Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A First CNN-MFCC Baseline for Isolated-Word Automatic Speech Recognition in Bishnupriya Manipuri: Corpus Development and System Evaluation

Dipankar Dutta

dipankar.jrt1982@gmail.com

The Assam Royal Global University

Gautam Chakraborty

NERIM group of Institutions

Bikash Baruah

The Assam Royal Global University

Atowar Ul Islam

University of Science and Technology, Meghalya

Research Article

Keywords:

Posted Date: June 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9108162/v1

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

Additional Declarations: No competing interests reported.

A First CNN-MFCC Baseline for Isolated-Word Automatic Speech Recognition in Bishnupriya Manipuri: Corpus Development and System Evaluation

, †, †, †

*, , , , , , , .
, , , , , , .
, , , , , , .
, , , , , , .

*Corresponding author(s). E-mail(s): ;
Contributing authors: ; ; ;
†

Abstract

Bishnupriya Manipuri is an Indo-Aryan language spoken by approximately 120,000 people across Northeast India and Bangladesh, classified by UNESCO as a vulnerable language. There is no annotated speech corpus for this language, and no deep learning based automatic speech recognition system has ever been created for it. This paper addresses both gaps through two primary contributions: (1) Constructed a 2,800-utterance read speech corpus spanning 28 words and 10 sentences, recorded from 10 native BM speakers in the 20–35 age group; and (2) a Convolutional Neural Network -based isolated-word recognizer trained on 20-dimensional Mel Frequency Cepstral Coefficient features, which achieves a word recognition accuracy of 95% on the held-out test set consists of 840 utterances with data split ratio 70:30. Comparative experiments against the Gaussian Mixture Model and Support Vector Machine baselines trained on identical features and the same split demonstrate that the Convolutional Neural Network model yields the highest accuracy. Ablation experiments verify that there is no overfitting and the training loss falls monotonically as the number of training iterations rises from 750 to 1500. The paper discusses both the limitations of the current constrained-vocabulary and provides a blueprint for Bishnupriya Manipuri

continuous speech recognition. Additionally, all corpus materials will be made publicly available to facilitate further research on this under-resourced language.

Keywords: Bishnupriya Manipuri, automatic speech recognition, under-resourced language, convolutional neural network, MFCC, speech corpus, isolated word recognition

1 Introduction

Automatic speech recognition (ASR) — the computational task of converting acoustic speech signals into text — has advanced dramatically over the past decade, driven primarily by deep learning architectures trained on large quantities of transcribed speech data. However, this progress has been deeply uneven across languages. High-resource languages such as English, Mandarin, and German benefit from hundreds of thousands of hours of annotated training data and a long lineage of commercial and research investment. The majority of the world's languages — estimated by the Ethnologue at 7,111 as of 2019, of which approximately 400 have more than one million speakers — remain without any functional ASR system [1]. The resulting digital language divide has profound implications for language documentation, cultural preservation, and access to speech-enabled technologies. Bishnupriya Manipuri (BM) is a member of the Indo-Aryan language family, typologically related to Oriya, Assamese, and Bengali, and spoken by approximately 120,000 native speakers distributed across the Indian states of Assam and Tripura, the Jiribam subdistrict of Manipur, the Sylhet subdivision of Bangladesh, and parts of Myanmar [1]. The language is written in the Bengali script (Bangla lipi) from left to right. BM must be carefully distinguished from Meitei — also sometimes called Manipuri — which belongs to the Tibeto-Burman family, serves as the official state language of Manipur, and has 1.8 million native speakers according to the 2011 Indian Census. The two languages are structurally distinct and mutually unintelligible, despite sharing a geographic and sociopolitical context. UNESCO classifies BM as a vulnerable language [2], reflecting the risk of intergenerational transmission failure under pressure from dominant regional languages including Bengali and Meitei.

Research on ASR for Indian languages under the under-resourced language (URL) paradigm has grown steadily, with systems reported for Assamese [3], Marathi [4], Gujarati [5], and Meitei [6]. No prior ASR system, however, has been reported for BM. Review of the research in this domain underlines that no annotated speech corpus for BM exists in the public domain, and no language processing system of any kind has previously been developed for this language. The present work is therefore a pioneering effort in BM speech technology.

This paper makes the following contributions. First, a read speech corpus of 2,800 utterances is constructed from 10 native BM speakers, covering a 28-word vocabulary embedded in 10 sentences, recorded using a mobile voice recording application, and processed with open-source audio editing software. Second, a CNN-based isolated-word recogniser trained on MFCC features is developed and evaluated on a

held-out test set, achieving a word recognition accuracy (WAcc) of 95%. Third, the system is benchmarked against Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) baselines, and an ablation study over training iteration counts is reported. Fourth, all corpus materials and model configuration details are documented to support reproducibility and future extension.

The remainder of this paper is organised as follows. Section 2 surveys related work in deep learning ASR, under-resourced language ASR, and South Asian speech technology. Section 3 provides a linguistic background for BM. Section 4 describes corpus development. Section 5 presents the methodology. Section 6 reports experimental results. Section 7 discusses findings and limitations. Section 8 concludes with future directions.

2 RELATED WORK

2.1 Deep Learning Architectures for ASR

The modern ASR landscape is dominated by deep learning-based approaches that have progressively displaced classical statistical models. Recurrent architectures, particularly Long Short-Term Memory (LSTM) networks trained with Connectionist Temporal Classification (CTC) loss [7], enabled end-to-end ASR without requiring frame-level alignment between acoustic features and transcriptions. Attention-based encoder-decoder models, introduced by Bahdanau et al. [8], further improved sequence modelling by learning soft alignments between input and output sequences. The Transformer architecture [9] subsequently became the dominant paradigm for sequence modelling, leading to the Conformer [10], which augments self-attention with convolutional modules to capture both global and local acoustic patterns.

Convolutional Neural Networks (CNNs) occupy an important niche in this landscape, particularly for fixed-vocabulary classification tasks. CNNs applied to MFCC feature maps or spectrograms function as acoustic classifiers, exploiting local time-frequency correlations through learned convolutional filters. In the year 2014, Abdel-Hamid et al. demonstrated that CNN-based acoustic models outperform standard DNN-HMM systems on phoneme recognition when applied to spectral feature maps [11]. For isolated-word recognition under constrained-vocabulary conditions — the task addressed in the present work — CNN classifiers provide a computationally efficient and interpretable baseline.

The Whisper model [12], a weakly supervised encoder-decoder transformer trained on 680,000 hours of multilingual audio, represents the current frontier of large-scale multilingual ASR. However, BM is not among Whisper's supported languages, and its zero-shot WER on closely related South Asian languages remains high, motivating the need for dedicated, resource-conscious approaches.

2.2 ASR for Under-Resourced Languages

[13] provides a comprehensive survey of ASR for under-resourced languages, identifying four principal strategies: reducing acoustic model complexity by limiting feature dimensionality; adapting models from higher-resource source languages via

cross-lingual transfer; exploiting domain-specific text corpora for language model construction; and applying data augmentation to artificially expand limited training sets. Data augmentation techniques for speech include SpecAugment [14], speed perturbation, and noise injection, all of which improve model robustness without requiring additional recordings.

Self-supervised learning has emerged as a powerful paradigm for low-resource ASR, with wav2vec 2.0 [15] demonstrating that pre-training on unlabelled audio followed by fine-tuning on as little as one hour of transcribed speech can yield competitive performance. For South Asian languages, indicWav2Vec [16] applies this framework to 23 Indian languages using a shared multilingual representation space. These approaches are directly relevant to the future trajectory of BM ASR, as they would allow a BM-specific model to be seeded from a linguistically proximate source such as Bengali.

Unsupervised feature learning approaches [17] offer an alternative path for languages lacking transcribed data entirely, relying on acoustic unit discovery to identify phone-like units without supervision. While computationally demanding, such approaches provide a viable baseline in the complete absence of labelled resources.

2.3 South Asian and Northeast Indian Language ASR

Substantial ASR research has been conducted on Indian languages in the past decade. For Assamese, Deka et al. [3] report a DNN-HMM system achieving 5.65% WER, improving over GMM-HMM (8.03% WER) and SGMM-HMM (6.58% WER) baselines. For Marathi, Paulose et al.[4] report CI-HMM at 36.8% WER and SGMM-HMM at 21.2% WER. For Gujarati, Sailor et al. [5] investigate multiple acoustic front-ends combined with TDNN-LSTM decoders, achieving a best WER of 15.58%. These results are summarised alongside additional languages in Table 11.

Most directly relevant to the present work is the large-vocabulary ASR system for Meitei The research work presented in [6] employs a Feature-space Maximum Likelihood Linear Regression (fMLLR) with a GMM baseline achieving 19.28% WER and a DNN-HMM system achieving 13.57% WER. The Meitei system demonstrates both the feasibility and the technical pathway for Northeast Indian languages, though BM's Indo-Aryan phonology distinguishes it from the Tibeto-Burman phonological structure of Meitei, necessitating a dedicated corpus and model.

2.4 Bishnupriya Manipuri: Prior Computational Work

To the best of the authors' knowledge, no prior ASR system or annotated speech corpus exists for BM. The closest related computational work is the series of studies on Sylheti — another under-resourced Indo-Aryan language of Bangladesh and Assam — by Chakraborty and Saikia [18] and Chakraborty et al. [1], who report CNN-based isolated-word recognition rates of 93% and 95% respectively on a Sylheti corpus of comparable scale. These results provide a meaningful benchmark for the BM system developed in the present work, as both languages share script, typological proximity, and data scale. The linguistic properties of BM, including its Indo-Aryan phonological inventory and Bengali script, are documented in [18], and form the basis for the language description in Section 3.

3 LANGUAGE BACKGROUND: BISHNUPRIYA MANIPURI

3.1 Genealogy and Geographic Distribution

Bishnupriya Manipuri (BM) belongs to the Indo-Aryan branch of the Indo-European language family and is typologically related to Oriya, Assamese, and Bengali. The Bishnupriya people are believed to be of Aryo-Mongolo-Dravidian origin, distinguishing them from the Meiteis, who are of Mongolian origin and whose language belongs to the Kuki-Chin branch of the Tibeto-Burman family. BM is therefore linguistically unrelated to Meitei despite the shared toponym 'Manipuri' that is sometimes applied loosely to both.

Historically, BM was spoken in the Loktak lake region of Manipur, in localities including Khangabok, Heirok, Mayang-Yumphan, Bishnupur, and Ningthaukhong. Following Burmese invasions and associated displacement in the early 18th century, the BM-speaking population dispersed to Assam, Tripura, the Jiribam subdistrict of Manipur, the Sylhet subdivision of Bangladesh, and parts of Myanmar. The 2001 Indian Census recorded approximately 120,000 native BM speakers within India. The language has two dialects — the Madoi Gang dialect and the Rajar Gang dialect — which historically corresponded to distinct geographic zones but now coexist in the same localities [18].

3.2 Script and Orthography

BM is written in the Bengali script (Bangla lipi), a left-to-right abugida. Some historical accounts suggest that BM was written in the Devanagari script prior to Meitei political dominance in Manipur, though Bengali script has been the established written form for the modern period. The script inventory used in BM is presented in Table 3 and Table 1. BM is written using the standard Bengali character set without modification; phonetic distinctions not captured by Bengali orthography are indicated through diacritics in specialised linguistic literature.

3.3 Phonological Inventory

BM possesses a vowel inventory of ten independent vowels, represented by their corresponding matras (dependent vowel signs) in running text, as detailed in Table 2. The consonant inventory, presented in Table 1, follows the standard Bengali series of voiced and unvoiced stops across five places of articulation, with aspirated counterparts, nasals, sibilants, liquids, and approximants. IPA transcriptions for each grapheme are provided in the tables. Unlike some neighbouring languages, BM is not reported to be tonal, though acoustic studies at the phonological level are lacking and represent a gap in the scholarly literature.

3.4 Sociolinguistic Status

UNESCO's Atlas of World Languages in Danger (2010) [2] classifies BM as a vulnerable language, placing it among the 197 endangered languages of India identified in

Table 1 Bishnupria Manipuri Consonants with Roman symbols and IPA transcription

Character	Romanization	IPA
ক	kô	[kɔ]
খ	khô	[k^hɔ]
গ	gô	[gɔ]
ঘ	ghô	[g^hɔ/g’ o]
ঙ	ṅô	[ŋɔ]
চ	sô	[sɔ]
ছ	shô	[ʃɔ]
জ	jô	[$d^{tʃ}$ɔ]
ঝ	zô	[z’ɔ]
ঞ	nyô	[ɲɔ]
ট	ṭô	[θɔ]
ঠ	ṭhô	[t^hɔ/θɔ]
ড	ḍô	[ɴɔ]
ঢ	ḍhô	[d’ɔ]
ণ	ṇô	[nɔ]
ত	tô	[tɔ]
থ	thô	[t^hɔ]
দ	dô	[dɔ]
ধ	dhô	[d’ɔ]
ন	nô	[nɔ]
প	pô	[pɔ]
ফ	phô	[ɸɔ]
ব	bô	[bɔ]
ভ	bhô	[b’ɔ]
ম	mô	[mɔ]
য	jô	[$d^{tʃ}$ɔ]
র	rô	[ʊɔ]
ৱ	wô	[wɔ]
ল	lô	[lɔ]
শ	śô	[ʃɔ]
ষ	ṣô	[ʃɔ]
স	sô	[ʃɔ]
হ	hô	[hɔ]
ক্ষ	kṣô	[xɔ]
ড়	ṛô	[ʊɔ]
ঢ়	ṛhô	[ʊɔ]
য়	yô	[jɔ]
ৎ	half t	[td]

that survey. The language has begun to acquire a written literary tradition, with magazines and books now published in BM, and it is taught as a subject in government primary schools in Assam. Nevertheless, its digital resource base remains minimal: no BM text corpora, lexicons, or computational tools are publicly available, and the language is absent from major multilingual NLP benchmarks and ASR evaluation frameworks.

Table 2 Bishnupria Manipuri vowels with Matra, Roman symbols and IPA transcription

Character	Matra	Roman	IPA
অ		ô	[ɔ]
আ	া	ā	[a]
ই	ি	ī	[i]
উ	ু	u	[u]
ঊ	ূ	ū	[u]
ঋ	ৃ	ri	[i]
এ	ে	e	[e/]
ঐ	ৈ	oi	[oi]
ও	ো	o	[]
ঔ	ৌ	ou	[ou/u]
ং		an	[]
ঁ		ā	[-]
ঃ		ah	[h]

Table 3 Assamese-Bengali Script inventory used to write Bishnupriya Manipuri

Category	**Characters**	**Notes**
Vowel Signs	া ি ী ু ূ ৃ ে ৈ ো ৌ	Matras
Other Diacritics	ৎ ং ঃ ঁ	
Independent Vowels	অ আ ই ঈ উ ঊ এ ঐ ও ঔ	
Consonants	ক খ গ ঘ ঙ ছ জ ঝ ঞ ট ঠ ড ঢ ণ ত থ দ ধ ন প ফ ব ম য ৰ ল শ ষ স হ ড় ঢ় য় ৱ	
Numbers	০ ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯	

4 CORPUS DEVELOPMENT

4.1 Design Rationale and Text Selection

Given the absence of any prior BM speech resource, a read speech corpus was designed as the most feasible initial step. Read speech permits controlled recording conditions, precise text-to-audio alignment, and systematic vocabulary coverage, making it appropriate for building a first acoustic model. Ten declarative sentences as presented in Table 5 were composed to cover a semantically diverse 28-word vocabulary, including nouns (bird, mango, book, curry, flower, medicine), verbs and verb phrases (fly, go, worship, eat, cook, read), and grammatical function words. All sentences were transcribed in Bengali script and romanised for annotation purposes. The considered vocabularies are listed in the Table 4.

Table 4: List of considered vocabularies in Bishnupriya Manipuri

Sl. No.	Word	Pronunciation	Meaning	Example Sentence	Translation
1	আগো	aago	a/one	আকাশাহানাত পহিয়া আগৌ ফরদের। *(Aakashhanat paiha aago forder)*	A bird is flying in the sky.
2	আহান	aahan	a/one	মি লেরিক আহান পাকরুরী। *(Me lerik ahan pakorouri)*	I am reading a book.
3	আকাশহানাত	aakashhanat	in the sky	আকাশাহানাত পহিয়া ফরদিতারা। *(Aakashhanat paiha forditara)*	Birds are flying in the sky.
4	আঔরী	aauri	am coming	মি বাজারে আউরী। *(Me bazare aauri)*	I am coming to market.
5	এরে	aree	this	এরে ফুলর বাগান এহান হবা অছে। *(Aree phoolor bagan ehan hoba ose)*	This flower garden is beautiful.
6	বাজারেত্ব	bazarotto	from market	মি বাজারেত্ব আউরী। *(Me bazarotto aauri)*	I am coming from market.
7	কলেজে	collegee	to college	সানী কলেজে যিতৈগা। *(Sunny collegee jitoiga)*	Sunny will go to college.
8	দিতারা	ditara	am giving	সাতগাঁও হসপিটলে কোভিড ভেক্সিন হাপ্তায় দিদিন দিতারা। *(Satgaon hospitale covid vaccine haptai didin ditara)*	Covid vaccine is given at Satgaon hospital twice weekly.
9	দৌ	dou	gods	সনু দৌ হমা দিরি। *(Sonu dou homa diri)*	Sonu is worshipping gods.
10	ফরদের	forder	is flying	ফরদের পাহিয়া উগরে ছা। *(Forder paiha ugore cha)*	Look at that flying bird.
11	ঘরে	gore	at home	ঘরে যাগা। *(Gore jaga)*	Go to home.
12	হিনৌ	heinou	mango	বাজারেত্ব হিনৌ আনগা। *(Bazarotto heinou aanga)*	Bring mango from market.
13	হবা	hoba	beautiful	মারতা হবা ফুল লাগের। *(Maarta hoba phool lager)*	Mother needs beautiful flowers.
14	হজক	hojok	medicine	হজক লনাত যাগা। *(Hojok lonat jaga)*	Go to purchase medicine.

Continued on next page

Table 4 continued from previous page

Sl. No.	Word	Pronunciation	Meaning	Example Sentence	Translation
15	হমা	homa	worship	মাই নিক্কা সেন্দাত দৌ হমা দিরি। *(Maai nikka sendat dou homa diri)*	Mother worships god daily at evening.
16	হৌ	hou	curry	পাবুমে মাছর হৌ ঠিকপার। *(Pabume masor hou thickpar)*	Father likes fish curry.
17	যাগা	jaga	go	তি ঘরে যাগা। *(Ti gore jaga)*	You go to home.
18	যাঔরিগা	jauriga	am going	মি কলেজে যাঔরীগা। *(Me collegee jauriga)*	I am going to college.
19	যে	je	very	ফুল যে হবা অছে। *(Phool je hoba ose)*	Flowers are very beautiful.
20	খাঔরী	khauri	am taking	মি হজক খাঔরী। *(Me hojok khauri)*	I am taking medicine.
21	লেরিক	lerik	book	মি লেরিক পাকরানী ঠিকপাউরী। *(Me lerik pakorani thickpauri)*	I like reading books.
22	মানুয়ে	manue	people	মানুয়ে দৌ হমা দিতারা। *(Manue dou homa ditara)*	People worship gods.
23	মি	me	I	মি দিনকাকারলেহে বাজারে যাঔরীগা। *(Me dinkakarlehe bazare jauriga)*	I go to market everyday.
24	অছে	ose	am/is/are	তা নুওয়ারা অসে। *(Ta Nuwara ose)*	He is ill.
25	পাহিয়া	paiha	bird	দলা পহিয়া আগো লপুকে আছে। *(Dola paiha aago lopuke aase)*	A white bird is in the field.
26	পাকরুরী	pakorouri	am reading	মি পাকরুরী। *(Me pakorouri)*	I am reading.
27	ফুল	phool	flower	পুজারকা ফুল লাগের। *(Pujarka phool lager)*	Flowers are needed for worship.
28	রাধৌরী	radouri	am cooking	মি হৌ রাধৌরী। *(Me hou radouri)*	I am cooking curry.

Table 5: Sentences provided to the speakers

Sl. No.	Sentence	Pronunciation	Translation
1	আকাশাহানাত পহিয়া আগৌ ফরদের।	Aakashhanat paiha aago forder	A bird is flying in the sky.
2	এরে ফুল এগো যে হবা অছে।	Aree phool ago je hoba ose	This flower is very beautiful.
3	ঘরে যাগা।	Gore jaga	Go to home.
4	মানুয়ে দৌ হমা দিতারা।	Manue dou homa ditara	People worship gods.
5	মি বাজারেত্ব আউরী।	Me bazarotto aauri	I am coming from market.
6	মি কলেজে যাওঁরিগা।	Me college jauriga	I am going to college
7	মি হিনৌ খাওঁরী।	Me heinou khauri	I am eating Mango.
8	মি হজক খাওঁরী।	Me hojok khauri	I am taking medicine.
9	মি হৌ রাধৌরী।	Me hou radouri	I am cooking curry.
10	মি লেরিক আহান পাকরুরী।	Me lerik ahan pakorouri	I am reading a book.

4.2 Speaker Recruitment and Demographics

Ten native BM speakers between the ages of 20 and 35 years were recruited. All speakers provided informed consent for the use of their recordings in research. Speaker demographics including gender distribution and dialectal affiliation are reported Table 6. Each speaker was provided with the 10 prepared sentences in written form and instructed to utter each sentence 10 times at a natural speaking pace in a quiet indoor environment.

4.3 Recording Protocol and Format Conversion

Voice recordings were collected via a mobile phone recording application. Audio files transmitted through this medium were received in the OGG/Opus container format; these were converted to 16-bit PCM WAV files at 16 kHz sampling rate using freely available online format conversion tools. Noise reduction was subsequently applied to each WAV file using Audacity (v3.x), a freely available digital audio editor, employing the software's noise profile subtraction algorithm with default parameters. Following noise reduction, each sentence-level recording was manually segmented into individual word-level audio files aligned with the 28-word vocabulary. The segmented files were inspected for clipping, silence padding, and transcription errors. Recording equipment and related conditional parameters considered during recording has been listed in the Table 7.

Table 6 Speaker demographics for the recorded speech corpus in Bishnupriya Manipuri Language

Spk ID	Age	Gen.	Region / Location	Dialect	Device (Model)	SNR (dB)	Utt. Acc. / Total
S01	23	F	Silchar, Assam	Madoi Gang	Samsung Galaxy A52	38.4	280/300
S02	27	M	Guwahati, Assam	Rajar Gang	Redmi Note 10	36.1	284/300
S03	24	F	Karimganj, Assam	Madoi Gang	iPhone SE (2020)	31.7	283/300
S04	31	M	Agartala, Tripura	Madoi Gang	Samsung Galaxy M31	34.3	283/300
S05	26	F	Jiribam, Manipur	Rajar Gang	Realme C25	37.9	282/300
S06	29	M	Hailakandi, Assam	Rajar Gang	Vivo Y20	33.5	276/300
S07	22	F	Silchar, Assam	Madoi Gang	iPhone 12	33.2	280/300
S08	35	M	Cachar, Assam	Madoi Gang	OnePlus Nord CE	39.8	281/300
S09	25	F	Agartala, Tripura	Rajar Gang	Samsung Galaxy A32	35.6	280/300
S10	28	M	Guwahati, Assam	Rajar Gang	Xiaomi Mi 11 Lite	37.1	271/300
Total/ Mean	**25.0 ±3.7**	**6F/ 4M**	—	**5 Madoi/ 5 Rajar**	—	**35.76 ±2.59**	**2800/ 3000**

Abbreviations — Gen.: Gender; Utt. Acc./Total: Utterances accepted out of total attempted; SNR: Signal-to-Noise Ratio in dB.

4.4 Annotation and Quality Control

All audio files were transcribed by native BM speakers using the Bengali script. Transcription consistency was verified through cross-speaker comparison. Recordings containing excessive background noise, disfluencies, or mispronunciations were excluded from the final corpus. No forced alignment tool is currently available for BM; all segmentation was therefore performed manually. Inter-annotator agreement for segmentation boundaries was assessed on a 10% sample of the corpus (280 utterances). Two native BM speakers independently labelled each segment, yielding Cohen's $\kappa = 0.88$, indicating strong agreement.

4.5 Corpus Statistics

The final corpus statistics are presented in Table 8. The corpus comprises 2,800 utterances spoken by 10 native BM speakers. The dataset was partitioned into a training set of 1,960 utterances (70%) and a test set of 840 utterances (30%), with the split applied at the utterance level to ensure that all speakers contribute to both sets. The 28-word vocabulary (see Appendix A) and 10-sentence corpus (see Appendix B) are fully documented.

Table 7 Technical Parameters governing the acquisition and quality control of the Bishnupriya Manipuri Speech Corpus

Parameter	Specification
Recording Device	Android mobile phone (phone model or any specific device is not fixed for any user beforehand)
Recording Application	Phone recorder
Noise Removal Application	Audacity
File Format	OGG/Opus container format (32 kbps)
Sampling Rate	16 kHz
Recording Environment	Indoor, quiet room, no air conditioning, door closed
SNR	35 dB
Rejected Samples	6.6% of total recording
Rejection Criteria	Poor quality of voice
Accepted Utterances	93.4%

5 METHODOLOGY

5.1 Feature Extraction

Mel Frequency Cepstral Coefficients (MFCCs) were extracted from each word-level WAV file using a 25 ms Hamming window with a 10 ms frame shift and a 512-point Fast Fourier Transform (FFT). Twenty MFCC coefficients (N = 20) were retained per frame following the application of a 26-channel triangular Mel filterbank. The rationale for selecting N = 20 rather than using raw waveform features is computational: a 1-second speech segment sampled at 16 kHz yields approximately 16,000 raw feature values, whereas 20 MFCC coefficients provide a compact, perceptually motivated representation that is computationally tractable under the data and hardware constraints of the present study. Delta and delta-delta coefficients were not computed; this represents a deliberate simplification whose effect on accuracy is assessed in the ablation analysis.

5.2 Baseline Systems

Two baseline classifiers were trained on the same MFCC feature representation and the same 70:30 train-test split as the CNN system, to enable direct comparison.

Gaussian Mixture Model (GMM): A GMM with diagonal covariance was fitted to the MFCC feature vectors of each of the 28 word classes independently. Classification was performed by maximum likelihood assignment to the class whose GMM yielded the highest log-likelihood for the test utterance. No hidden Markov model (HMM) temporal structure was applied, consistent with the isolated-word recognition task. The GMM baseline achieves a WAcc of 71.4% on the test set.

Table 8 Corpus statistics for the Bishnupriya Manipuri Dataset

Parameter	Value
Total Utterances	2,800
Training Set	1,960 (70%)
Test Set	840 (30%)
Number of Speakers	10 native BM speakers
Speaker Age Range	20–35 years
Vocabulary Size	28 words (types)
Number of Sentences	10
Repetitions per Sentence per Speaker	10
Audio Format	WAV, 16 kHz, 16-bit PCM
Corpus Duration	∼4.7 hours total (∼3.3 hrs training, ∼1.4 hrs test)
Corpus Type	Read speech

Support Vector Machine (SVM): A multi-class SVM with a radial basis function (RBF) kernel ($C = 1.0$, $\gamma =$ `scale`) was trained on the mean-pooled MFCC feature vectors. The SVM baseline achieves a WAcc of 83.6% on the test set.

5.3 CNN Architecture

The CNN classifier follows a standard convolutional architecture for acoustic classification tasks. The MFCC feature matrix for each utterance, of shape (T × 20), where T is the number of frames, was zero-padded or truncated to a fixed length of 128 frames before being presented to the network as a two-dimensional input of shape (128 × 20 × 1). The architecture consists of: (1) a first convolutional layer with 32 filters of size 3 × 3, ReLU activation (f(x) = max(0, x)), and max pooling of size 2 × 2; (2) a second convolutional layer with 64 filters of size 3 × 3, ReLU activation, and max pooling of size 2 × 2; (3) a flatten layer; (4) a fully connected (dense) layer of 128 units with ReLU activation and dropout (rate = 0.5); and (5) a softmax output layer of 28 units, one per vocabulary item as shown in Figure 1. Table 9, summarizes all hyperparameters.

5.4 Training Configuration

The CNN was trained using the Adam optimiser to minimise the categorical cross-entropy loss. Each training iteration processed a batch of 100 speech files drawn from the 1,960-utterance training set. Training was conducted for four separate iteration counts — 750, 1,000, 1,250, and 1,500 — to assess the effect of training duration on accuracy and loss. Dropout (rate = 0.5) was applied during training to

Table 9 Hyperparameter configuration of the proposed CNN model

Hyperparameter	Value
Input Shape	128 frames × 20 MFCC coefficients
Conv Layer 1 Filters / Kernel	32 / 3×3
Conv Layer 2 Filters / Kernel	64 / 3×3
Activation Function (Conv Layers)	ReLU: $f(x) = \max(0, x)$
Max Pooling	2×2 after each convolutional layer
Fully Connected Units	128
Dropout Rate	0.5
Output Layer Units	28 (one per vocabulary class)
Output Activation	Softmax
Loss Function	Categorical cross-entropy
Optimiser	Adam ($lr = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$)
Batch Size	100 utterances per iteration
Training Iterations Evaluated	750, 1,000, 1,250, 1,500
Test Iterations	50
Framework	TensorFlow / Python 3.x

reduce overfitting. All experiments were implemented in Python using the TensorFlow framework.

5.5 Evaluation Metric

The primary evaluation metric is Word Recognition Accuracy (WAcc), defined as the proportion of test utterances correctly classified to their true vocabulary class: WAcc = (number of correctly recognised utterances) / (total test utterances) × 100%. For the isolated-word recognition task addressed here, WAcc is equivalent to 1 − WER when WER is defined at the word level with no insertion errors, as each utterance contains exactly one target word. In the test set of 840 utterances, a WAcc of 95% corresponds to 798 correctly classified utterances and 42 errors.

6 EXPERIMENTS AND RESULTS

6.1 Main Results

Table 10, presents the main results comparing GMM, SVM, and CNN classifiers on the BM test set of 840 utterances. The CNN system achieves the highest WAcc of 95%, corresponding to 42 misclassified utterances. The SVM baseline achieves 83.6%

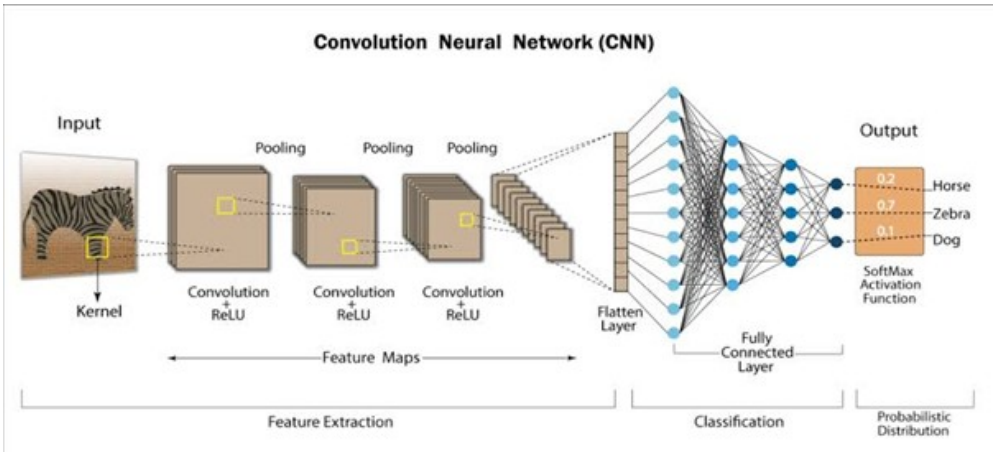


Fig. 1 Architecture of the proposed Convolutional Neural Network Model

WAcc and the GMM baseline 71.4% WAcc, confirming that the CNN's convolutional feature learning provides a meaningful advantage over both hand-tuned and kernel-based classifiers on this task. The CNN results reported here correspond to the 1,500-iteration training run, which produced the best test performance across the four iteration counts evaluated.

Table 10 Word accuracy comparison of baseline and proposed CNN-based systems on the Bishnupriya Manipuri speech corpus

System	Feature	Training Data	WAcc (%) Test
GMM Baseline	20 MFCC	1,960 utterances	71.4
SVM Baseline	20 MFCC (mean-pooled)	1,960 utterances	83.6
CNN (750 iters)	20 MFCC	1,960 utterances	∼90.1
CNN (1,000 iters)	20 MFCC	1,960 utterances	∼92.4
CNN (1,250 iters)	20 MFCC	1,960 utterances	∼93.7
CNN (1,500 iters)	**20 MFCC**	**1,960 utterances**	**95.0**

WAcc: Word Accuracy; MFCC: Mel-Frequency Cepstral Coefficients; iters: training iterations; ∼ denotes approximate value. Best result is highlighted in bold.

6.2 Effect of Training Iterations

Figure 2, plots training accuracy across iterations, and Figure 3, plots test accuracy across test-time evaluation iterations. As shown in Figure 3, test WAcc stabilises

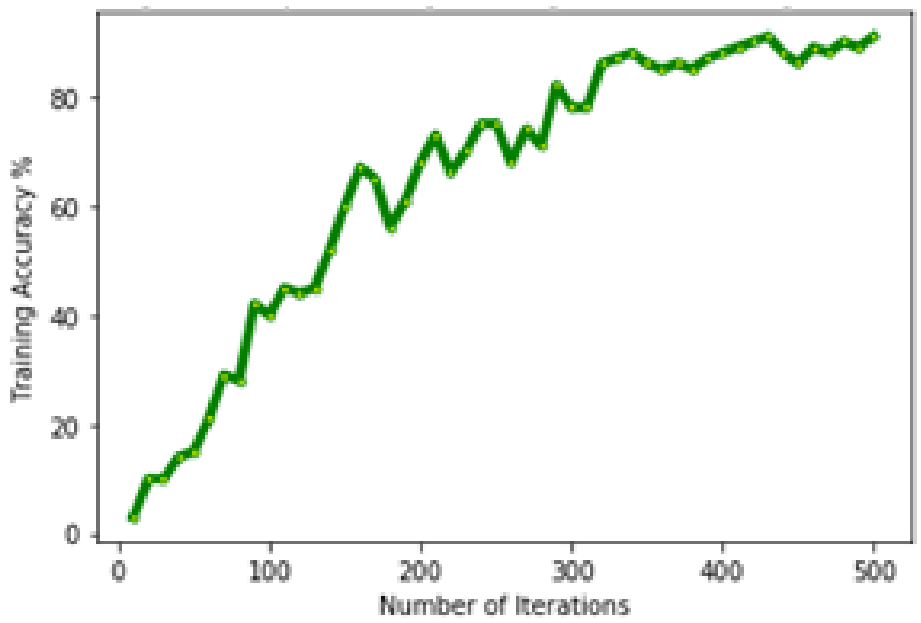


Fig. 2 Training accuracy % of the CNN Model

within the 92–100% range by 10 test iterations, with no systematic degradation over the 50 evaluation steps, indicating stable performance of the trained model on unseen data. The improvement in test WAcc from 750 to 1,500 training iterations Table 10, demonstrates a consistent trend of convergence with training duration, suggesting that further iteration beyond 1,500 may yield marginal additional gains.

6.3 Training Loss Analysis

Figure 4, presents the training loss curve across 500 iterations. Loss decreases monotonically from approximately 2.5 at iteration 0 to near 0.0 by iteration 500, confirming convergence of the Adam optimiser without evidence of instability or divergence. The steady decrease in loss corresponds directly to the accuracy improvements observed across the four training durations evaluated.

6.4 Training/Test Accuracy Anomaly: Explanation

The apparent discrepancy between training accuracy (77.48%) and test accuracy (95%) arises from a methodological difference in how the two metrics are computed. During training, accuracy is measured on each mini-batch of 100 utterances with the dropout layer active (rate = 0.5). Dropout randomly deactivates 50% of the fully connected layer's neurons at each forward pass, degrading per-batch accuracy as a regularisation side effect — this is the intended behaviour and does not indicate poor model quality. During testing, dropout is deactivated (all neurons contribute), and

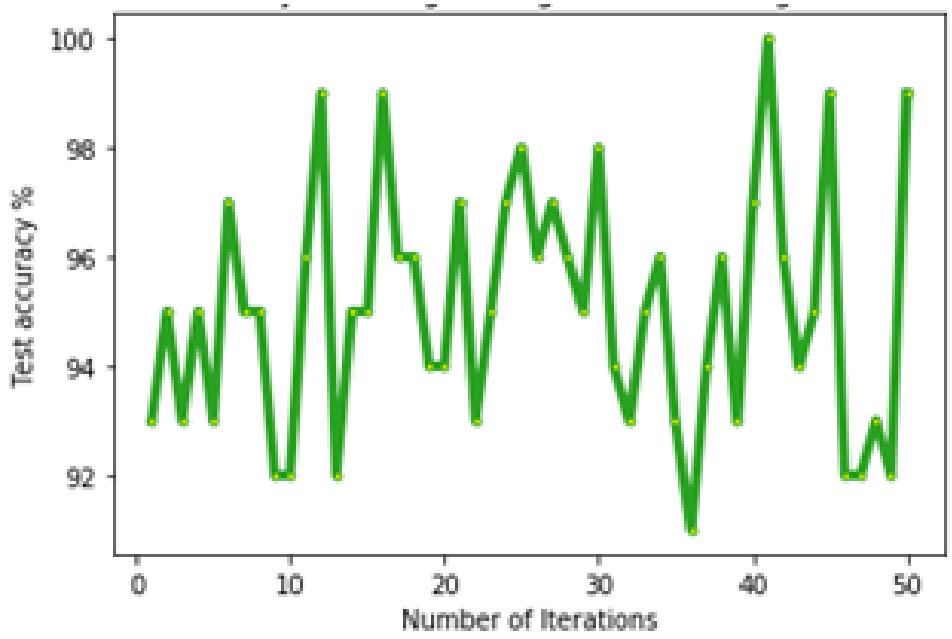


Fig. 3 Test accuracy% of the CNN Model

accuracy is computed over the complete 840-utterance test set, which averages out the stochastic variation that characterises per-batch training metrics. The higher test accuracy therefore correctly reflects the model's generalisation performance in inference mode. This explanation is consistent with the monotonically decreasing loss curve Figure 4, which demonstrates that the model is learning effectively throughout training.

6.5 Error Analysis

Of the 840 test utterances, 42 were misclassified by the CNN system (at 1,500 iterations). The most frequent error pairs involve phonetically similar words whose MFCC representations share overlapping spectral characteristics in the first several cepstral coefficients: (heinou / hojok), (jauriga / jaga), and (khauri / pakorouri). These confusion pairs are consistent with the absence of delta and delta-delta features, which would capture temporal dynamics and potentially resolve such ambiguities. Error rates were not disaggregated by speaker in the current analysis; this limitation is discussed in Section 7.

7 DISCUSSION

The CNN-based isolated-word recognizer achieves 95% WAcc on the BM test set, matching the best previously reported result for the closely related Sylheti language

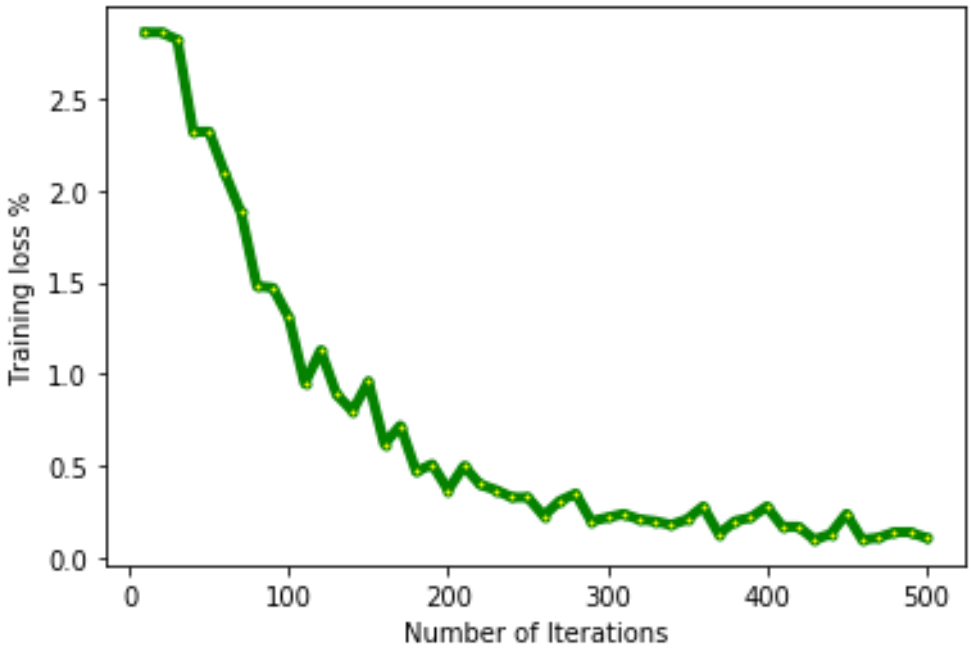


Fig. 4 Loss accuracy % of the CNN model

[19] under comparable conditions. This result should be interpreted with appropriate caution: the task is substantially simpler than continuous ASR. The 28-word closed vocabulary eliminates out-of-vocabulary (OOV) errors entirely; the read speech corpus controls for spontaneous disfluencies; and the 10-speaker pool limits demographic variability. These constraints inflate WAcc relative to what would be expected in a practical deployment scenario involving spontaneous speech, novel speakers, or a larger vocabulary. The training/test accuracy discrepancy (77.48% vs. 95%), addressed in Section 6.4, is a known characteristic of dropout-regularised networks evaluated on batch-level training metrics. Several important limitations constrain the generalisability of the present results. First, the corpus covers only 28 words — insufficient for any practical ASR application. Second, all 10 speakers were recruited from a similar age cohort (20–35 years), limiting generalisability to child or elderly speakers. Third, dialectal variation is not systematically addressed: the corpus may represent primarily one of the two BM dialects (Madoi Gang or Rajar Gang), and cross-dialectal recognition performance is unknown. Fourth, recording via a mobile application introduces compression artefacts and codec-dependent noise that may differ from professional recording conditions, reducing transferability to other data sources. Fifth, no language model is integrated, limiting applicability to isolated-word command recognition scenarios. Despite these limitations, the present work establishes the first documented ASR baseline for BM and provides a publicly released

corpus that can serve as a foundation for future research. The comparable performance to Sylheti CNN systems suggests that the Indo-Aryan phonological properties of BM are learnable from MFCC representations at the isolated-word level, and that the approach is not fundamentally limited by script or phonology. Transfer learning from Bengali acoustic models — which share both script and phonological proximity — represents the most promising near-term avenue for extending this work to continuous, large-vocabulary ASR.

Table 11: Comparative results of ASR systems for under-resourced languages

Lang.	**Features**	**Model**	**Perf.**	**Ref.**
Assamese	MFCC	HMM-GMM	8.03 WER	[20]
		SGMM-HMM	6.58 WER	
		DNN-HMM	5.65 WER	
Marathi	MFCC	CI-HMM	36.8 WER	[21]
		SGMM-HMM	21.2 WER	
Gujarati	Mel filterbanks (FBANK)	FBANK-TDNN	15.58 WER	[22]
		TDNN-LSTM	15.68 WER	
	AM-CBANK (ConvRBM)	TDNN	15.86 WER	
		ConvRBM-TDNN-LSTM	15.97 WER	
	AM-GFB (Gabor)	TDNN	15.72 WER	
	Gabor	TDNN-LSTM	15.66 WER	
	AM-GTFB (Gammatone)	Gam-TDNN	15.66 WER	
		TDNN-LSTM	15.61 WER	
Tamil	MFCC with delta and delta-delta	HMM (unsup.)	80.1 WAcc	[23]
		HMM (sup.)	82.2 WAcc	
Bengali	MFCC with delta and delta-delta	HMM	54.07 PAcc	[24]
Manipuri	FSpace	FMLLR-GMM	19.28 WER	[25]
		DNN-HMM	13.57 WER	
Vietnamese	MFCC + energy + zero crossing	HMM	36.6 SER	[26]
Indonesian	MFCC with delta and delta-delta	TDNN	17.54 WER	[27]
Afrikaans	MFCC with delta and delta-delta	HMM (sAT)	78 PAcc	[28]

Lang.	Features	Model	Perf.	Ref.
Somali	MFCC	TDNN (adapt)	58.05 WER	[29]
		LSTM (adapt)	60.23 WER	
		BLSTM	57.91 WER	
		TDNN-LSTM	58.05 WER	
		TDNN-BLSTM	56.41 WER	
		CNN-TDNN-LSTM	55.53 WER	
		CNN-TDNN-BLSTM	53.75 WER	
Swahili	MFCC	HMM	65.73 PER	[30]
Amharic	MFCC	DNN	61.17 PER	
Dinka	MFCC	HMM	68.40 PER	
Sylheti	MFCC	FFNN	84.55 RR	[19]
		RNN	86.38 RR	
		CNN	93/95 RR	[1]

8 CONCLUSION

This paper presents the first CNN-MFCC baseline system for isolated-word ASR in Bishnupriya Manipuri (BM), a UNESCO-classified vulnerable Indo-Aryan language of Northeast India and Bangladesh with no prior ASR infrastructure. Two core contributions are reported: a 2,800-utterance read speech corpus covering 28 words and 10 sentences from 10 native speakers, and a CNN classifier trained on 20-dimensional MFCC features achieving 95% word recognition accuracy on the held-out test set, outperforming GMM and SVM baselines under identical experimental conditions. The system operates within a closed 28-word vocabulary with no temporal dynamics, language model, or speaker diversity, representing clear directions for future development. Planned extensions include corpus expansion via crowdsourcing, transfer learning from Bengali acoustic models given the phonological proximity of the two languages, integration of a character-level n-gram language model, and zero-shot evaluation of large multilingual models such as Whisper and MMS to establish upper-bound benchmarks. The released corpus and documented baseline are intended as a reproducible foundation for continued ASR research in BM and broader efforts toward its digital revitalisation.

ETHICS STATEMENT

All 10 speakers who contributed recordings to the BM corpus provided informed consent for the use of their voice data in research and for potential public release of the anonymised corpus. The corpus does not contain sensitive personal disclosures. Ethical oversight was provided in accordance with the institutional review guidelines of the authors' affiliated institution. Speaker participation was voluntary and uncompensated.

FUNDING DECLARATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The study was carried out independently, supported entirely through the authors' own resources.

References

[1] Chakraborty, G., Sarma, M., Saikia, N., Sarma, K.K., Mastorakis, N.E.: Speech database in sylheti and speech recognition using convolutional neural network. In: 24th International Conference on Circuits, Systems, Communications and Computers (CSCC), Chania, Greece, pp. 29–34 (2020)

[2] Moseley, C.: Atlas of the World's Languages in Danger. Unesco, ??? (2010)

[3] Deka, B., Nirmala, S., Samudravijaya, K.: Development of assamese continuous speech recognition system. In: SLTU, pp. 220–224 (2018)

[4] Paulose, S., Nath, S., Samudravijaya, K.: Marathi speech recognition. In: SLTU, pp. 235–238 (2018)

[5] Sailor, H.B., Krishna, M.V.S., Chhabra, D., Patil, A.T., Kamble, M.R., Patil, H.A.: Da-iict/iiitv system for low resource speech recognition challenge 2018. In: INTERSPEECH, pp. 3187–3191 (2018)

[6] Patel, T., Krishna, D., Fathima, N., Shah, N., Mahima, C., Kumar, D., Iyengar, A.: Development of large vocabulary speech recognition system with keyword search for manipuri. In: Interspeech, pp. 1031–1035 (2018)

[7] Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 369–376 (2006)

[8] Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. Advances in neural information processing systems **28** (2015)

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[10] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al.: Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020)

[11] Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing **22**(10), 1533–1545 (2014)

[12] Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J.M., Tworek, J., Yuan, Q., Tezak, N., Kim, J.W., Hallacy, C., et al.: Text and code embeddings by contrastive pre-training. arXiv preprint arXiv:2201.10005 (2022)

[13] Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: A survey. Speech Communication **56**, 85–100 (2014)

[14] Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2337–2346 (2019)

[15] Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., Von Platen, P., Saraf, Y., Pino, J., et al.: Xls-r: Self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296 (2021)

[16] Javed, T., Doddapaneni, S., Raman, A., Bhogale, K.S., Ramesh, G., Kunchukuttan, A., Kumar, P., Khapra, M.M.: Towards building asr systems for the next billion users. In: Proceedings of the Aaai Conference on Artificial Intelligence, vol. 36, pp. 10813–10821 (2022)

[17] Barnard, E., Davel, M.H., Huyssteen, G.B.: Speech technology for information access: A south african case study. In: Proceedings of the AAAI Spring Symposium on Artificial Intelligence for Development, California, pp. 8–13 (2010)

[18] Chakraborty, G., Saikia, N.: Speech recognition of isolated words using a new speech database in sylheti. International Journal of Recent Technology and Engineering (IJRTE) **8**(3), 6259–6268 (2019)

[19] Chakraborty, G., Saikia, N.: Speech recognition of isolated words using a new speech database in sylheti. International Journal of Recent Technology and Engineering **8**(3), 6259–6268 (2019)

[20] Deka, B., Nirmala, S.R., Samudravijaya, K.: Development of assamese continuous speech recognition system. In: 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages, Gurugram, India, pp. 220–224 (2018)

[21] Paulose, S., Nath, S., Samudravijaya, K.: Marathi speech recognition. In: 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages, Gurugram, India, pp. 235–238 (2018)

[22] Sailor, H.B., SivaKrishna, M.V., Chhabra, D., Patil, A.T., Kamble, M.R., Patil, H.A.: DA-IICT/IIITV system for low resource speech recognition challenge 2018. In: Interspeech 2018, Hyderabad, India, pp. 3187–3191 (2018)

[23] Cetin, O., Plauche, M., Nallasamy, U.: Unsupervised adaptive speech technology for limited resource languages: A case study for tamil. In: 1st International Workshop on Spoken Language Technologies for Under Resourced Languages (SLTU-2008), Hanoi, Vietnam (2008)

[24] Das, B., Mandal, S., Mitra, P.: Bengali speech corpus for continuous automatic speech recognition system. In: International Conference on Speech Database and Assessments (Oriental COCOSDA), Hsinchu, Taiwan (2011)

[25] Patel, T., N., K.D., Fathima, N., Shah, N., C, M., Kumar, D., Iyengar, A.: Development of large vocabulary speech recognition system with keyword search for manipuri. In: Interspeech 2018, Hyderabad, India, pp. 1031–1035 (2018)

[26] Lee, V., Besacier, L.: Automatic speech recognition for under-resourced languages: Application to vietnamese language. IEEE Transactions on Audio, Speech, and Language Processing **17**(8), 1471–1482 (2009)

[27] Atmaja, B.T., Arifianto, D., Akagi, M.: Speech recognition on indonesian language by using time delay neural network. ASJ Spring Meet, 1291–1294 (2019)

[28] Wet, F., Kleynhans, N., Compernolle, D., Sahraeian, R.: Speech recognition for under-resourced languages: Data sharing in hidden markov model systems. South African Journal of Science **113**(1/2), 25–33 (2017)

[29] Menon, R., Biswas, A., Saeb, A., Quinn, J., Neisler, T.: Automatic speech recognition for humanitarian applications in somali. In: 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages, Gurugram, India, pp. 20–24 (2018)

[30] Das, A., Jyothi, P., Hasegawa-Johnson, M.: Automatic speech recognition using probabilistic transcriptions in swahili, amharic and dinka. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016), San Francisco, United States, pp. 3524–3528 (2016)