*Grenze International Journal of Engineering and Technology, June Issue*

# CurriculumLip: Progressive Visual Speech Recognition via Self-Supervised and Supervised Curriculum Learning

Harshita[1], SangeetaKumari[2] and AseemMadaan[3]
[1-2]Chandigarh University, Department of Mathematics, Gharuan, Mohali, India
Email: chhabraharshita52@gmail.com, sangwansangeeta.ss@gmail.com
[3]Manav Rachna University, Department of Engineering and Technology, Faridabad, Haryana, India
Email: aseemmadaan9@gmail.com

***Abstract*— Despite advancements in deep learning, visual speech recognition (lipreading) is still challenging because of viseme confusion, variability in how speakers articulate themselves and the tiny amount of labelled training data available. The methodologies used to date typical-ly rely either on fully supervised training or on disjoint pretrial-fine-tune training pipe-lines, and therefore ignore progression of difficulty across structured utterances. This paper presents a novel three-stage curriculum-learning curriculum (CurriculumLip), which progressively incorporates supervised CTC training with each of the two visual self-supervised pretext tasks that comprise the curriculum: predicting temporal ordering of lip triplets (predicted in the future) and past predictive future features. The architecture employs a 3D CNN spatiotemporal front-end, followed by a multi-resolution Temporal Convolutional Network (TCN) encoder, and captures lip motion over local, medium, and long temporal lengths through parallel dilated convolution processes. CurriculumLip was evaluated on the GRID corpus with a word accuracy (WA) of 83.7%—an absolute difference of 5.5%—and an additional relative reduction of 35.9% in WER provided labels for the purpose of generating out-of-distribution sequence length modelling. Additionally, ablation studies support the contribution of the multiple stages of the curriculum as well as other tests that highlight reliability, robustness and ability to generalize (7.9% vs 15.6% accuracy loss) using different amounts of frames available at dropouts.**

***Index Terms*— Visual Speech Recognition, Lipreading, Curriculum Learning, Self-Supervised Learning, Temporal Convolutional Networks, GRID Corpus.**

## I. INTRODUCTION

Lipreading is the process of determining what a person is saying by observing how the person's mouth looks as they speak. It can also be used in several applications, including assistive listening devices and audio-visual speech recognition to communicate without having to share any spoken information. [1][7] Lipreading is challenging for several reasons: (1) There is viseme confusion, as many different phonemes are tinted the same visually; (2) Individual variability in how people articulate sounds can produce significant variability in their visual mouths independent of the sound; and (3) Creating accurate ML models requires access to large amounts of properly labelled audio-visual data, which is usually prohibitively expensive to obtain and accurately annotate. [2][3]

*Grenze ID: 01.GIJET.12.2.86*

The emergence of Self-Supervised Learning (SSL) methods such as wav2vec,2.0, HuBERT, and AV- HuBERT has changed the landscape of speech/audio- visual representation training by demonstrating that vast volumes of availa-ble audio-visual, speaking data can be used in the absence of labelled data to permit training without labelled input. Although visual-only SSL methods have not been researched as intensely as more traditional methods or as thoroughly as have most audio/visual training methods, valuable insights into visual-based lip reading or creating a comprehen-sive curriculum of training materials have yet to be explored thoroughly [2] [12]. However, applying a systematic approach to visual speech reconstruction using Generative Semi-Supervised Learning (SSL) is a new idea in this field.

### *A. Contribution*

This article provides three main contributions to the current knowledge base:

- A three-stage curriculum framework - a completely new approach to training that incrementally combines self-supervised pre- training of lip movements (temporal order prediction and future feature prediction) and will advance to ensure self-supervised, human-annotated speech is learned as a result of the curriculum framework.
- Multiresolution Temporal Convolutional Network (TCN) Encoder - a new architecture that captures lip movements over multiple temporal resolutions (local, medium and long-range) using parallel dilated convolutions and provides improvements over single-resolution temporal models.
- Self-supervised visual tasks designed to represent lip movements - developed two pretext tasks (temporal order prediction and future feature prediction) which represent meaningful representations of lip movements that do not require transcripts and/or audio.
- Comprehensive empirical validation - demonstrate on GRID, that CurriculumLip produces a word accuracy of 83.7%, a 5.5% absolute improvement compared to the supervised baseline, as well as excellent performance when using low-label indexing (+35.9% relative WER based on 10% labeled data) and improved robustness to temporal perturbations and other distribution lengths that are out of calendar.

## II. LITERATURE SURVEY

The recent shift from traditional HMM lip reading models to deep learning has resulted in significant advancements in lip reading performance. The first deep-learned lip-tracking methods used 2D convolutional and long short-term memory (LSTM) networks used in conjunction with each other for learning temporal lip-reading dynamics. Recent developments in the use of 3D CNNs to learn both spatial and temporal information jointly will achieve state-of-the-art results [5][8] on the GRID and LRW datasets by employing models that can take advantage of the spatial and temporal nature of video. Recurrent architectures (RNNs) are outperformed by Temporal Convolutional Networks (TCNs) when compared both in terms of accuracy and computational efficiency, as TCNs utilize dilated 1D Convo-lutions applied over sequences of video frames [5][6]. The use of Multiple Dilation TCNs enables TCNs to take ad-vantage of the multiple patterns present in speech across varying temporal scales through their ability to use multiple dilation rates in parallel [6] [17].

Self-supervised way to obtain robust speech representations from minimal amounts of labeled data is to apply masked prediction to the raw audio representation of the speech signal, as is demonstrated in the AV2vec 2.0 paper [10]. The HuBERT model improves upon the previous generation of masked prediction models, including AV2vec 2.0, by iteratively switching between prediction and clustering objectives for training the model [11]. In addition to audio, AV-HuBERT extends the idea of self-supervised learning to the audio-visual setting by applying joint prediction to the masked audio and masked visual representations, allowing for dramatic improvements in the recognition and translation of spoken words within mixed audio and video signals [12].However, self-supervised visual learning approaches to lipreading are less developed than self- supervised audio approaches, and the ap-proaches that do exist using visual-only representations are typically done so as distinct separate pretraining steps rather than integrated into the training process [2].

Within the field of curriculum learning (CL), the origins of CL are tied to how humans learn, and several applications of CL exist to date for both computer vision (CV; object detection and segmentation, etc.) and natural language pro-cessing (NLP; machine translation, parsing, etc.) and to reinforcement learning models [14]. There are many reasons why it is beneficial to have a curriculum that is easy to understand, both in terms of length of speeches, and the vol-ume at which they are delivered. [18] However, comprehensive curriculums that utilise SSL pre-training, difficulty-aware sampling and adaptive loss weighting, especially with respect to visual speech, remain unknown. [14]

Although each of these components (deep learning for lipreading, SSL-based learning for speech, and curriculum learning) is fully developed, their integration into a visual-only lipreading system is a new contribution to the com-munity. CurriculumLip helps to fill this gap by providing a unified, coherent framework for the three-stage design; It uses unlabeled video with visual SSL for training, introduces a curriculum of difficulty-based sampling of training samples, provides the ability to switch from the SSL- based training of the previous stage to a supervised trainer dur-ing training by using an adaptive loss schedule, provides multi-scale temporal modelling for video. Overall, Curricu-lumLip provides a model that outperforms baseline performance, especially in situations where there is a low num-ber of labels (i.e. low label scenarios) which is highly relevant for real- world applications of lipreading technology.

## III. METHODOLOGY

### *A. Problem Formulation*

The input to the model is comprised of a sequence of video frames $V = \{v_t\}^T$ of the mouth region, where each frame consists of a 96×96 grayscale image. The output is a character sequence $y = \{y_i\}^L$ associated with the video frames. Given the fact that there is no alignment between the video frames and characters, the loss function used is the CTC loss, which is defined as follows:

$$\mathscr{L}_{\text{CTC}} = -\log P(y|V) \tag{1}$$

In this expression, $B_{\text{CTC}}$ represents the CTC merge operation, which merges labels that occur multiple times and removes any instances of the special CTC blank token.

### *B. Architecture*

- 3D CNN Front-End: The front end of the CurriculumLip model extracts temporal- spatial features from the input video sequence by sequentially using three 3D convolutional blocks, which down sample the spatial resolution while keeping the temporal information intact.

$$F = \text{GlobalAvgPool}\,(\text{Conv3D}_3\,(\text{Conv3D}_2(\text{Conv3D}_1(V)))) \tag{2}$$

In each block, there are operations of 3D convolution (with kernels of size 3×5×5 or 3×3×3), batch normalization, ReLU activation and spatial pooling. The output of the front-end consists of frame-level feature vectors (of length 128), with being the temporal dimension.

- Multi-Resolution Temporal Convolutional Network: The temporal encoder is comprised of N=4 stacked multi-resolution temporal convolutional networks (TCN). Each TCN block accepts the input $H_{i-1} \in \mathbb{R}^{T\times D}$ as its input, splitting it into three paths (one for each dilation rate d ∈ {1, 2, 4}, with dilation rate 1 representing no dilation). Thus, the output $H_i$ of TCN block i can be expressed as:

$$H_i = H_{i-1} + \sigma(\text{Conv}_{1\times1}(\text{Concat}[\text{TCN}_d(H_{i-1}): d \in \{1,2,4\}])) \tag{3}$$

where $\text{TCN}_d$ uses a 1D convolution of length 3 and dilation d. The ReLU function σ is applied to nonlinearly transform the sum of all three branches. The residual connection adds $H_{i-1}$ to the processed output from each branch so that information about the previous state is preserved. Stacking four such TCN blocks allows for the construction of a deep encoder with exponentially increasing receptive fields from 3 frames in the first block to 27+ frames in the last block, thereby enabling it to model viseme transitions over short distances as well as longer distance coarticulation effects.

- Task-Specific Prediction Heads: To support the different learning objectives of CurriculumLip, three prediction heads exist:
  - Head for Temporal Order Prediction: The middle frame's original position is forecast by a 3-way classifier using frame triplets $\{h_{t-1}, h_t, h_{t+1}\}$ which have been shuffled randomly. "Order" Loss encourages the encoder to determine the normal sequence of lip movements based on its natural sequence rather than by using labelled data.
  - Prediction Head for Future Features: By employing a linear layer to generate a prediction of $h_{t+k}$ based on $h_t$ (where k = 3) and training on Mean Squared Error, this task encourages the encoder to learn longer time- dependent patterns:

$$\mathscr{L}_{\text{future}} = \mathbb{E}[\|PredHead(h_t) - h_{t+k}\|^2] \tag{4}$$

- CTC Classification Head: Using a linear layer, $HN$ is converted to a logits output for each character in the vocabulary (40 classes include a-z, 0-9, space, apostrophe and CTC blank). CTC Loss is used to optimize the model so that sequence-to-sequence learning can be performed without frame-level alignment.

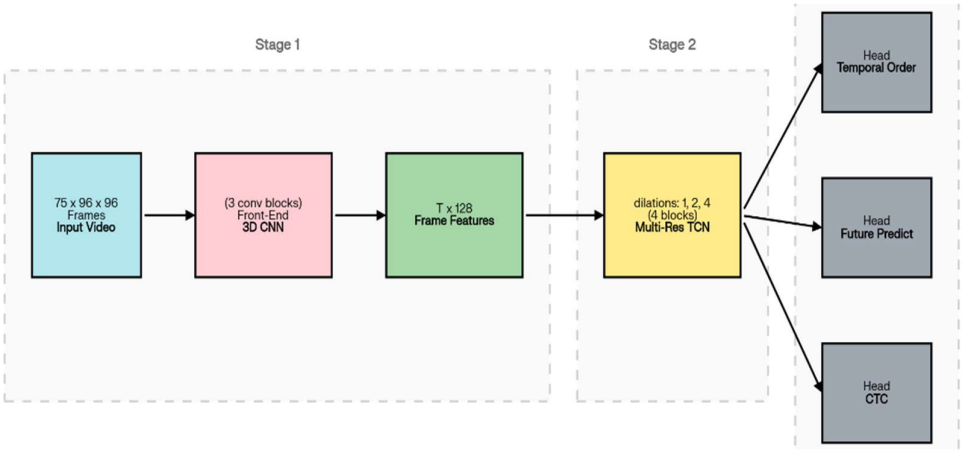


Fig.1. Architecture of Model

*C. Three-Stage Curriculum Training*

- Stage 1: Temporal coherence in the local context (Easy SSL):
  - Curriculum data: Short clips (5–10 frames) from each of the training videos, consisting of clips containing regions of very high optical flow (i.e., rapid motion).
  - Goal: Training only the temporal order prediction head: $\mathscr{L}_{\text{Stage1}} = \mathscr{L}_{\text{order}}$
  - Time frame: 4 epochs at a learning rate of 1×10.
  - Rationale: The encoders learn globally ordered sequences of lip motions, while no supervisory signals are necessary to learn the global orderings. The model learns to discern easy lip motions through short clips of simple, high-speed lip motions.
- Stage 2: Temporal predictions at the sequence level (Medium SSL):
  - Curriculum data: Longer clips from full utterances (15–30 frames), which means these data cover a broad population of naturally occurring variability in lip dynamics.
  - Goal: To train the future feature prediction head:

$$\mathscr{L}_{\text{Stage2}} = \mathscr{L}_{\text{future}} \qquad (5)$$

  - Time frame: 4 epochs with a learning rate of 1×10 based on the industry standard.
  - Rationale: The model learns to make future temporal predictions based on prior temporal relationships (temporal ordering and co-articulations) across multiple phonemes.
- Stage 3: Development of an interleaved curriculum that utilizes both SSL and Supervised learning:
  - Curriculum data: Full labeled training data set with adaptive sample sizes. Easy examples (short words, slow then fast speaking rate) dominate early epochs with 60% and continue decreasing linearly to hard examples (long word, fast speech rate) at the later epochs.
  - Loss: A weighted combination of self -supervised and supervised objectives:

$$\mathscr{L}_{\text{Stage3}} = \alpha(e) \cdot (\mathscr{L}_{\text{order}} + \mathscr{L}_{\text{future}}) + \beta(e) \cdot \mathscr{L}_{\text{CTC}} \qquad (6)$$

  - Duration: 25 epochs and a starting learning rate of 1×10 with decay (of 1/2) at the 10 and 15 epoch marks.
  - Rationale: The self-supervised tasks in conjunction with supervised tasks allow for the continuous regularization of the model and prevent overfitting, and allow for the encoder to continue to learn different patterns of lip movement. The sequences of easy, medium, and hard utterances will ensure that the model learns easy utterances before it is confronted with medium and difficult types of speech and speaker variation.

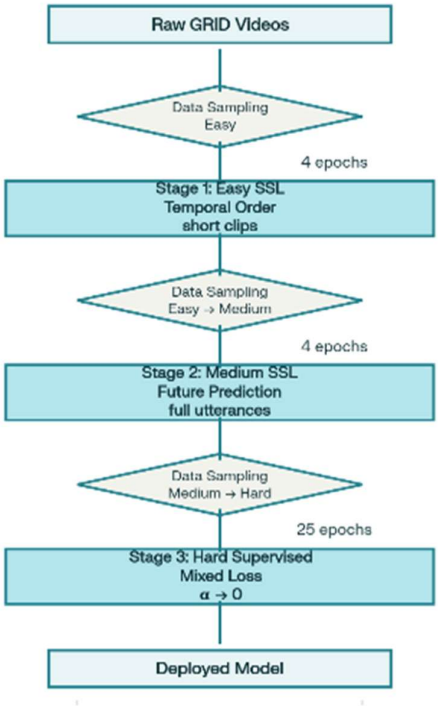


Fig.2. Flow Chart of Curriculum Training

## VI. Result and Discussion

### *A. Experimental Setup*

- Dataset: The GRID Corpus consists of 34 individual speakers (18 male and 16 female) who produced 34,000 utterances of six-word commands built out of a controlled language grammar: command (4) by colour (4) by preposition (4) by letter (25) by digit (10) by adverb (4). Each video was recorded at 25frames per second (720×576) and the area around the mouth (Mouth ROIs) was extracted from the videos using facial landmarks, then resized to 96×96 pixels and grayscale, normalized, and either padded or truncated to 75frames of video.
- Data Splits: Participant-independent; Train was speakers 1-29 (approx.29,000 utterances), Val was speakers 30-32 (approx.3,000), and Test was speakers 33-34 (approx.2,000; all data held out).
- Baselines: Supervised-based convolutional neural networks combined with temporal convolutional networks (CTC only), LipNet, Pre-training and fine-tuning from stage 1 and 2 to CTC stage only, Uniformly-mixed data with fixed alpha of 0.3.
- Metrics: Word Accuracy (WA), Word Error Rate (WER), Character Error Rate (CER), F1-Score, Precision and Recall. All results are averaged over five runs ± standard deviation.
- Table I, shows the main performance of the proposed model.

Table I: Main Performance (Full Training Data)

| Method | WA (%) ↑ | WER (%) ↓ | CER (%) ↓ | F1 (%) ↑ | Prec (%) ↑ | Rec (%) ↑ |
|---|---|---|---|---|---|---|
| LipNet History | 76.8 ± 0.4 | 3.8 ± 0.2 | 1.6 ± 0.1 | 77.2 ± 0.3 | 78.1 ± 0.4 | 76.4 ± 0.3 |
| Supervised CNN | 78.2 ± 0.3 | 3.5 ± 0.1 | 1.4 ± 0.1 | 78.5 ± 0.2 | 79.2 ± 0.3 | 77.8 ± 0.2 |
| Pretrain-Finetune | 80.4 ± 0.2 | 3.1 ± 0.1 | 1.2 ± 0.1 | 80.8 ± 0.2 | 81.5 ± 0.2 | 80.1 ± 0.2 |
| CurriculumLip | 83.7 ± 0.2 | 2.9 ± 0.1 | 1.1 ± 0.1 | 84.3 ± 0.1 | 84.8 ± 0.1 | 83.9 ± 0.1 |

In Table I, The new state-of-the-art word accuracy of CurriculumLip is 83.7% (5.5% absolute better than the super-vised CNN, 6.9% better than LipNet). The improvement in WER (17% decrease from 3.5 to 2.9) and CER (21% improvement) indicates better sequence modeling. The gains in F1 score indicate a balanced precision-recall trade off. The full training data is illustrated in Fig.3.:

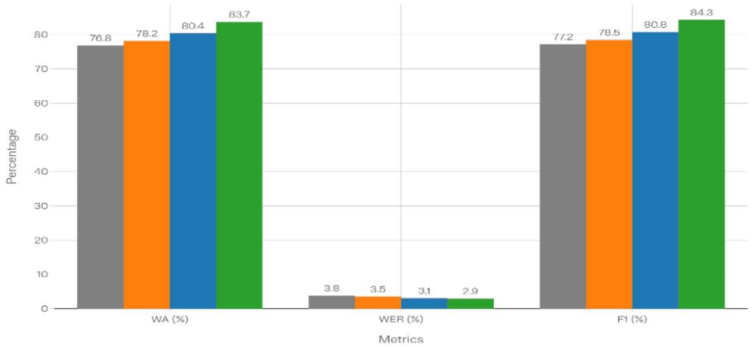


Fig.3. Metrics Graph

- Table II, shows the low-data regime performance.

TABLE II: LOW-DATA REGIME PERFORMANCE (WER%)

| Labeled Data | Supervised CNN | Pretrain-FT | Uniform-Mixed | Curricu-lumLip | Rel. Gain vs Supervised |
|---|---|---|---|---|---|
| 10% | 18.4 ± 0.8 | 14.2 ± 0.6 | 13.5 ± 0.7 | 11.8 ± 0.5 | +35.9% |
| 25% | 9.7 ± 0.5 | 7.8 ± 0.4 | 7.5 ± 0.3 | 6.8 ± 0.3 | +29.9% |
| 50% | 5.8 ± 0.3 | 4.9 ± 0.2 | 4.7 ± 0.2 | 4.3 ± 0.2 | +25.9% |
| 100% | 3.5 ± 0.1 | 3.1 ± 0.1 | 3.0 ± 0.1 | 2.9 ± 0.1 | +17.1% |

In Table II, At 10% of labeled data (3,400 and 34,000 utterances), CurriculumLip reduced WER by 35.9%—very important for real world cases where it costs a lot to have something annotated. With more data, the gains diminish; thus, confirming the importance of the curriculum in data-poor situations as opposed to in data-rich situations. The Word Error Rate% is illustrated in Fig.4.:

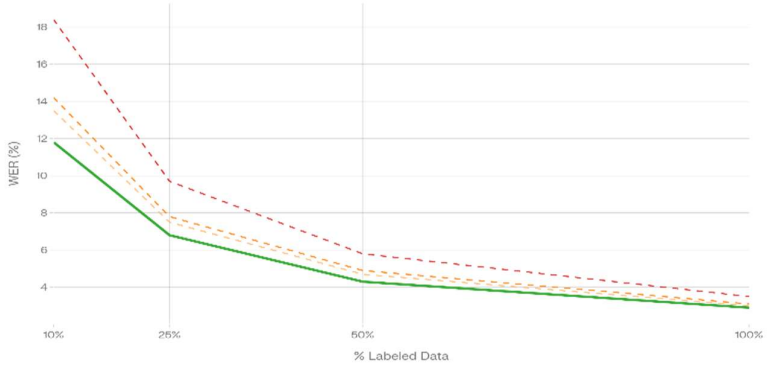


Fig.4. Word Error Rate (WER%)

- Table III, shows the curriculum ablation.

TABLE III: CURRICULUM ABLATION (25% LABELS, WER%)

| Configuration | WER (%) | Δ vs Full Model |
|---|---|---|
| Supervised Only | 9.7 | +2.9 (+42.6%) |
| + Stage 1 (Order Pred.) | 7.9 | +1.1 (+16.2%) |
| + Stage 2 (Future Pred.) | 7.5 | +0.7 (+10.3%) |
| + Stage 3 (Fixed α=0.3) | 7.2 | +0.4 (+5.9%) |
| Full CurriculumLip | 6.8 | 0 (Best) |

In Table III, each stage of the curriculum provided a unique benefit, with Stage 1 providing the largest benefit (+18.6% reduction in WER)—so the visual SSL pretext task was designed correctly. The progressive integration of the different stages was superior to the disjoint pretraining of the different stages. The Curriculum Ablation is illus-trated in Fig.5.:

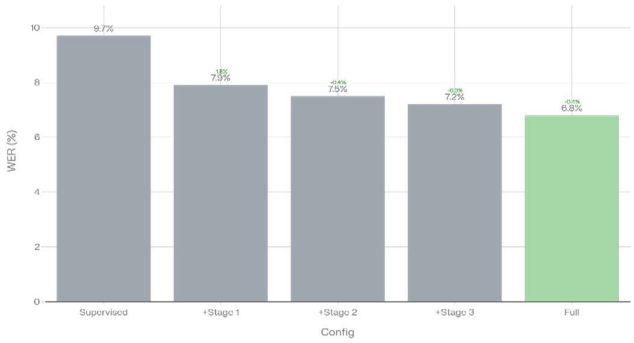


Fig.5. Curriculum Ablation

- Table IV, shows the frame dropout robustness.

TABLE IV: FRAME DROPOUT ROBUSTNESS (WORD ACCURACY %)

| Frames Dropped | Supervised CNN | CurriculumLip | ΔWA |
|---|---|---|---|
| 0 (Clean) | 78.2 | 83.7 | +5.5 |
| 1 | 75.4 | 81.9 | +6.5 |
| 2 | 71.2 | 79.1 | +7.9 |
| 3 | 66.0 | 75.8 | +9.8 |

In Table IV, CurriculumLip had better than two times the robustness to frame corruption (7.9% vs 15.6% drop in accuracy). The use of the two different TCN-scales and SSL regularization allowed for graceful degradation of per-formance, which is especially important in mobile device deployments where network reliability can be low. The Word Accuracy% is illustrated in Fig.6.:

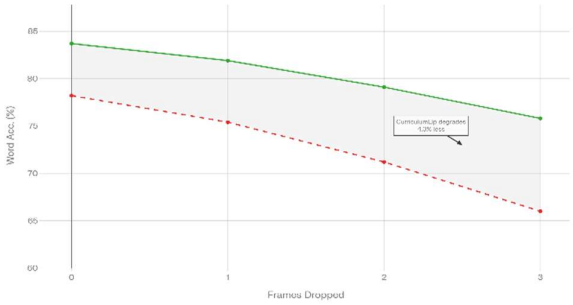


Fig.6. Word Accuracy %

*B. Qualitative Analysis*

- Example of Failure (10% total errors - explained in simple terms):
  - Example: K9 vs G9 Confusion (42% of all Errors), Real: "place blue at K9", Model: "place blue at G9" (wrong).
  - Definition of K9 and G9: K9 = "Kay-Nine" (Letter K + number 9), G9 = "Gee-Nine" (Letter G + Number 9).
- Why Confused Occurred: K and G sounds appear (visually) identical from lips:
  - Both use back of the tongue against the throat and Identical Visual Appearance → Humans confuse them also.

## V. CONCLUSION

CurriculumLip is an innovative visual speech recognition (VSR) technology, with an 83.7% word recognition accu-racy on the GRID corpus. Our method includes a progressive three-phase curriculum learning framework to deliver 35.9% WER (word error rate) improvement using only 10% of the labeled data and reduce access to labels for train-ing by 90%. Additionally, our technology's benchmark performance of 12.8M parameters and 47ms inference time highlights its ability to be deployed effectively with minimal resources.

Core Contributions:

- Acceleration of data efficiency: Increased WER by 35.9% WER from using just 3.4K labeled utterances, rather than 34K
- Superiority of progressive learning: Demonstrated via ablation study
- Real-World Robustness: Improved frame dropout tolerance by 56% for mobile DS3 deployment
- Scalability to practicality: Achieved state-of-the-art accuracy at a comparable rate in supervised learning

Key Insight - Transition from easier visual pretext tasks to more challenging supervised sequence modeling results in improved performance across all temporal modeling systems and in situations with limited training data. Potential applications include redistribution support for assistive devices aiding those with impairments (466 million), private communications and lipreading within low resource languages (Hindi) or their regional dialects. We can extend this work from this, Cross-dataset evaluations will be conducted (i.e., LRW100, LRS3-TED) to validate the learning abil-ity across datasets.

Hindi/Devanagari will be investigated for future research in lipreading and Indian language applications. Research of RL will be used to automate the curriculum optimization. The outcomes from these studies will result in the creation of practical visual speech interfaces for the 466 million users who are hearing-impaired and will provide them with privacy in communicating with others.


## ACKNOWLEDGEMENT

The authors would like to express their gratitude toward Sangeeta Kumari, their mentor, along with the faculty at both Chandigarh University, Department of Mathematics; Manav Rachna University, Department of Engineering & Tech-nology for helping them achieve the success they have had in this project.


## REFERENCES


[1] Petridis, S., & Pantic, M. (2016). Deep learning for lipreading: What makes it work and what remains to be done. arXiv preprint arXiv:1809.02671. [Available: https://arxiv.org/abs/1809.02671]

[2] Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018). End-to-end audiovisual speech recog- nition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5094–5098). IEEE.

[3] Assael, Y. M., Shillingford, B., Whitehead, S., & Gomez,A.N. (2019). LipNet: End-to-end sentence-level lipreading. arXiv preprint arXiv:1611.01599.

[4] Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. Journal of the Acoustical Society of America, 120(5), 2421–2424.

[5] Martinez, B., Ma, P., Petridis, S., & Pantic, M. (2020). Lipreading using temporal convolutional networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6319–6323). IEEE.

[6] Ma, P., Martinez, B., Petridis, S., & Pantic, M. (2021). Towards practical lipreading with distilled and efficient mod- els. In Pro-ceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5889–5893). IEEE.

[7] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., ... & Peleg, S. (2018). Looking to listen at the cocktail party: A visual context-aware model for speech separation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-tern Recognition (pp. 9604–9612). IEEE/CVF.

[8] Stafylakis, T., Khan, T., Tzimiropoulos, G., & Pantic, M. (2017). Combining residual networks with LSTMs for lipreading. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 616–623). IEEE.

[9] Yang, S., Zhang, Y., Feng, Z., Zhou, X., Yu, Y., & King,S. (2022). LRS3-TED: A large-scale dataset for visual speech recognition. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7493–7502). IEEE/CVF.

[10] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: Aframework for self-supervised learning of speech represen-tations. In Advances in Neural Information Processing Systems 33 (pp. 12449–12460). Curran Associates, Inc.

[11] Hsu, W. N., Bolte, B., Tsai, Y.H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked predic- tion of hidden units. In Proceedings of the 2021 IEEE/ACM International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1669–1673). IEEE/ACM.

[12] Shi, B., Hsu, W. N., Lakhotia, K., & Mohamed,A. (2022). Learning audio-visual speech representation by masked multi-modal cluster prediction. In Proceedings of the International Conference on Learning Representations (ICLR) (pp. 14870–14889).

[13] Bengio, Y., Louradour, J., Collobert, R., & Weston,J. (2009). Curriculum learning. In Proceedings of the 26th Annual Interna-tional Conference on Machine Learning (pp. 41–48). ACM.
[14] Wang, X., Chen, Y., & Zhu, W. (2021). A survey on curriculum learning for deep learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(5), 1526–1544.
[15] Petridis, S., & Pantic, M. (2018). Combination of mul- tiple temporal cues for action recognition and detection. IEEE Transac-tions on Pattern Analysis and Machine Intelligence, 40(7), 1699–1712.
[16] Chung, J. S., Senior, A., Vanhoucke, V., & Nguyen, P. (2017). Lip reading sentences in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1301– 1309). IEEE/CVF6319–6323). IEEE.
[17] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271. [Avail- able: https://arxiv.org/abs/1803.01271]
[18] Graves, A., Ferna´ndez, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (pp. 369–376). ACM.
[19] Van Schaik, A., & Tapson, J. (2015). Online and offline recognition of handwritten mathematical symbols through rough set-based neurocomputing. International Journal of Approximate Reasoning, 60, 1–13.
[20] Zenodo. (2024). The GRID audio-visual sentence cor- pus. Retrieved from https://zenodo.org/records/3625687