

Development of a Devanagari Speech Dataset for Audio CAPTCHA Using Mobile-Based Recording and Automated Segmentation

Dr. Sanjay E. Pate¹

¹ Nanasahab Yashwantarao Narayanrao Chavan Arts, Science & Commerce College, Chalisgaon, Dist. Jalgaon, Pin 424 101 (Maharashtra), India.

Article Info

Article History:

Published: 05 August 2026

Publication Issue:

Volume 3, Issue 8
August-2026

Page Number:

27-38

Corresponding Author:

Dr. Sanjay E. Pate

Abstract:

The development of a scalable and well-organized Devanagari speech database for audio CAPTCHA application is presented in this research. 536 recordings of 10 digits and 23 chosen Devanagari characters were gathered from a variety of speakers, including boys and girls, using mobile devices to produce the dataset. For audio format conversion (M4A to MP3), pause-based segmentation, and methodical arrangement of audio samples, a Python-based processing pipeline is put into place. A total of 10,675 verified audio samples are produced, comprising 3,567-digit and 7,108-character samples. By selecting properly spoken samples and eliminating incorrect or ambiguous recordings, manual verification is carried out to guarantee data quality. The proposed approach offers an affordable, scalable, and useful way to create speech datasets in language environments with limited resources. Applications like assistive technology, voice recognition, and audio CAPTCHA systems can use the created dataset.

The novel dataset presented in this work is accessible to the public on:

1. Mendeley Data at <https://data.mendeley.com/datasets/cwd42xmnp/1>
2. The IEEE data port is available at <https://ieee-dataport.org/documents/natural-devanagari-character-digit-audio-dataset-deva-audio>.

Keywords: Devanagari Speech Dataset, Audio CAPTCHA, Speech Processing

1. Introduction:

Audio CAPTCHA systems are widely used to provide accessibility and enhance security in web applications, particularly for visually impaired users and to prevent automated bot attacks [6]. In recent years, speech-based CAPTCHA mechanisms have gained importance due to their usability and effectiveness in real-world environments.

Significant research has been carried out in the development of speech datasets for English, with well-known corpora such as TIMIT and LibriSpeech providing high-quality and structured data for speech processing tasks [1], [2]. These datasets are typically recorded in controlled environments and are extensively used in Automatic Speech Recognition (ASR) research.

However, for Indian languages, especially those based on the Devanagari script, there is limited availability of structured and publicly accessible speech datasets [11]. Existing works primarily focus on word-level or sentence-level Hindi speech recognition systems [3], with very little emphasis on character-level datasets required for

applications such as audio CAPTCHA systems. This gap highlights the need for developing a reliable, diverse, and application-specific speech corpus for Devanagari.

In this work, we focus on the creation of a Devanagari speech dataset consisting of digits (0–9) and 23 carefully selected characters. The dataset is designed specifically to support audio CAPTCHA generation and related speech processing applications. The proposed dataset addresses the lack of character-level Devanagari resources and provides a practical solution for low-resource language scenarios.

2. Literature Review:

Speech dataset creation has been widely studied in the domains of Automatic Speech Recognition (ASR) and speaker verification [4], [7]. Early works primarily focused on English datasets such as TIMIT and LibriSpeech, which rely on controlled environments and professional recording setups [1], [2]. These datasets provide high-quality audio but are resource-intensive to develop.

In recent years, researchers have explored low-cost and scalable approaches using mobile devices for speech data collection [9]. Studies indicate that mobile-based recordings introduce variability in noise, pronunciation, and recording quality, which can improve robustness of machine learning models [12]. However, challenges such as inconsistent pauses, background noise, and segmentation errors remain significant.

For Indian languages, especially those using the Devanagari script, limited datasets are publicly available [11]. Existing works focus mainly on Hindi ASR [3], with less attention given to character-level datasets required for applications like audio CAPTCHA systems [6]. Some approaches utilize synthetic speech generation, but these lack natural variability.

Pause-based segmentation techniques using silence detection algorithms have been widely adopted for splitting continuous speech into smaller units [10]. Python libraries such as pydub and librosa are commonly used for audio processing tasks including format conversion, segmentation, and feature extraction [5].

Manual validation remains a critical step in ensuring dataset quality, especially in low-resource settings. While automated validation methods are emerging, human verification continues to provide higher accuracy for labelling and filtering speech data [8].

3. Dataset Design:

०	१	२	३	४	५	६	७	८	९
शून्य	एक	दोन	तीन	चार	पाच	सहा	सात	आठ	नऊ
shunya	ek	don	teen	char	pach	saha	sat	aatha	nau

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	अं	अः
a	aa	i	ii	u	uu	e	ai	o	am	ah
क	ख	ग	घ	च	छ	ज	झ	ट	ठ	ड
k	kh	ga	gha	ch	cha	j	jh	t	tha	d
ढ	ण	त	थ	द	ध	न	प	फ	ब	भ
dha	ana	t	tha	d	dha	na	p	pha	b	bha
म	य	र	ल	व	श	ष	स	ह	ळ	ज्ञ
m	y	r	l	va	sh	sha	sa	ha	la	dnya
क्ष	त्र									52
ksha	tra									

Table 1 : Devanagari Character Set

The Devanagari script comprises 52 alphabets; however, certain vowels and consonants exhibit similar phonetic characteristics, leading to potential ambiguity in speech-based systems.

Structure of Devanagari Alphabets: There are two main types Swar (स्वर) – Vowels (13 main ones) Vyanjan (व्यंजन) – Consonants (33 traditional ones) Articulation and Pronunciation. Understanding the articulation of each consonant is crucial for accurate pronunciation



Fig. 1 : Mouth Positions for Pronunciation

Anatomy of the Human Vocal Tract Used in Speech Production:

1. **Gutturals (कण्ठ्य)** : These sounds are produced at the back of the mouth. क (ka), ख (kha), ग (ga), घ (gha)
2. **Palatals (तालव्य)** : These sounds are articulated with the tongue against the hard palate. च (cha), छ (chha), ज (ja), झ (jha).
3. **Retroflexes (मूर्धन्य)** : Produced with the tongue curled back. ट (ta), ठ (ṭha), ड (ḍa), ढ (ḍha), ण (ṇa)
4. **Dentals (दंत्य)** : Articulated with the tongue against the upper teeth. त (ta), थ (tha), द (da), ध (dha), न (na)
5. **Labials (ओष्ठ्य)** : Produced with both lips. प (pa), फ (pha), ब (ba), भ (bha), म (ma)

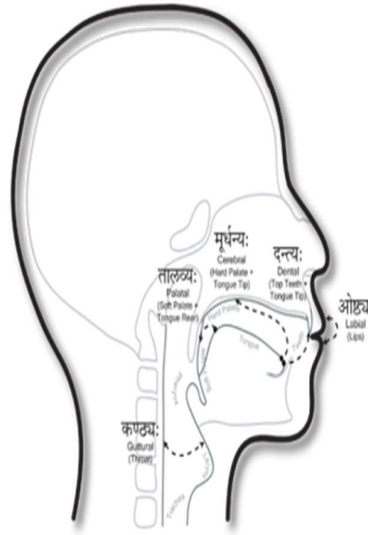


Fig. 2: Human Vocal Tract

Vowels (स्वर) with similar pronunciation

इ (i) and ई (ī) → Difference is mainly **short vs long sound** इ = short “i” (as in *bit*) , ई = long “ee” (as in *see*)

उ (u) and ऊ (ū) → Short vs long

उ = short “u” (as in *put*) ऊ = long “oo” (as in *food*)

ए (e) and ऐ (ai)

ऋ (ṛ) sometimes sounds close to रि (ri) in modern spoken Marathi

ओ, अं, अः

क (ka) and ख (kha) → Difference is **aspiration (air)**

ग (ga) and घ (gha) → Airy vs non-airy

च (cha) and छ (chha) ज (ja) and झ (jha)

ट (ṭa) and ठ (ṭha) ड (ḍa) and ढ (ḍha)

त (ta) and थ (tha) द (da) and ध (dha)

प (pa) and फ (pha) ब (ba) and भ (bha)

न (na) and ण (ṇa) → In fast speech, many speakers pronounce them similarly

ङ (ṅa), ञ (ña), ण (ṇa) → Mostly appear with specific consonants and sound similar to न

S sounds (very common confusion) - स (sa), श (śa), ष (ṣa) → In spoken Marathi, all often sound like “स”

To minimize phonetic ambiguity, characters with similar pronunciation were removed. The resulting dataset comprises 10 Devanagari digits (0–9) and 23 distinct Devanagari characters.

०	१	२	३	४	५	६	७	८	९
शून्य	एक	दोन	तीन	चार	पाच	स हा	सात	आठ	नऊ

अ	इ	उ	ए	क	ग	च	ज	ट	ड
त	द	न	प	ब	म	य	र	ल	व
श	ह	क्ष							23

Table II : Devanagari Digits and Selected 23 Characters

4. Methodology

The dataset preparation process consists of multiple stages:

A. Character Set Design : A character set consisting of 10 numerals (0–9) and 23 selected

Devanagari characters

B. Training of Recording Participants :

Recordings were collected from multiple speakers including: - Boys – Girls. This ensures variability in pitch, tone, and pronunciation.

Before data collection, all recording participants were trained with clear instructions regarding: Proper pronunciation of each character, Maintaining uniform speaking speed, introducing small pauses between consecutive characters, Recording in a relatively noise-controlled environment

C. Audio Recording Procedure

Each participant recorded the complete character set using a collar microphone connected to a mobile device to enhance audio clarity. Recording was performed using a mobile sound recorder application- Voice Recorder, which required internet connectivity for proper functioning.

D. Storage of Raw Recordings :

The recorded audio files were automatically stored in the internal storage of the mobile device in M4A format.

E. Data Transfer to Processing System :

All recorded audio files were transferred from mobile devices to a laptop using E-mail transfer, or OTG (On-The-Go) cable

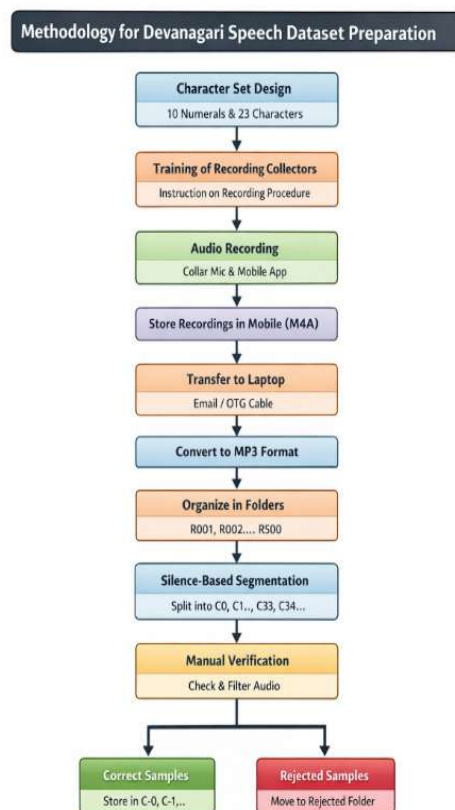


Fig. 3 : Dataset Design Framework

F. Format Conversion :

The collected audio files in M4A format were organized into a dedicated folder and converted into MP3 format using Python scripts.

G. Structured Storage of Recordings :

The converted MP3 files were systematically stored in separate folders labelled as:

R001, R002, ..., R500, where each folder represents one complete recording instance of the character set.

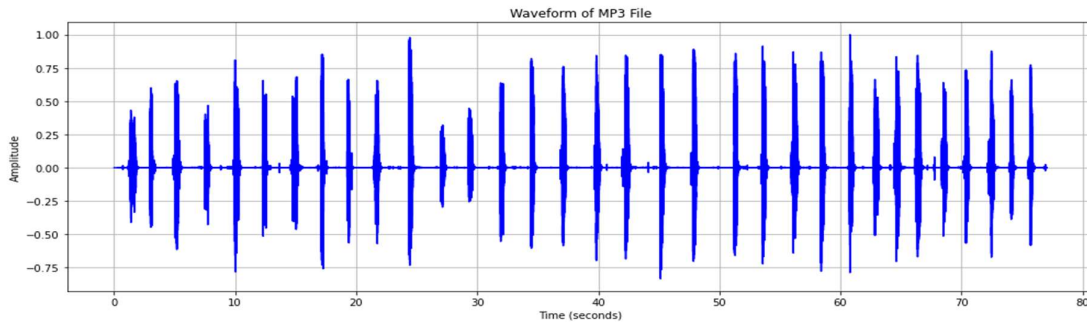


Fig. 4 : Waveform of Devanagari Digits and Characters Recording

The waveform represents the amplitude variation of the recorded MP3 audio signal over time, where high-energy peaks correspond to spoken Devanagari characters or digits and low-amplitude regions indicate silence. The distinct and regularly spaced speech segments reflect structured recording, while amplitude variations capture natural differences in pronunciation.

H. Silence-Based Segmentation

Each recording was processed using a **silence detection algorithm** to split the continuous audio into individual character segments. This resulted in segmented files labelled sequentially as:

C0, C1, ..., C33, with additional segments (C34–C44) generated due to extra pauses or recording variations.

After silence-based segmentation, each recording was divided into 33–40 individual audio segments. A representative waveform of a segmented sample is shown to illustrate the temporal and amplitude characteristics of the speech signal

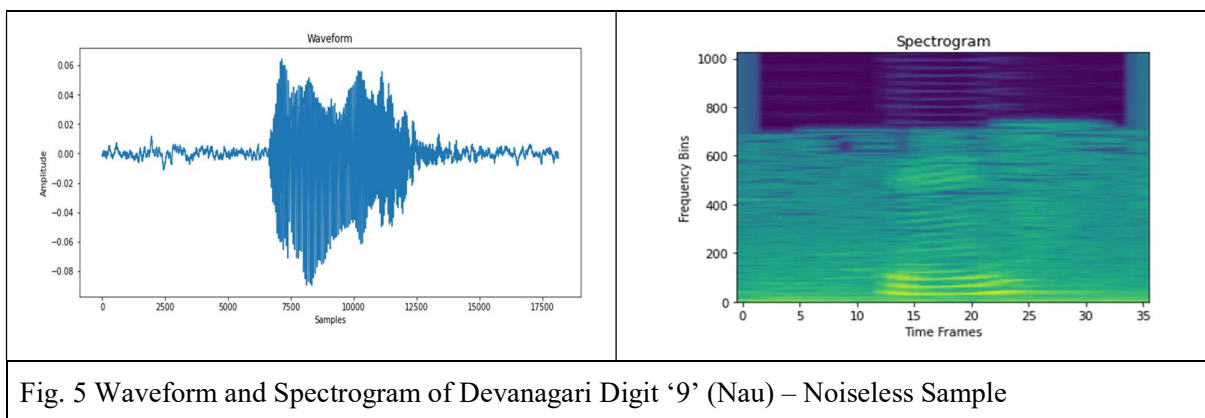
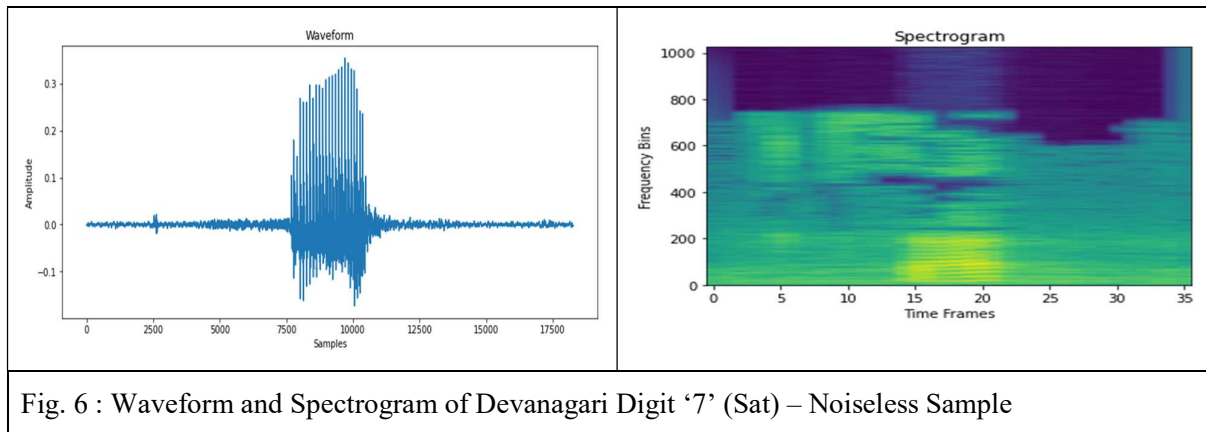


Fig. 5 Waveform and Spectrogram of Devanagari Digit ‘9’ (Nau) – Noiseless Sample



Representative waveform of a segmented Devanagari speech sample of 9 (nau) and 7 (sat) obtained after silence-based segmentation. The waveform illustrates amplitude variations over time, where high-energy regions correspond to speech segments and low-amplitude regions indicate silence

Representation of Spectrogram of a segmented Devanagari speech sample showing time–frequency representation of 9 (nau) and 7 (sat) . The intensity of color indicates signal energy, highlighting the distribution of frequency components during speech.

I. Organization of Segmented Files :

The segmented character files were stored within their respective recording folders.

For example: R001 → C0 to C33, R002 → C0 to C40 (in some cases due to extra segments)

This hierarchical structure ensured traceability of each segment to its original recording.

J. Manual Verification and Quality Control :

Each segmented audio file was manually reviewed to ensure:

Correct pronunciation, Audible clarity, Absence of noise or distortion

Based on verification: Correct samples were retained, Incorrect or unclear samples

were moved to a Rejected Folder

K. Final Dataset Organization:

After validation, all correctly identified character samples were grouped class-wise. For example: All instances of digit ‘0’ were stored in folder C-0, All instances of digit ‘1’ were stored in folder C-1, Similarly, all characters were organized into their respective folders

6. Results and Discussion:

Recording Collection: The speech data was collected from a diverse group of participants, including boys, girls, men, and women from both rural and urban backgrounds. The data collection process was carried out in four sessions,

resulting in a total of 536 recordings. The inclusion of speakers from varied age groups and geographical regions helps to introduce diversity in pronunciation, accent, and speaking style, thereby improving the robustness of the dataset.

Data Collection	Boys		Girls		Men		Women		Total
	Rural	Urban	Rural	Urban	Rural	Urban	Rural	Urban	
1	25	10	50	20	5	2	12	6	130
2	18	13	55	22	7	2	12	6	135
3	23	11	54	22	6	3	10	2	131
4	10	24	18	49	7	10	12	10	140
	76	58	177	113	25	17	46	24	536

Table III : Demographic-wise Distribution of Recorded Participants (Rural and Urban)

Symbol	Correct	Rejected	Total	C %	R %	Symb ol	Corre ct	Rejec ted	Total	C %	R %
०	425	111	536	79.29	20.71	ज	310	226	536	57.87	42.13
१	402	134	536	75.00	25.00	ट	310	226	536	57.87	42.13
२	375	161	536	69.91	30.09	ड	310	226	536	57.87	42.13
३	372	164	536	69.44	30.56	त	308	228	536	57.41	42.59
४	355	181	536	66.20	33.80	द	308	228	536	57.41	42.59
५	345	191	536	64.35	35.65	न	308	228	536	57.41	42.59
६	335	201	536	62.50	37.50	प	305	231	536	56.94	43.06
७	323	213	536	60.19	39.81	ब	305	231	536	56.94	43.06
८	320	216	536	59.72	40.28	म	305	231	536	56.94	43.06
९	315	221	536	58.80	41.20	य	303	233	536	56.48	43.52
अ	318	218	536	59.26	40.74	र	303	233	536	56.48	43.52
इ	319	217	536	59.59	40.41	ल	300	236	536	56.02	43.98
उ	344	192	536	64.25	35.75	व	300	236	536	56.02	43.98

ए	315	221	536	58.80	41.20	श	300	236	536	56.02	43.98
क	313	223	536	58.33	41.67	ह	300	236	536	56.02	43.98
ग	313	223	536	58.33	41.67	क्ष	298	238	536	55.56	44.44
च	313	223	536	58.33	41.67	TOTAL	10675	7013			

Table IV : Accuracy Analysis of Audio Recordings (Correct vs Rejected)

The table presents the distribution of correct and rejected audio samples for each Devanagari digit and character. A total of 10,675 correct samples and 7,013 rejected samples were obtained after manual verification.

It is observed that digits generally show higher correctness percentages compared to characters, with digit '०' achieving the highest accuracy of 79.29%. In contrast, several characters such as 'क्ष' and 'ल' exhibit lower correctness rates due to phonetic complexity and similarity with other characters.

The variation in correctness across symbols highlights the challenges in speech data collection, particularly for character-level recognition, where pronunciation differences and recording conditions significantly affect data quality.

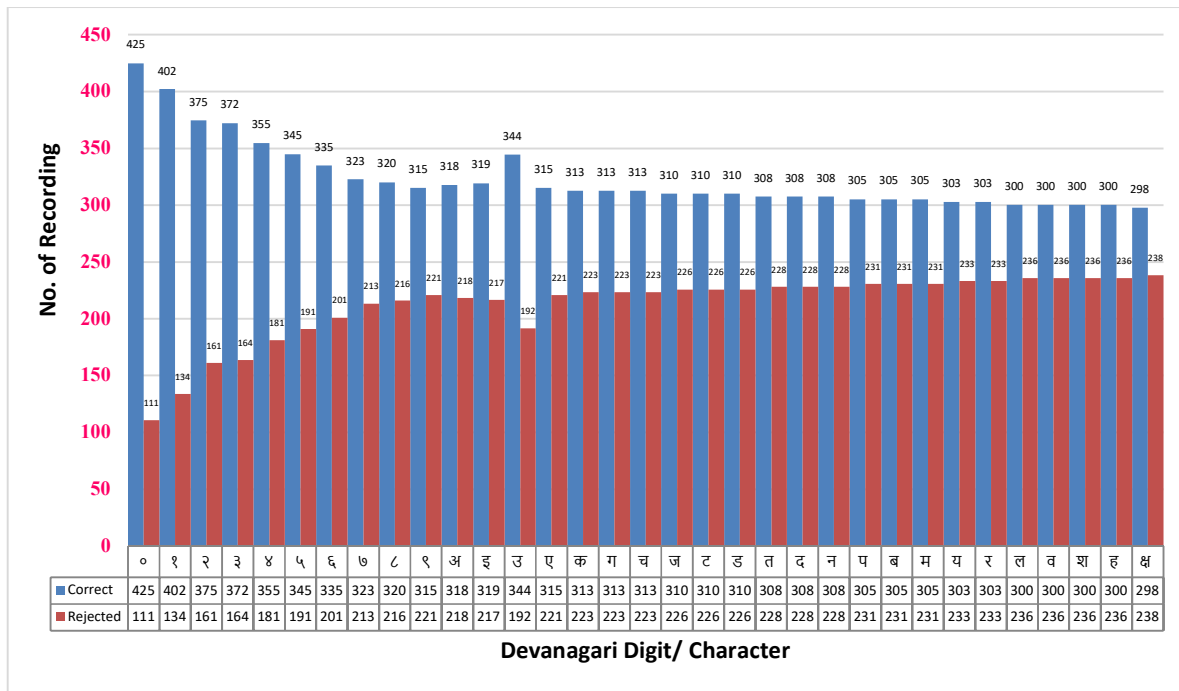


Fig. 7: Devanagari Digit and Characters, Correct vs Rejected Audio Recordings (%)

Character Recording	Digits Recording	Dataset Total Recording
7,108	3,567	10,675

Table V : Total Correct Character and Digits recordings

The final dataset consists of 10,675 validated audio samples, including 7,108 recordings of Devanagari characters and 3,567 recordings of digits. This balanced distribution ensures sufficient representation of both character and digit classes, making the dataset suitable for audio CAPTCHA and speech processing applications.

The proposed system successfully generates a structured and labeled dataset suitable for audio CAPTCHA applications. The use of mobile devices simplifies data collection but introduces variability, which enhances robustness.

Manual verification ensures high-quality data, though it requires human effort. The segmentation approach based on pauses proves effective for isolating characters.

7. Applications

The dataset can be used for: - Audio CAPTCHA generation - Speech recognition systems - Language learning tools - Assistive technologies

8. Limitations

- **Background Noise Issue:** Audio recorded using mobile devices may include background noise, which can affect the clarity of the speech samples.
- **Manual Effort Required:** Some steps like recording checking and audio segmentation are done manually, which is time-consuming.
- **Limited Dataset Size:** The current dataset includes a limited number of samples and speakers, so it may not cover all variations in pronunciation.

9. Future Work

- **Extension to Full Devanagari Set:** The dataset will be extended to include all Devanagari characters, including vowels, consonants, and symbols.
- **Automation Using Machine Learning:** Machine learning techniques will be used to automate validation and reduce manual effort.

Acknowledgment

The author gratefully acknowledges the financial support received from Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon under the Vice Chancellor Research Motivation Scheme (VCRMS), Ref.No. KBCNMU/11A/VCRMS-SCI-2024-25/152/2025 Dated: 11/03/2025. This support significantly contributed to the successful development of the Devanagari speech dataset.

10. Conclusion:

This paper presents a cost-effective and scalable approach to developing a Devanagari speech dataset. By combining mobile-based recording, automated processing, and manual validation, a high-quality dataset is achieved. Future work includes automation of validation and expansion of the dataset with more speakers and environmental diversity.

References

- [1] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, 1993.
- [2] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2015.
- [3] A. Kumar and R. Singh, "Hindi Speech Recognition System Using HTK," International Journal of Computer Applications, 2014.
- [4] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980.
- [5] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in Proc. SciPy Conference, 2015.
- [6] J. Robert, "Audio CAPTCHA Systems and Security," IEEE Security & Privacy, 2012.
- [7] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in Proc. IEEE ICASSP, 2013.
- [8] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in Proc. ACM Multimedia, 2010.
- [9] S. Kumar, P. Singh, and R. K. Sharma, "Mobile-Based Speech Data Collection for Low-Resource Languages," IEEE Access, 2020.
- [10] R. Sharma and M. Verma, "Silence Detection Techniques for Speech Segmentation Using Python Tools," International Journal of Signal Processing, 2019.
- [11] P. Joshi and S. Kulkarni, "Devanagari Script Processing and Applications in Speech Systems," Journal of Indian Language Technology, 2021.
- [12] K. Patel and A. Shah, "Audio Data Preprocessing Techniques for Machine Learning Applications," IEEE Transactions on Audio, Speech, and Language Processing, 2018.
- [13] S. E. Pate, "Devanagari Natural Character Audio Dataset," Mendeley Data, vol. 1, 2026. [Online]. Available: <https://data.mendeley.com/datasets/cwd42xmnp/1>.
- [14] S. E. Pate, "Natural Devanagari Character and Digit Audio Dataset (DEVA-AUDIO)," IEEE DataPort, 2026. [Online]. Available: <https://ieee-dataport.org/documents/natural-devanagari-character-digit-audio-dataset-deva-audio>.