National Research Journal of Information Technology & Information Science
Volume No: 13, (January) Year: 2026 (Special Issue)
PP: 831-839
ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

SPEECH-DRIVEN LIP-SYNC GENERATION USING GANs

Pawan Gupta

Correspondence, Chandigarh University, NH-05, Ludhiana - Chandigarh State Hwy, Punjab

Ankit Kumar

Chandigarh University, NH-05, Ludhiana- Chandigarh State Hwy, Punjab

Utsav kumar

Chandigarh University, NH-05, Ludhiana- Chandigarh State Hwy, Punjab

Devansh Singh

Chandigarh University, NH-05, Ludhiana- Chandigarh State Hwy, Punjab

Er. Priyanka Devi

Chandigarh University, NH-05, Ludhiana- Chandigarh State Hwy, Punjab

ABSTRACT

Facial expressions during speech involve three key elements: personality, emotional state, and articulation patterns. Creating realistic virtual agents requires modeling the sophisticated relationships between these factors. Our research introduces a Conditional Sequential Generative Adversarial Network (CSG) that automatically learns to associate speech patterns with corresponding emotional expressions. This novel framework generates naturalistic facial animations by using acoustic-prosodic features extracted directly from speech, eliminating the need for text transcripts. Experimental results confirm our model's superior performance over three state-of-the-art alternatives across both quantitative and qualitative measures. We developed two emotion-specific variants: CSG-Emo-Adapted (fine-tuned on emotional datasets) and CSG-Emo-Aware (emotion-conditioned generation). Quantitative analysis reveals CSG-Emo-Adapted produces facial trajectories closer to natural recordings. In perceptual tests, this adaptation proves particularly effective for generating happy expressions, demonstrating significant improvements over the baseline CSG model.

Keywords— Lip motions, expressive and realistic lip movements, generative adversarial networks, and speech-driven models, orofacial.

INTRODUCTION

In many applications, such as deepfake technology, animated movies, and virtual assistants, lip-syncing—the synchronization of lip movements with speech—is an essential technology. Conventional techniques depended on statistical or rule-based methodologies, which frequently lacked realism. Significant advancements in producing incredibly lifelike lip movements triggered by speech input have been made possible by the development of deep learning, especially GANs. The variance in the orofacial region is caused by a number of causes. The articulatory movements forced through the vocal area activate the orofacial muscles.The emotional cues conveyed in the communication influence the relationship between lip movements and phonetic content. Individual differences also have an impact on this link between lexical and emotional elements.In the or ofacial region, these elements are intricately integrated [1, 2]. To prevent speaker fluctuations, the majority of earlier research on lip movement synthesis has relied on recordings from a single participant [3, 4, 5]. It is crucial that the models are able to accurately represent speaker variability because multimodal emotional corpora typically comprise numerous speakers with little data per subject [6]. If these differences aren't careful consideration, the model might forecast trajectory. Traditional approaches to lip motion synthesis often generate overly smoothed movements by averaging natural variations in facial expressions. Most existing systems rely heavily on linguistic representations (phonemes, triphones) [7-9] or combinations of phonetic and emotional labels [5,10,11], creating practical limitations due to their dependence on text transcriptions. We propose a novel transcription-free approach that learns the dynamic relationships between speech signals, emotional content, and corresponding facial movements directly from audio data.

Speech carries both linguistic and paralinguistic information that significantly influences visual facial expressions. As emotional states are prominently conveyed through vocal characteristics [12], our audio-driven framework captures the nuanced nonverbal cues present in natural communication. The proposed system synthesizes lip movements directly from speech features, eliminating the need for intermediate phonetic representations.Our solution employs a conditional Generative Adversarial Network (cGAN) architecture with Long Short-Term Memory (LSTM) components, forming a Conditional Sequential Generative (CSG) network. This framework learns the conditional distribution of orofacial movements given acoustic input features. The adversarial training process involves competing discriminators that enforce

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

temporal synchronization between generated lip motions and the driving speech signal, resulting in realistic, expressive facial animations.

2. GENERATIVE ADVERSARIAL NETWORKS FOR LIP-SYNC GENERATION

GANs consist of a generator and a discriminator trained in an adversarial manner. Speech-driven lip-sync GANs typically employ variations such as

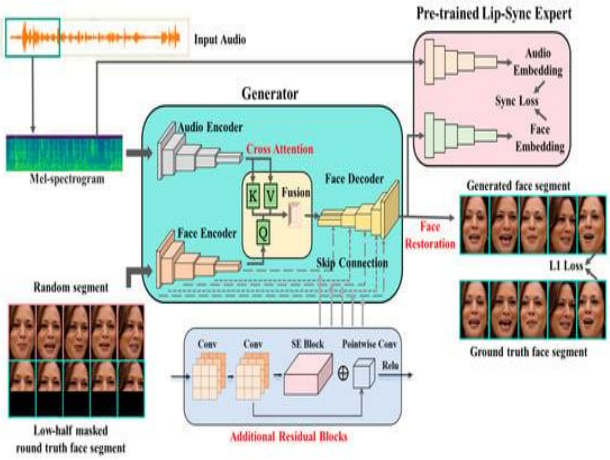

Fig1.The architecture of VividWav2Lip

2.1 GANs with conditions (cGANs)

Conditional Generative Adversarial Networks (cGANs) enhance traditional GAN architectures by incorporating auxiliary inputs, such as speech features, to guide the synthesis process. This approach has proven particularly effective for speech-driven facial animation systems. Our proposed Conditional Sequential GAN (CSG) extends this framework by integrating Long Short-Term Memory (LSTM) networks to explicitly model temporal relationships in sequential data. The evolution of Generative Adversarial Networks has led to their widespread adoption across numerous applications. While standard GAN architectures demonstrate powerful generative capabilities, their unconditional formulation provides limited control over output characteristics. The original GAN framework generates images without explicit constraints on content, making it less suitable for applications requiring precise output specifications. On the other hand, conditional GANs (cGANs) were developed in response to the necessity for precise control over the output generated by several real-world applications. which use explicit conditioning to direct the generation process. In order to generate samples that meet that particular criterion, cGANs expand the initial framework by adding new information (conditions). Different approaches to conditioning have been put forth, and they vary in how they incorporate the conditioning data into the discriminator and generator networks. In this study, we examine the many conditioning techniques that have been suggested for GANs, examining their distinct features and theoretical underpinnings. Additionally, we compare these approaches and assess how well they perform on different picture datasets. By use of these evaluations, we want to shed light on the advantages and disadvantages of different conditioning strategies, directing further study and use in generative modeling.

2.2 Conditional Sequential GAN (CSG)

A basic GAN is not the same as our suggested model [15].

Our goal is to use acoustic elements to drive the lip movements. As a result, we suggest using a conditional GAN model, in which the generator and discriminator constraints are Acoustic characteristics A window of speech characteristics and random noise interspersed across the frames make up the input for our model. The model transforms the original lip movement distribution based on speech characteristics into a noise distribution based on speech attributes that change over time. A conditional sequential GAN (CSG) is the name given to our model. Previous studies Previous studies have built a variety of sequential GAN models to capture dynamics in motion pictures [16,17]. Sequential GANs, however, have previous conditional static conditions that are linked throughout the input sequence [17]. The key aspect of our CSG model is that it uses a time-varying signal, or speech characteristics, as the input variable to condition the GAN models.

We employ two layers of BLSTMs to construct our cGANs in order to understand the link between time-continuous inputs, including lip movements and speech. For the generator, we take into consideration a linear output layer that is connected to every frame. Every frame has a sigmoid layer that we take into consideration for the discriminator. The properties of the speech (x) input are used to condition the discriminator and generator. Our learning method trains the discriminator using two types of fake samples: real samples including lip motion and voice data from various recordings, as well as samples from the generator. The approach builds upon Reed et al.'s [18] matching-aware discriminator framework, originally

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

designed for text-to-image synthesis. To produce lifelike lip movements, the generator must achieve two key objectives: first, minimizing the discrepancy between synthetic and real lip trajectories, and second, accurately capturing the temporal alignment between speech and corresponding lip motion. The discriminator plays a crucial role in this process by learning to identify two distinct types of flawed samples—mismatched lip sequences and those with incorrect audio-visual synchronization.While a standard conditional GAN could implicitly learn these errors from randomly generated outputs (Type 1), training efficiency improves significantly by explicitly exposing the discriminator to artificially desynchronized audio-lip pairs (Type 2). This strategy reinforces the model's ability to recognize and enforce the temporal coherence between speech and lip movements, which is essential for generating convincing results.The suggested CSG model must provide lip movement trajectories that are smooth. The noise, z, is therefore applied to each input frame. It symbolizes how different conditional lip motions are found all around the world. The time-varying speech elements that are shown across the frames are used as evidence by the CSG to record the dynamics of the orofacial movements. In respect to the speech features, the generator's generated sequence must have realistic dynamics, and every orofacial configuration generated at each frame must seem realistic [17].

We so apply fake/real labels to each intermediate frame from the discriminator, in addition to the final frame. We can minimize the loss function by employing this technique. for each of the sequence's intermediate frames as well as the final frame. emphasizes that the outputs from each frame of the sequence are taken into account by the discriminator. We found that this method speeds up learning in our initial tests. Remember that when we train our model, we consider a predetermined window length for both the discriminator and generator.

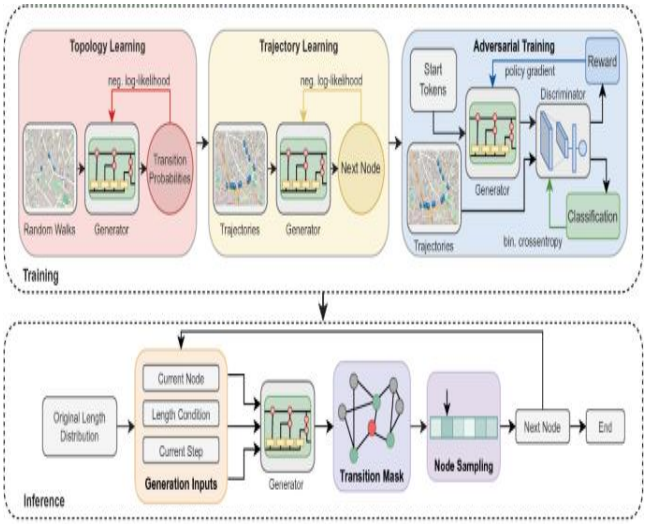

Fig2 CondTraj-GAN: Using Conditional Sequential GAN to Create Artificial Vehicle Pathways

3. CONNECTED WORKS

This section provides a summary of earlier research on lip motion generation. These investigations are divided into three groups: unit selection, techniques based on hidden Markov models (HMMs), and approaches based on deep neural networks (DNNs).

3.1 Unit Selection

The fundamental idea behind unit selection is that by choosing suitable sub-word units from a database of genuine speech, we may create new utterances that seem natural.

For such a system to function, a number of requirements must be fulfilled. Let's look at this fundamental idea of unit selection. The space of the issues will be illustrated by generalizing this specific instantiation, even if it originates from [1].

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) \quad (1)$$

In [1], as well as in subsequent and older methods of unit selection [2],

Formally speaking, the target cost C^t can be defined as the weighted sum of the relevant feature differences.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Numerous characteristics, usually representing prosodic, metrical, and phonetic context, have been proposed. We may define continuity cost as a weighted sum of feature differences in addition to choosing depending on goal cost.

$$C^c(u_{i-1}, u_i) = \sum_{k=1}^{q} w_k^c C_k^c(u_{i-1}, u_i) \quad (2)$$

These two costs must then be optimized in order to find the string of units from the database that minimise the overall cost.

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) + \quad (3)$$

$$C^c(S, u_1) + C^c(u_n, S) \quad (4)$$

where $C^c(S, u_1)$ and $C^c(u_n, S)$ handle the circumstances at the beginning and end of the utterance, and S indicates quiet. A significant amount of effort has been and will continue to be put into determining which features should be utilized and how to weight them. The secret to consistently high-quality synthesis will be to get the algorithms, metrics, and weights just right. Despite our achievements, we still have a lot to accomplish when we consider the volume of research and tests conducted in the comparably complicated subject of voice recognition. It's noteworthy to note that while theoretical benefits may be found when comparing existing algorithms, it's unclear if they hold up in practice because databases vary.

3.2 HMM-based Modeling

An extension of the Markov process, the concealed Markov explain (HMM) is used to explain processes in which the states are latent or concealed but nonetheless produce observations. For example, the states in a voice recognition system, such as a speech-to-text converter, are concealed and do not immediately reflect the text words that need to be predicted.Instead, you only need to use the speech (audio) signals that correlate to each word to infer the states.

Similar to this, POS tagging involves looking at the words in a phrase but hiding the POS tags. Hence, a Hidden Markov simulate may be used to simulate the POS tagging job, where the hidden states stand in for POS tags that produce observations, or words. A certain chance exists for the concealed states to emit observations. The likelihood that a certain state would emit a given observation is therefore represented by the emission probabilities of the Hidden Markov Model. HMMs are modeled using these emission probabilities in addition to the transition and beginning state probabilities.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

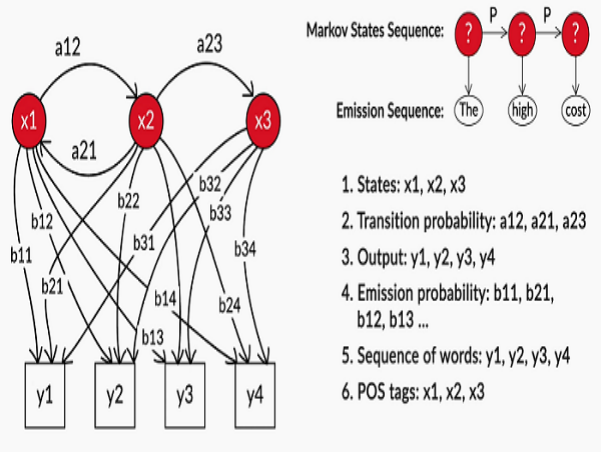

Fig.3 Hidden Markov Model (HMM)

3.3 Methods of Deep Learning

Deep neural networks (DNNs) and GANs outperform traditional methods by learning complex relationships between speech and visual movements. Machine learning, which is a branch of artificial intelligence, has given rise to deep learning or hierarchical learning. The goal of artificial intelligence is to enable robots to think and carry out intellectual activities that are typically completed by people. AI is a traditional programming paradigm in which the machine generates the answers after humans create rules for the data. There were concerns about whether a machine could automatically pick up data processing rules just by examining the data. This question led to the development of a new programming paradigm called machine learning. Data and solutions were put into machine learning so that the machines could create the rules. Instead of being explicitly programmed, a machine learning model is trained. Machine learning and deep learning emerged as a response to the need to handle ambiguous and complex problems including language translation, picture categorization, and audio recognition. Deep learning and machine learning are essentially about understanding how to represent the available data to get the intended outcomes. It is possible to use deep learning models to learn intricate relationships between inputs and outputs. Many processing layers are used in deep learning to reveal the intricate structure of the data with multiple levels of abstraction.

The term "deep" in deep learning refers to progressively higher levels of representation. A deep learning model's many nonlinear hidden layers are parameterized by weights The weights of each layer in a network must be set appropriately for the network to accurately translate inputs to targets in a deep learning model.

4. EVALUATION METRICS

Lip-sync models are evaluated using both objective and subjective metrics Both factual and subjective assessments are used to compare the models

The procedures and metrics that are regularly employed in the experimental evaluation are explained in this section.

4.1 Objective Assessment

To provide objective evaluations of the results generated by GANs, a distribution is usually fitted to the generated samples, and the probability of the test samples in that distribution is calculated.[15]. The degree to which the produced samples' distribution resembles that of the actual samples is shown by this figure. Density estimations based on the Parzen window [15] are used. We supply the test set's input characteristics since we employ conditional GANs. The generator will then supply the samples. We consider each frame as a separate sample for the purposes of estimating the distributions. Avoid falling victim to the curse of dimensionality. In order to increase the original samples' inaccuracy In order to decrease their dimensions from 45 to 15, we employ principal component analysis, or PCA. Parzen mathematics. Over 95% of the diversity in individual orofacial data is preserved in a 15D vector. We use cross validation on the generator's output samples to determine the bandwidth of the Parzen estimator. The log-likelihood of the test samples is calculated using the derived distribution, which shows the average values and standard deviations, and the estimated distribution (Sec. 3.1). However, there are always more examples of CSG models available online. We only generate one trajectory for every speaking turn by sampling several values, as the baseline systems can only supply one value per voice input.

4.2 Descriptive Measures

Our created trajectories must be regarded as realistic in addition to having a distribution that is comparable to the original sequences. To evaluate the results, we conducted subjective assessments using relative comparisons rather than absolute scoring, as prior research suggests this leads to more consistent human judgments [23]. Participants were shown pairs of

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

sequences generated using different methods and asked to select the one that appeared more natural. Additionally, they were questioned about the emotional impact of each clip.

The evaluation interface (Figure 5) allowed assessors to indicate their preference along a confidence spectrum, ranging from "definitely Video 1" to "definitely Video 2." These responses were converted into percentage-based scores for analysis. For instance, if 75% of participants leaned toward "moderately Video 1" while 25% strongly favored "definitely Video 2," the data was statistically analyzed using a two-sample z-test. The null hypothesis assumed no preference between models (i.e., a 50-50 split).

To further refine the analysis, soft preference labels were transformed into binary votes (Equation 3), where each video was assigned a single vote in cases of a tie. This approach ensured a balanced comparison while accounting for varying degrees of user confidence.

$$p = \frac{\sum_{i=1}^{i=n} 1\,(e_i \geqslant 50)}{n + \sum_{i=1}^{i=n} 1\,(e_i = 50)} \quad (5)$$

For subjective evaluation, we employed Amazon Mechanical Turk (AMT) to assess video sequences generated from different methods. The comparison included three baseline models (SWDNN, BLSTM-MSE, BLSTM-CCC) and three variants of our proposed CSG approach (CSG, CSG-Emo-Adapted, CSG-Emo-Aware), alongside the original motion-capture data. To ensure a balanced assessment, we randomly selected five speech turns per emotion, resulting in a total of 20 videos for evaluation.

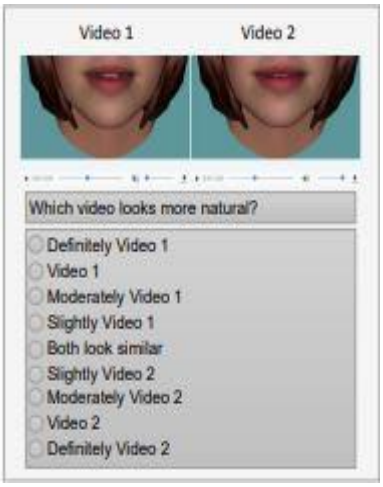

Fig. 4 Our subjective assessments are conducted using the AMT interface.

For the identical phrase. Throughout each job, the films' locations and the couples' orders are changed at random. Twenty comparisons are made for each human intelligence task (HIT). To lessen the possibility that evaluators may respond to the questions before seeing the movies, the question is displayed after the annotator has finished playing the two videos. Cronbach's alpha is used to measure the degree of agreement among assessors. Only those who have done well in our prior crowdsourcing assessments are allowed to serve as annotators [23,24,25].

5 RESULTS

5.1 Noise Dimension

The noise dimension is a crucial component of our model. We employ an m-dimensional Gaussian noise with zero mean and a diagonal covariance matrix. We changed m ∈ {1, 10, 40, 80, 150} by applying the CSG model to the noise dimension in order to choose the noise dimension. Subjective evaluations were applied to ten videos from the validation set that were created by each model. A comparison between the videos created using the original lip motion sequences is done.The outcomes make it possible to compare the models with varying noise dimensions in an indirect manner. We employ the AMT procedure outlined in Section 4.4.2.

For this review, we enlisted 15 evaluators to compare 10 pairs of videos, resulting in three assessments for each movie. Figure 5 presents the findings. $\alpha 1 = 0.72$, $\alpha 10 = 0.65$, $\alpha 40 = 0.78$, $\alpha 80 = 0.50$, and $\alpha 150 = 0.73$ are the annotators' Cronbach's alphas (the noise dimension is indicated by the subscript of α). We followed suit. Cronbach's alpha for HIT with additional raters was less than zero after two raters' pairwise average judgments were eliminated. As anticipated, average preferences for the original sequences have shifted. The results indicate that the most competitive models (i.e., the

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

bars are displaced toward the center) are m = 10 and m = 80. For the remainder of the experimental assessment, we choose m = 10 as the noise dimension.

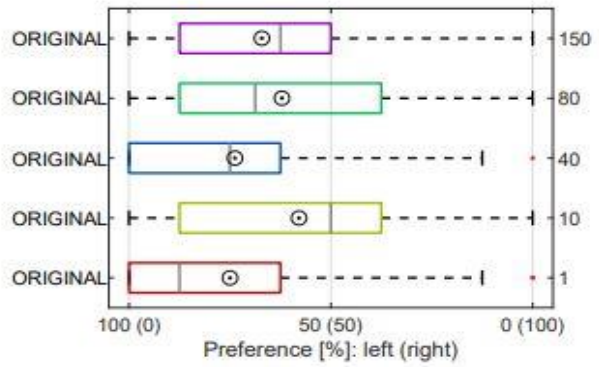

Fig. 5 compares the performance of the CSG model across different noise levels. The box plots display the interquartile range (first to third quartile) as bars, with median values marked by vertical gray lines. Mean values are represented by circular markers, while dashed lines extend to the minimum and maximum observed values. Outliers, shown as red dots, highlight data points falling outside the typical distribution range.

5.2 Comparative Analysis of the CSG Model Against Baseline Approaches

This section evaluates the proposed CSG model against three baseline methods: SWDNN, BLSTM-MSE, and BLSTM-CCC, using both objective and subjective assessments.

Objective Evaluation: We analyze the synthetic sample distributions generated by each model using the Parzen window density estimator. A total of 555K samples are produced for comparison. The results demonstrate that the CSG model significantly outperforms all baselines, with statistical significance confirmed by z-tests ($p < 0.0001$). Notably, the CSG model achieves higher log-likelihood scores, indicating better sample quality. Additionally, the BLSTM-CCC model surpasses BLSTM-MSE, highlighting the effectiveness of the CCC (Concordance Correlation Coefficient) loss function in improving performance.

Subjective Evaluation: In the subjective assessment, animations generated by the CSG, SWDNN, BLSTM-MSE, and BLSTM-CCC models were compared against ground-truth motion-captured videos. Sixteen evaluators participated, with each annotating 20 video pairs (four raters per comparison).

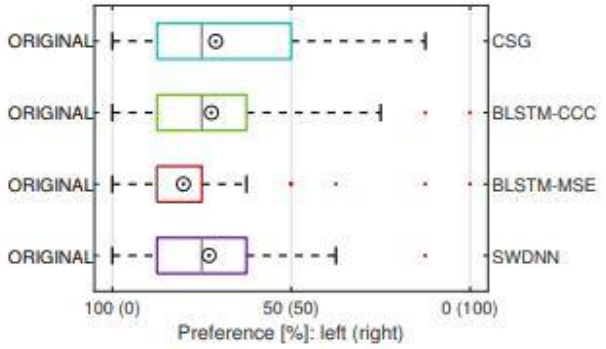

Figure 5. As the subscript of α specifies the dimension of noise, the annotators' Cronbach's alphas are α1 = 0.72, α10 = 0.65, α40 = 0.78, α80 = 0.50, and α150 = 0.73. Two raters whose average pairwise Cronbach's alpha was less than zero had their assessments removed, and we then ran the HIT with more raters. As planned,

Figure 6 shows the outcomes of these comparisons, with αSWDNN =0.88, αBLSTM−MSE =0.91 and αBLSTM−CCC =0.83, and αCSG =0.82 as the evaluators' agreement as determined by Cronbach's alpha. The p-Wert is less than 1e-12, according to the z-Test, which indicates that the original sequences are preferred because the average of all these scores is not equal to 50%. Because these approaches are speech-driven models that are not dependent on scripts, it should be emphasized that the lip synchronization is not perfect. The evaluators are therefore expected to prefer films that have

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

distinct paths. Equation 3 is used for each of these models to get the percentage of preference for the original Motion-Capture-Daten.

The SWDNN model is chosen 87% of the time, followed by the BLSTM-MSE model 92% of the time, the BLSTM-CCC model 78% of the time, and the CSG model 76% of the time.

5.3 Conversation

Overall, the experimental evaluations show that the suggested CSG models outperform the competing baselines employed in the study. Using log-likelihood, objective evaluations of the models show that the expression-aware CSG models outperform the CSG model. These evaluations also demonstrate how adaptable the suggested framework is to include expressive lip movements. The emotion classifier's results also demonstrate that the emotion expression-aware CSG models may produce lip-motion videos that portray emotional The findings also imply that the model may be enhanced.Although there is a discernible trend in the subjective assessments across emotional classes, only pleasure shows a statistically significant preference for the expression-aware CSG models. One significant finding is that certain emotions could affect the orofacial region more strongly than others. For instance, there is a definite correlation between lip shape and happiness. The link can be more nuanced for different emotions. We speculate that the lip parametrization employed in this and Xface's lack of expressiveness study are the primary causes of the subjective evaluation's lack of more conclusive findings. Xface is a basic set of tools.

This simplifies our modeling environment by allowing us to use motion capture data to parametrize the lip shape and create an animation. Considering that the conscious manifestation of emotions The evaluators' unclear perception of CSG models indicates the need for a more advanced rendering toolbox. Moreover, the outer lip markers are the only ones that describe the lip shape because the IEMOCAP database lacks data for the inside portion of the lips. As a result, our framework could miss crucial lip nuances that are necessary to express the desired mood. We're working to solve these issues right now. Despite these drawbacks, the study amply illustrates the CSG framework's modeling capabilities, opening up intriguing possibilities for speech-driven lip motion creation techniques.

6 CONCLUSIONS

We explored the generation of realistic and expressive lip movements driven by speech using Conditional Sequential Generative Adversarial Networks (CSG). Unlike traditional methods that rely on predefined transcriptions or limited speaker data, our model effectively captures the temporal relationships between speech and lip motion, ensuring greater realism. By leveraging bidirectional LSTMs within a cGAN framework, our approach improves the synchronization of lip movements with speech while preserving speaker variability. The integration of both generator-produced and mismatched samples during training enhances the model’s ability to generate authentic lip trajectories. Our findings indicate that incorporating temporal dependencies and conditioning on dynamic speech features leads to more natural and expressive animations. Future research can further refine these models by improving generalization across diverse speakers and enhancing real-time inference capabilities. Overall, our work contributes to advancing speech-driven lip-syncing technologies, making them more applicable to fields like virtual assistants, deepfake detection, and animated film production. The benefits of the suggested CSG models were shown by the experimental assessment, creating new avenues for model improvement. The present investigation centers on the orofacial region, since this region exhibits a more robust interaction between expressive and affective content. Generating facial expressions throughout the face, where the emotion can be regulated by defining the target category, is a straightforward extension of the suggested architecture. Increasing the resolution of the characteristics that characterize the lips is a second expansion of the procedure. There are no inner mouth indicators in the IEMOCAP corpus. including a more thorough illustration within the Lip arrangement will enable us to produce animations that are more realistic and expressive. Similarly, Xface isn't a particularly expressive toolbox. By using improved rendering toolkits, we hope to produce better animations. Lastly, the framework's present iteration does not require phonetic information; it is solely driven by speech. This is among the main components of our strategy. Nevertheless, by restricting the models to the lexical Finally, if the intended use requires better synchronization between the phonetic information and lip movements, the current architecture may be extended. For example, by adding phonemes as additional constraints to the CSG models, we may supply lexical content. We shall investigate these possibilities in our further research.

REFERENCES:

1. S. Mariooryad and C. Busso, “Feature and model level compensation of lexical content for facial emotion recognition,” in IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013), Shanghai, China, April 2013, pp. 1–6.
2. C. Busso and S. Narayanan, “Interplay between linguistic and affective goals in facial expression during emotional utterances,” in 7th International Seminar on Speech Production (ISSP 2006), Ubatuba-SP, Brazil, December 2006, pp. 549–556.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

3. S. Taylor, A. Kato, I. Matthews, and B. Milner, "Audioto-visual speech conversion using deep neural networks," in Interspeech 2016, San Francisco, CA, USA, September 2016, pp. 1482–1486.
4. T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," ACM Transactions on Graphics (TOG), vol. 36, no. 4, p. 94, July 2017.
5. R. Anderson, B. Stenger, V. Wan, and R. Cipolla, "Expressive visual text-to-speech using active appearance models," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013), Portland, OR, USA, June 2013, pp. 3382–3389. [6] X. Li, Z. Wu, H. Meng, J. Jia, X. Lou, and L. Cai, "Expressive speech driven talking avatar synthesis with DBLSTM using limited amount of emotional bimodal data," in Interspeech 2016, San Francisco, CA, USA, September 2016, pp. 1477–1481.
6. S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," ACM Transactions on Graphics (TOG), vol. 36, no. 4, p. 93, 2017.
7. Y. Xu, A. W. Feng, S. Marsella, and A. Shapiro, "A practical and configurable lip sync method for games," in Motion in Games (MIG 2013), Dublin, Ireland, November 2013, pp. 131–140.
8. Z. Deng, J. Lewis, and U. Neumann, "Synthesizing speech animation by learning compact speech coarticulation models," in Computer Graphics International (CGI 2005), Stony Brook, NY, USA, June 2005, pp. 19– 25.
9. Y. Cao, W. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," ACM Transactions on Graphics, vol. 24, no. 4, pp. 1283–1302, October 2005. [11] J. Parker, R. Maia, Y. Stylianou, and R. Cipolla, "Expressive visual text to speech and expression adaptation using deep neural networks," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. New Orleans, LA, USA: IEEE, March 2017, pp. 4920–4924
10. C.-C. Lee, J. Kim, A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Speech in affective computing," in The Oxford Handbook of Emotional Intelligence, [170–183] in R. Calvo, S. D'Mello, J. Gratch, and A. Kappas, eds. Oxford University Press, New York, NY, USA, December 2014.
11. https://www.mdpi.com/2079-9292/13/18/3657
12. https://arxiv.org/abs/2408.156
13. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems (NIPS), 2014, pp. 2672–2680.
14. C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in Advances In Neural Information Processing Systems, 2016, pp. 613–621.
15. S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," arXiv preprint arXiv:1707.04993, 2017
16. S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," Proceedings of The 33rd International Conference on Machine Learning (ICML), 2016.
17. https://www.google.com/url?sa=i&url=https%3A%2F%2Flink.springer.com%2Fchapter%2F10.1007%2F978-3-031-33377-4_7&psig=AOvVaw1SKa3oRZrz3bGsArMDts4R&ust=1741102656376000&source=images&cd=vfe&opi=89978449&ved=0CBcQjhxqFwoTCIiL3J6f7osDFQAAAAAdAAAAABA
18. https://www.cs.cmu.edu/~awb/papers/IEEE2002/allthetime/node1.html
19. https://wisdomml.in/hidden-markov model- hmm-in-nlp-python/#Hidden_Markov_Model
20. https://onlinelibrary.wiley.com/doi/10.1155/2021/5541134
21. G. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in International Conference on Affective Computing and Intelligent Interaction (ACII 2017), San Antonio, TX, USA, October 2017
22. A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," IEEE Transactions on Affective Computing, vol. 7, no. 4, pp. 374–388, OctoberDecember 2016.
23. N. Sadoughi, Y. Liu, and C. Busso, "Meaningful head movements driven by emotional synthetic speech," Speech Communication, vol. 95, pp. 87–99, December 2017.
24. R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," IEEE Transactions on Affective Computing, vol. To appear, 2017.