Visual Speech Recognition: Exploring the Potential of Lip Reading Techniques

Liza Arora
Sage University, Bhopal (India)

***Abstract*—This study introduces a new system that reads lips automatically using Python and TensorFlow, making it easier for people with hearing difficulties to understand spoken words. We use Python because it has great tools, and TensorFlow helps us teach computers to understand lip movements. First, we process video clips in Python to pick out important lip parts. Then, with TensorFlow, we build a smart system combining Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN) to catch detailed lip movements. We fine-tune the system to be really good at predicting what words are being said. The cool part is that it works in real-time, predicting words instantly from live videos using Python. Even though there are challenges like different ways people speak and varying lighting, we keep improving the system by trying out new ideas. This research shows that by using Python and TensorFlow, we can make lip reading systems better, helping people with hearing challenges and finding more ways to use this technology in different situations.**

Keywords—Lip reading, TensorFlow, Python, Deep Learning, Convolution Neural Networks, Recurrent Neural Networks, Real-time Prediction, Communication Accessibility.

I.INTRODUCTION

Lip reading, an essential modality for communication, has witnessed transformation progress with the integration of Python and TensorFlow within the realm of deep learning. This abstract provides a concise overview of the methodology, innovations, and implications of a lip reading system developed using these powerful technologies. This research leverages Python's versatile libraries and TensorFlow's robust framework to address the complexities of automated lip reading. The datasets, comprising video sequences capturing diverse speech patterns, undergoes preprocessing, involving efficient lip region extraction and data augmentation through Python's rich ecosystem. The neural network architecture, implemented using TensorFlow, integrates Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), optimizing the system for capturing both spatial and temporal features in lip movements.

Training the model involves meticulous fine-tuning of hyperparameters, and the evaluation showcases the system's accuracy in predicting spoken words. Real-time deployment of the trained model in a Python environment underscores the practicality of the lip reading system, allowing for instantaneous word prediction from live video streams.

.

II. LITERATURE REVIEW

The integration of TensorFlow in the development of lip reading systems has marked a transformative shift in the field, as evidenced by recent literature. Early approaches, rooted in rule-based systems, struggled with the intricacies of diverse speech patterns. The advent of deep learning, particularly TensorFlow, has catalyzed a shift towards data-driven models, with studies consistently demonstrating the framework's efficacy in architecting robust lip reading systems. Architectural innovations, notably the fusion of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), enabled by TensorFlow's flexibility, have emerged as a cornerstone in extracting nuanced spatial and temporal features crucial for accurate lip reading. Researchers also grapple with data set challenges, emphasizing the importance of large and diverse datasets for training TensorFlow-based models. Moreover, the literature highlights the increasing interest in multi-modal approaches, combining lip movements with audio signals using TensorFlow, to improve system performance in real-world, noisy environments. Transfer learning and fine-tuning, facilitated by TensorFlow, have further enhanced model efficiency. While challenges persist, the reviewed literature collectively underscores TensorFlow's pivotal role in advancing the precision and real-world applicability of automated lip reading systems, contributing significantly to communication accessibility for individuals with hearing impairments.

III.MATERIALS & METHODS

Creating a lip reading system using TensorFlow requires a combination of essential hardware and software components. For hardware, a computer with a GPU is recommended, as TensorFlow can leverage the parallel processing capabilities of a graphics card to significantly accelerate the training of deep learning models. In terms of software, Python serves as the primary programming language, and TensorFlow, an open-source machine learning framework, is a crucial library for building neural networks. NumPy is utilized for efficient numerical operations, while OpenCV is indispensable for processing video frames and extracting relevant lip regions. Keras, a high-level neural networks API that runs on top of TensorFlow, simplifies the model development process. An annotated lip reading dataset, containing video sequences and corresponding transcriptions, is fundamental for training and evaluating the system. Additionally, access to a webcam or other video input source is beneficial for real-time testing and deployment. These components, combined with an appropriate integrated development environment (IDE) and, if necessary, GPU drivers, form the comprehensive material foundation for the development of a lip reading system using TensorFlow.

METHODOLODY:

Data Collection: The first phase of developing a lip reading system involves meticulous data collection. The selection or creation of a dataset is paramount, emphasizing diversity in speakers, speech patterns, and environmental conditions. Datasets such as LRW or GRID are commonly chosen, with a keen focus on capturing a broad range of lip movements and spoken words. Annotations play a critical role, ensuring that each video sequence is accurately transcribed to facilitate supervised learning. Data quality assurance measures include verifying clear and well-lit video sequences, diverse facial expressions, and accurate transcriptions aligned with the spoken content. Ethical considerations, including participant consent and privacy safeguards, are integral to responsible data collection. The dataset is systematically organized, and a strategic split into training, validation, and test sets ensures the model's robustness and prevents over fitting.

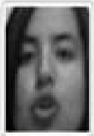
Data Preprocessing: Following data collection, preprocessing steps are applied to prepare the dataset for model training. Computer vision techniques, particularly OpenCV, are employed to extract relevant lip regions from video frames. Data augmentation strategies, such as rotation, flipping, and scaling, enhance dataset diversity and improve model generalization. Normalization is crucial for maintaining consistent pixel values across the dataset, contributing to optimal model convergence during training. The processed dataset is carefully documented, detailing the applied preprocessing techniques and any considerations that might impact subsequent model development.

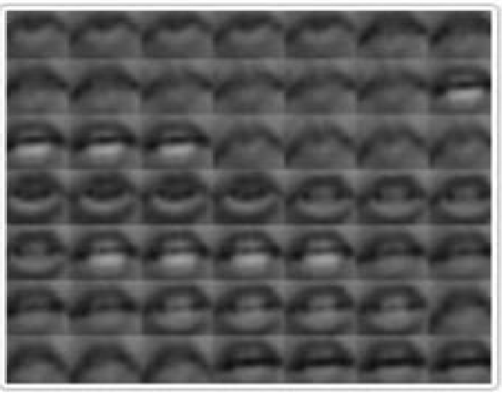
color_001.jpg

color_002.jpg

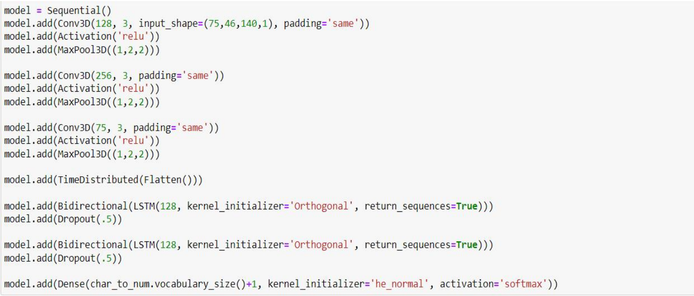
color_003.jpg

color_004.jpg


color_005.jpg

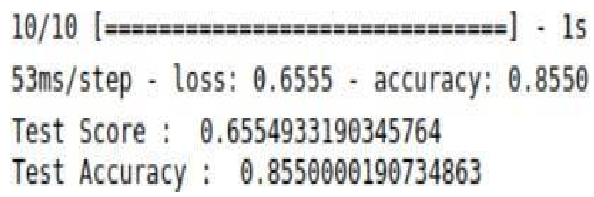


Model Development: The design and implementation of the neural network architecture constitute the core of model development. A thoughtful combination of Convolutional Neural Networks (CNN) for spatial feature extraction and Recurrent Neural Networks (RNN) for capturing temporal patterns is essential. TensorFlow, coupled with the high-level neural networks API Keras, serves as the backbone for model implementation. The flexibility of TensorFlow allows seamless integration with Keras, facilitating an efficient development process. This phase involves configuring the model for lip reading tasks, specifying input and output layers, and defining the loss functions and optimization algorithms.



```
model = Sequential()
model.add(Conv3D(128, 3, input_shape=(75,46,140,1), padding='same'))
model.add(Activation('relu'))
model.add(MaxPool3D((1,2,2)))

model.add(Conv3D(256, 3, padding='same'))
model.add(Activation('relu'))
model.add(MaxPool3D((1,2,2)))

model.add(Conv3D(75, 3, padding='same'))
model.add(Activation('relu'))
model.add(MaxPool3D((1,2,2)))

model.add(TimeDistributed(Flatten()))

model.add(Bidirectional(LSTM(128, kernel_initializer='Orthogonal', return_sequences=True)))
model.add(Dropout(.5))

model.add(Bidirectional(LSTM(128, kernel_initializer='Orthogonal', return_sequences=True)))
model.add(Dropout(.5))

model.add(Dense(char_to_num.vocabulary_size()+1, kernel_initializer='he_normal', activation='softmax'))
```

Training and Evaluation: With the model architecture in place, training begins using the preprocessed dataset. Hyper parameter tuning, including adjustments to learning rate, batch size, and optimizer choice, fine-tunes the model for optimal performance. Model training is monitored on the training set, with periodic validation on a separate set to prevent over fitting. Evaluation metrics such as accuracy, precision, recall, and F1-score provide quantitative insights into the model's performance. The training process is iterative, with adjustments made based on evaluation results to enhance the model's predictive capabilities.



```
10/10 [==============================] - 1s
53ms/step - loss: 0.6555 - accuracy: 0.8550
Test Score :  0.6554933190345764
Test Accuracy :  0.8550000190734863
```

Real-time Deployment: Upon successful training and evaluation, the trained model is saved for future use. Real-time deployment involves the integration of the model into a Python script using TensorFlow. This script allows the lip reading system to make predictions from live video streams, leveraging the practical utility of the developed model. Integration with a webcam or other video input source enables real-world applications, showcasing the system's effectiveness in dynamic settings.

Optimization and Fine-tuning: To further enhance the system's efficiency, researchers explore transfer learning techniques and consider feedback from real-world deployment. Transfer learning leverages pre-trained models or knowledge from related tasks, while the feedback loop allows for fine-tuning based on experiences in real-time scenarios. This iterative optimization phase aims to address any challenges encountered during deployment and continually improve the system's performance.

Documentation and Reporting: Comprehensive documentation throughout the methodology ensures transparency and replicability. Details about data sources, preprocessing steps, model architecture, and training specifics are documented. The final step involves generating a comprehensive report summarizing findings, including model performance metrics, challenges faced, and recommendations for future work. This documentation serves as a valuable resource for researchers and practitioners interested in the lip reading system's development process.

IV. Application Area

Lip reading systems, propelled by advancements in deep learning and computer vision, find diverse applications across several domains where traditional spoken communication may pose challenges or limitations. One significant application lies in enhancing accessibility for individuals with hearing impairments. Real-time communication is vastly improved as lip reading systems provide visual cues for spoken words, offering a valuable alternative or complement to traditional audio-based communication methods. Beyond accessibility, these systems have practical implications in security and surveillance, where the ability to analyze verbal content in noisy or covert environments can bolster efforts in identifying potential threats. In the realm of education and language learning, lip reading systems aid language learners by providing insights into pronunciation and improving speech recognition skills.

Additionally, these systems contribute to assistive technology, empowering individuals with motor disabilities to control smart devices through a combination of lip movements and spoken commands. The integration of lip reading in automated transcription services has applications in business meetings, conferences, and lectures, making content accessible and searchable. Moreover, lip reading systems play a role in human-robot interaction, healthcare communication with patients facing speech challenges, and entertainment industries by facilitating accurate dubbing and subtitling. In law enforcement, forensic analysis benefits from the ability to analyze lip movements in surveillance footage, complementing traditional audio evidence. From virtual assistants in customer service to applications in public safety, emergency response, and smart home integration, the versatility of lip reading systems continues to expand, offering innovative solutions to improve communication, accessibility, and user experience across a wide range of domains.

V. DIFFICULTIES AND CHALLENGES OF LIPREADING

Developing a lip reading system poses a myriad of difficulties and challenges rooted in the complexity of interpreting visual cues from lip movements. One of the foremost challenges lies in the variability of lip movements observed during speech. Lips exhibit a diverse range of shapes and motions, making it intricate to devise a model that can accurately generalize across different individuals and speaking styles. This necessitates the construction of a system that can effectively capture and interpret the nuanced variations in lip dynamics, a task that demands a comprehensive and diverse dataset for training.

Another significant challenge arises from the intricacies of speech variations and co-articulation. Co-articulation, where the pronunciation of one phoneme affects adjacent phonemes, introduces complexity into lip reading. Accommodating variations in speech speed, style, and linguistic contexts further complicates the accurate interpretation of lip movements. Crafting a model that can adeptly handle these intricacies poses a substantial challenge, requiring a nuanced understanding of phonetic variations and a flexible architecture capable of accommodating diverse speech patterns.

The limited availability of annotated lip reading datasets represents an ongoing challenge. Such datasets are often constrained in size and may not adequately cover the full spectrum of possible lip movements and spoken words. Training a robust model with a limited dataset can lead to over fitting, diminishing the system's capacity to generalize effectively to unseen data. Expanding and diversifying available datasets is a persistent difficulty in ensuring the development of models that can comprehensively capture the complexities of lip reading.

Noise and environmental factors introduce additional hurdles. Noisy surroundings and fluctuations in lighting conditions can compromise the quality of video frames, affecting the system's ability to accurately extract relevant lip features. Robust preprocessing techniques are essential to mitigate these challenges, necessitating careful consideration and innovative solutions to handle noise and lighting variations effectively.

The integration of lip reading with other modalities, such as audio signals, presents a significant challenge. Aligning and fusing information from different sources to improve overall accuracy demands careful consideration. Achieving seamless integration while preserving the complementary nature of visual and audio cues is a complex task requiring sophisticated model architectures and integration strategies.

Real-time processing capabilities for live video streams represent yet another challenge. Optimizing the model for efficiency without compromising accuracy is a delicate balance that requires careful architectural design and optimization efforts. The computational demands of the model must align with the imperative for instantaneous predictions, underscoring the need for innovative solutions in model deployment and execution.

VI. CONCLUSION

In conclusion, the development of a lip reading system presents a tapestry of challenges rooted in the intricate nature of interpreting visual cues from lip movements. From the variability in lip dynamics to the complexities introduced by speech variations and co-articulation, each challenge underscores the need for sophisticated models capable of nuanced understanding. Limited datasets and the persistent influence of noise and environmental factors further underscore the intricacies of constructing a robust lip reading system. However, these challenges also represent opportunities for innovation and improvement, driving the need for expansive and diverse datasets, advanced preprocessing techniques, and ethical considerations in model development. As the field evolves, addressing these difficulties becomes imperative for creating effective and widely applicable lip reading systems that contribute meaningfully to accessibility, communication, and interaction across diverse domains. Despite the complexities, the pursuit of overcoming these challenges stands at the forefront of advancing technology to make lip reading systems more accurate, reliable, and inclusive in their applications.

VII. CONCLUSION

[1] H. Zhou, W. Zhou, Y. Zhou, and H. Li, ''Spatial-temporal multi-cue network for continuous sign language recognition,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Mar. 2020, pp. 13009–13016.

[2] W. H. Sumby and I. Pollack, ''Visual contribution to speech intelligibility in noise,'' *J. Acoust. Soc. Amer.*, vol. 26, no. 2, pp. 212–215, Mar. 1954.

[3] E. D. Petajan, ''Automatic lipreading to enhance speech recognition,'' in *Proc. Global Telecommun. Conf.*, 1984, pp. 265–272.

[4] A. J. Goldschen, O. N. Garcia, and E. Petajan, ''Continuous optical automatic speech recognition by lipreading,'' in *Proc. 28th Asilomar Conf. Signals, Syst. Comput.*, 2002, pp. 572–577.

[5] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, ''Continuous auto matic speech recognition by lipreading,'' in *Motion-Based Recognition*, M. Shah and R. Jain, Eds. Dordrecht, The Netherlands: Springer, 1997, pp. 321–343.

[6] G. Zhao, M. Pietikäinen, and A. Hadid, ''Local spatiotemporal descriptors for visual recognition of spoken phrases,'' in *Proc. ACM Int. Multimedia Conf. Exhib.*, Sep. 2007, pp. 57–66.

[7] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, ''Multimodal deep learning,'' in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.

[8] H. Lee, C. Ekanadham, and A. Y. Ng, ''Sparse deep belief net model for visual area V2,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 873–880.

[9] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, ''Lipreading using convolutional neural network,'' in *Proc. Conf. Int. speech Commun. Assoc.*, 2014, pp. 1149–1153.

[10] M. Wand, J. Koutnik, and J. Schmidhuber, ''Lipreading with long short term memory,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6115–6119.

[11] J. S. Chung and A. Zisserman, ''Lip reading in the wild,'' in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 87–103.

VOLUME 8, 2020

204539M. Hao *et al.*: Survey of Research on Lipreading Technology

[12] Y. M. Assael, B. Shillingford, S. Whiteson, and N.de Freitas, ''Lip Net: End-to-end sentence-level lipreading,'' 2016, *arXiv:1611.01599.* [Online]. Available: http://arxiv.org/abs/1611.01599

[13] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, ''LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild,'' in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognition. (FG)*, May 2019, pp. 1–8.

[14] A. Fernandez-Lopez and F. M. Sukno, ''Survey on automatic lip-reading in the era of deep learning,'' *Image Vis. Comput.*, vol. 78, pp. 53–72, Oct. 2018.

[15] T. Wark, S. Sridharan, and V. Chandran, ''An approach to statistical lip modelling for speaker identification via chromatic feature extraction,'' in *Proc. 14th Int. Conf. Pattern Recognition.*, vol. 1, 1998, pp. 123–125.

[16] T. W. Lewis and D. M. Powers, ''Lip feature extraction using red exclusion,'' in *Proc. Sel. Papers Pan-Sydney Workshop Visualisation*, vol. 2, Dec. 2000, pp. 61–67.

[17] E. Skodras and N. Fakotakis, ''An unconstrained method for lip detection in color images,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 1013–1016.

[18] V. E. C. Ghaleh and A. Behrad, ''Lip contour extraction using RGB color space and fuzzy c-means clustering,'' in *Proc. IEEE 9th Int. Conf. Cybernetic Intell. Syst.*, Sep. 2010, pp. 1–4.

[19] A. D. Gritzman, D. M. Rubin, and A. Pantanowitz, ''Comparison of colour transforms used in lip segmentation algorithms,'' *Signal, Image Video Process.*, vol. 9, no. 4, pp. 947–957, May 2015.

[20] H. Jun and Z. Hua, ''A real time lip detection method in lipreading,'' presented at the 26th Chin. Control Conf., Zhangjiajie, China, 2007. [Online]. Available: http://www.wanfangdata.com.cn/details/ detail.do?_type=conference&id=6431173.

[21] X. Fan, F. Zhang, H. Wang, and X. Lu, ''The system of face detection based on OpenCV,'' in *Proc. 24th Chin. Control Decis. Conf. (CCDC)*, May 2012, pp. 648–651.

[22] N. Puviarasan and S. Palanivel, ''Lip reading of hearing impaired per sons using HMM,'' *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4477–4481, Apr. 2011.

[23] M. Kass, A. Witkin, and D. Terzopoulos, ''Snakes: Active contour mod els,'' *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[24] Q. Dinh Nguyen and M. Milgram, ‘‘Multi features active shape models for lip contours detection,’’ in *Proc. Int. Conf. Wavelet Anal. Pattern Recognition.*, vol. 1, Aug. 2008, pp. 172–176.

[24] L. Rothkrantz, ‘‘Lip-reading by surveillance cameras,’’ in *Proc. Smart City Symp. Prague (SCSP)*, May 2017, pp. 1–6.

[33] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, ‘‘Audio-visual auto matic speech recognition: An overview,’’ in *Issues in Visual and Audio Visual Speech Processing*. Cambridge, MA, USA: MIT Press, 2004.

[25] C. Lee, E. Lee, S. Jung, and S. Lee, ‘‘Design and implementation of a real-time lipreading system using PCA and HMM,’’ *J. Korea Multimedia Soc.*, vol. 7, no. 11, pp. 1597–1609, 2004.

[26] J. Yao and Z. Kaifeng, ‘‘Evaluation model of the artist based on fuzzy membership to improve the principal component analysis of robust ken nel,’’ in *Proc. Int. Conf. Big Data Secur. Cloud*, Apr. 2016, pp. 322–326.

[27] G. Sterpu and N. Harte, ‘‘Towards lipreading sentences using active appearance models,’’ in *Proc. Int. Conf. Auditory-Vis. Speech Process*, 2017, pp. 70–75.

[28] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, ‘‘A comparison of model and transform-based visual features for audio-visual LVCSR,’’ in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Aug. 2001, pp. 825–828.

[29] S. S. Morade and S. Patnaik, ‘‘Lip reading using DWT and LSDA,’’ in *Proc. IEEE Int. Advance Comput. Conf. (IACC)*, Feb. 2014, pp. 1013–1018.

[30] J. He, H. Zhang, and J. Z. Liu, ‘‘LDA based feature extraction method in DCT domain in lipreading,’’ *Comput. Eng. Appl.*, vol. 45, no. 32, pp. 150–155, 2009.

[31] Y. Liang, W. Yao, and M. Du, ‘‘Feature extraction based on LSDA for lipreading,’’ in *Proc. Int. Conf. Multimedia Technol.*, Oct. 2010, pp. 1–4.

[32] I. Almajai, S. Cox, R. Harvey, and Y. Lan, ‘‘Improved speaker inde pendent lip reading using speaker adaptive training and deep neural networks,’’ in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2722–2726.

[33] G. Potamianos, J. Luettin, and C. Neti, ‘‘Hierarchical discriminant features for audio-visual LVCSR,’’ in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 2001, pp. 165–168.

[34] A. A. Shaikh, D. K. Kumar, W. C. Yau, M. Z. C. Azemin, and J. Gubbi, ‘‘Lip reading using optical flow and support vector machines,’’ in *Proc. 3rd Int. Congr. Image Signal Process.*, vol. 1, Oct. 2010, pp. 327–330.

[35] L. Cappelletta and N. Harte, ‘‘Viseme definitions comparison for visual only speech recognition,’’ in *Proc. 19th Eur. Signal Process. Conf.*, 2011, pp. 2109–2113.

[36] Z. Zhou, G. Zhao, and M. Pietikainen, ‘‘Towards a practical lipreading system,’’ in *Proc. CVPR*, Jun. 2011, pp. 137–144.

[37] G. Zhao and M. Pietikainen, ‘‘Dynamic texture recognition using local binary patterns with an application to facial expressions,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[38] A. Rekik, A. Ben-Hamadou, and W. Mahdi, ‘‘A new visual speech recognition approach for RGB-D cameras,’’ in *Image Analysis and Recognition*. Cham, Switzerland: Springer, 2014, pp. 21–28.

[39] X. Ma, L. Yan, and Q. Zhong, ‘‘Lip feature extraction based on improved jumping-snake model,’’ in *Proc. 35th Chin. Control Conf. (CCC)*, Jul. 2016, pp. 6928–6933.