

The First Character-Level Visual Speech Recognition for Bangla: Task, Protocol and Results

Abdul Hasib Uddin

hasib1536@cseku.ac.bd

Khulna University <https://orcid.org/0000-0003-1000-2600>

Abu Shamim Mohammad Arif

Khulna University

Research Article

Keywords: Bangla, Bengali, visual speech recognition, lip reading, character-level recognition, sequence transduction, LipBengal, speaker-independent evaluation

Posted Date: October 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11207165/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

The First Character-Level Visual Speech Recognition for Bangla: Task, Protocol and Results

Abdul Hasib Uddin* · Abu Shamim Mohammad Arif

Computer Science and Engineering Discipline, Khulna University, Khulna-9208, Bangladesh

*Corresponding author: hasib1536@cseku.ac.bd · shamim@cse.ku.ac.bd

Abstract

This report presents the first character-level visual speech recognition work for Bangla. Every Bangla lip-reading system published to date assigns a silent video one label from a fixed list, whether that label is a word, a detected word region or a vowel class. Character-level recognition asks a different question: generate the written characters of the utterance in order, with nothing restricting the output to a label list. The report surveys the existing Bangla work and shows that none of it poses that question, defines the character-level task for Bangla, fixes an evaluation protocol for it on the LipBengal corpus, and reports a completed speaker-independent result. On 14,431 test clips spoken by 30 speakers who appear in no training clip, a visual-only character transducer reaches $78.15 \pm 1.45\%$ charWER, the word error rate of a character-level output, and $63.48 \pm 0.79\%$ character error over Unicode code points, averaged over five training seeds, against a 99.79% majority-class floor. Measured over Bangla grapheme clusters rather than code points the character error is $71.73 \pm 0.98\%$, which is why the unit must be stated with any Bangla character error rate. The clip-level predictions behind all three rates, and a script that recomputes them from those predictions alone, are archived at <https://doi.org/10.5281/zenodo.23036345>. The absolute error is high and the task is far from solved; the purpose of this report is to fix the task, the protocol and a reference point so that later work has something definite to improve on, and to place that contribution on the public record.

Keywords: Bangla; Bengali; visual speech recognition; lip reading; character-level recognition; sequence transduction; LipBengal; speaker-independent evaluation.

1 The claim

Character-level visual speech recognition has not, to the authors' knowledge, been reported for Bangla before this report. Section 2 sets out the basis for that statement: every Bangla visual speech system the authors could locate, published or preprint, predicts a label from a closed set, and none generates written characters. What follows is therefore offered as the first such work — the task stated, an evaluation protocol fixed for it, and a completed speaker-independent result measured under that protocol.

The hedge “to the authors' knowledge” qualifies the search, not the result. A literature search cannot exclude work that is unpublished, embargoed, or released concurrently. Section 2 therefore records what was searched, where, and when, so that a reader can judge the claim rather than take it on trust, and so that any earlier work that surfaces later can be placed against a documented procedure.

2 Existing Bangla visual speech recognition, and the basis for the claim

2.1 What has been published

BenAV is a Bengali audio-visual corpus of 50 word classes recorded from 128 speakers, used to compare whole-word recognizers under two split arrangements [1]. Rahman and colleagues classify Bengali words with a convolutional encoder, a bidirectional recurrent network and a classifier on top [2]. A 2023 study

recognizes which Bengali vowel accompanies a consonant from still face and lip images using deep and ensemble classifiers, so its input is a still image rather than a video sequence [3]. Bhuiyan and Uddin locate and identify Bengali words from lip movement with Mask R-CNN followed by a generalized linear model, reporting 91.8% accuracy [4]. Mahjabin and Ahsan add lip landmarks to an Xception and multi-GRU pipeline over 100 word classes [5].

LipBengal, the corpus used here, is a dataset contribution [6]. It names word error rate and character error rate among the measures a user might adopt and reports 84.6% accuracy for a LipNet [11] word-identification experiment, which is a closed-set choice among the recorded labels. It does not report a character-sequence benchmark, and that 84.6% figure is not comparable with anything in this report, because it scores a different prediction problem. Its own comparison of publicly available lip-reading corpora places it as the only Bengali entry among twelve datasets spanning English, Spanish, Czech, German, French, Arabic, Japanese and Russian.

Two further Bangla lip-reading resources were deposited in 2024 without an accompanying study: LRW-B, a collection of isolated Bangla word recordings [7], and LRV-B, which covers six Bangla vowel sounds [8]. Both are organized around words or vowels as the unit of annotation. A further 2024 deposit sets out to address Bengali lip reading with a three-dimensional convolutional network, but reports evaluation on a private synthetic English set while the Bengali recordings are still being prepared, so it carries no completed Bengali result [9]. Taken with the corpora above, the available Bangla data has been built for classification targets throughout.

Independent corroboration comes from a 2026 cross-linguistic review of lip reading that consolidates 79 datasets and traces architectures from shallow feature extraction to global temporal modelling [10]. Bengali appears in it twice: once as LipBengal in the dataset inventory, and once in a sentence grouping Bengali with Greek and other languages that together account for 21.5% of the corpora surveyed. It records no Bengali system at all.

Table 1. Bangla visual speech recognition work and data located by the search of Section 2.2, and what each predicts.

Work	Year	What it predicts	Output space
BenAV [1]	2021	one of 50 word classes from video	closed list
Rahman et al. [2]	2022	one word class from a fixed set, from video	closed list
Uddin and Arif [3]	2023	one vowel class from still face and lip images	closed list
Bhuiyan and Uddin [4]	2024	one Bengali word, detected and identified	closed list
LRW-B [7]	2024	dataset of isolated Bangla words; no system reported	—
LRV-B [8]	2024	dataset of six Bangla vowels; no system reported	—
Ahmed et al. [9]	2024	in progress; evaluated on synthetic English only	—
Mahjabin and Ahsan [5]	2025	one of 100 word classes from video and landmarks	closed list
LipBengal [6]	2025	dataset; LipNet word identification baseline	closed list
This report	2026	a Bangla character string, generated symbol by symbol	unrestricted

The pattern in Table 1 is uniform. Bangla visual speech recognition has been posed as classification throughout: the system is handed a list and chooses from it, and its accuracy states how often the choice is right. No published Bangla system emits characters, so none can be partly correct, none can be scored with a character error rate, and none can produce a string that is not on the list. That is the gap this report fills.

Three of the entries above are the authors’ own earlier work: the vowel study [3], and the LRW-B and LRV-B deposits by the first author [7, 8]. None of them generates characters either. The gap this report closes is therefore one that the authors’ own prior work left open.

2.2 What was searched

The claim is bounded to work that is available online, either peer reviewed or as a public preprint. Within that scope the search covered Google Scholar, IEEE Xplore, SpringerLink, ScienceDirect and the ACM Digital Library, the preprint servers arXiv, TechRxiv, Research Square and Preprints.org, and the data repositories Zenodo and Mendeley Data, in September 2026. Query terms combined “Bangla” and “Bengali” with “lip reading”, “lipreading”, “visual speech recognition”, “visual-to-text”, “character-level”, “sequence-to-sequence” and “transducer”, and the reference lists of the works in Table 1 were followed. Proceedings that are not indexed online are outside this scope and outside the claim.

Two classes of near neighbours turn up repeatedly and should be set aside explicitly. Bangla automatic speech recognition, including character- and sequence-level systems, works from audio and does not address the visual problem. Bengali speech-to-sign and grapheme-mapping systems generate symbols but take speech, not silent video, as input. Neither maps video to written characters.

3 The task

The input is a silent sequence of preprocessed mouth-region frames covering one spoken utterance. The output is a string of Bangla characters, produced symbol by symbol. The system is given no candidate list at test time and may emit any sequence its tokenizer can express, including strings that are not Bangla words.

The difference from classification is not cosmetic, and it survives the fact that each clip here holds a single word. A classifier is right or wrong; a character-level system can be partly right, and how wrong it is can be measured character by character. A classifier cannot produce a string outside its label set; a character-level system can, and does, which is informative about what it has actually learned. The two settings are not interchangeable, and a number from one cannot be compared with a number from the other.

What makes the task character-level is the unit the model emits. The text vocabulary is built with SentencePiece in its character mode [14] and is fitted only to the training references, never to validation or test text; for the full set it holds 58 units, reserved symbols included. Each unit is a single Bangla character observed in training, so a word is produced one character at a time and any string composed of those characters can be emitted, including words that never occur in training. This is what separates the setting from subword or whole-word modelling, where the emission alphabet already carries word-sized pieces. The 58 units are the alphabet, not the label set, and should not be confused with the 509 reference strings of Section 4.1.

Two boundaries of the task should be stated plainly, because they limit what any score on it means. First, the clips are isolated utterances, so this is not continuous sentence recognition. Second, every label used at

test time also occurs in training, so the evaluation measures generalization to new speakers rather than to new words. A system that scores well here has not been shown to spell a word it never saw.

4 Evaluation protocol

4.1 Speakers and clips

The corpus is partitioned by speaker so that no person contributes clips to more than one split. Training uses speakers p1–p90, validation p91–p120 and testing p121–p150. Table 2 gives the resulting clip counts after preprocessing.

Table 2. Speaker-disjoint partition of the processed LipBengal clips.

Split	Speakers	Clips
Training (p1–p90)	90	43,749
Validation (p91–p120)	30	14,730
Test (p121–p150)	30	14,431
Total	150	72,910

The label set contains 509 distinct recorded reference strings. Two of them contain a space and are two-word compounds, ছোট বোন and ছোট ভাই; the remaining 507 are single words. The exact recorded strings are scored as they stand, with no orthographic normalization applied at evaluation time. Any later study that normalizes them must say so before comparing its numbers with these.

4.2 Measures

Three quantities are reported, all computed over the whole test set by summing numerators and denominators rather than by averaging per-clip rates.

charWER is the standard word-level rate, $(S + D + I) / N$, with the decoded string of a clip treated as a single item: the word error rate of a system whose output units are characters, and the same quantity that a whole-string evaluation calls utterance or sentence error. A clip counts as an error unless every character of the prediction matches its reference, so charWER is exactly 100% minus exact-match accuracy. Treating each reference as one item keeps the definition uniform across the two compound labels; a conventional whitespace-split word error rate differs from it by less than 0.1 percentage point on every seed, and both give 78.15% averaged over the five seeds.

Character error rate is the Levenshtein distance between prediction and reference divided by the reference length. The primary figure counts Unicode code points, including spaces. A second figure, written CER-g, counts extended grapheme clusters as defined in Unicode Annex #29 [16]. The distinction is not cosmetic in Bangla: ধ্য is three code points and one letter to a reader, and বিশ্ববিদ্যালয় is fourteen code points and six letters. Both the numerator and the denominator change with the unit, so the two figures are not convertible into one another and a Bangla character error rate quoted without its unit is ambiguous.

For orientation, a predictor that always answers the single most frequent test label makes an error on 99.79% of test clips, and uniform guessing among the 509 labels would give 0.20% exact-match accuracy. These are floors, not competitive systems; they exist so a reader can see whether a reported number reflects learning at all.

4.3 Reporting

A run is one training seed. Every reported figure is the mean over five seeds with the population standard deviation, and the number of seeds is stated with it. Within a seed, the checkpoint with the lowest greedy-decoding validation character error is selected, and that checkpoint alone is decoded on the test set with a beam width of four. Selection therefore uses greedy validation decoding and measurement uses beam test decoding; the two are different conditions and are not mixed.

5 A first system and its result

The system is visual-only. Preprocessed grayscale mouth-region frames at 96×96 pixels enter a three-dimensional convolution followed by a residual image encoder, whose per-frame features are modelled through time by a Zipformer encoder [12]; a prediction network and joiner emit the next text unit or a blank. The architecture family follows a sentence-level Vietnamese system that joined a three-dimensional residual front end to Zipformer [13], retrained here from scratch on Bangla video with a different tokenizer and training recipe. The text units are the character-level SentencePiece vocabulary described in Section 3. Training minimizes a pruned recurrent neural network transducer loss with an auxiliary connectionist temporal classification branch [15] at weight 0.3; the auxiliary branch is not used for decoding. No audio, no external language model and no pretrained visual or language weights enter the runs.

Table 3. First reported character-level result for Bangla, from the work of this report, on the 14,431 test clips of Table 2. Mean $\pm$ population standard deviation over five training seeds.

Measure	Value
charWER (exact string), %	78.15 $\pm$ 1.45
Exact-match accuracy, %	21.85 $\pm$ 1.45
Character error, Unicode code points, %	63.48 $\pm$ 0.79
Character error, grapheme clusters (CER-g), %	71.73 $\pm$ 0.98
Majority-class floor, charWER, %	99.79

The system is well clear of the majority-class floor on every seed and the spread across seeds is small, so the task is learnable from this corpus under a speaker-disjoint split. Exact-match accuracy is 21.85%, so it is also far from solved. Table 3 is offered as the reference point that later work on this task can be measured against.

6 Limitations

The evaluation covers new speakers saying words that appear in training. It says nothing about words held out of training, about continuous speech, or about recordings made in another setting. The five seeds give a useful picture of run-to-run stability but are too few for strong distributional claims, and no significance test is offered here.

Two sources of variation are distinct and are not interchangeable. Variation across training seeds describes what happens when the same recipe is run again. Variation across the identity of the test speakers describes what happens when different people are recorded. A figure that captures one of them does not bound the other.

The claim of Section 1 rests on a search, and a search has a date. It was carried out in September 2026 over the sources named in Section 2.2. Work that was unpublished, embargoed or released concurrently would not have been found.

7 Data and code availability

LipBengal is available from its authors [6]. The clip-level prediction files behind Table 3 — reference string, decoded string, speaker identifier and clip identifier for each of the 14,431 test clips, for each of the five seeds — the per-seed training and validation record, a run report stating the parameter count, the training environment and the metric definitions, and a self-contained scoring script are archived at <https://doi.org/10.5281/zenodo.23036345>, together with the training notebook and its recorded outputs. The script recomputes charWER, CER and CER-g from the prediction files alone, so every rate in Table 3 can be checked without access to the environment that produced it. The processed video arrays and the trained checkpoints are not distributed there: the arrays are derived from LipBengal and are the dataset authors’ to release. Both will accompany the extended study described in Section 8.

8 Relation to other work by the authors

This report is the record of the task, the evaluation protocol and one completed result, and of the date on which they were made public. An extended study by the same authors, covering additional model configurations, a comparative analysis of architectural choices, uncertainty estimates and a broader error analysis, is in preparation for peer-reviewed publication. That study will cite this report, and the material here will appear in it in revised and expanded form. Readers, reviewers and editors should treat the two as one line of work by the same authors, not as separate contributions.

9 How to cite this report

Work that adopts the task definition of Section 3, uses or modifies the evaluation protocol of Section 4, compares against the result of Table 3, or refers to character-level Bangla visual speech recognition as a defined problem should cite this report. Where a later publication by the authors supersedes it, both may be cited; this report remains the record of first disclosure and carries the earlier date.

The identifier that the preprint server assigns on posting is shown on this report’s record page and should be added to the citation below. The accompanying data and code archive already carries a permanent identifier and can be cited as it stands.

Plain text:

Uddin, A. H., and Arif, A. S. M. The First Character-Level Visual Speech Recognition for Bangla: Task, Protocol and Results. Preprint, September 2026.

Uddin, A. H., and Arif, A. S. M. (2026). The First Character-Level Visual Speech Recognition for Bangla: data and code. Zenodo. <https://doi.org/10.5281/zenodo.23036345>

BibTeX:

```
@misc{uddin2026banglachar,
  title   = {The First Character-Level Visual Speech Recognition for Bangla:
            Task, Protocol and Results},
  author  = {Uddin, A. Hasib and Arif, Abu Shamim Mohammad},
  year    = {2026},
  note    = {Preprint}
}
```

```
@dataset{uddin2026banglachardata,
  title      = {The First Character-Level Visual Speech Recognition for Bangla:
               data and code},
  author     = {Uddin, A. Hasib and Arif, Abu Shamim Mohammad},
  year      = {2026},
  publisher  = {Zenodo},
  doi       = {10.5281/zenodo.23036345}
}
```

Declarations

Funding: this research was funded by the Division of Information and Communication Technology (ICT), Ministry of Posts, Telecommunications and Information Technology, Government of the People’s Republic of Bangladesh, tracking number 24FS37893. Competing interests: the authors declare none. Ethics: no new data were collected from people for this work; it uses only the publicly released LipBengal corpus [6], under the terms set by its authors. Copyright remains with the authors. The archived prediction files and training record are released under CC BY 4.0 and the scoring code under the MIT licence.

References

- [1] Pondit, A., Rukon, M. E. A., Das, A., and Kabir, M. A. (2021). BenAV: A Bengali audio-visual corpus for visual speech recognition. In *Neural Information Processing, LNCS 13109*, pp. 526–535. https://doi.org/10.1007/978-3-030-92270-2_45
- [2] Rahman, M. M., Tanjim, M. R., Hasan, S. S., Shaiban, S. M., and Khan, M. A. (2022). Lip reading Bengali words. *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence*. <https://doi.org/10.1145/3579654.3579677>
- [3] Uddin, A. H., and Arif, A. S. M. (2023). Deep learning and ensemble machine learning for Bengali vowel recognition from consonant pronunciation using face and lip images. *2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. <https://doi.org/10.1109/ICICT4SD59951.2023.10303591>
- [4] Bhuiyan, A. B., and Uddin, J. (2024). Bengali word detection from lip movements using Mask RCNN and generalized linear model. *Indonesian Journal of Electrical Engineering and Informatics*, 12(2), 373–383. <https://doi.org/10.52549/ijeei.v12i2.5088>
- [5] Mahjabin, W., and Ahsan, S. M. M. (2025). Lip landmark aided Bengali visual speech recognition leveraging Xception and Multi-GRU architecture. *2025 IEEE 2nd International Conference on Computing, Applications and Systems*. <https://doi.org/10.1109/COMPAS67506.2025.11381839>
- [6] Sahed, M. T. R., Aronno, M. T. I., Nyeem, H., Wahed, M. A., Ahsan, T., Islam, R. R., et al. (2025). LipBengal: Pioneering Bengali lip-reading dataset for pronunciation mapping through lip gestures. *Data in Brief*, 58, 111254. <https://doi.org/10.1016/j.dib.2024.111254>
- [7] Uddin, A. H. (2024). Lip Reading Word-Bangla (LRW-B) dataset. Mendeley Data, V1. <https://doi.org/10.17632/k5d598277b.1>
- [8] Uddin, A. H. (2024). Lip Reading Vowel-Bangla (LRV-B) dataset. Mendeley Data, V1. <https://doi.org/10.17632/sdhjyzkxsd.1>
- [9] Ahmed, T., Shormi, L. S., Alvi, A. M., and Zaman, K. M. A. U. (2024). Automated lip reading from video dataset with 3D convolutional neural network. *Zenodo*. <https://doi.org/10.5281/zenodo.10826306>
- [10] Hu, Q., Kipli, K., Zulcaffle, T. M. A., Liu, Y., and Cao, J. (2026). A review of lip-reading: from datasets and architectures to cross-linguistic performance gaps. *Advances in Technology Innovation*, 11(3), 277–303. <https://doi.org/10.46604/aiti.2026.16377>
- [11] Assael, Y. M., Shillingford, B., Whiteson, S., and de Freitas, N. (2016). LipNet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*. <https://arxiv.org/abs/1611.01599>

- [12] Yao, Z., Guo, L., Yang, X., et al. (2024). Zipformer: A faster and better encoder for automatic speech recognition. International Conference on Learning Representations. <https://arxiv.org/abs/2310.11230>
- [13] Huu, P. N., and Sy, T. H. (2025). A novel sentence-level visual speech recognition system for Vietnamese language using ResNet3D and Zipformer. Modelling and Simulation in Engineering, article 2087573. <https://doi.org/10.1155/mse/2087573>
- [14] Kudo, T., and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. EMNLP System Demonstrations, pp. 66–71. <https://doi.org/10.18653/v1/D18-2012>
- [15] Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning, pp. 369–376. <https://doi.org/10.1145/1143844.1143891>
- [16] Unicode Consortium. Unicode Standard Annex #29: Unicode Text Segmentation. <https://www.unicode.org/reports/tr29/>