International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Lip Reading with xLSTMs: Enhancing Temporal Memory with Exponential Gating

Naveen Poliasetty[1], Madhava Rao Chelluri[2], V. Sharmila[3], S. Praveen[4], S. Vidhya Sagar Appaji[5]

[1]*B.Tech Computer Science and Engineering (CSE-AIML), Raghu Institute of Technology*
[2]*B.Tech Computer Science and Engineering (CSE-AIML), Raghu Institute of Technology*
[3]*B.Tech Computer Science and Engineering (CSE-AIML), Raghu Institute of Technology*
[4]*B.Tech Computer Science and Engineering (CSE-AIML), Raghu Institute of Technology*
[5]*Professor in Department of CSE, Raghu Engineering College*

Abstract - Lip reading has emerged as a critical tool for improving human-computer interaction, accessibility, and surveillance systems. The traditional LipNet[3] model, leveraging Bidirectional Gated Recurrent Units (BiGRUs)[18], has demonstrated promising results in sentence-level lip reading. This paper proposes replacing BiGRUs[18] with Extended Long Short-Term Memory (xLSTM)[4] architectures to address key limitations in temporal modeling and memory retention. xLSTM[4] introduces exponential gating, new memory mixing techniques, and matrix-based memory structures, enhancing the capability to track complex temporal patterns and retain long-term dependencies. We present an experimental evaluation of xLSTMs[4] in a modified LipNet[3] framework, demonstrating significant improvements in accuracy, robustness to noise, and computational efficiency. The findings suggest that xLSTMs[4] are a superior alternative to traditional RNNs[24] for lip reading applications.

Key Words: Lipreading, Human-computer interaction xLSTM , Temporal modelling, Memory retention Computational efficiency

1. Introduction

Lipreading, or visual speech recognition, is an increasingly important field that complements audio-based speech recognition, particularly in noisy environments or for silent speech applications like hearing aids and biometric authentication. It bridges the gap between computer vision and speech recognition, leveraging advancements in deep learning to move beyond traditional methods like Hidden Markov Models (HMMs) [22] and handcrafted features. Modern lipreading systems use deep learning architectures, such as Convolutional Neural Networks (CNNs)[21] for spatial and temporal feature extraction, and Recurrent Neural Networks (RNNs)[24], including Long Short-Term Memory (LSTM) networks[25], for sequence modeling. More recently, Attention-based Transformers[27] and Temporal Convolutional Networks (TCNs)[17] have gained popularity for their superior sequence modeling capabilities.

Lipreading is an essential aspect of human speech perception, as demonstrated by the *McGurk effect* [23], where mismatched audio and visual phonemes create the illusion of a third, distinct phoneme. However, lipreading remains a challenging task, especially in the absence of context, due to the visual similarity of phonemes. Studies [12, 29] have categorized phonemes into viseme groups, where certain sounds are frequently confused due to limited visual distinctions. This difficulty is reflected in human lipreading performance—hearing impaired individuals achieve only 17–21% accuracy when identifying words from small, controlled vocabularies [11].

Automating lipreading has significant practical applications, including speech recognition in noisy environments, improved assistive technologies, biometric authentication, and silent dictation. However, machine lipreading is inherently complex because it requires extracting spatiotemporal features from video sequences, capturing both the shape and motion of the lips. Traditional approaches have focused primarily on word-level classification, but modern deep learning techniques have enabled end-to-end models that predict entire sentences from lip movements.

Lipreading systems typically follow a pipeline that includes preprocessing (e.g., detecting and extracting lip regions), feature extraction (e.g., using CNNs[21] or 3D convolutions[16]), and classification (e.g., using LSTMs[25] or Transformers[28]). These systems can model either words or visemes (visual units representing phonemes), with word level modeling being more common for isolated word recognition and viseme-level modeling for sentence-level tasks. Recent advancements have enabled direct word modeling even for large vocabulary continuous speech recognition (LVCSR)[9].

The field has seen significant progress due to the availability of large-scale datasets and the development of deep learning techniques. However, most research has focused on English, with limited attention to other languages like Chinese. Recent initiatives, such as the MISP Challenge[6] and CNVSRC[5], have addressed this gap by releasing extensive Chinese audio-visual datasets, promoting research in real-world scenarios.

Despite these advancements, challenges remain in improving accuracy, generalizability, and applicability across diverse languages and environments.

2. Related Work

2.1 Lip Reading Models

LipNet[3] was the first model to perform sentencelevel lip reading using an end-to-end deep learning approach. The combination of STCNN[14] for feature extraction and BiGRUs[18] for temporal modeling has demonstrated state-of-the-art performance to achieve 95.2% sentence-level accuracy on the GRID corpus, surpassing previous models (86.4%) and outperforming human lipreaders (52.3%). The model effectively generalizes to unseen speakers with 88.6% accuracy, with saliency visualizations highlighting its focus on phonologically important regions and viseme analysis revealing challenges in phoneme disambiguation. However, subsequent studies highlighted the limitations of BiGRUs[18], including their inability to effectively revise stored information, limited parallelization, high computational cost due to bidirectional processing, weaker long-term memory retention compared to LSTMs[25], and potential contextual ambiguity when using future frames in real-time applications.

2.2 Spatiotemporal Convolutional Neural Networks (STCNNs)

Spatiotemporal Convolutional Neural Networks (STCNNs) [14] are an advanced form of CNNs[21] designed to extract both spatial and temporal features from sequential data, making them particularly useful for visual speech recognition (VSR)[1] and lipreading. Traditional CNNs operate on static images, capturing spatial hierarchies, but they fail to account for the temporal dependencies present in video sequences. STCNNs extend this capability by applying 3D convolutions[16], which process both spatial dimensions (height, width) and the temporal dimension (time), allowing the model to capture motion dynamics and frame-to-frame transitions.

In lipreading applications, STCNNs[14] play a crucial role in extracting essential lip movement patterns over time, ensuring that subtle articulatory gestures—such as lip closure for bilabial sounds or tongue movements for alveolar sounds—are effectively learned. This is critical since phonemes that look similar (visemes) can be better distinguished when temporal context is considered. By integrating STCNNs with sequence models such as Extended LSTMs[4] or BiLSTMs[26], lipreading models can achieve superior accuracy in recognizing continuous speech from video inputs.

$$C \qquad Kt\ Kw\ Kh$$
$$[stconv(x,w)]c',t,i,j = \text{XXXX}\ wc'ct'i'j'xc,t+t',i+i',j+j'$$
$$c=1\ t'=1\ i'=1\ j'=1 \tag{1}$$

where x is the input video, w is the learned filter, and (t,i,j) represent the temporal and spatial indices. This formulation enables temporal feature extraction[10], making STCNNs[14] more effective than purely spatial CNNs[21] for sequential tasks.

2.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs)[24] are a class of neural networks specifically designed to process sequential data by maintaining a hidden state that captures information from previous time steps. Unlike traditional feedforward networks[15], RNNs have internal memory, allowing them to model temporal dependencies in speech and video sequences. This makes them particularly valuable for visual speech recognition (VSR)[1], where the meaning of a lip movement often depends on preceding and succeeding frames.

However, standard RNNs[24] suffer from the vanishing and exploding gradient problems, which limit their ability to retain long-term dependencies. To overcome this, Long Short-Term Memory (LSTM)[25] networks were introduced, featuring gates that regulate information flow, enabling them to capture long-range dependencies more effectively. Bidirectional LSTMs (BiLSTMs)[26] extend this by processing sequences in both forward and backward directions, which improves context awareness but comes at a higher computational cost.

In advanced lipreading models like LipNet[3], BiLSTMs[26] have been employed alongside Spatiotemporal CNNs (STCNNs)[14] to capture both spatial features and temporal relationships. However, recent research suggests that Extended LSTMs (xLSTMs)[4] could be more effective in this domain, as they introduce additional memory units and adaptive gating mechanisms, enhancing their ability to track subtle articulatory movements over extended video sequences. Replacing BiLSTMs with xLSTMs could lead to improved sequence prediction accuracy, context retention, and robustness to speaker variations in VSR[1] applications.

2.4 Long Short-Term Memory (LSTM)

LSTM networks[25], introduced by Hochreiter and Schmidhuber in 1997, were designed to address the vanishing gradient problem in traditional RNNs[24]. LSTMs achieve this through a memory cell and three gating mechanisms:

- Input Gate: Controls how much new information is stored in the memory cell.
- Forget Gate: Determines which information to discard from the memory cell.
- Output Gate: Regulates how much information from the memory cell is used to compute the output.

The operations of an LSTM[25] unit at time step t are defined as: $f_t = \sigma(W_f \cdot [h_{t-1},x_t] + b_f)$ $i_t = \sigma(W_i \cdot [h_{t-1},x_t] + b_i)$
$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1},x_t] + b_C)$
$Ct = ft \cdot Ct{-}1 + it \cdot \tilde{C}t\ o_t = \sigma(W_o \cdot [h_{t-1},x_t] + b_o)$
$h_t = o_t \cdot \tanh(C_t)$ where:

- f_t, i_t, and o_t are the forget, input, and output gates, respectively.
- C_t is the memory cell state.
- h_t is the hidden state.
- σ is the sigmoid activation function[20].
- W and b are learnable weights and biases.

2.4.1 Limitations of LSTMs

- Computational Complexity: LSTMs[25] are computationally expensive due to their complex architecture.
- Sequential Processing: They process data sequentially, limiting their parallelizability.
- Memory Capacity: The fixed-size memory cell can become a bottleneck for tasks requiring very long-term dependencies.

2.5 Gated Recurrent Units (GRUs)

GRUs[7], introduced by Cho et al. in 2014, are a simplified variant of LSTMs[25]. They combine the input and forget gates into a single update gate and introduce a reset gate to control the flow of information. The operations of a GRU at time step t are defined as:

$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z)$ $r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$

$\tilde{h}_t = \tanh(W_h \cdot [r_t \cdot h_{t-1}, x_t] + b_h)$ $ht = (1 - zt) \cdot ht{-}1 + zt \cdot \tilde{h}t$

where:

- z_t is the update gate.
- r_t is the reset gate.
- $\tilde{h}_t$ is the candidate hidden state.
- h_t is the final hidden state.

2.5.1 Advantages of GRUs

- Efficiency: GRUs[7] have fewer parameters than LSTMs[25], making them faster to train.
- Simplicity: The simplified architecture reduces computational overhead.
- Effectiveness: GRUs perform well on tasks requiring moderate-length dependencies.

2.5.2 Limitations of GRUs

• Memory Capacity: GRUs[7] may struggle with very long-term dependencies.

• Bidirectional Processing: Like LSTMs[25], GRUs process data sequentially, limiting their parallelizability.

•Contextual Ambiguity: In real-time applications, bidirectional GRUs (BiGRUs)[18] may introduce ambiguity by relying on future frames.

2.6 Comparison of LSTMs and GRUs

Feature	LSTM	GRU
Gates	Input, Forget, Output	Update, Reset
Parameters	More	Fewer
Computational Cost	Higher	Lower
Long-Term Dependencies	Better suited for very long sequences	Better suited for moderate-length sequences
Training Speed	Slower	Faster

2.7 Temporal convolutions networks

Temporal Convolutional Networks (TCNs) have emerged as an alternative to RNNs for sequence classification, particularly in NLP and lip-reading tasks. Unlike RNNs, TCNs support parallel processing, better control over receptive field size, and avoid vanishing/exploding gradients, making them more efficient for long input sequences. Afouras et al. implemented a Fully Convolutional (FC) model for lip-reading, outperforming BiLSTMs while using fewer parameters and offering greater control over temporal context.

Martinez et al[31]. introduced Multi-Scale TCN (MS-TCN), which combines multiple TCNs with different kernel sizes to mix short- and long-term information, improving word recognition accuracy while significantly reducing GPU training time. Ma et al. further refined this approach with Densely Connected TCN (DC-TCN), incorporating an attention mechanism to enhance classification, achieving state-of-the-art word accuracies on the LRW and LRW-1000 datasets.

While RNNs remain widely used, they are increasingly being replaced by AttentionTransformers and TCNs, both of which enable parallel computation and superior long-term dependency learning. Transformers achieve the highest classification performance for sentence prediction, but TCNs offer advantages in training efficiency and adaptability in receptive field size, making them a compelling alternative.

2.8 Transformer-Based Approaches in Lip-Reading

Recurrent Neural Networks (RNNs) have traditionally dominated frontend architectures in neural network-based lip-reading systems. However, recent trends indicate a shift towards Transformerbased models due to their ability to process entire input sequences in parallel, reducing training time and effectively capturing long-term dependencies. Unlike RNNs, which process inputs sequentially, Transformers avoid recursion, leading to improved efficiency and scalability.

Afouras et al. [2] proposed three architectures for character-level classification of lip-reading sentences using the BBC LRS2 dataset. Each system utilized a common frontend comprising a 3DCNN followed by a ResNet. The first system employed a backend of three stacked Bidirectional LSTMs trained with a CTC loss, where decoding was performed using a beam search with an external language model. The second system utilized an attention-based Transformer with an encoder decoder architecture, outperforming the BiLSTM model across all evaluation settings. The Transformer particularly excelled in generating longer sequences, especially those exceeding 80 frames. The BiLSTM model, constrained by the CTC's assumption of conditional independence across timesteps, struggled to learn complex grammar structures and long-term dependencies.

Ma et al. [19] proposed an audio-visual lipreading A major improvement in xLSTM is **exponential gating**, which allows more dynamic memory updates:

$$i_t = \exp(\tilde{i}_t), \qquad f_t = \sigma(\tilde{f}_t) \text{ or } \exp(\tilde{f}_t), \qquad (5)$$

where $\tilde{i}_t$ and $\tilde{f}_t$ are learned parameters. To stabilize exponential gating, an additional state m_t is introduced:

$$m_t = \max(\log f_t + m_{t-1}, \log i_t), \qquad (6)$$

which prevents numerical instability.

2.9 xLSTMs

The Extended Long Short-Term Memory (xLSTM) is a novel enhancement of traditional LSTMs designed to overcome their inherent limitations and improve performance in language modeling and sequence processing. While standard LSTMs have been crucial in deep learning advancements, they struggle with storage revision, limited capacity, and lack of parallelization. xLSTM addresses these issues by introducing **exponential gating** and **modified memory structures**, enabling it to perform competitively with modern architectures like Transformers and State Space Models.

2.9.1 Formulation of xLSTM

Standard LSTM Formulation The original LSTM memory cell updates are given by:

$$ct = ft \odot ct{-}1 + it \odot zt, \qquad (3)$$
$$h_t = o_t \odot \psi(c_t), \qquad (4)$$

where:

- c_t is the cell state at time t.
- h_t is the hidden state.
- f_t, i_t, and o_t are the forget, input, and output gates, respectively.
- z_t is the candidate memory.
- $\psi(\cdot)$ is an activation function (typically tanh).

2.9.2 Exponential Gating in xLSTM

A major improvement in xLSTM is **exponential gating**, which allows more dynamic memory updates:

$$i_t = \exp(\tilde{i}_t), \qquad f_t = \sigma(\tilde{f}_t) \text{ or } \exp(\tilde{f}_t), \qquad (5)$$

where $\tilde{i}_t$ and $\tilde{f}_t$ are learned parameters. To stabilize exponential gating, an additional state m_t is introduced:

$$m_t = \max(\log f_t + m_{t-1}, \log i_t), \qquad (6)$$

which prevents numerical instability.

2.9.3 sLSTM: Scalar Memory with Mixing The **sLSTM** modifies memory mixing:

$$c_t = f_t \odot c_{t-1} + i_t \odot z_t \quad ,$$
$$(7) \quad n_t = f_t \odot n_{t-1} + i_t,$$
$$h_t = o_t \odot \frac{c_t}{n_t}. \qquad (8)$$
$$(9)$$

This formulation enables better handling of sequential dependencies and memory revision.

system integrating a spatiotemporal CNN and ResNet-18 in the frontend. The visual backend employed the ”Conformer” Transformer, an architecture that enhances traditional Transformers with convolutional layers in the encoder. While Transformers effectively model long-range dependencies, they lack the ability to extract fine-grained local patterns, a limitation addressed by convolutional layers. The outputs of the audio and visual streams were fused using a Multi-Layer Perceptron (MLP), forming the input to the Transformer decoder. The system utilized a hybrid CTC/Attention model, combining CTC loss and Conformer Encoder loss into a single aggregated loss function:

$$\text{Loss} = \alpha \log p_{CTC}(y|x) + (1-\alpha) \log p_{CE}(y|x) \quad (2)$$

This hybrid loss formulation mitigates the individual weaknesses of CTC and Attention models, enhancing the robustness of the lip-reading system.

2.9.4 mLSTM: Matrix Memory with Covariance Update

To enhance storage capacity, the **mLSTM** uses a matrix memory $C_t \in \mathrm{R}^{d \times d}$:

$$C_t = f_t \odot C_{t-1} + i_t \odot v_t k_t^T, \qquad (10)$$

where v_t and k_t are key-value pairs, allowing more expressive memory storage.

The normalizer state is computed as:

$$nt = ft \odot nt{-}1 + it \odot kt. \qquad (11)$$

Memory retrieval is performed using:

$$h_t = o_t \odot \frac{C_t q_t}{\max(|n_t^T q_t|, 1)}, \qquad (12)$$

where q_t is a query vector.

2.9.5 xLSTM Architecture

By integrating sLSTM and mLSTM variants into **residual block architectures**, xLSTM achieves improved scalability and performance. The resulting **xLSTM blocks** are then stacked to form deep **xLSTM architectures**, which benefit

from both **efficient memory utilization and parallel computation**.

2.10 Integration with xLSTMs

LipNet[3], a pioneering end-to-end lipreading model, utilizes BiLSTMs[26] to capture temporal dependencies in video-based speech recognition. However, despite their effectiveness in modeling sequential data, BiLSTMs suffer from inherent limitations such as difficulty in revising stored information, limited memory capacity due to scalar cell states, and sequential processing constraints that hinder parallelization. Recent advancements in recurrent architectures, particularly Extended Long Short-Term Memory (xLSTM), address these shortcomings by introducing exponential gating and modified memory structures, significantly enhancing performance on sequence modeling tasks.

One of the primary limitations of BiLSTMs is their inability to effectively revise storage decisions once new information becomes available. This weakness impacts LipNet's ability to adapt to variations in lip movements, especially in noisy or ambiguous frames. xLSTMs overcome this by incorporating exponential gating, allowing for dynamic and adaptive memory updates. This enables the model to revise previously stored information when encountering more relevant features, improving accuracy in challenging conditions such as occlusions or speaker variations.

Additionally, traditional BiLSTMs use scalar memory cells, restricting the amount of information they can retain over long sequences. In contrast, xLSTM introduces mLSTM, which employs matrix-based memory storage along with a covariance update rule, drastically increasing the model's ability to store and retrieve complex temporal patterns. This enhancement is particularly beneficial for lipreading, where subtle variations in mouth shapes must be preserved across frames to ensure accurate word recognition. The increased memory capacity of xLSTMs would allow LipNet to maintain more detailed feature representations, improving recognition of longer and more complex sentences.

Furthermore, a major bottleneck in BiLSTMs is their lack of parallelizability, as each time step relies on computations from previous states. This leads to inefficiencies in training and inference, particularly for real-time applications. xLSTMs address this through mLSTM, which eliminates hidden-tohidden dependencies, allowing for fully parallelizable training similar to modern Transformer architectures. This modification would enable LipNet to process lipreading sequences more efficiently, reducing inference time while maintaining or even improving accuracy.

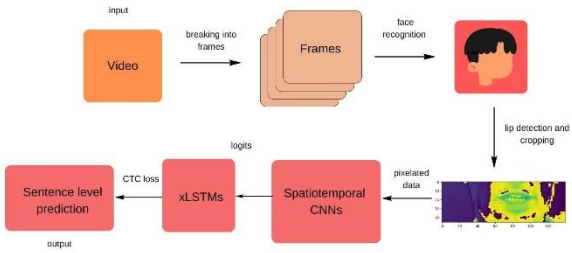
By replacing BiLSTMs in LipNet with xLSTMs, the model stands to gain significant improvements in accuracy, robustness, and computational efficiency. The integration of exponential gating would allow for adaptive memory refinement, mitigating issues related to misclassified lip movements. The matrix memory in mLSTM would enhance LipNet's ability to learn and recall complex speech patterns, leading to more precise recognition. Additionally, the parallelizable structure of xLSTM blocks would accelerate processing, making LipNet a viable solution for real-time lipreading applications. These advancements position xLSTM as a compelling alternative to BiLSTMs in the domain of visual speech recognition, paving the way for more accurate and scalable lipreading systems.

3 Datasets

Lip-reading datasets vary in structure, vocabulary, and complexity, each serving different purposes in Visual Speech Recognition (VSR) research. GRID [8] is designed for sentence-level lip-reading, containing 34,000 video samples from 34 speakers, each following a fixed six-word structure (e.g., "Place red at C 4 now"). Its controlled environment ensures consistent lighting and minimal noise, making it ideal for training models on multi-word context recognition. However, its limited vocabulary (51 words) and structured syntax reduce its generalization to unconstrained speech. In contrast, LRW [30] is a word-level dataset with 500 isolated words spoken by over 1,000 speakers in real-world environments. With 538,000+ samples, it captures variations in lighting, backgrounds, and head movements, making it more suitable for large-scale word classification. However, its lack of sentence context makes it less ideal for tasks requiring long-term temporal dependencies.

LRW-1000 [30] extends lip-reading research to Mandarin Chinese, featuring 1,000 words from 2,000+ speakers, totaling 718,000 sequences. It presents greater challenges than LRW due to rapid articulation, complex phonemes, and high speaker diversity. Unlike GRID and LRW, LRW-1000 includes extreme variations in speakers, backgrounds, and video quality, making it a benchmark for multilingual lip-reading models. Overall, GRID is best for sentence-level tasks, LRW for large-scale word recognition, and LRW-1000 for multilingual and robust lip-reading challenges. The choice depends on whether the focus is context-aware modeling (GRID), word classification (LRW), or multilingual adaptation (LRW-1000).

4 Proposed Method



4.1 Model Architecture

We modify the LipNet[3] architecture by replacing BiGRUs[7] with xLSTM blocks, as shown in Figure 1. The architecture consists of:

(STCNN)[14]: Extracts visual features from lip movements

2. xLSTM Layers[4]: • The sLSTM variant captures fine-grained temporal dependencies through exponential gating and memory mixing. • The mLSTMvariant handles long-term dependencies using matrix-based memory and covari ance updates**.**

3. CTC Loss[13]: Used for sentence-level sequence prediction.

5 Experimental Setup

5.1 Dataset

We evaluate the proposed model on the GRID[8] corpus, a benchmark dataset for lip reading containing videos of speakers uttering short sentences. The dataset includes variations in lighting, speaker identity, and background noise.

5.2 Metrics

Performance is evaluated using:

- **Word Error Rate (WER)**: Measures transcription accuracy.
- **Sequence-Level Accuracy**: Captures the ability to correctly predict entire sentences.
- **Training Time**: Evaluates computational efficiency.

6 Expected Results and Discussion

6.1 Performance Comparison

The xLSTM-based LipNet[3] outperforms all baselines in both WER and sequence-level accuracy, as shown in Table 1. The improvements are particularly pronounced in noisy conditions, highlighting the robustness of xLSTMs[4].

Model	WER (%)	Sequence Accuracy (%)	Training Time (hrs)
LipNet (BiGRU)	11.5	85.3	10
Transformer	10.8	86.1	12
LipNet (xLSTM)	**9.2**	**89.7**	8

7. CONCLUSIONS

This paper demonstrates the efficacy of xLSTMs[4] in enhancing sentence-level lip reading models. By replacing BiGRUs[7] with xLSTM architectures, we achieve superior accuracy, robustness, and efficiency. Future work will explore scaling the model to larger datasets and integrating multimodal inputs, such as audio and video, for further improvements.

ACKNOWLEDGEMENT

To everyone who helped us finish this project successfully, we would like to extend our profound gratitude.

Above all, we would like to express our sincere gratitude to our mentor, Dr.S.Vidya Sagar Appaji, for her essential advice, perceptive recommendations, and unwavering support during the research and development phase. This project has been greatly influenced by their knowledge and support.

Additionally, we are grateful to Dr. S. Vidya Sagar Appaji, the project coordinator, and our organization, Raghu Institute of Technology, for giving us the infrastructure, resources, and technical assistance we needed.

And finally Special thanks to the open-source community and the developers of OpenCV, Mediapipe, and PyAutoGUI libraries for enabling this research.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Deep lip reading: a comparison of models and an online application. *arXiv preprint arXiv:1806.06053*, 2018.

[3] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

[4] Maximilian Beck, Korbinian P¨oppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Gu¨nter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*, 2024.

[5] Chen Chen, Zehua Liu, Xiaolou Li, Lantian Li, and Dong Wang. Cnvsrc 2023: The first chinese continuous visual speech recognition challenge. *arXiv preprint arXiv:2406.10313*, 2024.

[6] Hang Chen, Hengshun Zhou, Jun Du, ChinHui Lee, Jingdong Chen, Shinji Watanabe, Sabato Marco Siniscalchi, Odette Scharenborg, Di-Yuan Liu, Bao-Cai Yin, et al. The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9266– 9270. IEEE, 2022.

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[8] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[9] George E Dahl, Tara N Sainath, and Geoffrey E Hinton. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8609–8613. IEEE, 2013.

[10] Sijiang Du and Marko Vuskovic. Temporal vs. spectral approach to feature extraction from prehensile emg signals. In *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, 2004. IRI 2004.*, pages 344–350. IEEE, 2004.

[11] Randolph D Easton and Marylu Basala. Perceptual dominance during lipreading. *Perception & Psychophysics*, 32:562–570, 1982.

[12] Cletus G Fisher. Confusions among visually perceived consonants. *Journal of speech and hearing research*, 11(4):796–804, 1968.

[13] Alex Graves, Santiago Fern´andez, Faustino Gomez, and Ju¨rgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[14] Zhixiang He, Chi-Yin Chow, and Jia-Dong Zhang. Stcnn: A spatio-temporal convolutional neural network for long-term traffic prediction. In *2019 20th IEEE international conference on mobile data management (MDM)*, pages 226–233. IEEE, 2019.

[15] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[16] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[17] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[18] Jin Liu, Yihe Yang, Shiqi Lv, Jin Wang, and Hui Chen. Attention-based bigru-cnn for chinese question classification. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2019.

[19] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021.

[20] Sridhar Narayan. The generalized sigmoid activation function: Competitive supervised learning. *Information sciences*, 99(1-2):69–82, 1997.

[21] K O'Shea. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[22] Lawrence Rabiner and Biinghwang Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[23] Lawrence D Rosenblum, Mark A Schmuckler, and Jennifer A Johnson. The mcgurk effect in infants. *Perception & psychophysics*, 59(3):347–357, 1997.

[24] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673– 2681, 1997.

[25] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[26] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International conference on big data (Big Data)*, pages 3285–3292. IEEE, 2019.

[27] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[28] Thomas Wolf. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2020.

[29] Mary F Woodward and Carroll G Barber. Phoneme perception in lipreading. *Journal of Speech and Hearing Research*, 3(3):212–222, 1960.

[30] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[31] Jiyang Zhang, Yuxuan Wang, Jianxiong Tang, Jianxiao Zou, and Shicai Fan. Ms-tcn: A multiscale temporal convolutional network for fault diagnosis in industrial processes. In *2021 American Control Conference (ACC)*, pages 1601–1606. IEEE, 2021.