Speech to Lip Sync generation using Deep learning Algorithm

P Hirishikesh
Department Of Networking and Communication
SRM Institute of Science and Technology
Chennai, India
hp5615@srmist.edu.in

MVS Yaswanth
Department Of Networking and Communication
SRM Institute of Science and Technology
Chennai, India
my2154@srmist.edu.in

Helen Victoria A
Department Of Networking and Communication
SRM Institute of Science and Technology
Chennai, India
helenvia@srmist.edu.in

2023 OITS International Conference on Information Technology (OCIT) | 979-8-3503-5823-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/OCIT59427.2023.10430635

Abstract— The emerging growth of artificial intelligence (AI) technology created a way for many innovative developments. Speech to lip synchronization (STLS) is one of the key requirement in applications related to film making, voice modulation, video creation etc. The identification of speaking face synchronized with the voice need to be done accurately to improve the quality of the video. Many existing systems face the problem while imposing the new audio into the existing video files. The movement of lips dynamically changes according to the speaking faces. To solve the missing synchronization problem of existing videos to retrieve the original quality from the video input, the proposed system is focused on creating accurate lip synchronization model using Deep SyncNet (DSN) using Deep learning convolution architecture. The timing accuracy of the video synchronization extensively improve the quality of the video and demands real time experience from the video footage. Detection of facial variations without extracting the features are particularly challenging. The proposed system considers the existing challenges like misclassification, delayed synchronization and evaluated the SyncNet achieved the accuracy of 95% on lip synchronization with labelled dataset.

Keywords— Deep learning, Audio analysis, Speech conversion, Data augmentation, Audio processing.

I. INTRODUCTION

Have you at any point watched a video and found it hard to follow the speaker on the grounds that the lip developments were in conflict with the words being expressed? Or, in a video game are movie, have you ever been impressed by how realistically a character's lips move? If this is the case, you have witnessed this technology's power[1]. In this discourse, the system will investigate how profound learning calculations and structures can be utilized to consequently synchronize discourse with lip developments, making practical and convincing recordings that precisely convey the expressed word. The system will highlight the difficulties associated with this technology and talk about the method's technical details. In addition, the system will demonstrate some of the applications of speech-to-lip sync generation, such as its use in virtual assistants, education, and the entertainment industry. You will have a better understanding of the exciting developments in this field and the potential impact of this

technology on our day-to-day lives by the time this speech is over [2]. Therefore, let's take a look at deep learning-based speech-to-lip sync generation. The ability of the machine learning, neural networks and deep learning algorithms to work on the features, in-depth pixels of the input images improve the performance of audio synchronization. With enough example accounts, for instance, it is conceivable to combine practical sound in anybody's voice. With enough example pictures, it is feasible to blend pictures of individuals who don't exist. Additionally, anyone can be portrayed in realistic videos by saying and doing whatever the creator desires. Naturally, these synthesized pieces of content, also known as "deep fakes," can be used for a wide variety of entertaining and useful purposes [3]. However, this content can also be used as a weapon; for generating video files misused the social contents and pornography. Fake generation of facial images changing the audio files are realized. Detection of such anomaly activity need to be reduced. Facial expressions are the direct feedback of internal emotions of the humans. In certain cases detection of facial expression and Lip sync learning is important to extract the deep emotions of the humans. Machine learning based 3 dimensional facial animation model with the help of 3-4 minute high quality video is discovered here. Deep learning neural network enabled facial animation detection is explored in existing article [4].

A case study conducted to synthesis the quality of video with video data of Obama president is considered. The specific voice pattern is collected and opted for testing the quality of footage. The mouth shape is extracted using recurrent neural network (RNN). The sparse version of mouth is synchronized with the audio visuals and finally the composite video is generated. One of the efficient audio processing filter is the Mel frequency cepstral coefficient (MFCC) model [5].

979-8-3503-5823-0/23/$31.00 ©2023 IEEE

A deep synchronized realistic architecture is created with the help of Long-short term memory (LSTM) model. Specific architecture is created with the help of Obama voice synchronization dataset [6]. A customized architecture uniquely created for Obama audio is explained in article [7].

- The proposed model considers the Wav2Lip dataset collected from Kaggle.com[8]. The data are pre-trained to create a model with deep learning convolutional neural network (DCNN).
- The data are pre-processed, by cleaning the dataset removing the unwanted labels. Removal of not clarity data from the database and transformed into labelled data.
- The data is split into training data of 80%, testing data of 20% to be applied with Deep SyncNet architecture (DSN). Based on the performance of the DSN model, further testing process getting involved with few randomly selected data.
- Data augmentation takes place within the DSN model that provides deeper extraction of image features. The performance of the system is evaluated with accuracy, precision and Recall values.

The rest of the journal is developed by detailed background study and exploration in Section II. In this journal, the next section discusses the challenges in existing developments, based on the summary the tool selection and architecture derivation is explored in Section III. Further the methodology is explored in Section IV. The results and discussions with analysis are discussed in Section IV. The journal is further concluded with future scope.

II. BACKGROUND STUDY

Chung et al. (2019) the author presented a system in which creation of videos on human faces talking with Image reference is discussed. The method considered images and sound clips towards creating lip sync for the target face synchronized with the sound element. The method utilized with convolutional neural network (CNN) based Encoder and Decoder architecture. The model was trained with synthetic video collected with respect to the talking faces. Based on the self-monitoring methods, that considers trained images are synced with the audio and video frames with multiple CNN models.

Abhishek et al. (2019) the author presented a system in which the translation of digital communication based speaker identification is done. An automated pipeline method is demonstrated here where the major problem in the existing methodology is face translation issues. the interpretation of face using generalized adverse real network is discovered here. the contact you analysis of Lip sync based LipGAN is evaluated and removes all the issues in the existing method of translation of face data and further enhance the proposed approach with the demo videos.

Vougioukas et al. (2019) the author presented a system in which voice guided facial animation technique is utilized. Generalized adversarial network (GAN) with 3 discriminators are utilized here. The proposed approach considers end to end model for creating the key images of person synchronized with the audio clips. The presented system also considered the unique feature of face expression such as Ice and eyelashes are considered as visual element for synchronizing the audio. The presented system build a quality of video with better accuracy and ability to create visual feel better than the existing platform.

Eskimez et al. (2019) in the presented system the author discuss about the inter model generation and orbital vocal speech based image synchronization. In existing development aggregation problem and smoke non smoothness of vocal elements are considered. The presented system developed with combining sound and image in order to synchronize the lip changes with respect to the speech. The presented system also considered element based data compared with various existing state of approaches to provide better synchronized result.

Chen, L., et al. (2019) the author presented generally that was in network enabled speech synchronization model with different facial expressions and sound data connected from publicly available database. To avoid pixel flickering issue in existing model the presented system created a dynamic tracking mechanism that uses the flickering problem and improve the quality of video with respect to the facial images and sounds synchronization as per the real state of hot approach.

- Various problems are persist with the existing model includes,
- Face translation is not synchronized with the reference images.
- Delay in face image frame with audio sequence
- Low accuracy when dealing with small dataset with unstructured data etc.

III. SYSTEM DESIGN

Considering various existing drawbacks in face synchronization problem. The proposed approach is developed with the help of python software. Deep learning architecture is implemented using the scikit.learn library. Further the model is trained with the pretrained dataset collected from Kaggle.com as wav2lip dataset[16]. The dataset contains various video samples with synchronized audio data with labelled information.

IV.METHODOLOGY

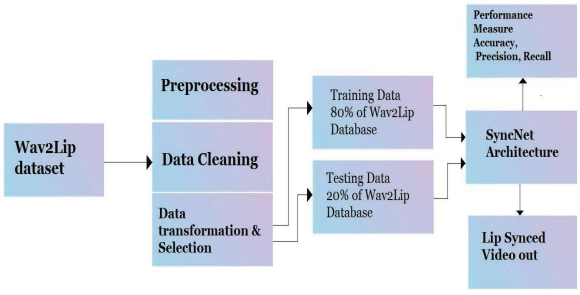

Fig 1. System architecture of proposed Deep SyncNet (DSN)

Fig 1. Shows the system architecture of Deep SyncNet (DSN) created for the analysis of pre-trained dataset collected from Wav2Lip. It acts as the data augmentation model for keeping the input data in a more precise manner.

The proposed architecture considered the wave to lip data set collected from publicly available website. The presented

convolutional neural network architecture is composed of synchronized network here represented as sync net is composed of generalized adverse real network (GAN) as the base architecture. It consists of a discriminated and generator model. The input data is preprocessed in which the noise he data and low quality data are removed. The images are collected from publicly available website hence the size of the images are different. hence resizing of images are implemented in the preprocessing stage.

The part of the images are considered for reference and utilized for ground truth comparison. The generated model keeps the reference images and generate a new pattern of face images. on the other hand the discriminator model compare the generator frames with respect to the ground true images and provide the binary cross entropy loss to provide the visual quality of the discover matter. So the generator frames are compared with the audio data which is nothing but the pre process data collected with the same database. The speech encoder is composed of 2D stacked convolutional architecture with input speech connected with the concatenated phase. The decoder is a collection of stacked convolutional layer along with sample data. The L1 reconstruction images lies between the generator image frames and the original ground truth image frames are calculator by the formula below.

$$L_G: L_{Recon} = 1/N\Sigma||L_g - L_G|| \quad (1)$$

The generator is composed of 2D convolutional neural network architecture hence the frames are independently compared with the synchronized image data. Further the performance of the architecture is evaluated using accuracy Precision and recall presented in the table 3.Performance Measure of Proposed SyncNet.

Synchronized network is a deep convolution network architecture in which the audio and video streams are compared and synchronized to each other. The model compass the features extracted from the audio data and compare with the combination of convolutional rural Network and recurrent neural network. The network consider the audio processing engine in which the audio sequences are sample and visually connected with the image frames. Embedding the captured images with the synchronized network has two parallel network handle the graphical images as well as the audio data at the same time. With highest scores in the synchronization then more correlation is presented. SyncNet is composed of large database with audio sure yes connected to the model of convolutional Neural network (CNN). Hence comparison of facial images with reference image provides ability to create audio on visual feature synchronization offset effectively.

IV.RESULTS AND DISCUSSIONS

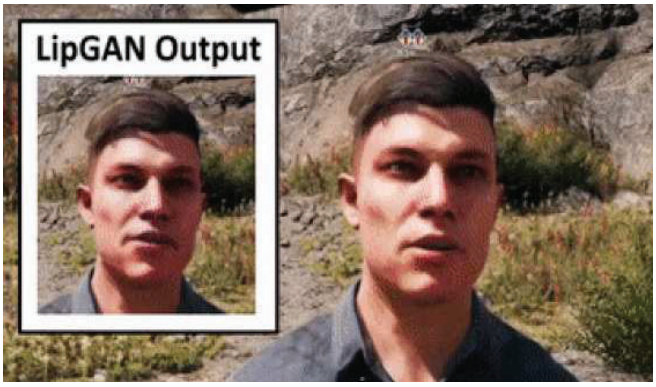

Fig 2. Output of LipGAN based synchronized face(KR. et al, (2019))

Fig 2. Shows the existing model developed using LIPGAN [13] where the synchronization achieved with mean value of 45.2, and Standard deviation of 23 approximately.

Fig 3. Lip Sync generated by SyncNet

Fig 3. Shows the proposed SyncNet architecture with achieved the sound phonics correlated with the lip movements.

Table 1. Lip Sync face translation performance *(KR et al. 2019)*

Method	Semantic Consistency	Overall Experience
Automatic Translated Subtitles	**3.45**	2.10
+ Automatic Dubbing	3.22	2.21
+ Automatic Voice Transfer	3.16	2.54
+ lip-sync	3.16	**2.96**
Manual dubbing	4.79	4.18
+ lip-sync	**4.80**	**4.55**

Table 2 Quantitative result on existing method using GRID dataset (chen et al. 2019)

Method	LRW			GRID		
	LMD	SSIM	PSNR	LMD	SSIM	PSNR
Chen [1]	1.73	0.73	29.65	1.59	0.76	29.33
Wiles [35]	1.60	0.75	29.82	1.48	0.80	29.39
Chung [3]	1.63	0.77	29.91	1.44	0.79	29.87
Baseline	1.71	0.72	28.95	1.82	0.77	28.78
ATVG-ND	**1.35**	0.78	30.27	1.34	0.79	30.51
ATVGnet	1.37	**0.81**	**30.91**	**1.29**	**0.83**	**32.15**

Table 3. Performance measure of proposed SyncNet

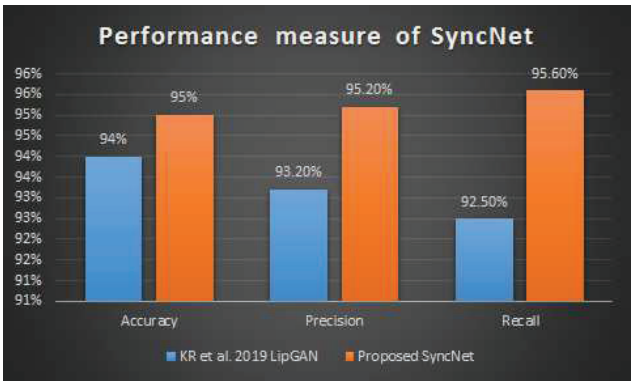

Fig 4. Performance measure of SyncNet

Fig 4. Shows the chart showing the performance of SyncNet in terms of accuracy, precision and Recall. Using LipGAN from Reference [13] Accuracy of 94% is achieved, precision of 93.20% and Recall of 92.5% is obtained. The proposed SyncNet achieved 95% accuracy, precision of 95.2% and Recall of 95.6% is achieved.

Challenges

- Images for horizontal flipping and random cropping. The data is limited only for specific facial images.
- Differences and lip movements of the recognition delay happens in the convolutional process.
- Audio visual synchronization is required to overcome the challenge of variation with existing non synchronization problem and wave to lip database created a labelled data that can able to sync with the audio data.
- The limitations of current model is the challenge with artificial neural network architecture that provides realistic images.

V. CONCLUSION

In the current technological development real world applications required automatic generation of audio visual products that can be helpful for creating videos related to Various domains such as Educational videos, news reading, emergency information transferring, alert messages. Automated instruction readers even more. the presented system consider the lips synchronization issues in audio visual products and consider the new architecture that compare the audio data with the video frames with the labelled free train to dataset. The goal of the presented study is to create a novel system that can able to create lip synchronized video with better accuracy. The presented system achieved 95% accuracy and compared with various existing state of hot approaches the Precision of 96% and recall of 95.2%. Further the presented system need to be enhanced by adding the stack and architecture such as cyclic GAN and stacked GAN to improve the quality of video and enhance the lips synchronization results.

Reference	Methodology	Accuracy	Precision	Recall
KR et al. 2019[13]	LipGAN	94%	93.2%	92.5%
Proposed	SyncNet	95%	95.2%	95.6%

REFERENCES

[1] A. Jamaludin, J. S. Chung and A. Zisserman, "You said that?: Synthesising talking faces from audio," International Journal of Computer Vision, vol. 127, no. 11-12, pp. 1767-1779, 2019.

[2] Y. Chen, W. Gao, Z. Wang, J. Miao and D. Jiang, "Mining audio/visual database for speech driven face animation," in 2001 IEEE International Conference on Systems, Man and Cybernetics. eSystems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236), 2001.

[3] S. Ogata, K. Murai, N. Satoshi and S. Morishima, "Model-based lip synchronization with automatically translated synthetic voice toward a multimodal translation system.," in IEEE International Conference on Multimedia and Expo, 2001.

[4] T. Karras, T. Aila, S. Laine, A. Herva and J. Lehtinen, "Audio-driven facial animation by joint end-toend learning of pose and emotion," ACM Transactions on Graphics (TOG), vol. 36, no. 4, pp. 1-12, 2017.

[5] S. Suwajanakorn, S. M. Seitz and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," ACM Transactions on Graphics (TOG), vol. 36, no. 4, pp. 1-13, 2017.

[6] R. Kumar, J. Sotelo, K. Kumar, A. d. Brébisson and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," arXiv preprint arXiv:1801.01442, 2017.

[7] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. 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 (MM '20). Association for Computing Machinery, New York, NY, USA, 484–492. https://doi.org/10.1145/3394171.3413532.

[8] Chung, S. W., Chung, J. S., & Kang, H. G. (2019, May). Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3965-3969). IEEE.

[9] Jha, Abhishek, et al. "Cross-language speech dependent lip-synchronization." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[10] Vougioukas, K., Petridis, S., & Pantic, M. (2019, June). End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs. In CVPR Workshops (pp. 37-40).

[11] Eskimez, S. E., Maddox, R. K., Xu, C., & Duan, Z. (2018). Generating talking face landmarks from speech. In Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Guildford, UK, July 2–5, 2018, Proceedings 14 (pp. 372-381). Springer International Publishing.

[12] Chen, L., Maddox, R. K., Duan, Z., & Xu, C. (2019). Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7832-7841).

[13] KR, P., Mukhopadhyay, R., Philip, J., Jha, A., Namboodiri, V., & Jawahar, C. V. (2019, October). Towards automatic face-to-face translation. In Proceedings of the 27th ACM international conference on multimedia (pp. 1428-1436).

[14] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, J. K. Andrew Tao and a. B. Catanzaro, "Video-to-video synthesis," rXiv preprint arXiv:1808.06601, 2018. 61

[15] S. Tulyakov, M.-Y. Liu, X. Yang and a. J. Kautz, "Mocogan: Decomposing motion and content for video generation," in In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018 .

[16] Wav2Lip dataset (kaggle.com)