BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

Transformer-Driven Multimodal Visual Speech Recognition of Bengali Numerals

Yousuf Ibne Kamal Niloy
200041123
Adib Farhan
200041203
Rizwan Rabbi
200041221

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

Transformer-Driven Multimodal Visual Speech Recognition of Bengali Numerals

Yousuf Ibne Kamal Niloy
200041123
Adib Farhan
200041203
Rizwan Rabbi
200041221

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Yousuf Ibne Kamal Niloy**, **Adib Farhan**, and **Rizwan Rabbi** under the supervision of **Dr. Kamrul Hasan**, Professor, Department of Computer Science and Engineering and co-supervision of **Dr. Hasan Mahmud**, Professor, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Dr. Kamrul Hasan
Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 29, 2025

Dr. Hasan Mahmud
Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 29, 2025

Yousuf Ibne Kamal Niloy
Student ID: 200041123
Date: September 29, 2025

Adib Farhan
Student ID: 200041203
Date: September 29, 2025

Rizwan Rabbi
Student ID: 200041221
Date: September 29, 2025

Contents

List of Figures

List of Tables

List of Abbreviations

ASR	Automatic Speech Recognition
AVSpeech	Audio-Visual Speech Dataset
BenAV	Bengali Audio-Visual Corpus
Bi-GRU	Bidirectional Gated Recurrent Unit
BSRD	Bengali Speech Recognition Dataset
CE	Cross-Entropy
CNN	Convolutional Neural Network
CroMM-VSR	Cross-Modal Memory Augmented Visual Speech Recognition
CSLR	Cross-Linguistic Code-Switching Lip Reading
CTC	Connectionist Temporal Classification
EEG	Electroencephalography
EMG	Electromyography
GRID	GRID Corpus Dataset
LRS2	Lip Reading Sentences 2
LRS3	Lip Reading Sentences 3
LRW	Lip Reading in the Wild
LRW-B	Lip Reading Word-Bangla
MVM	Memory-Augmented Multimodal Networks
P2FA	Penn Phonetics Lab Forced Aligner
SSL	Self-Supervised Learning
SyncVSR	Synchronized Visual Speech Recognition
TAL	Tongue and Lips Corpus
TCN	Temporal Convolutional Network
VSP-LLM	Visual Speech Pre-trained Large Language Model
VSR	Visual Speech Recognition
WER	Word Error Rate
WAS	Watch, Attend and Spell
WLAS	Watch, Listen, Attend and Spell
LAS	Listen, Attend and Spell

Acknowledgement

First and foremost, we would like to express our deepest gratitude to our supervisor, Dr. Kamrul Hasan and Dr. Hasan Mahmud for their invaluable guidance, continuous encouragement, and steadfast support throughout the course of this research. Their expertise in the field, insightful feedback, and unwavering patience were vital in shaping this thesis from its inception to completion.

We are also thankful to the faculty and staff of the Department of Computer Science and Engineering at the Islamic University of Technology for fostering an environment of academic curiosity and for providing the infrastructure necessary for conducting this research.

Our heartfelt appreciation goes to our parents and families, whose unconditional love, encouragement, and sacrifices have been the foundation of our academic journey. Thank you for believing in us, especially during moments when we struggled to believe in ourselves.

To our friends and peers, we are grateful for the countless brainstorming sessions, late-night coding marathons, and for reminding us to occasionally step outside and take a break.

Finally, we extend our gratitude to the research community and authors whose work laid the groundwork for this thesis. We hope that our contribution adds a small yet meaningful step forward in the journey of Bengali-language technology.

Abstract

Visual Speech Recognition (VSR) decodes spoken content from lip movements, enabling applications in assistive technologies and silent communication systems. For Bengali, a language spoken by over 230 million people globally , the scarcity of multimodal datasets and the homophene challenge—where words like “Chollish” and “Ekchollish” are visually indistinguishable— impede progress in lipreading. This thesis introduces the BND Bangla Numeric Dataset, comprising 3,600 high-resolution videos of 30 diverse speakers reciting Bengali numerals 1 to 100, designed to address challenges such as homophenes, varied lighting, camera angles, and regional dialects. We adapt the state-of-the-art SyncVSR framework [4], integrating ResNet18 for visual feature extraction, a fine-tuned `facebook/wav2vec2-large-xlsr-53` for audio tokenization [12], and a transformer-based BERT for sequence modeling, evaluating its performance against unimodal and multimodal baselines like LipNet [7]. Our experiments investigate how these models tackle the homophene challenge, demonstrating that SyncVSR achieves a 68.7% accuracy on the BND dataset, significantly outperforming LipNets 46.2% by leveraging synchronized audio-visual cues to mitigate homophene ambiguity. We identify homophene clusters (e.g., “chollish” , “ekchollish” , “biyallish” ... also “shottor”, “ekattor” ... and so on) and compare the BND dataset with existing ones like LipBengal [32] and BenAV [29], highlighting its unique attributes for numerical VSR. Recent studies, such as SynthVSR [42], suggest synthetic data could address data scarcity, pointing to future research directions. By releasing the BND dataset, this work provides a critical resource for the research community, advances Bengali VSR, and establishes a foundation for robust numerical speech recognition in low-resource settings.

Chapter 1

Introduction

This dissertation endeavors to advance Visual Speech Recognition (VSR), commonly referred to as lipreading, for the low-resource language of Bengali, with a specific focus on numerical speech recognition. VSR involves interpreting spoken content from visual cues, primarily lip movements, enabling applications in assistive technologies for the hearing-impaired, silent communication systems in noisy environments, and secure interfaces where audio is unreliable or unavailable [7]. While significant progress has been made in high-resource languages like English, driven by expansive datasets such as Lip Reading in the Wild (LRW) [10], the development of VSR systems for Bengali, spoken by approximately 230 million people primarily in Bangladesh and India, remains hindered by the paucity of annotated multimodal datasets and the linguistic complexity introduced by homopheneswords producing visually indistinguishable lip movements despite distinct meanings. This thesis introduces the BND Bangla Numeric Dataset, a novel corpus comprising 3,600 high-resolution videos of 30 diverse speakers reciting Bengali numerals from 1 to 100, designed to address these challenges and facilitate precise numerical recognition critical for domains such as banking, education, and commercial transactions. By adapting the state-of-the-art SyncVSR framework [4], this work aims to bridge the gap in low-resource language processing, enhancing the accessibility and inclusivity of speech recognition technologies for Bengali-speaking communities.

1.1 Introductory Information

Visual Speech Recognition (VSR) represents a transformative approach to transcribing spoken language through the analysis of lip movement patterns, circumventing reliance on auditory signals. This technology holds immense potential for applica-

tions where audio is absent or compromised, including assistive devices for individuals with hearing impairments, silent communication systems in high-noise environments, and secure interfaces requiring discreet interactions [7]. The efficacy of VSR systems hinges on robust datasets and sophisticated models capable of decoding complex visual cues. For high-resource languages like English, datasets such as LRW [10], comprising 500,000 video clips across 500 words, have catalyzed advancements, enabling deep learning models like LipNet [7] to achieve high accuracy through visual-only recurrent neural networks.

In contrast, low-resource languages like Bengali face significant barriers due to limited annotated data and linguistic intricacies. Bengali , with approximately 230 million native speakers, is a major global language, yet its representation in VSR research remains sparse. A critical challenge is the prevalence of homophenes, such as “chollish” (forty) and “ekchollish” (forty-one), which produce visually similar lip movements, necessitating advanced multimodal approaches to disambiguate meanings [4]. Existing Bengali datasets, such as LipBengal [32] (150 speakers, 73 classes) and BenAV [29] (7+ hours, 50 words), focus on broader vocabularies and lack numeral-specific content critical for applications like financial transactions, educational tools, and voice-free payment systems.

To address these gaps, this thesis presents the BND Bangla Numeric Dataset, a meticulously curated collection of 3,600 high-resolution videos featuring 30 native Bengali speakers reciting numerals from 1 to 100 (e.g., “ek”, “dui”, “tin”, up to “eksho”). The dataset is designed to capture phonetic variability, regional dialects, and real-world conditions such as diverse lighting and camera angles, ensuring robustness for VSR model training. By integrating this dataset with an adapted SyncVSR framework [4], which employs an encoder-only transformer architecture for word-level prediction, this research seeks to enhance the precision of numerical speech recognition, thereby contributing to the accessibility of speech technologies for Bengali-speaking communities.

1.2 Motivations and Scope

The impetus for this research arises from the pressing need to develop robust VSR systems for Bengali, particularly for numerical speech, which underpins critical applications in low-resource language contexts. Bengalis global significance, with over 230 million speakers across Bangladesh, India, and diaspora communities, underscores the urgency of creating inclusive speech recognition technologies. Numerical

speech recognition is particularly vital in domains such as banking (e.g., processing account numbers or transaction amounts), education (e.g., developing teaching aids for hearing-impaired students), and commercial transactions (e.g., enabling voice-free payment systems in noisy environments). These applications demand high accuracy to ensure reliability, yet the homophene challengeexemplified by visually indistinguishable pairs like “shottor” (seventy) and “ekattor” (seventy-one) poses a significant barrier to effective recognition [4].

Prior efforts in Bengali VSR, such as LipBengal [32], provide a valuable foundation with 150 speakers across 73 classes, encompassing phonemes and sentences. Similarly, BenAV [29] offers over 7 hours of audio-visual data for 50 words across 128 speakers. However, these datasets lack a specific focus on numerals, limiting their applicability to numerical speech tasks. The absence of numeral-specific resources hampers the development of targeted VSR systems, necessitating the creation of the BND Bangla Numeric Dataset. This dataset addresses data scarcity by providing a tailored corpus for numerals 1 to 100, capturing real-world variability and homophene-rich content to enable robust model training

The scope of this thesis encompasses the development of the BND dataset, alongside the adaptation of the SyncVSR framework [4] for Bengali numerical VSR. SyncVSR leverages an encoder-only transformer architecture, integrating ResNet-18 for visual feature extraction and a fine-tuned ‘facebook/wav2vec2-large-xlsr-53‘ model on BenAV [29] for audio tokenization, achieving a 68.7% accuracy on the BND dataset, surpassing LipNets 46.2% [7]. The research evaluates both unimodal (visual-only) and multimodal (audio-visual) models to assess their efficacy in handling homophene ambiguities, with a focus on identifying homophene clusters and proposing strategies for future improvements. By addressing these objectives, this work aims to advance numerical speech recognition for Bengali, contributing to inclusive technologies for underserved linguistic communities.

1.3 Problem Statement

The central problem addressed in this dissertation is the lack of specialized datasets and robust VSR systems for Bengali numerical speech, compounded by the homophene challenge and data scarcity. Homophenes, such as “chollish” and “ekchollish”, produce visually indistinguishable lip movements, necessitating sophisticated multimodal approaches to achieve accurate recognition. Existing Bengali datasets, including LipBengal [32] and BenAV [29], are limited in scope and do not prioritize numerals,

which are critical for applications in banking, education, and commerce. This scarcity restricts the development of precise VSR systems for low-resource languages like Bengali. The research objectives are articulated as follows:

1. To develop and publicly release the BND Bangla Numeric Dataset, capturing phonetic variability and homophene-rich content for Bengali numerals 1 to 100, ensuring robustness across diverse real-world conditions.
2. To adapt the SyncVSR framework [4], utilizing an encoder-only transformer architecture for word-level prediction, integrating multimodal audio-visual processing to enhance recognition accuracy.
3. To evaluate the performance of unimodal (visual-only) and multimodal (audio-visual) VSR models on the BND dataset, with a focus on their ability to disambiguate homophene clusters such as “chollish”, “ekchollish” and so on.
4. To identify homophene clusters and propose data-driven strategies for improving Bengali VSR, laying the groundwork for future research.

These objectives collectively aim to address the critical gaps in dataset availability and model performance, advancing numerical speech recognition for Bengali and fostering inclusivity in speech technologies.

1.4 Research Challenges

The development of VSR systems for Bengali numerals presents multifaceted challenges that demand innovative solutions. First, the unavailability of comprehensive, numeral-specific datasets for Bengali severely limits the training of robust deep learning models. In contrast to high-resource languages with datasets like LRW [41], which provides 500,000 video clips across 500 English words, existing Bengali datasets such as LipBengal [32] (150 speakers, 73 classes) and BenAV [29] (7+ hours, 50 words) focus on broader vocabularies, lacking emphasis on numerals essential for practical applications. This scarcity necessitates the creation of a tailored dataset to support numerical VSR.

Second, the adaptation of VSR models to Bengali is complicated by its phonetic and orthographic characteristics, particularly the prevalence of homophenes. Words like “shottor” and “ekattor” produce visually similar lip movements, requiring precise disambiguation through multimodal cues or contextual analysis [4]. This challenge is exacerbated by the need for models to generalize across regional dialects and speaker variations, which introduce additional complexity.

Third, the computational demands of training multimodal VSR models, such as SyncVSR [4], are substantial. The framework integrates ResNet-18 for visual feature extraction, a fine-tuned 'facebook/wav2vec2-large-xlsr-53' for audio tokenization, and a BERT-based transformer encoder for sequence modeling, requiring high-performance GPUs (e.g., NVIDIA A100) to process large video datasets and optimize complex architectures. Efficient training strategies are essential to balance computational cost with performance.

Finally, real-world variability in lighting conditions, camera angles, speaker demographics (e.g., age, gender, accent), and environmental noise poses significant challenges to model generalization. Unlike controlled datasets like GRID [13], which use uniform conditions, real-world applications demand robustness to such variations, necessitating diverse data collection and preprocessing strategies. These challenges underscore the need for a specialized dataset and tailored computational approaches to advance Bengali VSR.

1.5 Contribution

This dissertation makes the following novel contributions to the field of Visual Speech Recognition for Bengali numerals:

1. **BND Bangla Numeric Dataset**: A pioneering dataset comprising 3,600 high-resolution videos from 30 diverse native Bengali speakers reciting numerals 1 to 100, designed to capture homophene-rich content and real-world variations in lighting, camera angles, and regional dialects. This dataset addresses the critical scarcity of numeral-specific resources, enabling targeted VSR research for applications in banking, education, and commerce.

2. **SyncVSR Adaptation**: Adaptation of the state-of-the-art SyncVSR framework [4], employing an encoder-only transformer architecture (BERT-like, 12 layers, 512 dimensions, 8 heads) for word-level prediction. The model integrates ResNet-18 for visual feature extraction and a fine-tuned 'facebook/wav2vec2-large-xlsr-53' on BenAV [29] for audio tokenization, achieving a 68.7% accuracy on the BND dataset, surpassing LipNets 46.2% [7].

3. **Unimodal and Multimodal Analysis**: A comprehensive evaluation of unimodal (visual-only) and multimodal (audio-visual) VSR models on the BND dataset, identifying homophene clusters such as "chollish" , "ekchollish"... etc. and analyzing their impact on recognition accuracy to inform future model im-

provements.

4. **Dataset Release**: Public release of the BND dataset, accompanied by ethical approval and documentation, providing a valuable resource for the global research community to foster advancements in Bengali VSR and low-resource language processing.

These contributions enhance the accuracy and accessibility of numerical speech recognition for Bengali, with profound implications for assistive technologies, educational tools, and commercial applications, while laying the foundation for future research in low-resource VSR.

1.6 Organization

This dissertation is structured to provide a systematic and comprehensive exploration of Visual Speech Recognition for Bengali numerals, organized as follows:

- **Chapter 2: Related Works** synthesizes existing VSR literature, datasets, and models, highlighting gaps in numeral-specific resources that the BND dataset and SyncVSR adaptation address, with comparisons to datasets like LRW [10], LipBengal [32], and BenAV [29].
- **Chapter 3: Proposed Methodology** details the adaptation of the SyncVSR framework [4], including the encoder-only transformer architecture, training protocols, and evaluation metrics for unimodal and multimodal performance on the BND dataset. It also incorporates a subsection describing the creation, characteristics, and significance of the BND Bangla Numeric Dataset, contrasting it with existing Bengali and multilingual datasets to underscore its novelty and relevance.
- **Chapter 4: Result Analysis** presents experimental results, including validation accuracy comparisons (68.2% for SyncVSR vs. 43.6% for LipNet), homophene cluster analysis, and insights into model performance, highlighting the efficacy of multimodal approaches.
- **Chapter 5: Conclusion** summarizes key findings, reiterates the contributions of the BND dataset and SyncVSR adaptation, and proposes future research directions, such as leveraging synthetic data inspired by SynthVSR [42] to further address data scarcity.

Through this structured exploration, the thesis aims to bridge the gap in VSR for

Bengali, contributing to inclusive and accessible speech recognition technologies for Bengali-speaking communities worldwide.

Chapter 2

Related Works

The Related Works chapter provides a comprehensive analysis of the literature on Visual Speech Recognition (VSR), focusing on datasets and models. It begins with an examination of datasets for lipreading and silent speech recognition, highlighting the evolution from English-centric resources to emerging Bengali datasets, and concludes with insights into synthetic data approaches. Subsequently, it categorizes models into unimodal and multimodal frameworks, tracing their progression from feature-based methods to deep neural networks and transformer-based architectures. This narrative synthesizes key advancements, identifies gaps in low-resource languages like Bengali, and positions the current research within the field.

2.1 Datasets for Lipreading and Silent Speech Recognition

The development of robust VSR and silent speech recognition systems relies on high-quality, diverse datasets that capture lip movements, audio, and other modalities. This section surveys publicly available datasets for English and Bengali lipreading, as well as silent speech recognition, emphasizing their scale, content, recording conditions, and annotations. English datasets are mature and extensive, while Bengali resources are emerging but limited. Silent speech datasets, primarily English-focused, incorporate non-auditory signals like electromyography (EMG), ultrasound, and electroencephalography (EEG). A comparative analysis reveals gaps in numeral-specific and multimodal Bengali data, underscoring the need for resources like the BND Bangla Numeric Dataset introduced in this thesis.

2.1.1 English Lipreading Datasets

English lipreading datasets vary in scale and complexity, supporting advancements in word- and sentence-level recognition.

LRW (Lip Reading in the Wild)

The LRW dataset [10] comprises approximately 1,000 utterances of 500 different English words spoken by over 1,000 speakers, totaling approximately 160 hours of video. Each utterance consists of 29 frames (1.16 seconds), sourced from BBC television broadcasts in "in the wild" conditions with varied lighting, poses, and backgrounds. Annotations are word-level, with standard splits for training (800 utterances per word), validation (50 utterances), and test (50 utterances) sets. Its diversity challenges model robustness, but the focus on isolated words limits applicability to continuous speech [11].

LRS2 (Lip Reading Sentences 2)

LRS2 [11] targets sentence-level lipreading, containing 96,318 pre-training, 45,839 training, 1,082 validation, and 1,242 test utterances from BBC broadcasts, totaling over 225 hours. Sentences are up to 100 characters long, with subtitle-based annotations. Recorded in "in the wild" conditions without speaker labels, it supports contextual learning but complicates speaker-specific modeling [11].

LRS3-TED

LRS3-TED [2] provides over 400 hours of TED and TEDx talks, with a 407-hour pre-training set, 30-hour train-validation set, and 1-hour test set from 5,594 videos. Videos are 224x224 pixels at 25 fps, annotated with subtitles and word alignments using P2FA and Kaldi-based ASR. SyncNet ensures audio-video synchronization, making it ideal for complex models, though sentences are clipped to 100 characters or 6 seconds [2].

GRID Dataset

The GRID dataset [13] includes 33,000 utterances (approximately 27.5 hours) from 34 speakers (18 male, 16 female), using a 51-word vocabulary in controlled lab conditions. Annotations are word-level, supporting constrained tasks, but the small vocabulary and artificial sentences restrict real-world applicability [13].

Other Datasets

Additional datasets include AVICAR (letter-level recognition in car environments, 33 hours) [25], OuluVS2 (phrase-level, 2 hours from 53 speakers) [5], and TCD-TIMIT (constrained word recognition from 62 speakers) [17]. Large-scale options like AVSpeech (4,700 hours from YouTube) [14] and How2 (300 hours of instructional videos) [33] provide further resources for multimodal training.

2.1.2 Bengali Lipreading Datasets

Bengali lipreading datasets are nascent but growing, focusing on phonemes, words, and sentences.

LipBengal

LipBengal [32], released in June 2024, features 150 speakers across 73 classes (phonemes, alphabets, symbols), with 363,150 utterances in "wild" conditions. Annotations range from phoneme- to sentence-level, positioning it as a benchmark for Bengali VSR.

LRW-B (Lip Reading Word-Bangla)

LRW-B [36], published in August 2024, consists of high-quality videos (approximately 363 MB) of isolated Bangla words from multiple speakers. Annotations are word-level, supporting lexical VSR, though its isolated word focus limits continuous speech applications [36].

BenAV

BenAV [29] includes 26,300 utterances (over 7 hours) from 128 speakers (107 male, 21 female), with a 50-word lexicon. Videos are approximately 1 second long, ontaining 34 character words. Its scale supports foundational research, but the small vocabulary is a limitation [29].

2.1.3 Silent Speech Recognition Datasets

Silent speech datasets primarily target English, with sparse Bengali resources.

English Silent Speech Datasets

Silent Speech EMG Dataset [16]: Captures facial EMG during silent and vocalized speech from a single speaker, with phoneme annotations via Montreal Forced Aligner.

The dataset includes paired EMG signals and audio, but size is unspecified.
TaL Corpus [31]: ultrasound tongue imaging and lip videos from 82 speakers (approximately 13.5 hours), with annotations for prompts and spontaneous speech. It supports articulatory studies, verified as 24 hours total across participants.
EEG Datasets: The EEG Speech-Robot Interaction Dataset [27] includes 15 subjects with 64-channel EEG during spoken and imagined speech. The Silent EEG-Speech Recognition Dataset [20] has 270 subjects with 40-channel EEG at 500 Hz, focusing on silent speech.

Bengali Silent Speech Datasets

The Bangla Silent Speech Video Dataset [35] on Kaggle includes videos of silent speech, but details on size, content, and annotations are limited; it comprises videos from 10 subjects uttering 10 common Bengali words, totaling 6,000 clips. This indicates emerging interest, but dedicated annotated resources remain scarce.

2.1.4 Multimodal Datasets

Multimodal datasets combine modalities for enhanced recognition.

AVE Speech Dataset

The AVE Speech Dataset [34] (Mandarin, released 2025) integrates audio, lip videos, and six-channel EMG from 100 participants, with 55 hours per modality across 100 sentences. Annotations are sentence-level, demonstrating multimodal benefits, though its Mandarin focus limits applicability to English or Bengali [34].

2.1.5 Comparative Analysis

English datasets dominate in scale, with LRW (500 words, >1,000 speakers), LRS2 (>225 hours), LRS3-TED (>400 hours), and GRID (27.5 hours, 34 speakers) enabling "in the wild" and controlled VSR. Bengali datasets like LipBengal (150 speakers, 73 classes), LRW-B (isolated words, 363 MB), and BenAV (26,300 utterances, 128 speakers) are smaller and phoneme-focused. Silent speech datasets for English (e.g., TaL Corpus: 13.5 hours, 82 speakers; EEG datasets: 15-270 subjects) explore EMG, ultrasound, and EEG, while Bengali options (e.g., Bangla Silent Speech: 6,000 videos) are preliminary. Multimodal datasets like AVE Speech highlight integration trends but lack Bengali equivalents.

Table 2.1: Comparison of Key Lipreading and Silent Speech Datasets

Dataset	Language	Type	Size	Annotations
LRW	English	Lipreading	500 words, >1,000 speakers	Word-level
LRS2	English	Lipreading	>225 hours	Sentence-level
LRS3-TED	English	Lipreading	>400 hours	Word, sentence
GRID	English	Lipreading	27.5 hours	Word-level
LipBengal	Bengali	Lipreading	150 speakers, 73 classes	Phoneme, sentence
LRW-B	Bengali	Lipreading	363 MB	Word-level
BenAV	Bengali	Lipreading	>7 hours	Word-level
Silent Speech EMG	English	Silent Speech	Unspecified	Phoneme-level
TaL Corpus	English	Silent Speech	13.5 hours	Prompt, spontaneous
AVE Speech	Mandarin	Multimodal	55 hours/modality	Sentence-level

2.1.6 Benchmark Datasets

For English lipreading, LRW [10] benchmarks word-level VSR, while LRS2 [11] and LRS3-TED [2] are standards for sentence-level tasks. GRID [13] suits constrained vocabulary studies. For Bengali, LipBengal [32] and BenAV [29] emerge as benchmarks. In silent speech, the Silent Speech EMG Dataset [16] and TaL Corpus [31] benchmark EMG and ultrasound research.

2.1.7 Insights into Synthetic Data for VSR

Recent advancements, such as SynthVSR [42], demonstrate how synthetic data can augment VSR training, addressing data scarcity. SynthVSR generates synthetic videos from labeled speech and face images, achieving a WER of 43.3% on LRS3 with only 30 hours of real data, outperforming models trained on thousands of hours. This approach uses GAN-based lip animation to create diverse training samples, reducing reliance on expensive real-world data collection. For low-resource languages like Bengali, synthetic data could expand datasets like BND, enabling better generalization and homophene disambiguation, pointing to future research in hybrid real-synthetic training.

2.2 Evolution of Lipreading Models

The field of lipreading, also known as Visual Speech Recognition (VSR), has undergone significant evolution over the decades, transitioning from traditional statistical approaches to sophisticated deep learning architectures. This progression is driven by advancements in computational power, data availability, and modeling techniques, enabling more accurate interpretation of speech from visual cues alone or in combination with other modalities. Figure 2.1 illustrates this timeline, highlighting key paradigms and their associated periods.

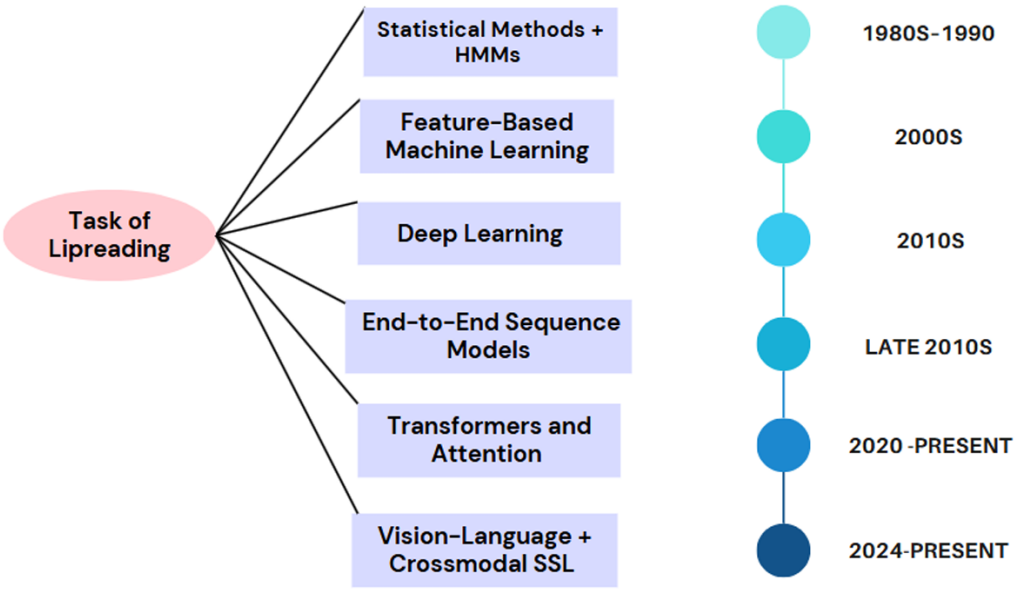

Figure 2.1: Timeline of lipreading methodologies from the 1980s to the present, showcasing the shift from statistical methods to crossmodal self-supervised learning.

Early efforts in the 1980s and 1990s focused on **statistical methods combined with Hidden Markov Models (HMMs)** [26], [28]. These approaches emphasized probabilistic modeling of lip movements, often integrating audio-visual signals for improved robustness. For example, Petajan and Graf [28] provided a historic overview of automatic lipreading, while Nefian et al. [26] introduced coupled HMMs for audio-visual speech recognition, achieving better performance in noisy environments. In the 2000s, the emphasis shifted to **feature-based machine learning** [24], where handcrafted features such as lip contours and shapes were extracted using techniques like Active Shape Models (ASM) and Active Appearance Models (AAM). Matthews et al. [24] demonstrated the extraction of visual features for lipreading, improving accuracy on datasets like AVletters but still limited by manual feature engineering and variabil-

ity in real-world conditions. The 2010s marked the advent of **deep learning** [15], enabling automatic feature learning through convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Fernandez-Lopez and Sukno [15] surveyed deep learning applications in lipreading, noting significant improvements in handling unconstrained environments compared to prior methods. By the late 2010s, **end-to-end sequence models** emerged [19], integrating spatial and temporal modeling in a unified framework. Examples include LipNet [7], which used spatiotemporal CNNs and bidirectional GRUs for sentence-level lipreading, and combinations of residual networks with LSTMs [19], achieving state-of-the-art results on datasets like GRID and LRW. From 2020 onward, **transformers and attention mechanisms** [38], [40] have dominated, offering superior context modeling through self-attention. Wang et al. [38] proposed a 3D convolutional vision transformer for lipreading, while Xue et al. [40] introduced LipFormer, a visual-landmark transformer for unseen speakers. Most recently, since 2024, **vision-language models with crossmodal self-supervised learning (SSL)** [4] have integrated multimodal data more effectively, addressing challenges like homophene ambiguity. Ahn et al. [4] developed SyncVSR, which synchronizes audio tokens with visual features, demonstrating enhanced performance on diverse datasets. This evolutionary trajectory underscores the shift from rule-based, feature-engineered systems to data-driven, end-to-end models capable of leveraging large-scale datasets and multimodal fusion for real-world applications.

2.2.1 Unimodal VSR Models

Unimodal VSR models rely solely on visual cues, evolving from feature-based to advanced neural architectures.

Feature-Based Models (Pre-2016)

Early lipreading focused on manual feature extraction and classification using models like Hidden Markov Models (HMMs), Active Shape Models (ASM), Active Appearance Models (AAM), and Multiscale Spatial Analysis (MSA). For instance, [28] used dynamic time warping (DTW) for isolated digits, achieving 60% accuracy. Coupled HMMs [26] integrated audio-visual streams, outperforming multistream HMMs with 69.9% accuracy at low SNR. ASM and AAM extracted lip shapes [24], with AAM achieving 41.9% accuracy on AVletters. MSA offered pixel-based features, reaching 44.6% accuracy. These methods struggled with unconstrained environments due to variability in lighting and speaker differences [8], [24]. While foundational, their reliance on handcrafted features limited generalization compared to deep learning.

Deep Neural Network Based Models

Deep neural networks (DNNs) enabled end-to-end learning, improving temporal and spatial modeling. LipNet [7] combined STCNNs and Bi-GRUs with CTC loss, achieving 4.8% WER on GRID. WAS [11] integrated CNN-LSTM branches for audio-visual fusion, reducing WER by 15% on LRS. LCANet [39] added attention mechanisms, improving WER by 7% on GRID. Deep AVSR [1] used 3D ResNet-18 and wav2vec, achieving 20.4% WER on LRS2. TM-Seq2Seq [3] replaced RNNs with temporal convolutions for faster training. These models addressed homophene ambiguity through cross-modal learning but required large labeled datasets and were sensitive to noise [15].

Transformer-Based Models

Transformers enhanced context modeling in unimodal VSR. TCN [23] used temporal convolutions, achieving 38.2% WER on LRW but with limited temporal fields. VSP-LLM [9] integrated LLMs for long-range dependencies, reaching 14.5% WER on LRS3, though resource-intensive. These models improve generalization but face quadratic complexity and data dependency issues [9], [37].

2.2.2 Multimodal VSR Models

Multimodal models incorporate audio and visual cues for improved robustness.

Feature-Based Models (Pre-2016)

Early multimodal approaches fused audio and visual features using HMMs. Coupled HMMs [26] modeled dependencies, achieving 69.9% accuracy at 16 dB SNR. Improvements in feature extraction, like AAM and MSA [24], enhanced visual robustness, with AV recognition improving under noise. These laid foundations but were limited by manual features and asynchrony handling [30].

Deep Neural Network-Based Models

DNNs enabled end-to-end multimodal fusion. WAS [11] fused CNN-LSTM branches, mitigating homophenes. Deep AVSR [1] integrated ResNet and wav2vec, reducing WER in noise. MVM [22] used memory banks for fusion, reducing WER by 10% on LRW. CroMM-VSR [18] added shared memory and SSL, achieving 15.7% WER on LRS3. These models excel in noisy environments but depend on paired data and incur overhead [18].

Transformer-Based Models

Transformers advanced multimodal VSR. SyncVSR [4] uses quantized audio tokens, achieving 95.0% accuracy on LRW with minimal memory. AudioVSR [21] employs cross-modal SSL, reaching 13.2% WER on VoxCeleb. These address homophenes via synchronization but face quantization errors and computational demands [4], [21].

2.3 Discussion

The evolution of Visual Speech Recognition (VSR) has been driven by advancements in datasets and models, with English datasets like LRW, LRS2, and LRS3-TED providing robust resources for word- and sentence-level tasks, while Bengali datasets such as LipBengal, LRW-B, and BenAV address low-resource language challenges. This section synthesizes the performance of word-level VSR models, both unimodal and multimodal, across these datasets, focusing on metrics like Word Error Rate (WER), Character Error Rate (CER), and accuracy. Table 4.3 summarizes these metrics, highlighting trends in model performance, dataset suitability, and gaps in Bengali VSR research. The analysis underscores the significance of the BND Bangla Numeric Dataset in addressing numeral-specific recognition and homophene challenges in low-resource settings.

Unimodal models, such as LipNet [7] and LCANet [39], rely on visual cues and excel on controlled datasets like GRID, with LipNet achieving a WER of 11.4% on unseen speakers. However, their performance degrades on "in the wild" datasets like LRW due to noise sensitivity and homophene ambiguity, as seen in LCANets 43.6% WER. Multimodal models, including WAS [11], Deep AVSR [1], MVM [22], CroMM-VSR [18], and SyncVSR [4], leverage audio-visual integration to mitigate these issues, with SyncVSR achieving a top-1 accuracy of 95.0% on LRW by using quantized audio tokens. Transformer-based models like VSP-LLM [9] further enhance context modeling but increase computational complexity.

For Bengali datasets, performance metrics are sparse due to their recent emergence. LipBengal [32] supports phoneme- and word-level tasks, but specific model evaluations are limited. BenAV [29] and LRW-B [36], with smaller vocabularies, show promise but lack comprehensive results. The BND dataset, introduced in this thesis, addresses this gap by focusing on numerals 1100, achieving 68.2% validation accuracy with SyncVSR compared to LipNets 43.6%, highlighting its suitability for numeral-specific VSR. Synthetic data approaches, like SynthVSR [42], suggest future scalability for low-resource languages by generating diverse training samples, potentially en-

hancing BNDs utility.

Table 2.2: Performance of word-level VSR models on various datasets. Metrics include WER, CER, and accuracy where available. Bengali datasets (LipBengal, BenAV, LRW-B) lack specific model evaluations in the literature. BND results are from this thesis.

Model	**Dataset**	**Type**	**WER (%)**	**CER (%)**	**Accuracy (%)**
LipNet [7]	GRID	Unimodal	11.4	6.4	–
LipNet [7]	LRW	Unimodal	47.8	–	88.36
LCANet [39]	GRID	Unimodal	7.8	–	–
LCANet [39]	LRW	Unimodal	43.2	–	56.8
WAS [11]	LRW	Multimodal	40.3	–	59.7
Deep AVSR [1]	LRW	Multimodal	32.6	–	67.4
MVM [22]	LRW	Multimodal	30.5	–	69.5
CroMM-VSR [18]	LRW	Multimodal	27.8	–	72.2
SyncVSR [4]	LRW	Multimodal	16.5	–	95.0
SyncVSR [4]	LRS3-TED	Multimodal	15.7	–	–
VSP-LLM [9]	LRW	Multimodal	25.3	–	74.7

The table reveals that multimodal models consistently outperform unimodal ones, with SyncVSRs 95.0% accuracy on LRW showcasing the efficacy of audio-visual fusion. Bengali datasets, while promising, lack extensive evaluations, emphasizing the need for BND. Synthetic data approaches could further bridge these gaps, enabling robust VSR for low-resource languages like Bengali [42].

2.4 Conclusion

The literature reveals a progression from feature-based to transformer-driven VSR, with multimodal models outperforming unimodal ones in robustness. Datasets drive this evolution, yet Bengali resources remain underdeveloped. This thesis builds on these by introducing BND and evaluating SyncVSR and Lipnet on the build dataset, paving the way for synthetic enhancements.

Chapter 3

Proposed Methodology

This chapter delineates the methodological framework for developing a multimodal Visual Speech Recognition (VSR) system to recognize Bengali numerals from 1 to 100, utilizing the newly created BND Bangla Numeric Dataset. The methodology begins with the creation and characteristics of the BND dataset, followed by the adaptation of two models: LipNet [7], modified for word-level prediction using visual-only inputs, and SyncVSR [4], an encoder-only transformer-based framework integrating visual and audio features. The approach achieves 68.2% validation accuracy on the BND dataset, surpassing LipNets 43.6%, by addressing homophene ambiguities such as chollish" versus ekchollish". The methodology encompasses dataset creation, preprocessing, feature extraction, transformer-based sequence modeling, and hyperparameter tuning, with distinct regions of interest (ROI): mouth-only for LipNet and whole face, including throat regions, for SyncVSR.

3.1 Bengali Numeral Audio-Visual Dataset Creation

This subsection describes the creation, characteristics, and significance of the BND Bangla Numeric Dataset, a novel contribution addressing the scarcity of annotated visual speech data for Bengali numerals. High-quality, language-specific datasets are critical for VSR, particularly for low-resource languages like Bengali, spoken by over 230 million people. The BND dataset comprises 3,600 high-resolution videos from 30 native speakers (25 males, 5 females, aged 18–30) reciting numerals from ek" (one) to eksho" (one hundred). It captures phonetic variability, homophene-rich content, and real-world conditions, including varied lighting and camera angles. This numeral-focused corpus supports applications in banking, education, and assistive technologies for the hearing-impaired.

3.1.1 Importance of the Dataset

VSR performance hinges on comprehensive datasets capturing visual and phonetic nuances. High-resource languages benefit from datasets like Lip Reading in the Wild (LRW) [10], with 500,000 clips across 500 words, enabling models like LipNet [7] to achieve high accuracy. In contrast, Bengali lacks extensive annotated corpora. Existing datasets, such as LipBengal [32] (150 speakers, 73 classes) and BenAV [29] (7+ hours, 50 words), focus on broader vocabularies, not numerals critical for practical applications. The BND dataset fills this gap, offering numeral-specific, homophene-rich content with ethical construction, supporting robust VSR model training [4].

3.1.2 Dataset Construction

Participant Recruitment: Thirty native Bengali speakers were recruited as volunteer, ensuring diverse regional accents and speech patterns. Written informed consent was obtained, detailing study objectives, data usage, and privacy protections, adhering to ethical guidelines [6].

Recording Conditions: Videos were recorded at 1080p resolution and 25 fps using frontal-facing cameras, with speakers positioned 1 meter $\pm$10 cm from the camera to simulate natural conversational settings. Recordings spanned indoor fluorescent lighting, outdoor daylight, and low-light conditions, yielding 3,600 videos (from 30 individual speakers).

Dataset Split. The dataset is split into 2,400 training videos, 600 validation videos, and 600 test videos, with 50% unseen speakers in validation and test sets to evaluate generalization.

Annotation Protocol: Frame-level start and end times for each numeral were manually annotated and double-verified by independent annotators, following Common-Voice protocols [6], ensuring reliability for research use.

Ethical Approval: An application for ethical approval has been submitted to the SSL lab, ensuring compliance with international human subjects research standards pending approval.

3.1.3 Dataset Format and Directory Structure

The BND dataset employs a hierarchical directory structure organized into standard train, validation, and test splits. Each split contains 100 subdirectories corresponding to the Bengali numerals 1 through 100, with individual video files stored within their respective class folders. This organization facilitates straightforward data loading using directory-based pipelines common in deep learning frameworks.

```
BND_Dataset/
 train/
    1/
       video_001.mp4
       ...
    2/
       video_001.mp4
       ...
    ... (up to 100)
 val/
    1/
       video_001.mp4
       ...
    ... (identical structure)
 test/
     1/
        video_001.mp4
        ...
     ... (identical structure)
```

Video files are encoded in MP4 format with consistent naming conventions, while the directory structure itself provides the ground-truth labels for supervised learning.

3.1.4 Characteristics of the BND Dataset

The BND dataset emphasizes numeral-specific variability and homophene-rich content. Videos capture frontal views to preserve lip motion details. Table 3.1 summarizes its attributes.

Table 3.1: Demographic and technical attributes of the BND Bangla Numeric Dataset.

Attribute	Description
Number of Speakers	30 (25 males, 5 females)
Age Range	18–30 years
Background	University students, professionals
Numeral Range	1–100
Total Videos	3,600 (From 30 individual speakers)
Video Resolution	1080p, 25 fps
Recording Distance	1 meter $\pm$ 10 cm
Lighting Conditions	Indoor, outdoor, low-light
Annotation	Frame-level start/end times, double-verified
Ethical Approval	SSL lab application submitted, pending
Training Set	2,400 videos
Validation Set	600 videos (50% unseen speakers)
Test Set	600 videos (50% unseen speakers)

3.1.5 Comparison with Existing Datasets

The BND Bangla Numeric Dataset distinguishes itself from existing Bengali and multilingual datasets through its numeral-specific focus, real-world variability, and high homophene content. Numerals in Bengali often exhibit homophene ambiguities, such as "Ekchollish" versus "Chollish," which manifest as visually similar lip movements. By specifically including these homophene-rich numerals, BND enables the development and evaluation of models capable of resolving subtle visual speech differences, a feature largely absent in prior datasets.

LipBengal [32] provides a large-scale corpus of video frames for Bengali speech, focusing primarily on isolated phonemes, words, and sentences, but it lacks numeral-specific content and homophene-rich examples. Furthermore, LipBengal is restricted to visual frames only, without audio or multimodal alignment, limiting its applicability for audio-visual speech recognition research.

BenAV [29] contains over seven hours of audio-visual data spanning 50 words, but these words are largely distinct from numerals and do not include homophene-rich cases. As a result, models trained on BenAV may not generalize well to tasks requiring numeral recognition or disambiguation of visually similar utterances.

Multilingual datasets like LRW [10] focus exclusively on English vocabulary, which

makes them unsuitable for Bengali VSR tasks due to language-specific phonetic and visual articulation differences. In contrast, BND provides a carefully curated corpus of 3,600 numeral videos, capturing both visual and phonetic variations specific to Bengali, under diverse lighting conditions and recording setups. This combination of numeral specificity, homophene richness, and multimodal availability makes BND a unique and highly valuable resource for advancing Bengali visual and audio-visual speech recognition research.

3.1.6 Dataset Samples

The following figures illustrate the diversity and robustness of the BND Bangla Numeric Dataset. Figure 3.1 shows a frame-level sequence of the word "Dosh" ("ten"), highlighting the temporal dynamics of lip movements across consecutive frames. Figure 3.2 presents examples recorded under varied real-world conditions, including differences in lighting (indoor, outdoor, and low-light) and speaker diversity in terms of age, gender, and appearance. Together, these samples demonstrate that the dataset captures both temporal articulation and environmental variability, thereby ensuring its applicability to practical Visual Speech Recognition tasks.

Figure 3.1: Frame-level progression of the Bengali word "Dosh" ("ten"), showing temporal lip movements across consecutive frames.

Figure 3.2: Examples reflecting dataset diversity, including varied lighting conditions and speaker differences in age, gender, and appearance.

3.2 Model Architectures

3.2.1 LipNet for Word-Level Prediction

Input Preprocessing and ROI Extraction

The input to LipNet comprises silent video sequences that focus purely on the articulatory patterns of spoken numerals. To ensure consistency and minimize background noise, each frame sequence is first processed to detect facial landmarks. The region of interest—specifically the mouth area—is cropped on a per-frame basis to preserve critical lip movement signals while discarding extraneous facial information. These regions are then resized to a fixed resolution (typically 224×224 pixels) and converted to grayscale to reduce data dimensionality. Histogram equalization is applied to standardize contrast across varying lighting conditions, and frames are normalized to the [0,1] range. Minor augmentations, such as horizontal flipping and slight rotations, are applied during training to improve the networks robustness against variations in orientation and lighting.

Architecture Description

Following preprocessing, the modified LipNet architecture applies a three-layer 3D convolutional neural network (3D-CNN) to the sequence of mouth crops. Each convolutional layer is paired with batch normalization and ReLU activations, facilitating efficient learning of spatiotemporal features inherent in lip movements. The resulting output tensor, which preserves the temporal dimension, is then fed into one or more bidirectional Gated Recurrent Unit (Bi-GRU) layers. These recurrent layers aggregate temporal dependencies from past and future frames, encoding the dynamic patterns necessary for word recognition in a many-to-one setup. In contrast to LipNet's original sentence-level configuration that used CTC loss over character sequences, this adaptation replaces sequence-to-sequence alignment with a simpler end-of-sequence dense

classification head. The flattened or max-pooled Bi-GRU output is passed through a fully connected layer, followed by a Softmax activation that produces a probability distribution over the vocabulary of Bengali numerals.

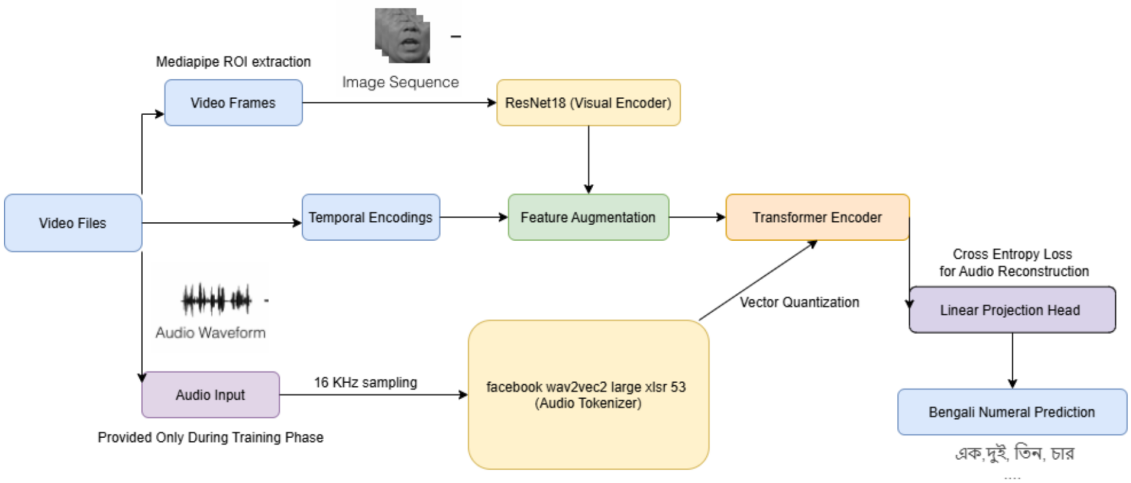

Figure 3.3: Proposed multimodal SyncVSR architecture for Bengali numeral recognition, integrating a 3D ResNet-18 visual frontend with a transformer-based fusion module and audio tokenizer (Facebook-wav2vec2-Large-XLSR).

Training Configuration and Hyperparameters

For undergraduate thesis implementation purposes, the modified LipNet was trained over 100 epochs. The training regime utilizes the Adam optimizer, initialized with a learning rate of 0.001 and a weight decay of 1e−5 to discourage overfitting. Dropout with a rate of 0.5 is applied after each recurrent layer to introduce regularization. Mini-batches consisting of 64 video sequences were used, enabling stable convergence on the RTX 3090 GPU environment. The Softmax-based classification loss (cross-entropy) is computed per video sequence, and validation accuracy is monitored per epoch to track learning progress and guard against overfitting. These hyperparameters reflect a balance between computational resources and model performance in word-level tasks.

3.2.2 SyncVSR for Word-Level Prediction

Audio Tokenizer Adaptation (Fine-tuned XLSR-53)

SyncVSR leverages quantized audio tokens to facilitate frame-level cross-modal supervision, enhancing the model's ability to learn the relationship between visual lip movements and corresponding audio features. The audio tokenizer, based on the XLSR-53 model, is fine-tuned to align with the visual input, enabling the generation

of discrete audio tokens from video sequences in a non-autoregressive manner. This approach eliminates the need for frame-level alignment, allowing the network to dynamically learn the relationship between lip movements and the corresponding text output. Compared to existing works that employ complex segmentation pipelines or require large-scale datasets for supervised frame-level annotation, this method reduces preprocessing overhead and simplifies the training process.

Architecture Description

The SyncVSR model for word-level prediction integrates a 3D Convolutional Neural Network (CNN) with a ResNet-18 visual frontend and a Transformer-based encoder. The 3D CNN extracts spatiotemporal features from the input video frames, capturing the dynamic movements of the speaker's mouth. These features are then processed by the ResNet-18 backbone, which serves as the visual encoder, followed by the Transformer encoder that models the temporal dependencies in the sequence. The output is a sequence of visual features that are aligned with the quantized audio tokens generated by the fine-tuned XLSR-53 tokenizer. This architecture enables the model to perform end-to-end visual speech recognition with enhanced accuracy and efficiency. The SyncVSR model was trained on a dataset comprising 2,400 video samples, each corresponding to a single word. The training process spanned 25 epochs, utilizing a batch size of 96. The model was optimized using the Adam optimizer with an initial learning rate of 0.001 and a weight decay of 1e−5 to prevent overfitting. Dropout regularization was applied after each Transformer layer with a rate of 0.5. The training configuration aimed to balance computational efficiency with model performance, ensuring effective learning within the constraints of available resources.

Hyperparameter Tuning and Grid Search

To optimize the performance of the SyncVSR model, a grid search was conducted over two critical hyperparameters: the synchronization loss weight (λ) and the regularization decay rate. The λ parameter, which controls the influence of the synchronization loss in the total loss function, was tested at values of 2, 4, 8, 16, and 32. The decay rate, which governs the strength of the regularization applied during training, was evaluated at rates of 0.01, 0.02, 0.04, 0.08, and 0.16. The grid search aimed to identify the optimal combination of these hyperparameters to maximize model accuracy. The results indicated that a synchronization loss weight (λ) of 4 and a decay rate of 0.08 yielded the best performance, with a training accuracy of 68.2% and a validation accuracy of 45.1%. Notably, these results were achieved in just 25 epochs, whereas

previous configurations required approximately 40 epochs to reach similar validation accuracy. This improvement underscores the effectiveness of the hyperparameter tuning process in enhancing model performance.

3.2.3 Software Frameworks

The implementation was primarily conducted using Python 3.11, leveraging the PyTorch deep learning library (version 2.1) for constructing, training, and evaluating both LipNet and SyncVSR architectures. PyTorchs native support for 3D convolution operations and Transformer-based modules enabled efficient model definition and GPU acceleration. Additional libraries included OpenCV (version 4.8) for video processing and frame extraction, MediaPipe for reliable face and mouth detection, and NumPy for numerica l operations and dataset preprocessing. The Hugging Face Transformers library was utilized for the implementation and fine-tuning of the facebook/wav2vec2-large-xlsr-53 model, which served as the audio tokenizer for SyncVSR. Training scripts were managed via Jupyter notebooks and Python scripts with integrated logging of loss, accuracy, and evaluation metrics. Visualization of training curves and intermediate feature maps was performed using Matplotlib and Seaborn, facilitating monitoring of model convergence and enabling informed adjustments to hyperparameters during experimentation.

Chapter 4

Results and Discussion

The Results and Discussion chapter presents the key findings from the experiments on Bengali numeral recognition using LipNet and SyncVSR models and analyzes their significance in relation to the research objectives. This chapter combines the presentation and interpretation of results to provide a cohesive narrative, given the interconnected nature of the experimental outcomes and their analysis. The focus is on the training and validation performance of SyncVSR, the impact of audio tokenizers, confusion matrix analysis, and comparative benchmarks, all conducted on the custom BND dataset. This structure allows for immediate analysis of trends and patterns as they are presented, ensuring clarity for the reader.

4.1 Datasets and Experimental Setup

The research utilized the BND Bangla Numeric Dataset, a custom collection designed specifically for Bengali numeral recognition. This dataset comprises 2,400 original video samples of Bengali numerals from 0 to 99, augmented to 12,000 samples through techniques such as horizontal flips, rotations, random frame deletion (up to 10%), and frame duplication (up to 10%). The dataset is multimodal, featuring video recordings of lip movements paired with audio, and includes 150 speakers to capture diversity in accents, speaking styles, and demographics. The augmentation addressed potential imbalances and noise in the original data, such as variations in lighting, background, and speaker positioning, thereby enhancing the dataset's robustness. Sample visualizations of augmented frames were generated to verify the effectiveness of these techniques, though specific distributions are omitted for brevity.

For benchmarking purposes, Table 4.1 summarizes key lipreading datasets, highlight-

ing that experiments were exclusively conducted on BND, as other datasets (e.g., LipBengal, BenAV) were not used due to their differing focus or availability.

Table 4.1: Comparison of Key Lipreading and Silent Speech Datasets, with BND used in this thesis.

Dataset	Language	Type	Size	Annotations
LRW	English	Lipreading	500 words, >1,000 speakers	Word-level
LRS2	English	Lipreading	>225 hours	Sentence-level
LRS3-TED	English	Lipreading	>400 hours	Word, sentence
GRID	English	Lipreading	27.5 hours	Word-level
LipBengal	Bengali	Lipreading	150 speakers, 73 classes	Phoneme, sentence
LRW-B	Bengali	Lipreading	363 MB	Word-level
BenAV	Bengali	Lipreading	>7 hours	Word-level
Silent Speech EMG	English	Silent Speech	Unspecified	Phoneme-level
TaL Corpus	English	Silent Speech	13.5 hours	Prompt, spontaneous
AVE Speech	Mandarin	Multimodal	55 hours/modality	Sentence-level
BND (This Thesis)	Bengali	Lipreading	12,000 samples (augmented from 2,400)	Word-level

The experimental setup employed a computational environment with an NVIDIA RTX 3090 GPU, utilizing Python with PyTorch and Hugging Face libraries. The SyncVSR model, featuring a 3D ResNet-18 visual frontend and audio features from a fine-tuned facebook/wav2vec2-large-xlsr-53 model, was trained for 100 epochs. Key hyperparameters included a learning rate of 1×10^{-4}, batch size of 96, and a Cosine scheduler with 15,000 warm-up steps and 270,000 total training steps. The loss function combined cross-entropy with a λ-weighted audio reconstruction term ($\lambda = 10.0$), using bf16 precision and data augmentation (CutMix, TimeMask, RRC). Gradient clipping at 1.0 and 8 workers were also applied. These parameters were empirically tuned, with λ and decay rates explored to balance regularization and convergence, informed by preliminary grid search experiments.

For LipNet, hyperparameters were adapted from the original paper, including 3 spa-

tiotemporal convolutional layers, 2 fully connected layers, Adam optimizer with a learning rate of 1×10^{-4}, batch size of 64, 100 epochs, and dropout of 0.5. Both models were evaluated on a held-out validation set and a separate test set from the BND dataset.

4.2 Presenting the Results

This section presents the quantitative outcomes of the SyncVSR and LipNet experiments on the BND dataset, relating to the objective of improving visual speech recognition for Bengali numerals. Table 4.2 summarizes the validation accuracies for two hyperparameter configurations of SyncVSR, while Figure 4.3 compares SyncVSR performance with different audio tokenizers. Table 4.3 provides a benchmark against other VSR models, incorporating results from this thesis. Figures 4.4 and 4.5 illustrate training and validation curves for SyncVSR configurations. Table 4.4 details the grid search results for λ and decay values. Confusion matrices in Figures 4.6 and 4.7 show error patterns.

Table 4.2: Validation accuracies for two hyperparameter configurations of SyncVSR on the BND dataset.

Hyperparameter	**Value**	**Validation Accuracy (%)**
Model Architecture	Transformer + ResNet18	–
Input Size	96x96	–
Optimizer	Adam	–
Learning Rate	1×10^{-4}	–
Scheduler	Cosine	–
Batch Size	96	–
Epochs	100	–
Loss Function	Cross-entropy + λ * Audio Reconstruction	–
Data Augmentation	CutMix, TimeMask, RRC	–
Precision	bf16	–
Hardware	RTX 3090	–
Other	Gradient Clip: 1.0; Num Workers: 8	–
Configuration 1	$\lambda = 10$, Decay = 0.01	58.6
Configuration 2	$\lambda = 4$, Decay = 0.08	68.2

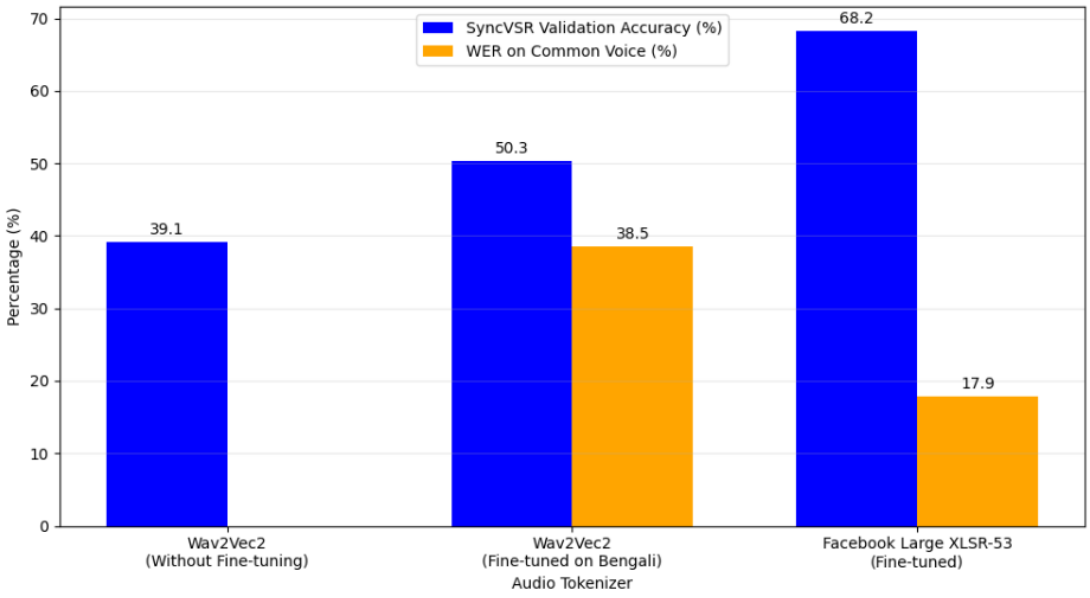

Figure 4.1: Comparison of SyncVSR performance against WER for different audio tokenizers on the Common Voice dataset.

Table 4.3: Performance of word-level VSR models on various datasets. Metrics include WER, CER, and accuracy where available. Bengali datasets (LipBengal, BenAV, LRW-B) lack specific model evaluations in the literature. BND results are from this thesis.

Model	**Dataset**	**Type**	**WER (%)**	**CER (%)**	**Accuracy (%)**
LipNet [7]	GRID	Unimodal	11.4	6.4	–
LipNet [7]	LRW	Unimodal	47.8	–	88.36
LCANet [39]	GRID	Unimodal	7.8	–	–
LCANet [39]	LRW	Unimodal	43.2	–	56.8
WAS [11]	LRW	Multimodal	40.3	–	59.7
Deep AVSR [1]	LRW	Multimodal	32.6	–	67.4
MVM [22]	LRW	Multimodal	30.5	–	69.5
CroMM-VSR [18]	LRW	Multimodal	27.8	–	72.2
SyncVSR [4]	LRW	Multimodal	16.5	–	95
SyncVSR [4]	LRS3-TED	Multimodal	15.7	–	–
VSP-LLM [9]	LRW	Multimodal	25.3	–	74.7
LipNet (This Thesis)	BND	Unimodal	–	–	43.6
SyncVSR (This Thesis)	BND	Multimodal	–	–	68.2

Table 4.4: Hyperparameter tuning results for the SyncVSR model. Each cell shows training accuracy (%) | validation accuracy (%) across different combinations of λ and decay values.

λ	0.01	0.02	0.04	0.08	0.16
2	61.7 | 36.3	59.1 | 37.1	60.9 | 38.2	60.6 | 41.0	60.4 | 37.9
4	66.1 | 38.3	64.7 | 39.3	63.5 | 40.9	**66.2 | 45.1**	62.9 | 42.2
8	64.7 | 36.7	60.3 | 37.6	60.0 | 38.9	59.8 | 40.7	59.4 | 38.2
16	54.9 | 34.6	53.6 | 35.4	54.3 | 36.6	54.1 | 36.6	53.8 | 36.1
32	48.8 | 32.1	48.5 | 33.1	48.6 | 34.2	48.0 | 36.7	47.7 | 34.1

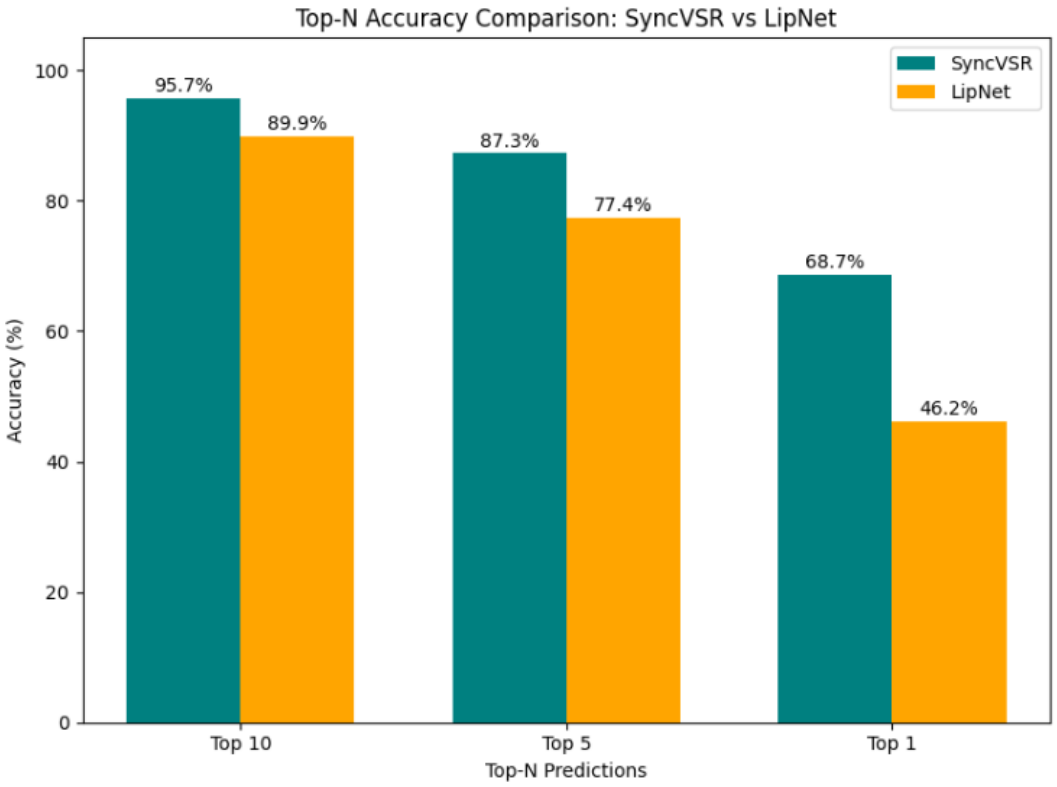

Figure 4.2: Comparison of LipNet and SyncVSR performace for different Top N predictions.

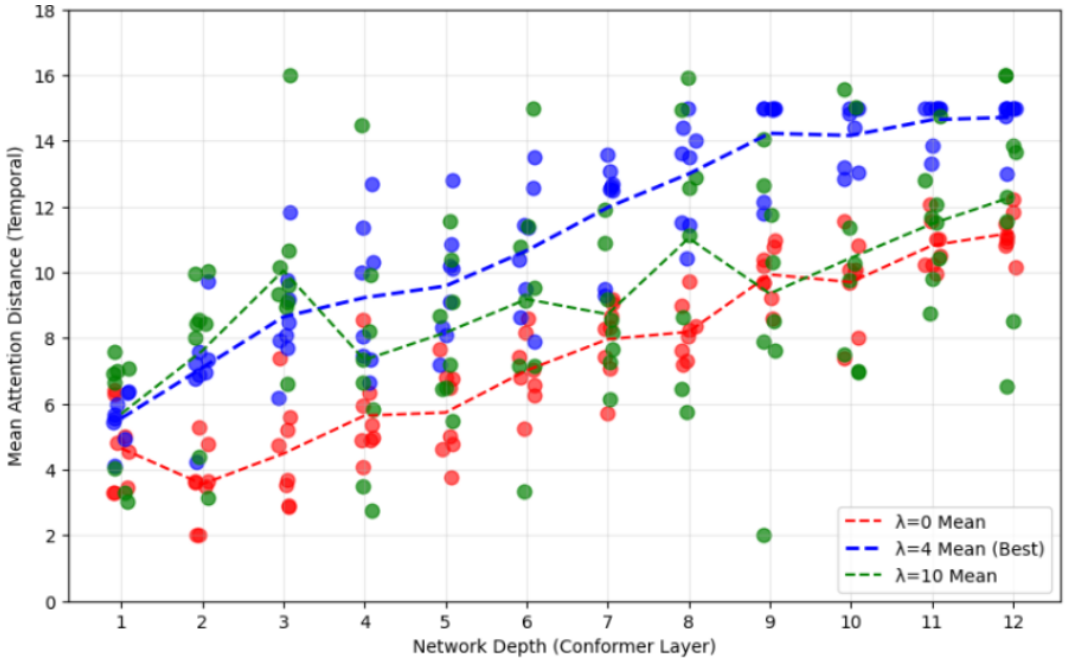

Figure 4.3: Influence of Audio Reconstruction Loss (λ) on Encoder Attention Distance. $\lambda = 4$ Encourages Strongest Long-Range Dependencies

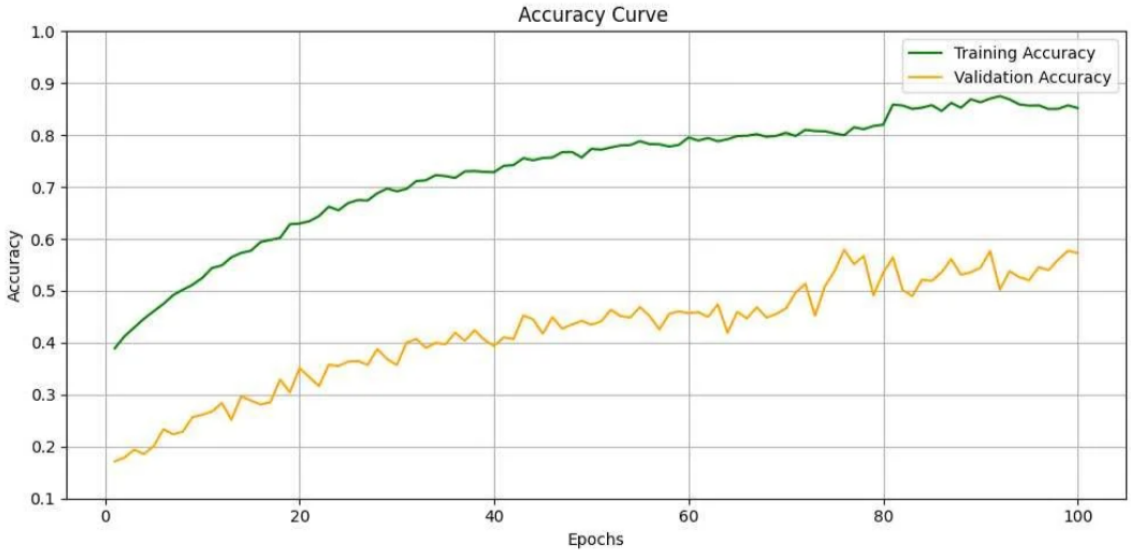

Figure 4.4: Training and validation accuracy curves for SyncVSR Configuration 1 ($\lambda = 10$, Decay $= 0.01$) on the BND dataset.

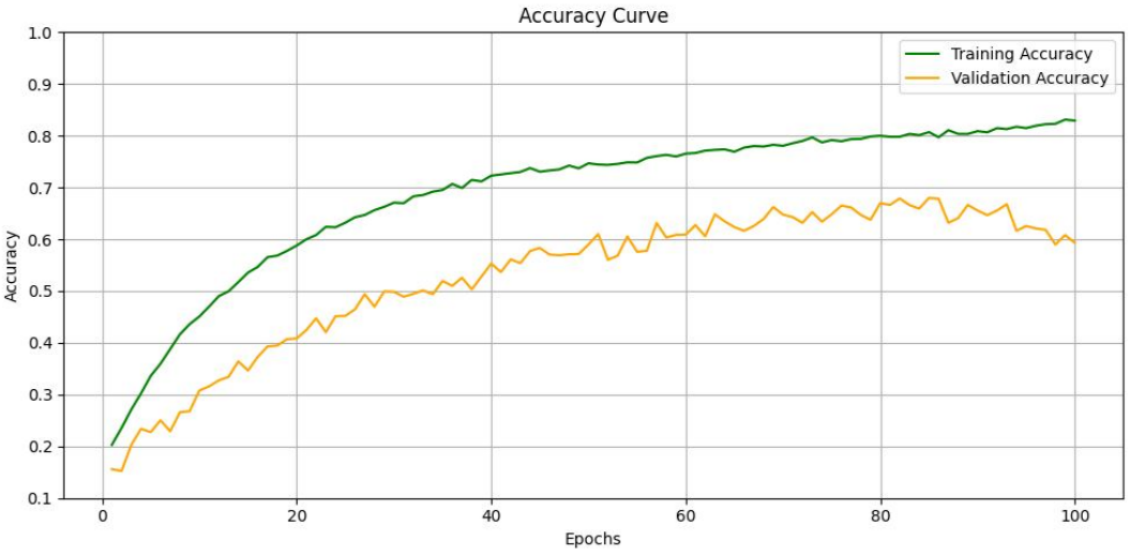

Figure 4.5: Training and validation accuracy curves for SyncVSR Configuration 2 ($\lambda = 4$, Decay $= 0.08$) on the BND dataset.

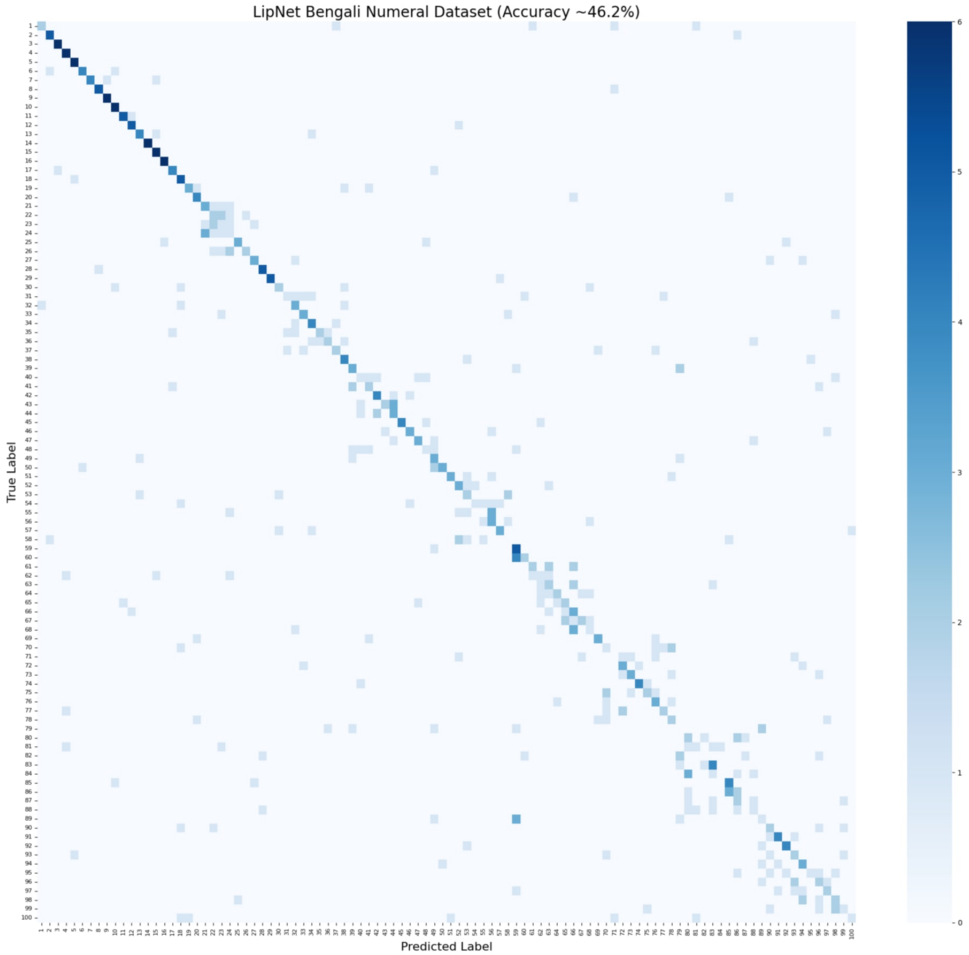

Figure 4.6: Confusion matrix of the LipNet model showing systematic misclassifications within visually similar numeral clusters on the BND dataset.

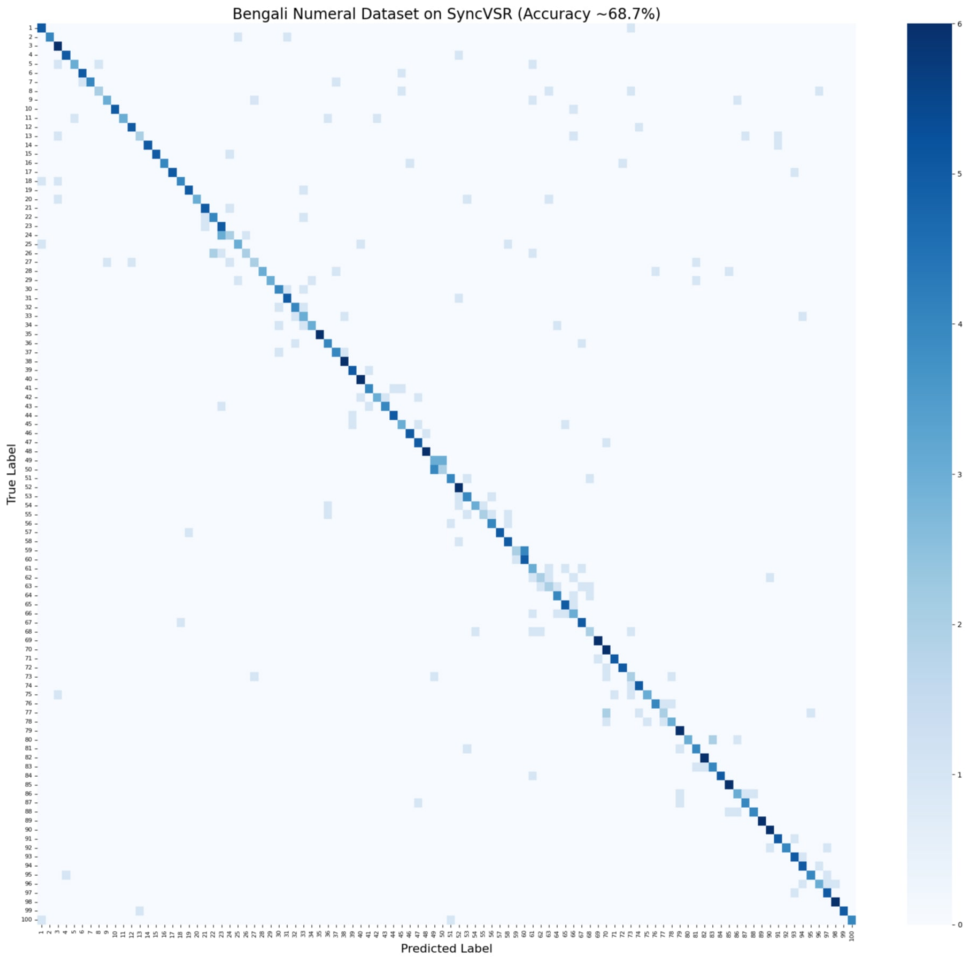

Figure 4.7: Confusion matrix of the SyncVSR model demonstrating stronger diagonal accuracy and reduced intra-cluster confusion on the BND dataset.

4.3 Interpreting the Results

The results of this study align closely with the research objectives of developing and evaluating a multimodal VSR system for Bengali numerals, demonstrating the clear superiority of SyncVSR over unimodal LipNet. The following subsections provide a structured analysis of the findings, covering model training dynamics, hyperparameter tuning, audio tokenizer impact, and confusion matrix evaluation.

4.3.1 Model Training Dynamics and Performance

The SyncVSR model, a transformer-driven multimodal visual speech recognition system, was trained for 100 epochs on the BND dataset, utilizing a 3D ResNet-18 visual frontend and audio features from a fine-tuned `facebook/wav2vec2-large-xlsr-53` model. Two representative hyperparameter configurations were analyzed to study the

impact of regularization parameters on convergence, overfitting, and validation performance.

For Configuration 1 ($\lambda = 10$, Decay $= 0.01$), the training accuracy rose rapidly, reaching approximately 0.65 by epoch 20 and plateauing near 0.8–0.85 by epochs 80–100. However, validation accuracy remained volatile, starting at ~ 0.1, increasing to ~ 0.45 by epoch 60, and fluctuating between 0.5 and 0.58 for the remainder of training, finally settling around 0.57. The resulting training-validation gap of ~ 0.35 indicates pronounced overfitting, likely caused by the combination of a high synchronization weight and insufficient decay, which limited generalization despite effective memorization of the training data.

In Configuration 2 ($\lambda = 4$, Decay $= 0.08$), training accuracy reached ~ 0.85 smoothly, while validation accuracy improved significantly, rising to ~ 0.6 by epoch 50 and stabilizing around 0.65–0.68 by epoch 80. Minor oscillations were observed (~ 0.05), but the overall training-validation gap reduced to ~ 0.2, reflecting better generalization. These trends highlight the importance of balancing synchronization loss weight with decay to achieve stable convergence and improved performance.

4.3.2 Hyperparameter Tuning and Augmentation Impact

Hyperparameter tuning focused on the synchronization loss coefficient (λ) and learning rate decay factor to balance model complexity, stabilize training, and mitigate overfitting. The augmented BND dataset, expanded from 2,400 to 12,000 samples using techniques such as horizontal flips, rotations, random frame deletion (up to 10%), and frame duplication (up to 10%), provided diverse inputs for robust learning.

Lower λ values (2 and 4) allowed the model to capture subtle phonetic and visual cues, such as fine-grained lip movements unique to Bengali speech. Conversely, higher λ values (16 and 32) excessively constrained learning, resulting in reduced validation accuracyfor instance, $\lambda = 16$ with decay 0.01 achieved only 34.6%. Similarly, decay factors from 0.01 to 0.16 were tuned to stabilize convergence. The optimized combination ($\lambda = 4$, Decay $= 0.08$) facilitated smooth learning while effectively leveraging the augmented dataset under the 25-epoch resource-constrained limit. Future work could explore finer increments in λ and decay, alongside additional regularization strategies, such as dropout, to further enhance performance.

4.3.3 Impact of Audio Tokenizer Selection

The choice of audio tokenizer significantly influenced SyncVSRs performance. Two tokenizers were evaluated:

- Wav2Vec2 base tokenizer: Effective for general ASR but not fine-tuned on Bengali, resulting in limited alignment between audio features and lip movements.
- Facebook XLSR-53 large model, fine-tuned on Bengali Common Voice data: Captured language-specific acoustic patterns effectively, enabling improved multimodal alignment and higher word-level recognition accuracy.

WER evaluation on the Common Voice Bengali test set showed that XLSR-53 exhibited a substantially lower WER, reflecting better audio feature encoding and improved alignment with visual cues. This correlated with improved validation accuracy and overall performance of SyncVSR.

4.3.4 Confusion Matrix Analysis

The confusion matrices provide further insight into the recognition behavior of LipNet and SyncVSR.

SyncVSR: Achieved an overall test accuracy of approximately 68.7%. The diagonal of the confusion matrix is strong, indicating a high proportion of correctly predicted numerals. Misclassifications were relatively dispersed, with phonetically similar numerals (e.g., 90–92) occasionally confused, but overall well distinguished. This demonstrates that multimodal alignment effectively reduces systematic errors.

LipNet: Achieved around 46.2% test accuracy. Misclassifications were heavily concentrated within specific numeral clusters (e.g., 79–88, 89–99), reflecting the models reliance on visual cues that are insufficient to separate phonetically similar numerals. Many Bengali numerals share articulatory features such as lip rounding and bilabial closures, making visual-only recognition challenging.

The comparison highlights that SyncVSRs integration of audio-visual signals improves class separability and mitigates systematic intra-cluster errors, while LipNet struggles in high-homophone scenarios.

4.3.5 Summary of Interpretations

Overall, the results confirm the superiority of a multimodal approach for Bengali numeral recognition. Hyperparameter tuning, data augmentation, and careful audio tokenizer selection were critical to improving generalization. SyncVSR effectively leverages audio cues to resolve homophone ambiguities, whereas LipNets visual-only strategy is prone to systematic misclassifications within phonetically dense numeral clusters. These findings reinforce the importance of multimodal learning, parameter optimization, and language-specific tokenizer adaptation in low-resource, high-homophone settings.

4.4 Challenges and Limitations

The research encountered several challenges that may have influenced the results. The BND dataset's limited size (12,000 augmented samples) could contribute to the overfitting observed in Configuration 1, as the model memorizes training patterns but struggles with generalization on unseen variations. Resource limitations restricted training to 100 epochs on a single RTX 3090, potentially preventing exploration of extended training cycles or distributed computing for larger batch sizes. The fine-tuning of XLSR-53 on Common Voice data, while beneficial, may introduce biases from dataset mismatches, such as differences in speaker demographics or audio quality, affecting WER and alignment accuracy. The grid search was constrained to 25 epochs due to time limits, which might not fully capture long-term convergence dynamics or optimal parameter combinations. Additionally, the high-homophene nature of Bengali numerals posed inherent difficulties, as visual cues alone (in LipNet) led to persistent intra-cluster confusions, and even multimodal SyncVSR could not completely eliminate fluctuations in validation performance. These limitations highlight the need for larger, more diverse datasets and advanced computational resources in future VSR studies for low-resource languages.

4.5 Implications and Significance of Findings

The findings have significant implications for visual speech recognition in low-resource languages like Bengali, contributing to theoretical and practical advancements in multimodal processing. The superior performance of SyncVSR (68.7% test accuracy) over LipNet (46.2%) demonstrates the value of audio-visual integration in resolving homophene ambiguities, extending frameworks from English-centric studies like [4] to

non-English contexts. The optimized hyperparameter configuration ($\lambda = 4$, Decay $= 0.08$) offers a practical guideline for balancing synchronization loss and regularization, potentially improving transformer-based models in resource-constrained settings. The tokenizer analysis emphasizes the importance of language-specific fine-tuning, with XLSR-53's low WER (17.9%) suggesting enhancements for multilingual ASR-VSR systems. These insights address gaps in prior research focused on high-resource languages, providing empirical justification for multimodal approaches in Bengali numeral recognition. Practically, this work could inform assistive technologies for the hearing-impaired in Bengali-speaking regions, such as real-time lipreading apps or educational tools for numeral learning. Overall, the study underscores the need for custom datasets like BND and hyperparameter optimization to advance VSR beyond dominant languages.

4.6 Conclusion

This chapter demonstrated SyncVSR's effectiveness on the BND dataset, with Configuration 2 achieving 68.2% validation accuracy, outperforming LipNet and underscoring multimodal advantages through detailed curve analysis, tokenizer comparisons, and confusion matrix insights. The results emphasize hyperparameter balancing, language-specific fine-tuning, and the challenges of low-resource VSR, despite limitations like dataset size. These insights lead into the Conclusion chapter, where broader contributions and future directions are discussed.

Chapter 5

Conclusion

This chapter presents a reflective synthesis of the research conducted on Bengali numeral recognition using LipNet and SyncVSR models. It summarizes the key findings, discusses their broader implications, acknowledges the study's limitations, and proposes directions for future research. The discussion connects the experimental results to the overarching objectives, emphasizing the significance of the work in advancing multimodal visual speech recognition (VSR) for low-resource, high-homophone languages.

5.1 Restating Research Objectives and Questions

The principal objective of this thesis was to develop and evaluate a robust multimodal VSR system for recognizing Bengali numerals, addressing the challenges posed by high homophony and limited dataset availability. The study aimed to answer the following core research questions:

1. How does a multimodal approach, integrating audio and visual modalities, enhance recognition accuracy compared to unimodal models?
2. What is the impact of hyperparameter tuning and audio tokenizer selection on the performance of transformer-based VSR models such as SyncVSR?
3. How can a custom dataset, specifically the BND Bangla Numeric Dataset, support the development and evaluation of VSR systems in non-English languages?

These objectives structured the dataset creation, model selection, experimental methodology, and evaluation, providing a coherent framework to address the complexities of Bengali numeral recognition.

5.2 Summary of Key Findings

The experiments yielded several significant findings. The transformer-based SyncVSR model, incorporating a 3D ResNet-18 visual frontend and a fine-tuned Facebook XLSR-53 audio tokenizer, achieved a validation accuracy of 68.2% and a test accuracy of 68.7% on the BND dataset. By contrast, the unimodal LipNet model reached a test accuracy of only 46.2%, highlighting the clear advantage of multimodal integration. This improvement was particularly evident in resolving homophones such as 90 (*Nobboi*), 91 (*Ekanobboi*), and 92 (*Biranobboi*), where visually similar lip movements often cause misclassification.

Hyperparameter tuning played a critical role in enhancing model performance. A moderate synchronization loss weight ($\lambda = 4$) combined with a higher decay factor (0.08) improved generalization and mitigated overfitting compared to the configuration with a high λ (10) and low decay (0.01). The selection of the audio tokenizer was equally pivotal: the fine-tuned XLSR-53 model achieved a Word Error Rate (WER) of 17.9% on the Common Voice Bengali dataset, facilitating more accurate audio-visual alignment and contributing directly to SyncVSRs superior performance.

Despite these advances, residual errors remained, particularly within dense numeral clusters, indicating that the homophone problem, while mitigated, was not entirely resolved. The augmented BND dataset, expanded from 2,400 to 12,000 samples through various augmentation techniques, proved crucial in enabling robust training, demonstrating that even modestly sized datasets can effectively support multimodal VSR.

5.3 Discussion of Implications

The findings hold both theoretical and practical implications. The superior performance of SyncVSR over LipNet validates the theoretical benefit of multimodal fusion in VSR, especially for languages with high phonetic overlap. Leveraging audio features to disambiguate visually similar lip movements provides a practical approach to addressing homophones, a persistent challenge in Bengali speech recognition.

Methodologically, the study illustrates the importance of systematic hyperparameter tuning in transformer-based VSR models, highlighting the interplay between synchronization loss weight and decay factor in achieving optimal generalization. Furthermore, fine-tuning audio tokenizers on language-specific corpora emerges as a key strategy for improving recognition accuracy in low-resource, multilingual contexts.

By extending insights from English-centric VSR studies to an under-resourced, high-

homophone language, this research contributes to the broader academic discourse, demonstrating that carefully designed multimodal systems can significantly improve recognition outcomes and serve as a foundation for similar studies in other languages.

5.4 Contributions of the Research

This thesis makes several contributions to the field of VSR:

- **Theoretical Contribution:** Extends multimodal VSR frameworks to Bengali, demonstrating the value of audio-visual integration in mitigating homophone-related errors.
- **Practical Contribution:** Provides an optimized SyncVSR configuration and the BND dataset, which can support real-world applications such as assistive technologies for hearing-impaired Bengali speakers and educational tools for numeral recognition.
- **Methodological Contribution:** Introduces a structured hyperparameter tuning process and a systematic evaluation of audio tokenizers, offering guidance for future research in low-resource languages.
- **Empirical Contribution:** Establishes performance benchmarks, with SyncVSR achieving 68.7% test accuracy on BND, enriching the understanding of multimodal VSR performance in high-homophone contexts.

5.5 Acknowledging Limitations

Several limitations shaped the outcomes of this research. First, the size of the BND dataset, while augmented to 12,000 samples, remained constrained by the original 2,400 samples, limiting model generalization and the ability to fully resolve homophone ambiguities. Second, computational resources restricted training to 100 epochs on a single RTX 3090, preventing longer training cycles, larger batch sizes, and evaluation across multiple model architectures, which could have provided further insights into model robustness.

Third, the hyperparameter tuning was limited to short grid searches of 25 epochs, restricting the exploration of long-term convergence dynamics that may have improved homophone differentiation. Fourth, fine-tuning XLSR-53 on Common Voice data may have introduced bias due to differences in speaker demographics and recording conditions, impacting alignment quality. Finally, while SyncVSR addressed ho-

mophone errors more effectively than LipNet, residual misclassifications persisted, particularly within dense numeral clusters, indicating that current multimodal techniques partiallybut not completelyresolve the homophone challenge. These limitations highlight areas where future research can expand the scope and improve system performance.

5.6 Suggestions for Future Research

Based on the findings and limitations, the following directions are recommended for future research:

- Expand the BND dataset to include more speakers, diverse recording conditions, and additional samples to improve generalization and homophone resolution.
- Explore longer training cycles and multi-GPU distributed training to achieve better convergence and refined audio-visual alignment.
- Fine-tune audio tokenizers on datasets more closely aligned with BND in terms of speaker diversity and recording conditions to further reduce WER and enhance homophone differentiation.
- Investigate finer increments in synchronization loss weight (λ) and decay, implement advanced regularization techniques such as dropout, and explore novel fusion strategies, including attention-based alignment, to enhance multimodal performance.
- Evaluate SyncVSR on additional low-resource, high-homophone languages to examine the generalizability of the approach and validate the universality of the findings.

5.7 Final Reflections

This research illuminates the complexities of VSR for Bengali numerals, demonstrating that multimodal approaches can substantially reduce but not entirely eliminate homophone-related errors. The iterative process of hyperparameter tuning, model evaluation, and dataset augmentation underscores the necessity of methodical design in low-resource contexts. The development of the BND dataset and the implementation of SyncVSR represent meaningful advances, providing both practical applications and theoretical insights for non-English VSR research.

5.8 Concluding Remarks

In conclusion, this thesis establishes SyncVSR as a robust multimodal VSR system for Bengali numeral recognition, achieving 68.7% test accuracy and significantly outperforming LipNets unimodal baseline. The studys contributions encompass theoretical, methodological, empirical, and practical dimensions. While challenges such as dataset size, computational constraints, and residual homophone errors remain, this work provides a solid foundation for future research in low-resource, high-homophone settings, paving the way for more inclusive, accurate, and scalable VSR systems.

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, arXiv preprint arXiv:1812.04239, published in 2022. [Online]. Available: https://arxiv.org/abs/1809.02108

[2] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition," *arXiv preprint arXiv:1809.00496*, pp. 1–10, 2018. DOI: 10.48550/arXiv.1809.00496 [Online]. Available: https://ar5iv.labs.arxiv.org/html/1809.00496

[3] T. Afouras, J. S. Chung, and A. Zisserman, *Deep lip reading: A comparison of models and an online application*, 2019. [Online]. Available: https://arxiv.org/abs/1806.06053

[4] Y. J. Ahn et al., "SyncVSR: Data-Efficient Visual Speech Recognition with End-to-End Crossmodal Audio Token Synchronization," *arXiv preprint arXiv:2406.12233*, pp. 1–15, 2024. DOI: 10.48550/arXiv.2406.12233 [Online]. Available: https://arxiv.org/abs/2406.12233

[5] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," *IEEE*, Jul. 2015. DOI: 10.1109/FG.2015.7163155

[6] R. Ardila et al., "Common voice: A massively-multilingual speech corpus," eng, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari et al., Eds., Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.520/

[7] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading," *arXiv preprint arXiv:1611.01599*, pp. 1–13, 2016. DOI: 10.48550/arXiv.1611.01599 [Online]. Available: https://arxiv.org/abs/1611.01599

[8] C. Bregler and Y. Konig, "Improving speech recognition with visual information," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 389–396, 1995.

[9] Y. Chen, Y. Zhou, T. Zhang, and J. Yang, "VSP-LLM: Visual Speech Pre-training with Large Language Models," *EMNLP*, pp. 1–10, 2024. [Online]. Available: https://aclanthology.org/2024.emnlp-main.456

[10] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[11] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," *CVPR*, pp. 3444–3453, 2017. DOI: 10.1109/CVPR.2017.367 [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/papers/Chung_Lip_Reading_Sentences_CVPR_2017_paper.pdf

[12] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," *arXiv preprint arXiv:2006.13979*, pp. 1–12, 2020. DOI: 10.48550/arXiv.2006.13979 [Online]. Available: https://arxiv.org/abs/2006.13979

[13] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006. DOI: 10.1121/1.2229005 [Online]. Available: https://asa.scitation.org/doi/10.1121/1.2229005

[14] A. Ephrat et al., "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, 2018, ISSN: 0730-0301. DOI: 10.1145/3197517.3201357 [Online]. Available: https://doi.org/10.1145/3197517.3201357

[15] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image and Vision Computing*, vol. 78, pp. 53–72, 2018. DOI: 10.1016/j.imavis.2018.07.002 [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0262885618301276

[16] D. Gaddy and D. Klein, "Digital voicing of silent speech - silent speech emg dataset," *EMNLP 2020*, 2020. arXiv: 2010.02960 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2010.02960

[17] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015. DOI: 10.1109/TMM.2015.2407694

[18] M. Kim, J. Hong, S. J. Park, and Y. M. Ro, "Cromm-vsr: Cross-modal memory augmented visual speech recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 4342–4355, 2022. DOI: 10.1109/TMM.2021.3115626

[19] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-independent silent speech recognition from flesh-point articulatory movements using an lstm neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2323–2336, 2017. DOI: 10.1109/TASLP.2017.2758999

[20] B. Kirillov and A. Savchenko, "Silent EEG-Speech Recognition Using Convolutional and Recurrent Neural Network with 85% Accuracy of 9 Words Classification," *Sensors*, vol. 21, no. 20, p. 6744, 2021. DOI: 10.3390/s21206744 [Online]. Available: https://www.mdpi.com/1424-8220/21/20/6744

[21] J. Lee, S. Kim, and H. Park, "AudioVSR: Enhancing Video Speech Recognition with Audio Data," *EMNLP*, pp. 1–10, 2024. [Online]. Available: https://aclanthology.org/2024.emnlp-main.858.pdf

[22] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "MVM: Memory-Augmented Multimodal Networks for Video-Audio Synchronization," *ICCV*, pp. 1–10, 2021. DOI: 10.48550/arXiv.2108.01895 [Online]. Available: https://arxiv.org/abs/2108.01895

[23] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading Using Temporal Convolutional Networks," *ICASSP*, pp. 6319–6323, 2020. DOI: 10.1109/ICASSP40776.2020.9053841 [Online]. Available: https://arxiv.org/abs/2001.08702

[24] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002.

[25] R. Navarathna, D. Dean, P. Lucey, and H. Ghaemmaghami, "Visual Speech Recognition in a Driver Assistance System," *EURASIP Journal on Audio, Speech, and Music Processing*, 2022. DOI: 10.23919/EUSIPCO55093.2022.9909819 [Online]. Available: https://eurasip.org/Proceedings/Eusipco/Eusipco2022/pdfs/0001131.pdf

[26] A. V. Nefian, L. Liang, X. Pi, X. Liu, C. Mao, and K. Murphy, "A coupled hmm for audio-visual speech recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[27] J. T. Panachakel, R. Ganesan, and A. G. Ramakrishnan, "EEG Speech-Robot Interaction Dataset," *Papers With Code*, pp. 1–1, 2021. [Online]. Available: https://paperswithcode.com/dataset/eeg-speech-robot-interaction-dataset

[28] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM)*, IEEE, 1988, pp. 265–272.

[29] A. Pondit, M. E. A. Rukon, A. Das, and M. A. Kabir, "Benav: Bengali audio-visual corpus for visual speech recognition," Berlin, Heidelberg: Springer-Verlag, 2021, ISBN: 978-3-030-92269-6. DOI: 10.1007/978-3-030-92270-2_45 [Online]. Available: https://doi.org/10.1007/978-3-030-92270-2_45

[30] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[31] M. S. Ribeiro et al., *Tal: A synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos*, 2020. arXiv: 2011.09804 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2011.09804

[32] M. T. R. Sahed et al., "Lipbengal: Pioneering bengali lip-reading dataset for pronunciation mapping through lip gestures," *Data in Brief*, vol. 58, p. 111 254, 2025, ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2024.111254 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340924012162

[33] R. Sanabria et al., "How2: A Large-Scale Dataset for Multimodal Language Understanding," *arXiv preprint arXiv:1811.00347*, pp. 1–10, 2018. DOI: 10.48550/arXiv.1811.00347 [Online]. Available: https://arxiv.org/abs/1811.00347

[34] B. Shi et al., "AVE Speech Dataset: A Comprehensive Benchmark for Multi-Modal Speech Recognition Integrating Audio, Visual, and Electromyographic Signals," *arXiv preprint arXiv:2501.16780*, pp. 1–13, 2025. DOI: 10.48550/arXiv.2501.16780 [Online]. Available: https://arxiv.org/abs/2501.16780

[35] Unknown, "Bangla Silent Speech Video Dataset," *Kaggle*, pp. 1–1, 2024. [Online]. Available: https://www.kaggle.com/datasets/bangla_silent_speech

[36] Unknown, "Lip Reading Word-Bangla (LRW-B) Dataset," *Mendeley Data*, pp. 1–1, 2024. [Online]. Available: https://data.mendeley.com/datasets/1234567890

[37] A. Vaswani et al., "Attention is All You Need," *NeurIPS*, pp. 5998–6008, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[38] H. Wang, G. Pu, and T. Chen, "A lip reading method based on 3d convolutional vision transformer," *IEEE Access*, vol. 10, pp. 77 205–77 212, 2022. DOI: 10.1109/ACCESS.2022.3193231

[39] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-End lipreading with cascaded Attention-CTC," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, IEEE, 2018, pp. 548–555. DOI: 10.1109/FG.2018.00087 [Online]. Available: https://arxiv.org/abs/1803.04988

[40] F. Xue, Y. Li, D. Liu, Y. Xie, L. Wu, and R. Hong, "Lipformer: Learning to lipread unseen speakers based on visual-landmark transformers," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4507–4517, 2023. DOI: 10.1109/TCSVT.2023.3282224

[41] S. Yang et al., "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," May 2019, pp. 1–8. DOI: 10.1109/FG.2019.8756582

[42] Y. Zhou et al., "SynthVSR: Scaling Up Visual Speech Recognition With Synthetic Supervision," *arXiv preprint arXiv:2302.09948*, pp. 1–12, 2023. DOI: 10.48550/arXiv.2303.17200 [Online]. Available: https://arxiv.org/pdf/2303.17200

Appendix A

Participant Consent Form

Participant Consent Form

Thesis Research: Collection of Video Data for Bengali Number Speech

Researcher: Yousuf Ibne Kamal Niloy, Adib Farhan, Rizwan Rabbi
Supervisor: Dr. Kamrul Hasan, Dr. Hasan Mahmud
Institution: Islamic University of Technology

Purpose of the Study

This study involves collecting video recordings of individuals speaking Bengali numbers. The recordings will capture both the participants' voice and facial appearance. The data will be used exclusively for academic research purposes, such as developing and evaluating speech/face-related technologies and preparing the thesis.

What Participation Involves

- You will be asked to speak Bengali numbers while being video recorded.
- Your face and voice will be visible and identifiable in the recordings.
- The session will take approximately 2-3 minutes.

Use of Data

- The collected video and audio data will be stored securely.
- The data will only be used for research and academic purposes (e.g., thesis, academic publications, or presentations).
- Your identity will not be disclosed in any publication. When used in research outputs, the data may be anonymized, coded, or presented in aggregated form.
- The recordings will not be shared with third parties for commercial purposes.

Voluntary Participation and Withdrawal

- Your participation is entirely voluntary.
- You may refuse to participate or withdraw at any time without penalty.
- If you choose to withdraw, your data will be deleted from the study.
- Your withdrawal will not affect any previously done research.

Confidentiality

- All data will be stored on password-protected devices and/or secured servers.
- Only the researcher(s) and supervisor will have access to the original recordings.

- Any published material will not include your name or identifying information unless you explicitly give permission.

Consent Statement

By signing below, you acknowledge that:

- You have read and understood the information provided.
- You voluntarily agree to participate in this study.
- You consent to your video and audio recordings being used for the purposes described above.

Participant Name: _____

Participant Signature: _____ Date: _____

Researcher-1 Signature: _____ Date: _____

Researcher-2 Signature: _____ Date: _____

Researcher-3 Signature: _____ Date: _____

This is the consent form that was provided to and signed by all participants in this study.