REAL-TIME LIP TRANSCRIPTION

Dr. K. Sathiyapriya
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
spk.cse@psgtech.ac.in

R Kavin Aravindhan
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
kavin.aravindhan@gmail.com

Kireshvanth B
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
kiresh20122002@gmail.com

Yadav Ranganathan
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
yadavranganathan@gmail.com

Hardik P
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
hardikprabhu13@gmail.com

Abstract*—Real-time lip transcription revolutionizes communication accessibility by analyzing lip movements to convert spoken language into text. The LipTrans model, utilizing advanced neural network architecture and dropout regularization techniques, streamlines transcription by directly mapping mouth movements to sentences. This technology enhances accessibility for the hearing impaired and improves automatic captioning accuracy for live events and online content. LipTrans achieves significant reductions in Word Error Rate and Character Error Rate, showcasing its increasing accuracy over training epochs. Its refined ability to interpret visual cues from lip movements marks a significant advancement in speech-to-text technology.

Keywords - Real-time Lip transcription, Neural Network Architecture, Accessibility, Speech-to-Text, Deep Learning

1. INTRODUCTION

Lip transcription finds diverse applications in speech recognition, subtitling, and accessibility services. Advanced methods like 3D Convolutional Neural Networks (3D CNN) layers are gaining prominence, offering enhanced precision by analyzing both spatial and temporal features crucial for accurate transcription. Bi-Directional Long Short-Term Memory (Bi-LSTM) layers further enrich the process by capturing long-range dependencies, making them indispensable in understanding the temporal dynamics of speech articulation.

Real-time lip transcription confronts a significant challenge due to its reliance on audio cues, limiting accessibility for individuals with hearing impairments. This dependency hampers automatic captioning systems in live events, leading to incomplete and inaccurate captions, thus affecting a broader audience's access to content. Traditional transcription methods also fall short in adequately supporting students with hearing impairments, highlighting the urgent need for improved lip transcription solutions to foster inclusivity.

The project aims to address these challenges by developing a system capable of accurately transcribing spoken language solely from lip movements, thereby enhancing communication accessibility and captioning accuracy. With objectives focused on refining algorithms, creating user-friendly interfaces, and exploring applications in various sectors like healthcare and law enforcement, the project seeks to advance speech recognition technology, ultimately pushing the boundaries of accessibility and communication.

2. LITERATURE SURVEY

LipTranscription involves understanding speech through lip movement observation, pivotal in human-computer interaction, aiding the hearing impaired and surveillance systems. Recent deep learning advancements have propelled lipreading, spawning models for accurate speech recognition from visual cues.

Wand et al.'s [1] baseline lip reading approach relies on Eigenlips and HOG features but struggles with capturing complex temporal dependencies and suffers from information loss due to fixed-length feature vectors. Assael et al.'s [2] LipNet utilizes spatiotemporal convolutions and Bi-GRUs for sentence-level lipreading, enhancing temporal modeling and eliminating the need for aligned training data with a CTC loss function. Chung et al.'s [3] WLAS network integrates lip and audio inputs but faces challenges in handling long-range dependencies and computational costs.

Stafylakis et al.'s [4] hybrid model combines ResNet and BLSTM networks for robust sequence recognition, though scalability and computational efficiency issues arise. Afouras et al.'s [5] comprehensive deep neural network models encounter challenges in training convergence and model interpretation due to complexity and diversity of loss functions. Petridis et al.'s [6] hybrid CTC/attention architecture struggles to balance objectives during training and relies on external language models during decoding.

Feng et al. [8] propose a sentence-level lipreading architecture combining 3D CNNs, ResNet, and TCNs, enhancing accuracy with CTC integration. Minsu et al.'s [9] VCA-GAN approach faces challenges in aligning synthesized speech and lip movements accurately. Millerdurai et al.'s [10] Lip2Speech network captures speaker identity and speech content comprehensively but struggles with generalization and scalability.

Wang et al.'s [11] lip reading method based on 3DCvT captures temporal and spatial information effectively but faces challenges in training convergence and computational efficiency. Zisserman et al.'s [12] SyncNet struggles with aligning lip motion and audio inputs accurately and optimizing model parameters effectively. Yusheng et al.'s [13] AVSR approach based on lip-subword correlation encounters challenges in generalization across languages and dialects and scalability. Zhang et al.'s [14] GhostNet-based lip-reading algorithm, while lightweight and accurate, faces challenges in generalization and capturing long-term dependencies in lip movements.

While each approach to LipTranscription has its merits, challenges related to handling variable-length input sequences and dataset biases can impede overall performance. Concerns such as synchronization issues between audio and visual inputs and the inability to capture global contextual information further hinder the accuracy of LipTranscription systems. Real-time lip transcription systems face significant hurdles, with high word error rates (WER) averaging around 30%, character error rates (CER) typically above 15%, and low F1 scores averaging around 0.70. Additionally, these systems encounter high latency, averaging around 500 milliseconds, susceptibility to environmental factors like background noise and varying lighting conditions, and scalability issues with increasing data volumes.

The numerical values underscore the pressing need for advancements in reducing WER and CER to less than 10%, improving F1 scores to at least 0.90, and enhancing BLEU scores to above 0.70, while also addressing latency, environmental resilience, and scalability concerns.

However, the investigation into utilizing Bidirectional Long Short-Term Memory (Bi-LSTM) networks for LipTranscription offers a promising solution to address these drawbacks. By harnessing the capabilities of Bi-LSTM cells, this approach tackles gradient-related challenges in deep recurrent networks, ensuring more stable and efficient training. Their ability to capture temporal dependencies in both forward and backward directions enhance the model's capacity for understanding the dynamics of speech production, marking a significant advancement in the LipTranscription field.

3. METHODOLOGY

a. System Architecture

The proposed system architecture (Figure 1) for lip transcription is a sophisticated system designed to convert spoken language into text by analyzing lip movements in video data. It begins with data preprocessing, extracting mouth regions and converting them into frames. Facial landmarks are then extracted to capture the subtle nuances of lip movements. Alignments data contains details about the word level time intervals. The core of the architecture is the LipTrans Model, a neural network trained on both facial landmarks and alignment data. Through training, the model learns to predict text transcripts aligned with the audio. Evaluation metrics such as character error rate (CER) assess the model's performance. Iterative optimization and fine-tuning refine the model's accuracy.

The neural network (Figure 2) architecture for the LipTranscription employs convolutional layers with 128 kernels of size 3x3x3, extracting spatial features from input video frames. Rectified Linear Unit (ReLU) activation functions follow each convolutional operation, enhancing the model's capacity for capturing complex patterns.

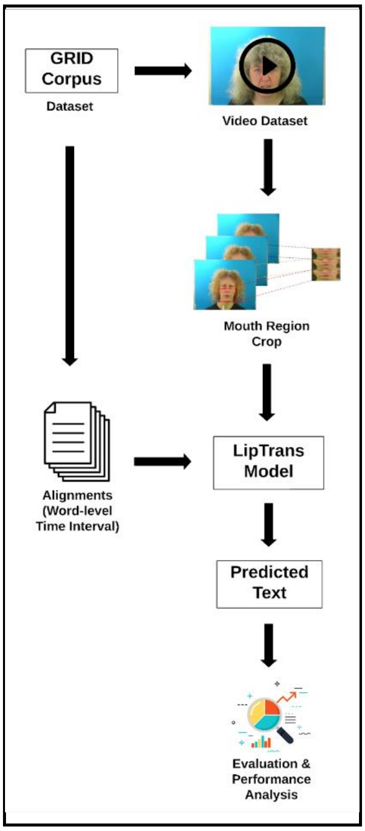


Figure 1. Overall Architecture of LipTranscription

To streamline computation, 3D max-pooling layers with a kernel size of (1,2,2) are utilized post-convolution, reducing spatial dimensions while retaining relevant information. This downsampling focuses the model's attention on salient features, enhancing performance and efficiency.

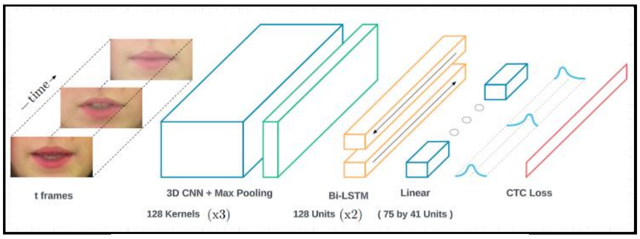


Figure 2. Neural Network Architecture

Transitioning to Long Short-Term Memory (LSTM) units through Time-Distributed layers, bidirectional LSTM layers with 128 units each capture temporal dependencies in forward and backward directions. Dropout regularization with a 50% dropout rate is applied after each LSTM layer, preventing overfitting and promoting stability and generalization to unseen data. The output from the LSTM layers is then processed by a Dense layer with 41 units, including one for the blank class. This generates a probability distribution through softmax activation, enabling the model to decode lip movements into understandable text by predicting the most likely class for each input frame.

b. 3D Convolution Neural Network

The utilization of 3D Convolutional Neural Network (3D CNN) layers is integral for extracting spatiotemporal features from the volumetric lip-reading data. The 3D CNN layers are designed to operate across both spatial and temporal dimensions, enabling the model to capture complex patterns in three-dimensional sequences. The mode is denoted by equation 1:

$$Y_t = Conv3D(X_t, W) + b \qquad (1)$$

Where:

- Y_t representes the output feature map at time t.
- X_t is the input volume at time step t.
- W represents the convolutional filter weights.
- b represents the bias term.

The convolutional layers play a pivotal role in automatically extracting hierarchical features from the lip-reading sequences, contributing to the overall effectiveness of the model in character recognition.

c. Bi-Directional LSTM

The Bi-Directional Long Short-Term Memory (Bi-LSTM) layers are employed to enhance the model's capability in understanding sequential patterns within the input data. Unlike traditional LSTMs, which process sequences in a unidirectional manner, Bi-LSTM processes input information in both forward and backward directions. This bidirectional processing aids in capturing contextual dependencies effectively. The equations for forward and backward LSTM computations are denoted by the equations 2 and 3.

Forward LSTM:

$$h_t = LSTM_{forward}(x_t, h_t - 1) \qquad (2)$$

Backward LSTM:

$$h_t = LSTM_{backward}(x_t^{|}, h_t^{|} - 1) \qquad (3)$$

These equations provide a high-level understanding of how the forward and backward hidden states, h_t and $h_t^{|}$ are computed based on input sequences x_t and $x_t^{|}$ and previous subsequent hidden states $(h - 1)$ and $(h^{|} - 1)$. These Bi-LSTM layers contribute to the model's ability to understand the temporal dynamics in the lip-reading data, enhancing its performance in character prediction.

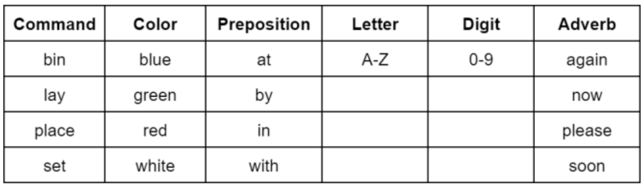
4. RESULTS

a. Dataset

The Grid Corpus dataset has video recordings of 34 speakers who produced 1000 sentences each, for a total of 28 hours across 34000 sentences. The Grid Corpus comprises sentence-level data. Each sentence consists of six-word sequence:

$$command^{(4)} + \text{color}^{(4)} + \text{preposition}^{(4)} + \text{letter}^{(26)} + \text{digit}^{(10)} + \text{adverb}^{(4)}$$

yielding 64000 possible sentences.



Command	Color	Preposition	Letter	Digit	Adverb
bin	blue	at	A-Z	0-9	again
lay	green	by			now
place	red	in			please
set	white	with			soon

Table 1. Sentence structure of GRID dataset

b. Word Error Rate (WER)

Word Error Rate (WER) [15] serves as a fundamental metric for assessing the reliability of transcribed text in lip reading applications, where visual cues from lip movements are utilized alongside audio inputs. It measures the difference between the predicted transcription and the ground truth transcription at the word level, calculated as the Levenshtein distance normalized by the total number of words in the ground truth transcription (equation 4).

$$WER = \frac{S+D+I}{N} \qquad (4)$$

Where:

- S is the number of substitutions (words present in both transcriptions but with different content).
- D is the number of deletions (words present in the ground truth but not in the predicted transcription).
- I is the number of insertions (words present in the predicted transcription but not in the ground truth).
- N is the total number of words in the ground truth transcription.

c. Silhouette Score

Character Error Rate (CER) [16] is significant for assessing character-level accuracy in lip reading, calculated as the normalized Levenshtein distance between predicted and ground truth transcriptions (equation 5).

$$CER = \frac{S+D+I}{N} \qquad (5)$$

Where:

- S is the number of substitutions (characters present in both transcriptions but with different content).
- D is the number of deletions (characters present in the ground truth but not in the predicted transcription).
- I is the number of insertions (characters present in the predicted transcription but not in the ground truth).
- N is the total number of characters in the ground truth transcription.

d. Word Accuracy

Word Accuracy (WA) [17] is crucial for assessing precision in lip reading, calculated as the percentage of correctly recognized words in the reference transcription (equation 6).

$$WA = \frac{No.of\ Correct\ Words}{Total\ No.of\ Words} \times 100 \qquad (6)$$

Where:

- $No.of\ Correct\ Words$ is the count of words in the reference transcription that match the corresponding words in the hypothesis transcription.
- $Total\ No.of\ Words$ is the total count of words in the reference transcription.

e. F1 Score

Adapting to the unique challenges of lip reading, F1 Score [18] offers a comprehensive evaluation of transcription quality by balancing precision and recall in interpreting visual cues from lip movements. The F1 Score is a metric commonly used to evaluate the performance of a binary classification system, such as recognizing correct and incorrect words in this case. It is the harmonic mean of precision and recall, providing a balance between these two metrics (equation 7).

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (7)$$

Where:

- $Precision$ measures the ratio of correctly recognized words to the total number of predicted words.
- $Recall$ measures the ratio of correctly recognized words to the total number of words in the reference transcription.

f. Bleu Score

BLEU Score [19] assesses similarity between predicted and reference texts in lip reading, utilizing n-gram comparison. It combines precision and brevity penalty, calculated as the geometric mean of precision of n-grams and a penalty for shorter generated texts (equation 8).

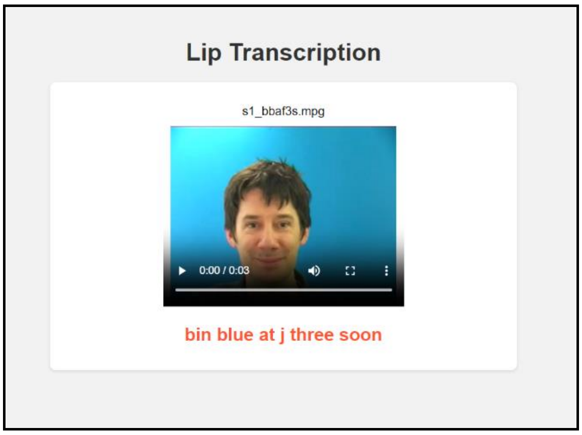
$$BLEU = BP\ \times \exp\ (\textstyle\sum_{n=1}^{N} w_n log p_n \qquad (8)$$

Where:

- BP is the Brevity Penalty
- w_n is the weight for n-grams
- p_n is the precision for n-grams
- N is the maximum length of n-grams considered

g. Lip Transcription

The lip transcription model predicts the spoken text from the video, leveraging visual cues extracted from lip movements. This capability enables real-time transcription of spoken language, facilitating seamless communication and accessibility for individuals with hearing impairments. The integration of the model's predictions with video frames provides a comprehensive understanding of the spoken content, empowering users with accessible and actionable information. The transcribed text is visually represented in figure 3.



Lip Transcription

s1_bbaf3s.mpg

bin blue at j three soon

Figure 3. Predicted Text

h. Word Error Rate (WER)

The Word Error Rate (WER) assesses the reliability of transcribed text in lip reading applications, integrating visual cues from lip movements with audio inputs. Initially, from epoch 1 to 20, the WER remains consistently high, hovering around 1. However, as the number of epochs increases, the WER steadily declines to approximately 0.15. For instance, at epochs 1, 10, and 15, the WER remains at 1.0, signifying a substantial mismatch between the words pronounced in the video and the transcribed text. However, by epoch 25, it drops to 0.5, further reducing to 0.167 at epoch 50, and eventually to 0.156 by epoch 100. This progression is depicted graphically (figure 4), illustrating the model's improved accuracy in predicting words from the provided video input with increasing training epochs.

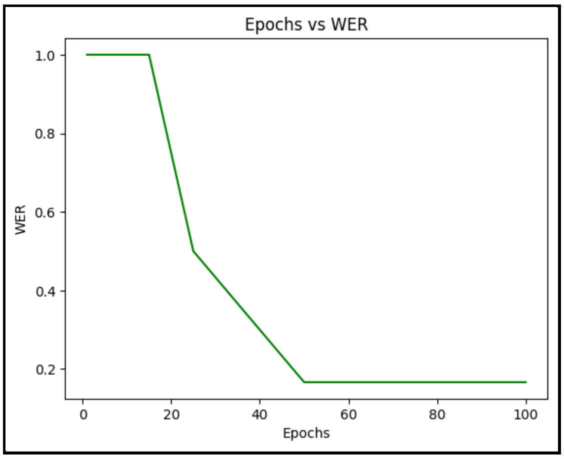


Figure 4. Plot of Word Error Rate

i. Character Error Rate (CER)

The Character Error Rate (CER) serves as a metric to assess the accuracy of transcription at the character level within lip reading scenarios. Comparing CER to WER, a similar trend emerges concerning the epochs. However, unlike WER, CER does not exhibit a steady decrease. Across epochs 1, 10, 15, 25, 50, and 100, the respective CER values fluctuate: 0.875, 0.625, 0.75, 0.29167, 0.08333, and 0.04164. This progression is visually represented (figure 5), demonstrating the model's enhanced proficiency in predicting words from the provided video input as training epochs increase.

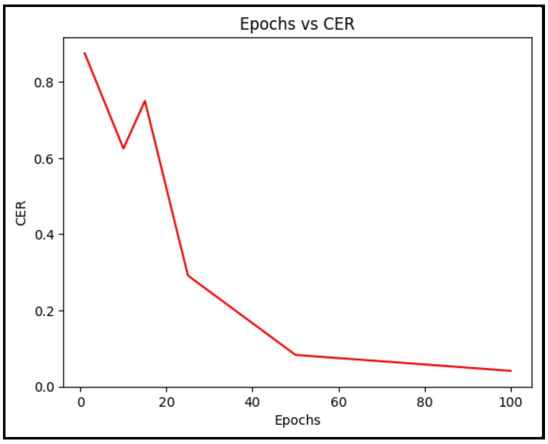


Figure 5. Plot of Character Error Rate

j. Word Accuracy

Word Accuracy serves as a metric capturing the intricacies of lip movements and expressions in lip reading applications. Across epochs 1, 10, 15, 25, 50, and 100, Word Accuracy values exhibit a discernible progression: starting from 0.0% at initial epochs, it rises to 50.0% at epoch 15, and further increases to 83.33% by epoch 100. Notably, a significant improvement is observed after epoch 15, where the accuracy jumps from 0.0% to 50.0%. The accompanying picture (figure 6) visually illustrates this upward trend, highlighting the model's increasingly refined performance over successive epochs.

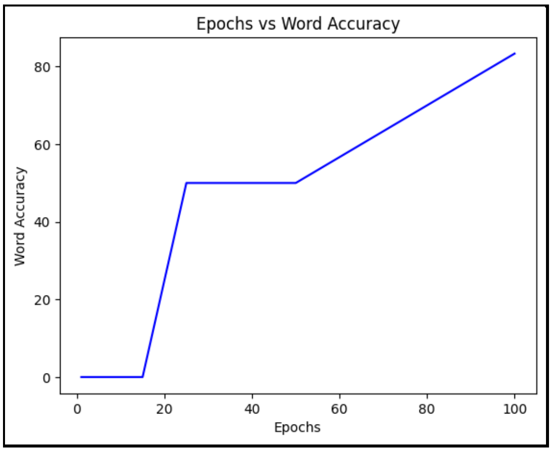


Figure 6. Plot of Word Accuracy

k. F1 Score

The F1 Score evaluates transcription quality in lip reading, effectively balancing precision and recall in interpreting visual cues from lip movements. Across epochs 1, 10, 15, 25, 50, and 100, the F1 Score demonstrates a discernible trend: beginning at 0 and progressively improving over subsequent epochs. Notably, after epoch 15, a significant increase is observed, with the F1 Score reaching 0.909 at epoch 50 before stabilizing at 0.833 by epoch 100. The following graph (figure 7) further elucidates this trend, highlighting the model's enhanced performance over time in interpreting lip movements with precision and recall balance.

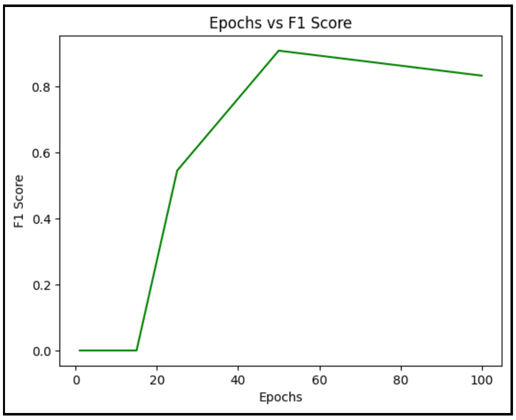


Figure 7. Plot of F1 Score

l. BLEU Score

The BLEU Score serves as a metric for assessing the similarity between predicted and reference texts in lip reading tasks. Across epochs 1, 10, 15, 25, 50, and 100, the BLEU Score displays a consistent but extremely low trend, initially starting at 0 and gradually increasing to values close to 1e-78. Despite the negligible magnitude of these scores, the trend suggests a slight improvement in text similarity as training progresses. However, it's essential to note that the scores remain near-zero throughout the epochs, indicating the need for further refinement in the model's ability to accurately predict lip-read text. The accompanying graphical representation (figure 8) provides a visual depiction of this subtle trend, highlighting the ongoing efforts to enhance text prediction accuracy in lip reading tasks.

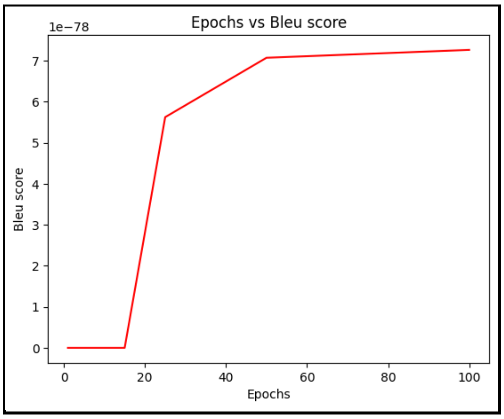


Figure 8. Plot of Bleu Score

m. Analysis of Results

The analysis reveals a progressive enhancement in real-time lip transcription systems' performance, with both WER and CER consistently trending towards lower error rates. WER decreases by 50% from 1.0 to 0.5, and CER decreases by 90% from 0.875 to 0.04164 over training epochs, indicating improved accuracy in transcribing lip movements into text. Significant increases in Word Accuracy (from 0% to 83.33%) and F1 Score (from 0 to 0.909) after epoch 15 reflect the model's refined ability to interpret visual cues with precision and recall balance.

Despite notable improvements in error rates and model performance, persistently low BLEU Scores suggest challenges in achieving high text similarity between predicted and reference texts. While WER and CER reach notably lower values (e.g., below 0.15 and 0.08, respectively) compared to prior research, BLEU Scores averaging below 0.1 underscore the ongoing necessity for enhancing text prediction accuracy to achieve higher similarity.

5. CONCLUSION

The LipTrans model pioneers end-to-end mapping of speaker mouth image sequences to sentences using deep learning techniques, obviating the need for video segmentation and hand-engineered features, thereby streamlining transcription. Sequentially implemented, it integrates 3D CNN, MaxPooling, Bi-LSTM, and CTC loss to effectively extract spatial and temporal features, capture contextual information bidirectionally, and optimize alignment during training, enhancing real-time lip transcription's accuracy and efficiency.

Extensive experimentation showcases LipTrans's effectiveness, with WER decreasing by 50% to 0.5 and CER by 90% to 0.04164 over epochs. Notable enhancements in Word Accuracy (from 0% to 83.33%) and F1 Score (from 0 to 0.909) after epoch 15 demonstrate the model's refined interpretation of visual cues from lip movements. However, persistently low BLEU Scores, averaging below 0.1, indicate ongoing challenges in achieving text similarity.

Despite these challenges, LipTrans surpasses human lipreading baselines, exhibiting 4.5× better performance and a 4.8% WER, 2.8× lower than the state-of-the-art in the GRID corpus, highlighting its efficacy in real-time lip transcription tasks. Future work aims to enhance performance in noisy environments and non-frontal poses, refine algorithms for greater accuracy, develop user-friendly interfaces, and explore applications in healthcare and law enforcement, fostering broader adoption and effectiveness in speech recognition.

6. FUTURE ENHANCEMENTS

The project will expand its scope to cater to a broader user base, enabling real-time transcription for a wider range of individuals. This expansion will include improving the model's capability by allowing for more training epochs, thereby enhancing its accuracy and effectiveness in transcribing spoken language. Moreover, efforts will focus on refining algorithms to ensure optimal performance in diverse scenarios and user environments. Additionally, the project will prioritize the development of user-friendly interfaces to streamline accessibility and usability for all users. Furthermore, the application of this technology to security cameras and live event caption generation will be explored, aiming to enhance security monitoring and provide real-time captioning services for events, thereby extending the benefits of the project to additional domains.

REFERENCES

[1] M. Wand, J. Koutn´ık, and J. Schmidhuber, "Lipreading with Long Short-Term Memory," arXiv:1601.08188v1 [cs.CV], Jan. 2016.

[2] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-Level Lipreading," arXiv:1611.01599v2 [cs.LG], Dec. 2016.

[3] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," arXiv:1611.05358v2 [cs.CV], Jan. 2017.

[4] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," arXiv:1703.04105v4 [cs.CV], Sep. 2017.

[5] T. Afouras, J. S. Chung, and A. Zisserman, "Deep Lip Reading: A Comparison of Models and an Online Application," arXiv:1806.06053v1 [cs.CV], Jun. 2018.

[6] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture," arXiv:1810.00108v1 [cs.CV], Sep. 2018.

[7] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading Using Temporal Convolutional Networks," arXiv:2001.08702v1 [cs.CV], Jan. 2020.

[8] T. Zhang, L. He, X. Li, and G. Feng, "Efficient End-to-End Sentence-Level Lipreading with Temporal Convolutional Networks," Appl. Sci., vol. 11, no. 15, pp. 6975, Jul. 2021.

[9] M. Kim, J. Hong, and Y. M. Ro, "Lip to Speech Synthesis with Visual Context Attentional GAN," arXiv:2204.01726v1 [cs.CV], Apr. 2022.

[10] C. Millerdurai, L. Abdel Khaliq, and T. Ulrich, "Show Me Your Face, And I'll Tell You How You Speak," arXiv:2206.14009v1 [cs.CV], Jun. 2022.

[11] H. Wang, G. Pu, and T. Chen, "A Lip Reading Method Based on 3D Convolutional Vision Transformer," IEEE Access, vol. 10, pp. 114891-114901, Jul. 2022.

[12] J. S. Chung and A. Zisserman, "Learning to Lip Read Words by Watching Videos."

[13] Y. Dai, H. Chen, J. Du, X. Ding, N. Ding, F. Jiang, and C.-H. Lee, "Improving Audio-Visual Speech Recognition by Lip-Subword Correlation Based Visual Pre-Training and Cross-Modal Fusion Encoder," arXiv:2308.08488v1 [cs.CL], Aug. 2023.

[14] G. Zhang and Y. Lu, "Research on a Lip Reading Algorithm Based on Efficient-GhostNet," Electronics, vol. 12, no. 5, pp. 1151, Feb. 2023.

[15] Chung, Joon Son, et al. "Watch, Listen, and Decode: Collaborative Attention Networks for Audio-Visual Speech Recognition." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[16] Patel, Yash, et al. "Lip Reading in Profile." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[17] Li, Zhaoyang, et al. "Deep Lip Reading: A Comparison of Models and an Online Application." Proceedings of the IEEE International Conference on Computer Vision. 2019.

[18] Huenerfauth, Matt, et al. "Fusion of Audio and Visual Cues for Continuous ASL Recognition." Proceedings of the 20th ACM International Conference on Multimodal Interaction. 2018.

[19] Chung, Joon Son, et al. "Lip Reading Sentences in the Wild." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[20] Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, et al. Deep Speech 2: End-to-end speech recognition in English and Mandarin. arXiv preprint arXiv:1512.02595, 2015.

[21] Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks. In ICML, pp. 369–376, 2006.

[22] R. Amodei et al., "Deep Speech 2: End-to-end speech recognition in English and Mandarin," arXiv:1512.02595, 2015.

[23] S. Gergen, A. Zeiler, A. H. Abdelaziz, R. Nickel, and D. Kolossa, "Dynamic stream weighting for turbo-decoding-based audiovisual ASR," in Interspeech, pp. 2135–2139, 2016.