyes, and mouth. This step showcases the core strength of our FFmpeg-based expression overlay approach, allowing synchronized animation from static images. Prepare aligned PNGs for each facial part with transparent backgrounds:

- Eyebrows: eyebrow_neutral.png, eyebrow_angry.png, eyebrow_raised.png, eyebrow_sad.png
- Eyes: eyes_neutral.png, eyes_blinking.png, eyes_squinting.png, eyes_wide.png
- Mouth: mouth_neutral.png, mouth_happy.png, mouth_sad.png, mouth_angry.png

All images should be positioned to match the base head image (head_base.png). Below is a minimal example demonstrating a 12-second animation with multiple expressions:

```
ffmpeg \
 -loop 1 -i head_base.png \
 -loop 1 -i mouth_neutral.png \
 -loop 1 -i mouth_happy.png \
 -loop 1 -i mouth_angry.png \
 -loop 1 -i mouth_sad.png \
 -loop 1 -i eyes_neutral.png \
 -loop 1 -i eyes_blinking.png \
 -loop 1 -i eyes_squinting.png \
 -loop 1 -i eyes_wideopen.png \
 -filter_complex "
   [1:v]scale=120:120[m1];
   [2:v]scale=120:120[m2];
   [3:v]scale=120:120[m3];
   [4:v]scale=120:120[m4];
   [5:v]scale=140:60[e1];
   [6:v]scale=140:60[e2];
   [7:v]scale=140:60[e3];
   [8:v]scale=140:60[e4];
   [0:v][m1]overlay=x=240:y=360:enable='between(t,0,1)'[v1];
   [v1][m2]overlay=x=240:y=360:enable='between(t,1,2)'[v2];
   [v2][m3]overlay=x=240:y=360:enable='between(t,2,3)'[v3];
   [v3][m4]overlay=x=240:y=360:enable='between(t,3,4)'[v4];
   [v4][e1]overlay=x=300:y=280:enable='between(t,0,1)'[v5];
   [v5][e2]overlay=x=300:y=280:enable='between(t,1,2)'[v6];
   [v6][e3]overlay=x=300:y=280:enable='between(t,2,3)'[v7];
   [v7][e4]overlay=x=300:y=280:enable='between(t,3,4)'[v]
 " \
 -map "[v]" \
 -t 4 \
 -pix_fmt yuv420p \
 result_multi.mp4
```

This demonstration confirms that multiple static expression images can be composited into a synchronized animated head entirely using FFmpeg, without requiring training data or 3D rigging, making it suitable for quick prototyping and demo purposes.

IV. Future Work

This section proposes future extensions to transform the current demonstration pipeline into a more expressive and intelligent singing head animation system. Two primary research directions are identified: AI-assisted generation of singing mouth shapes and enhanced animation fluidity using FFmpeg-based frame interpolation.

A. AI-Assisted Singing Mouth Shape Generation

Recent advances in image-to-image translation and diffusion-based generative models have made it possible to automatically synthesize intermediate mouth shapes that correspond to phonetic or emotional variations in singing. Instead of manually preparing discrete sprite images (e.g., neutral, happy, sad, angry), AI-assisted generation can interpolate natural transitions between visemes or vowel forms, producing context-aware mouth poses aligned with musical phrasing and vocal dynamics.

This enhancement enables the overlay framework to achieve continuous mouth articulation without requiring explicit 3D rigging or extensive training on video datasets. By integrating pre-trained generative tools such as ControlNet or FaceFusion, users can automatically produce multiple, stylistically consistent mouth frames from a single base image. The generated mouth shapes can then be exported as transparent PNGs and composited within the same FFmpeg pipeline, maintaining reproducibility while substantially improving visual expressiveness and realism.

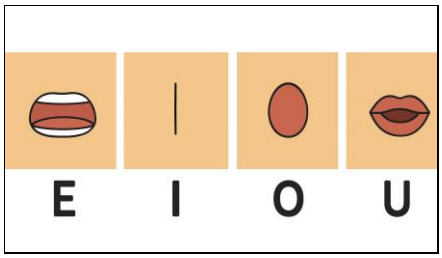

Fig. 3. Example of AI-assisted singing mouth shapes for vowels voice-synthesis applicationsE, I, O, and U.

A schematic illustration of this concept is presented in Fig. 3, which depicts how AI-assisted mouth-shape generation feeds directly into the FFmpeg compositing stage, forming a streamlined workflow for expressive singing animation.

B. Animation Smoothness Enhancement via FFmpeg

The current demonstration employs discrete frame switching between static sprites, which can produce perceptible “jump cuts” during expression transitions. To improve temporal coherence and realism, future versions will incorporate FFmpeg’s motion-compensated frame interpolation. By applying a command such as:

-filter:v minterpolate='mi_mode=mci:mc_mode=aobmc:vsbmc=1'

the system can automatically synthesize intermediate frames between two expressions, creating a visually smooth transition without manual animation curves or external post-processing. This method effectively enhances motion fluidity for eyes, eyebrows, and mouth overlays, producing perceptually continuous singing sequences even when source images remain static.

The combination of AI-assisted expression generation and FFmpeg interpolation introduces a hybrid workflow: AI contributes nuanced frame creation, while FFmpeg ensures real-time compositing and playback smoothness. Together, these advancements pave the way for an accessible yet expressive 2D singing animation system.

C. Resources and Replicability

All related resources, including example expression datasets, overlay scripts, and demonstration videos, remain publicly accessible to support reproducibility and open experimentation:

- GitHub DEMO: https://github.com/shungch-code/Singing-Head-Animation-Using-FFmpeg-Expression-Overlay
- Demo Video & Blog: https://shungch.blogspot.com/p/singing-face.html

V. Conclusion

This study presents a lightweight and fully reproducible approach for generating 2D singing head animations using FFmpeg overlay commands. By sequencing static facial expression layers—particularly

mouth movements—the method enables rapid rendering of expressive animations without the need for 3D modeling, facial rigging, or deep-learning frameworks. Despite its simplicity, the approach effectively conveys basic emotional and rhythmic variations within seconds, making it highly suitable for real-time demonstrations, educational use, and rapid prototyping.

The primary advantage of the FFmpeg-based overlay framework lies in its execution efficiency and transparency. The entire animation can be synthesized almost instantaneously on standard hardware, allowing reproducible experimentation with minimal computational resources.

Future research will focus on enhancing visual realism and expressiveness while maintaining computational simplicity. Planned extensions include AI-assisted mouth-shape generation for singing vowels, frame interpolation for smoother transitions, and phoneme-based mapping to improve lip synchronization with musical phrasing. These improvements aim to bridge the gap between static overlay animation and real-time, perceptually natural singing performance, offering an accessible foundation for future multimedia and singing voice synthesis and multimedia applications.

REFERENCES

[1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical

[7] media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].