Enhancing Human-Machine Interaction: A Novel Approach to Emotion-Controlled Speech-to-Animation

Rebecca Mobbs, Dimitrios Makris, Demetris Lappas, Vasileios Argyriou
School of Computer Science and Mathematics, Kingston University, London, UK

ABSTRACT

This paper presents a novel framework for emotion-controlled speech-to-animation, addressing the issue of emotional mismatches between speech and facial expressions in existing methods. Our approach synchronises emotional expression across audio and facial animations using State-of-the-Art (SOTA) pretrained models, eliminating the need for costly custom training while ensuring adaptability. A key contribution of our framework is the creation of a novel a Speech-to-Speech (S2S) pipeline for emotional control over generated speech. In addition, we introduce a novel evaluation metric, the Emotion Distribution Divergence (EDD), to assess our models ability to modify the emotions in the original videos. Experimental results demonstrate significant improvements in emotional expressiveness and realism over existing methods, establishing our approach as a major advancement in human-machine interaction, virtual assistants, and emotion-aware IoT applications.

1. INTRODUCTION

Recent advancements in deep learning and generative models have significantly improved emotion generation across facial expressions, speech, and text. These developments enable applications in smart home systems and automotive technologies, enhancing human-device interactions through nuanced emotional understanding and responsiveness. Emotion-adaptive visual chatbots in IoT and smart devices could respond appropriately to users' emotional states. While current systems enable emotionally responsive communication through audio, animated facial expressions could improve engagement. Context-aware mechanisms enhance personalisation, tailoring interactions to user needs. Additionally, these chatbots improve accessibility for individuals with hearing impairments.[1] Emotion recognition and generation together enable proactive problem-solving by identifying emotional cues such as frustration or distress and addressing issues preemptively. These advancements foster empathy-driven interactions, encouraging trust and long-term adoption of IoT technologies.

Emotion-controlled visual chatbots in IoT devices support various applications. In smart homes, emotion analysis enables adaptive experiences like mood-based lighting and security adjustments. In autonomous vehicles, recognising driver emotions enhances safety by detecting fatigue, distraction, or stress and triggering appropriate interventions. Linking facial animations and emotion generation to IoT devices improves intuitive and effective human-device interactions. Despite these opportunities, current emotion control approaches are limited. Many focus on generating facial animations without incorporating speech, leading to a disconnect between expressions and vocal emotions. Additionally, these methods often require large datasets or are computationally intensive, limiting real-time applicability.

Existing speech-to-animation methods with emotion control frequently suffer from a significant disconnect between audio and facial expressions, resulting in animations that feel unnatural, inconsistent, and emotionally unconvincing. This mismatch disrupts viewer immersion, severely limiting the effectiveness of IoT applications, visual virtual assistants, and smart devices. Without precise synchronisation, these systems fail to convey emotions authentically, reducing user engagement, trust, and overall interaction quality in human-computer communication.

We propose a novel speech-to-animation framework which provides fine-grained control over both generated facial expressions and emotional modulation in speech. Our key contribution is an original speech-to-speech

Further author information: (Send correspondence to Rebecca Mobbs)
Rebecca Mobbs: E-mail: k2369889@kingston.ac.uk
Dimitrios Makris: E-mail: d.makris@kingston.ac.uk
Vasileios Argyriou: E-mail: vasileios.argyriou@kingston.ac.uk

pipeline, enabling precise emotion-driven adjustments in speech synthesis—an advancement which sets a new standard in expressive animation. This approach is the first to seamlessly integrate these modalities. Our method uses SOTA pretrained models, eliminating the need for custom training while ensuring exceptional performance and adaptability. This adaptability makes our pipeline future-proof, capable of incorporating emerging technologies as they develop. Additionally, we introduce a novel evaluation metric, the EDD, to quantitatively assess the change in emotional intensity between original and generated videos. Experimental results demonstrate our model's advantageous capability in modifying emotional expressions, pushing the boundaries of speech-driven animation.

2. RELATED WORK

2.1 Visual Virtual Assistant Chatbots in IoTs

A visual virtual assistant[2] was designed to integrate voice and text-based interactions with automation and advanced conversational artificial intelligence. User access begins with a secure registration process, and Firebase manages user authentication. The system's interface, developed using ReactJS, combines voice assistant functionalities with features such as weather forecasts, news updates, and live sports scores. By combining Dialogflow with Puppeteer for web automation, the system can handle tasks like booking cars and shopping online.

2.2 Emotion Generation methods

Facial reenactment methods[3] have advanced with talking-head models like VASA[4] and EMO-Live,[5] enabling fine control over facial expressions in animations. Speech-to-animation models generate facial expressions using audio, video, or text inputs,[5,6] focusing on mouth movements and overall realism. State-of-the-art models like Wav2Lip[7] and diffusion-based approaches[5] achieve high quality expression synthesis with control over intensity and duration. Speech Emotion Generation (SEG) techniques[8] transform vocal characteristics to modify emotional expressions. Such systems utilise deep learning to produce realistic synthetic voices while maintaining emotional nuance.[9,10] Techniques such as adversarial training[11] and latent diffusion models enable precise control over emotional and stylistic variations in speech.[12] Text Sentiment Generation involves generating emotionally nuanced text responses using Large Language Models (LLMs) like ChatGPT[13] and Gemini.[14] These models prioritise grammar and contextual understanding to produce human-like outputs. Recent advancements use GANs and transformer-based architectures, enabling models to dynamically focus on relevant textual inputs for emotionally aligned text generation.[15]

2.3 Emotion Control Methods

Generative models with emotion control have advanced significantly across modalities like audio, video, and text. Audio-driven face expression generation[5] utilises a comprehensive architecture that synchronises facial expressions with audio input. However, it is computationally intensive and reliant on high quality audio. Video-driven face expression generation methods[16] tackle challenges like zero-shot editing by incorporating emotion prompts for guided expression generation. However, these methods depend on pretrained models, which may introduce biases in generated outputs. In SEG, a previously presented architecture[10] employs state-of-the-art feature extractors like Wav2Vec 2.0[17] for emotion-specific speech synthesis. While the model improves voice fidelity and emotional expressiveness, its reliance on high quality speaker data limits its generalisation across diverse datasets.

2.4 Emotion Control Evaluation Methods

Existing evaluation methods for emotion control models predominantly focus on assessing either the classification accuracy of the generated emotional expressions[16] or the perceptual realism of the synthesised facial outputs.[5] Commonly employed metrics in this domain include Fréchet Inception Distance (FID), Structural Similarity Index (SSIM), and SyncNet, which collectively quantify various aspects of video quality and audio-visual synchronisation. Nonetheless, there remains a critical gap in current evaluation protocols: no existing metric explicitly measures a model's capacity to alter the emotional content of an input video or image.

3. PROPOSED METHOD

In this section, we introduce the Emotion-Driven Animation (EDA) framework (see Fig. 1), which synchronises emotional changes across audio and facial expressions in generated animations. The EDA model takes a video and an emotion prompt as input to generate a corresponding animation. It integrates a Speech-to-Speech (S2S) module and a Facial Expression Animation Module (FEAM) to enhance the animation process. Our S2S module begins with a speech-to-text component (S2T)[18] to extract the spoken content from the original video. The emotional content of the transcribed text is then changed by the Emotional Text Generation (ETG) component[13] to the specified emotion, before being processed by the Emotional Text to Speech component (ET2S)[9] to create emotional speech using the target voice. The FEAM module processes $N = 5$ random frames from the video and generates new frames of the face, with specified emotions, using the Emotional Face Generator component (EFG).[16] The final part of our pipeline is the EFA[19] which creates the final animations using the emotional speech and emotional target face images.

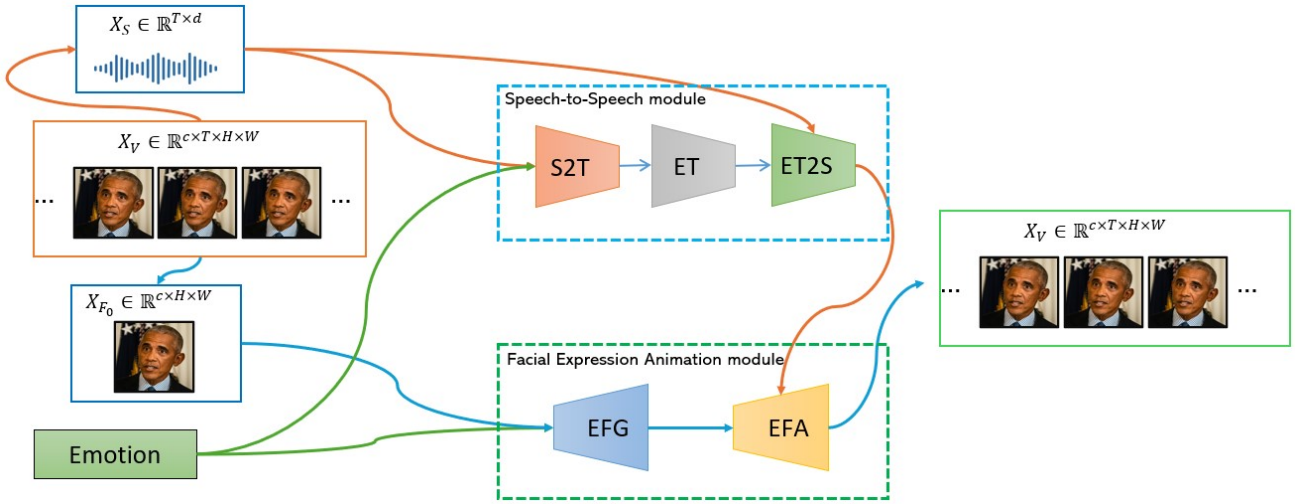

Figure 1: The input to our EDA model is a video and an emotion prompt. The audio from the video is processed by a Speech-to-Speech module which combines a Speech-to-text component (S2T), an Emotional Text Generation component (ETG) and an Emotional Text to Speech component (ET2S) to recreate the original audio in the prompted emotion. The secondary module is the Facial Expression Animation module (FEAM), which combines an Emotional Face Generator component (EFG) with an Emotional Face Animator component (EFA), this module regenerates the target face with the prompted emotion. The EFA component generates the final animation with synced emotions in the audio and face.

3.1 Speech-to-Speech Module

The Speech-to-Speech module transforms the original audio into an emotionally modified version through a three-stage pipeline involving speech-to-text, text-to-emotional text, and text-to-speech. This process ensures that the output speech aligns with the intended emotional tone while maintaining naturalness and intelligibility. After extracting the audio, the speech-to-speech model is initiated (see Fig. 1). The process begins with the S2T[18] which process the original audio $S_{\text{in}}(t)$ to generate a transcript T. This transcript is then processed by the ETG[13] along with the emotion prompt, represented as $I = [T, P]$, where T is the transcript, and P is the emotion modification prompt (e.g., "Make this text very very very happy"). The input is tokenized as $I_{\text{tokens}} = \text{Tokenizer}(I)$. The altered transcript is then processed by the ET2S.[9] ElevenLabs synthesises the new audio $S_{\text{out}}(t)$ using the target voice and the modified speech content, thereby producing an emotionally adapted version of the original audio.

3.2 Facial Expression Animation Module

The Facial Expression Animation module (FEAM) comprises two key components: Facial Expression Generator component (FEGC)[16] and Facial Expression Animation component (FEAC).[19] The FEGC model is utilised to

generate still images of a target face conditioned on a specified emotion prompt. This ensures that the facial expressions accurately reflect the intended emotional state before animation.

We use FEGC to generate the target face with specified prompted emotions. The model is trained using an emotion-agnostic pretraining stage that refines 3D latent keypoints to capture facial attributes, using datasets like MEAD and AffectNet to enhance expression diversity. An Audio-to-Expression Transformer (A2ET) maps audio features to keypoints, using PCA for dimensionality reduction to ensure computational efficiency. In the second stage, lightweight modules adapt the model for emotional tasks, integrating Emotional Prompts and an Emotional Deformation Network (EDN) to refine expression deformations. The Emotional Adaptation module (EAM) further enhances visual quality through learned scaling and shifting parameters. The model is trained with objectives including latent loss for keypoint alignment, synchronisation loss for audio-visual coherence, and CLIP[20] loss for text-image embedding alignment, ensuring high-fidelity and generalisable emotional expression generation.

These generated images serve as input to EFA, which synthesises a talking-head animation by integrating the emotionally conditioned facial images with corresponding audio. By utilising this two-stage approach, our system ensures both accurate emotion representation and natural motion dynamics. The EFA component is trained in three stages to generate realistic talking face animations from a single image and input speech audio. First, the model learns to predict accurate facial expressions from audio by distilling lip-only motion coefficients from a pretrained model[7] while using perceptual losses such as lip-reading loss and landmark loss to enhance accuracy. Second, a conditional variational autoencoder is trained to generate diverse identity-aware head motions by learning residual head pose changes and incorporating adversarial and KL-divergence losses for realism. Third, a 3D-aware face renderer is trained to map the generated 3D motion coefficients to an unsupervised 3D keypoint space, allowing natural video synthesis. During inference, the system takes input audio and a reference image, extracts motion coefficients, maps these coefficients to the learned 3D keypoints, and uses the face renderer to warp the image accordingly. This end-to-end process produces high-quality talking face videos with synchronized lip movements, natural head motions, and strong identity preservation, outperforming existing methods in video realism and expressiveness.

4. EXPERIMENTS

4.1 Dataset

The evaluation of our method was conducted using an existing dataset of edited YouTube videos.[6] The videos were edited to focus on one person, and are approximately 45 seconds long. We selected two videos, Al Pacino and Julia Roberts, to serve as test cases for our model. For additional consideration we used a YouTube clip of Robin Williams. We generated 520 animated samples, comprising 7 target emotions: Anger, Contempt, Disgust, Fear, Happiness, Neutral, and Sad. Due to cost restraints with using the E2TS component we generated only one sample per emotion, per subject, for a total of 21.

4.2 Experimental design

First, we evaluated our work for accuracy of the generated emotions in faces, text, and speech using pretrained emotion recognition models. The evaluation of the S2S components was conducted by XLM-Emo text sentiment recognition model[21] and SpeechBrain speech emotion recognition model.[22] The generation capabilities of the FEAM components were individually evaluated using EmoFAN facial expression recognition (FER) model.[23] Each of these emotion detection models evaluates different emotions, EmoFan evaluates 8 target emotions: Anger, Contempt, Disgust, Fear, Happiness, Neutrality, Sadness, and Surprise. However, due to a bias towards surprise in the training data for the FER model, we removed this emotion to allow the under-represented emotions neutral, fear, disgust, and contempt to be detected.[24] XLM-Emo evaluates anger, fear, happy, and sad, and SpeechBrain evaluates anger, happiness, sadness, and neutral. Secondly, we used EDD to assess the effectiveness of our model at generating emotions, based on the distance between the distribution of emotion in the original video and generated videos. In addition, we assessed our model using standard metrics CPBD, PSNR, FID, Head Motion, and Beat Align comparing our results to those obtained by the SOTA methods in our pipeline.

Emotion	Accuracy
Anger	100%
Fear	100%
Joy	100%
Sad	89%
Avg	**97%**

Table 1: ETG component emotion accuracy.

Emotion	Accuracy
Anger	50%
Neutral	25%
Happy	75%
Sad	50%
Avg	**50%**

Table 2: ET2S component emotion accuracy.

Method	CPBD↑	PSNR↑	FID↓
EFG[16]	21.8	11.52	5.3
EFA[19]	0.33	13.8	4.9

Table 3: Comparison of the SOTA components on standard metrics.

4.3 Evaluation of Speech-to-Speech Module Components

We performed emotion recognition on the components of our S2S model to evaluate whether the intended emotions were accurately generated. Table 1 presents the results for the ETG component, while Table 2 shows the performance of the ET2S component. Our results demonstrate that the method successfully generated the target emotion in 97% of cases for emotional text and 50% for emotional speech. However, some emotions were lost during the transition from text to speech, highlighting a limitation in the ET2S component's ability to convey emotions effectively in speech synthesis. Future work should explore the integration of emotionally expressive speech synthesis models or the fine-tuning of ET2S on emotion-labelled corpora to enhance affective fidelity.

4.4 Evaluation of Facial Emotion Animation Module Components

Figure 2: Comparison of frames from the original video (left) with EDA-generated frames across seven emotional expressions.

Fig.2 presents a comparative analysis of first frames extracted from the original video (left) and first frames generated using the EDA model, depicting seven distinct emotional expressions: disgust, anger, contempt, neutral, happy, fear, and sad. To evaluate the performance of individual components within the FEAM module, we conducted an emotion classification analysis on the frames generated by both the EFG component (see Fig.3) and the EDA model (see Fig.4). Our findings indicate a loss in emotion intensity on the final stage of our pipeline for some emotions. Emotions with strong facial distortions, such as Disgust and Fear, appear to be better preserved, whereas more subtle or nuanced emotions, such as Anger and Sadness, may degrade due to the constraints of the animation model.

4.5 Emotion Distribution Divergence Metric

Since our goal is not to reconstruct the original video but rather to generate a video that effectively portrays a user-defined emotion, conventional evaluation metrics such as emotion accuracy and FID are insufficient. A successful transformation should result in a video where the individual exhibits the intended emotion more prominently than in the original, even if the video appears visually different. To quantify this, we introduce the Emotion Depiction Distance (EDD) as our evaluation metric.

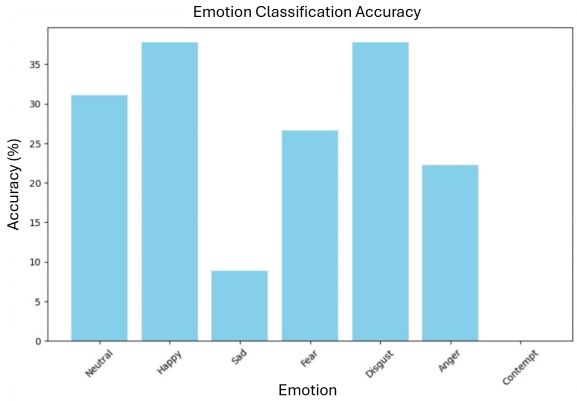

Figure 3: EFG component facial expression generation results evaluated using FER.

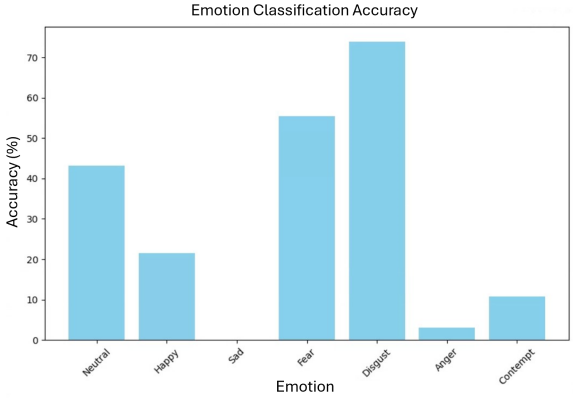

Figure 4: EDA facial animation generation results evaluated using FER.

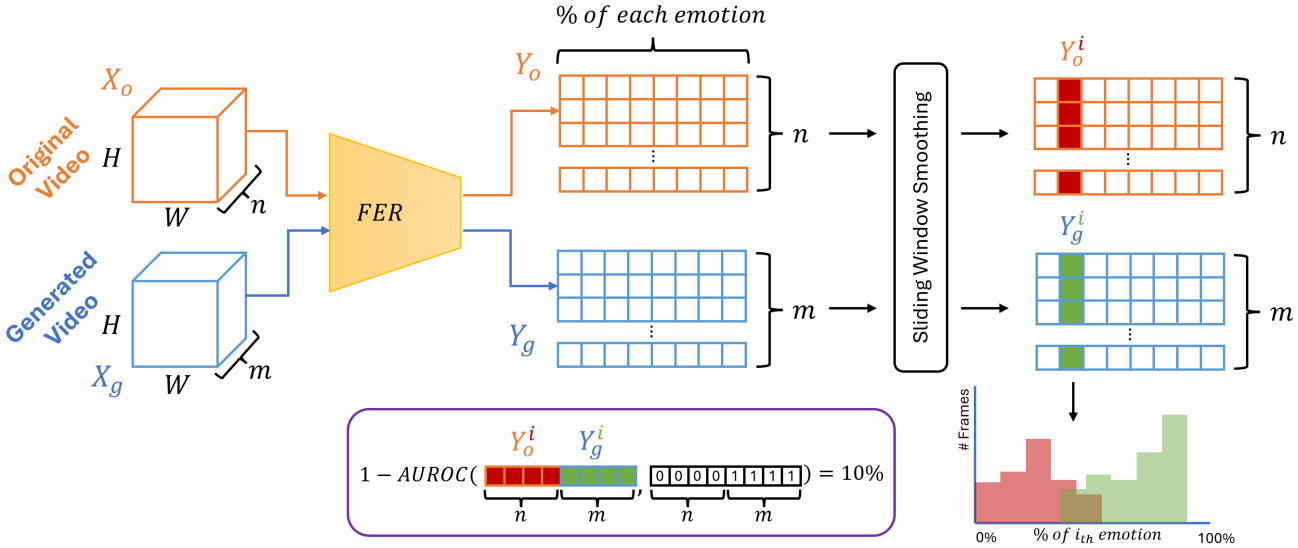

Figure 5: Visualisation of the EDD metric, illustrating how emotion probabilities are extracted, smoothed, and compared between the original and generated videos.

To assess whether our method successfully generates faces that express the target emotion (denoted as the i_{th} emotion), we compare the generated video with the original using a Facial Emotion Recognition (FER) model. Given an original video $X_o \in \mathbb{R}^{c \times n \times H \times W}$ and a generated video $X_g \in \mathbb{R}^{c \times m \times H \times W}$, where c is the number of channels, n and m are the respective frame counts, and $H \times W$ is the resolution, we pass both through FER, obtaining $Y_o \in \mathbb{R}^{n \times d}$ and $Y_g \in \mathbb{R}^{m \times d}$. Each row in these matrices represents the probability distribution of d possible emotions for a given frame.

As FER operates on individual frames without contextual awareness, temporal inconsistencies may arise. To mitigate this, we apply a sliding window averaging technique that smooths the emotion probabilities across a sequence of frames. From the smoothed outputs, we extract the values corresponding to the target emotion (i_{th} emotion), yielding two distributions: Y_o^i for the original video and Y_g^i for the generated one. A successful transformation should result in higher values in Y_g^i compared to Y_o^i, indicating that the generated video expresses the desired emotion more strongly.

To quantify the difference, we measure how well-separated these two distributions are using the Area Under the Receiver Operating Characteristic Curve (AUROC).[25] We construct a label vector V by concatenating a zero vector for the original frames and a one vector for the generated frames, while Y is formed by concatenating Y_o^i and Y_g^i. AUROC(Y, V) provides a measure of separation, with higher values indicating greater distinction

between the original and generated emotion distributions. The EDD metric is defined as:

$$EDD = 1 - AUROC(Y, V) = 1 - \frac{1}{|\mathcal{V}_1||\mathcal{V}_0|} \sum_{y_j \in \mathcal{Y}_1} \sum_{y_k \in \mathcal{Y}_0} \infty(y_j > y_k) \tag{1}$$

where $\mathcal{V}_1$ and $\mathcal{V}_0$ represent the indices of frames from the generated and original videos, respectively, while $\mathcal{Y}_1$ and $\mathcal{Y}_0$ contain their corresponding emotion scores. The indicator function $\infty(y_j > y_k)$ returns 1 if the generated frame exhibits a stronger presence of the target emotion than the original frame, and 0 otherwise. An overview of EDD can be seen in Fig.5.

4.6 EDD Results

Table 4 presents the average EDD scores for facial expressions in videos generated by our EDA framework, compared to those produced without the EFG component. The EDD score measures the overlap between the distribution of emotion classification results from frames in the generated video and those in the original video (see Fig.6 and Fig.7 for examples). For each model, we report the lowest score achieved on each emotion. For comparison, the EFA method[19] was evaluated using frames extracted from the original video. The EDD score quantifies the similarity between the two distributions by computing their statistical overlap.

A score of 1.0 indicates complete overlap, meaning the generated video retains the same emotional distribution as the original. In contrast, a lower score is better, as it signifies reduced overlap and greater emotional change in the generated video. This metric effectively evaluates how much the emotional content in the generated videos diverges from the original video, with lower scores indicating more successful emotional transformation.

Method	Anger	Contempt	Disgust	Fear	Happy	Neutral	Sad
SadTalker[19]	0.3521	0.2497	**0.0933**	0.0543	0.2503	0.3218	0.6032
EDA	**0.3168**	**0.1821**	0.0984	**0.0103**	**0.0314**	**0.0294**	**0.6291**

Table 4: EDD scores measuring the overlap between the emotion distributions of the original and generated videos. Lower scores indicate greater emotional transformation, meaning less overlap between the two distributions. The results show that our method, EDA, achieves consistently lower or comparable scores across most emotions compared to the baseline method, SadTalker,[19] demonstrating its effectiveness in altering emotional expressions.

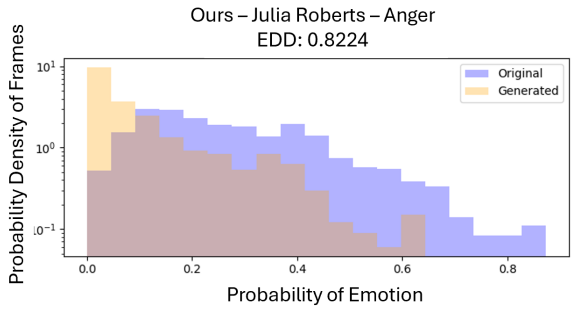

Figure 6: Example of a generated video with a low EDD score, indicating a significant change in emotional expression compared to the original video.

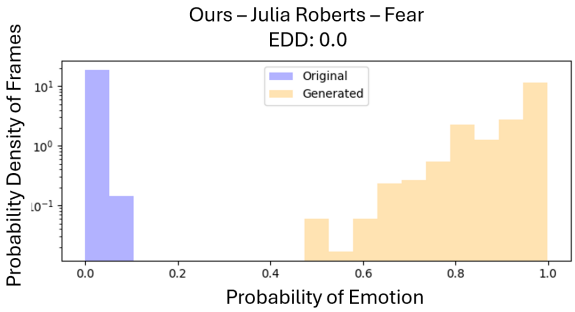

Figure 7: Example of a generated video with a high EDD score, indicating a large overlap in emotion distribution between the original and generated video.

Our method demonstrated superior performance in synthesising Contempt, Fear, Happy, and Neutral expressions, with the most notable improvements observed for Happy and Neutral, indicating enhanced refinement of smiles and subtle emotional states. However, it exhibited lower performance for Sad expressions, where EFA

marginally outperformed EDA, suggesting that the EFG component may introduce distortions that reduce clarity. Further analysis of the results revealed emotion bias present in the original videos influenced the outcomes (see Fig.7 where the original video scored higher on the emotion anger, than the generated video did).

5. COMPARATIVE ANALYSIS OF STANDARD METRICS

Table 3 presents a detailed comparative evaluation of the state-of-the-art (SOTA) components within the FEAM framework, specifically assessing EFG[16] and EFA.[19] The results demonstrate that EFA consistently outperforms EFG across all evaluated metrics. This superior performance emphasizes the critical need to integrate EFA, or an even more advanced SOTA method, into our pipeline to improve the realism and fidelity of the generated animations. The incorporation of a high performance method is essential for applications requiring accurate and expressive visual representations, such as virtual avatars, human-computer interaction, and affective computing. Achieving superior results across key evaluation metrics directly contributes to enhanced visual coherence, improved lip-sync accuracy, and more natural emotional expressiveness. By selecting the most effective SOTA component, we aim to minimize artifacts, improve temporal consistency, and produce animations that are both perceptually convincing and functionally robust.

6. CONCLUSION

This study introduced a novel methodology for emotion-controlled speech-to-animation, addressing the critical challenge of synchronising emotional expression across text, speech, and facial animations. By using SOTA pretrained models, our approach achieved high accuracy in emotion generation for text, demonstrated strong results in speech modulation, and effectively modified facial expressions to match the intended emotion. A key innovation of this work is the introduction of the Emotion Distribution Divergence (EDD) metric, based on AUROC, providing a novel and rigorous evaluation method for emotional consistency. Our results show that the model was able to affectively alter facial expressions to align with the desired emotions, validating the effectiveness of our approach. This metric offers valuable insights into model performance and sets a new standard for assessing emotional expressiveness in speech-driven animation. Another key contribution of this work is the Speech-to-Speech (S2S) module, which effectively altered the emotions in the original speech, enhancing the overall emotional coherence of the generated animations. This module plays a crucial role in ensuring that the emotional expressions in speech align with the intended visual representation. Despite certain limitations, including the constraints of pretrained models, our results underscore the significance of multimodal integration in developing emotionally expressive avatars. Future work should focus on refining synthesis techniques, using more diverse datasets, improving computational efficiency, and enhancing evaluation metrics to further push the boundaries of emotion generation in speech-to-animation.

REFERENCES

1. R. Valarmathi, P. Jaya Surya, P. Balaji, and K. Ashik, "Animated sign language for people with speaking and hearing disability using deep learning," in *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pp. 1–5, 2024.
2. H. Mauny, D. Panchal, M. Bhavsar, and N. Shah, "A prototype of smart virtual assistant integrated with automation," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 952–957, IEEE, 2021.
3. S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, and G. Tzimiropoulos, "Hyperreenact: One-shot reenactment via jointly learning to refine and retarget faces," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
4. S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo, "Vasa-1: Lifelike audio-driven talking faces generated in real time," *arXiv preprint arXiv:2404.10667* , 2024.
5. L. Tian, Q. Wang, B. Zhang, and L. Bo, "Emo: Emote portrait alive – generating expressive portrait videos with audio2video diffusion model under weak conditions," 2 2024.

6. F. P. Papantoniou, P. P. Filntisis, P. Maragos, and A. Roussos, "Neural emotion director: Speech-preserving semantic control of facial expressions in "in-the-wild" videos," 2021.
7. K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 484–492, 2020.
8. J. Yao, Y. Yang, Y. Lei, Z. Ning, Y. Hu, Y. Pan, J. Yin, H. Zhou, H. Lu, and L. Xie, "Promptvc: Flexible stylistic voice conversion in latent space driven by natural language prompts," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10571–10575, IEEE, 2024.
9. ElevenLabs, "ElevenLabs." https://www.elevenlabs.io/, 2024.
10. H. J. Park, S. W. Yang, J. S. Kim, W. Shin, and S. W. Han, "Triaan-vc: Triple adaptive attention normalization for any-to-any voice conversion," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* , 2023.
11. J. Yang, J.-S. Bae, T. Bak, Y. Kim, and H.-Y. Cho, "Ganspeech: Adversarial training for high-fidelity multi-speaker speech synthesis," *arXiv preprint arXiv:2106.15153* , 2021.
12. H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**(16), pp. 17862–17870, 2024.
13. J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774* , 2023.
14. M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530* , 2024.
15. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems* **27**, pp. 3104–3112, 2014.
16. Y. Gan, Z. Yang, X. Yue, L. Sun, and Y. Yang, "Efficient emotional adaptation for audio-driven talking-head generation," 2023.
17. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems* **33**, pp. 12449–12460, 2020.
18. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.
19. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8652–8661, 2023.
20. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020* , 2021.
21. F. Bianchi, D. Nozza, and D. Hovy, "Xlm-emo: Multilingual emotion prediction in social media text," 2022.
22. M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 6 2021.
23. A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nature Machine Intelligence* **3**(1), pp. 42–50, 2021.
24. A. P. Fard, M. M. Hosseini, T. D. Sweeny, and M. H. Mahoor, "Affectnet+: A database for enhancing facial expression recognition with soft-labels," *arXiv preprint arXiv:2410.22506* , 2024.
25. J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology* **143**(1), pp. 29–36, 1982.