Design and Development of Artificial Stuttered Speech Corpora

P.N. Deorukhakar [1] V.B. Waghmare [2] I.K. Mujawar [3] R.Y. Patil [4]

[1] Department of Computer Science, Shivaji University, Kolhapur.
[2, 3, 4] Department of Computer Science, Vivekanand College, Kolhapur.
pnd.rs.csd@unishivaji.ac.in
vishal.pri12@gmail.com
mujawarirfan@gmail.com
rajashrishendre@gmail.com

Abstract. This paper describes the design, development and testing process of an artificial Stuttered speech database. The acted stuttered speech samples are collected from Marathi movie and the similar speech samples are also collected by mimicking from Marathi speaking Male and Female speakers. The detailed and standardized procedure for collecting artificial stuttered speech samples for database development is presented in this paper. Also, this paper describes the denoising of collected stutter speech samples. After the development of the database, all samples were standardized. The developed speech database will be useful for researchers working in the field of stuttered speech recognition technologies for Marathi and other Languages.

Keywords: Corpora, PRAAT, Marathi, Stuttered Speech, Speech Disability.

1 Introduction

Human communication occurs through several forms including verbal, written, non-verbal, visual, and sign language. Among these, speech communication plays an important role in human–computer interaction (HCI) and intelligent systems. Modern speech technologies allow machines to record, interpret, and understand human speech, enabling the development of speech recognition systems and AI-based applications (Shrishirmal et al., 2015; Karray et al., 2008).

Marathi is an Indo-Aryan language spoken by approximately 71 million people and is one of the most widely spoken languages in India. Marathi also exhibits regional dialect variations such as Kolhapuri, Puneri, Kokani, Nagpuri, and, Khandeshi although the meaning of words generally remains the same (Waghmare et al., 2014). Despite its wide usage, research on stuttered speech recognition for Marathi remains limited and only a few speech datasets are available. Previous work on Marathi stuttered speech recognition has shown moderate performance using MFCC and LPC features (Waghmare et al., 2017).

The development of automatic stuttered speech recognition system in English, Polish, Spanish, German, Arabic, Dutch languages has achieved more recognition rate.
Therefore, this study focuses on the development of a structured artificial stuttered Marathi speech corpus collected from movie-based speech samples and additional mimicked recordings from male and female speakers. The developed corpus aims to support future research in stuttered speech recognition and speech disorder analysis. The paper is organized as Section 2 explains how data on movies have been selected and how the development process has been initiated. In section 3, the steps discuss developing the stuttered speech corpus discussed with denoising and annotations processes. Section 4 displays the process of collecting the artificial data, specification of speakers, recordings, and statistics of the data set. Section 5 specifies how the process of recording and about artificial database. The spectrogram analysis of the selected samples before and after the denoising is shown in section 6. After Section 7 talks about acknowledgements and Section 8 discusses conclusion.

1.1 Contributions of the Study

- Development of 3000 sample artificial stuttered Marathi corpus.
- Balanced gender distribution (50 male, 50 female).
- Structured disfluency labeling (R, P, B, I).
- Standardized recording and denoising protocol.
- Spectrogram-based validation of preprocessing.

2 Literature Review

The field of automatic speech recognition (ASR) has achieved significant development for fluent speech; however, recognition of disfluent or stuttered speech remains a much challenging research problem especially for Marathi language. Some of the studies have investigated about stuttered speech modeling using acoustic and temporal features. Waghmare et al. (2017) explored isolated stuttered Marathi speech recognition using MFCC and LPC features and reported a moderate recognition performance. Also Goldman et al. (2016) have developed a multilingual acted speech database that demonstrated the importance of controlled speech corpus design for speech technology research. Speech dataset development plays a significant role in improving speech recognition systems. Barry and Dalsgaard (1993) have emphasized the importance of structured annotation schemes for speech databases ensuring reproducibility and consistency among experiments. Assigning a systematic annotation of repetition, prolongation and block occurrence in stuttering is very essential for modeling stuttering behavior effectively. Additionally Yu et al. (2008) proposed time–frequency block thresholding methods for speech denoising, are widely used for improving signal quality prior to feature extraction. Rusz et al. (2021) provide compressive guidelines for speech recording standards, highlighting placement for microphone, controlling environmental as well as normalization techniques for ensuring high-quality recordings. With these advancements limited resources are available for

Marathi stuttered speech. Also the most existing datasets are small in size and lack balanced gender distribution. Therefore, the development of a structured artificial stuttered speech corpus is necessary to support future research in disfluent speech recognition. Although several studies have addressed speech corpus development and speech denoising techniques, resources for Marathi stuttered speech remain limited. Most available datasets contain a small number of samples and lack balanced gender representation or structured disfluency annotations. Therefore, the development of a structured artificial stuttered speech corpus is necessary to support reproducible experiments and future research in disfluent speech recognition. Table 1 presents an overview of the existing speech and stuttered speech corpora reported in the literature.

Table 1. Overview of Related Speech and Stuttered Speech Corpora.

Study	Year	Language	Speech Type	No of Speakers	Dataset Type	Dataset Size
Barry & Dalsgaard	1993	Multilingual	General speech	Not specified	Speech database framework	Not specified
Yu et al.	2008	English	General audio	–	Audio processing study	–
Waghmare et al.	2017	Marathi	Isolated words	Limited	Stuttered isolated speech	Small dataset
Goldman et al.	2016	Multilingual	Acted emphasized speech	Multiple	Acted speech database (SIWIS)	Medium-scale
Rusz et al.	2021	English (clinical)	Disordered speech	Clinical patients	Dysarthric speech study	Moderate
Present Work	2025	Marathi	Isolated words + sentences	100 (50M + 50F)	Artificial + Movie-based corpus	3000 samples

As shown in Table 1 existing studies focuses on general speech databases or clinical speech disorders, Although Marathi stutter speech resources are limited in scope and size. This research work addresses this gap by developing artificial Marathi stutter speech corpus contains 3000 samples with balanced gender representation. Table 2 represents a comparative analysis of annotation schemes, recording environments, feature extraction methods, key contributions, and limitations of related studies.

Table 2. Comparative Analysis of Annotation and Technical Characteristics.

Study	Annotation Scheme	Recording Environment	Features Used	Key Contribution	Limitations
Barry & Dalsgaard	Structured multilingual labeling	Controlled	Not feature-focused	Standardized annotation framework	No stutter modeling
Yu et al.	Not applicable	Laboratory signals	STFT, Time–frequency block thresholding	Advanced denoising method	Not speech-disorder specific

Waghmare et al.	Basic stutter labeling	Controlled	MFCC, LPC	Marathi stuttered ASR	Small dataset
Goldman et al.	Structured metadata	Studio setup	Acoustic features	Multilingual acted corpus	Not stutter-focused
Rusz et al.	Clinical guidelines	Medical setup	Acoustic analysis	Pathological speech standards	Not corpus-focused
Present Work	R, P, B, I disfluency labels	Controlled + movie extraction	Suitable for MFCC, LPC, STFT	Balanced, annotated Marathi corpus	Artificial (acted) stuttering

As shown in Table 2 previous studies focused on either speech processing techniques or speech database development. While the proposed corpus focused on structured disfluency annotations, standardized recording procedures, and balanced speaker distribution.

3 Selection of Data and Further Development

At the beginning, the Marathi movies containing stuttering characters were identified. In the research study it is found that a Marathi Indian movie entitled “Habaddi” released in the year 2022, having lead stuttering character and many stuttered speech scenes. So, it was decided to select the ‘Habaddi’ movie for the stuttered Marathi speech database development. Table 3 provides detailed specifications of the stuttered speech samples extracted from the movie.

Table 3. Technical Specifications of Extracted Movie Speech Samples.

Name	Details
Format String	PCM
Format setting	Little/Signed
Bit rate Mode	Constant
Bit rate	1536 Kb/s
Channels	2 channels
Sampling rate	48.0 kHz
Bit Depth	16 Bits
Stream Size	1.11 Mib

As shown in Table 3 the extracted speech samples were available in PCM format having sampling rate of 48 kHz and 16-bit resolution, which ease subsequent preprocessing and denoising operations. As every Indian movie has background music, this movie has a background music too which provides texture and aural interest that will engage audience and capture their attention. It was found that every collected sample

have a background music so it was decided to denoise the collected samples (Sutton & Austin, 2015).

3.1 Feature Representation and Potential Research Applications

The stuttered speech dataset is developed to support a wide range of acoustic feature extraction and modeling techniques. The recorded speech samples can be processed using conventional and advanced feature extraction techniques which commonly used in speech recognition and speech disorder analysis. To capture the short-term spectral envelope and formant characteristics of speech the standard spectral features such as Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) can be extracted. For identifying disfluency patterns such as repetitions and prolongations particularly Time–frequency representations such as the Short-Time Fourier Transform (STFT) enable analysis of temporal variations in speech signals. Also acoustic descriptors including Zero Crossing Rate (ZCR) and Spectral Contrast provide complementary information related to signal periodicity and frequency distribution. The structured annotation adopted in this corpus includes repetition (R), prolongation (P), block (B), and interjection (I) provides a labeled metadata that enables supervised machine learning techniques for stutter detection and classification. These annotations enables the development of models for binary classification (fluent vs. disfluent speech) as well as multi-class disfluency recognition. Further the balanced gender distribution and standardized recording conditions enhances the suitability of the dataset for training, validation, and benchmarking of speech recognition and speech disorder modeling systems. Therefore beyond corpus development, the dataset provides a foundation for future research in stuttered speech recognition, acoustic analysis of disfluency patterns, and intelligent speech-based diagnostic systems.

4 Steps Followed for Development

Firstly, the movie was searched and after searching process the movie watched carefully from start to end and understood the motto and story line behind the movie. After that the communication among actors in the movie was identified and shortlisted each and every scene from movie in which the stutter character was performing. It is found that a 10-year-old boy lead character ‘Manya’ and a ‘puppet’ in the movie are stuttering and also other actors are stuttering in some scenes. Again, the words, compound words and some sentences as a stuttered speech sample are found. Then the stuttered speech samples with respect to time frame was selected. In this step the whole movie was imported in the software called Audacity then all the samples which are selected on the basis of time frame were trimmed one by one and exported to the directory. All the samples were in .mp3 format and samples having bit rate of 1536 kbps. Then all sample were converted to the .wav format. Audio denoising is the process of removing noises from a speech sample without affecting the quality of the speech samples (Yu et al., 2008). At last, all the collected stuttered speech samples were denoised accordingly.

4.1 Details of Movie Database

The developed Dataset from movie contains total 68 stuttered speech samples. There are total 46 utterances of isolated words, 11 utterances of compound words and 11 utterances of continuous spoken sentences included in the dataset. All stuttered speech samples were denoised and filtered.

4.2 Annotation Details

The collected speech samples are manually annotated for identification of different types of stuttering events present in the dataset. For the annotation process the samples are first categorized into isolated words, compound words, and continuous spoken sentences. The annotation scheme used in this study consisting four primary disfluency categories that are repetition (R), prolongation (P), block (B), and interjection (I). Repetition refers to the repetition of a syllables, sounds, or words (e.g., "ma-ma-ma"). Prolongation represents the long extension of phoneme duration (e.g., "sssspeech"). Block indicates a silent pause or interruption before the initiation of speech due to the difficulty in production of speech and interjection refers to occurrence of filler expressions such as "uh" or "um" before or during speech. Each speech sample was listened and carefully analyzed with observation of temporal characteristics to determine the appropriate disfluency label. The predetermined labeling guidelines are followed during annotation process for maintaining the consistency across dataset. The general principles of speech database annotation described in previous studies (Barry & Dalsgaard, 1993) are followed. Then to improve annotation reliability each speech sample was reviewed through multiple listening checks and multiple cases were re-evaluated for ensuring assigned label accurately reflected the observed stuttering behavior. Table 4 provides a summary of the categories adopted for labeling stuttering events in the dataset.

Table 4. Disfluency Annotation for Dataset

Name	Details
Block	Present Gaps or Pause in stuttered speech sample.
Prolongation	Prolongation of syllable of Stuttered Speech sample.
Repetition	Repeated Syllable in Stuttered Speech Sample
Interjection	Present occurrence of filler expressions

These annotation labels provide a structured framework to identify and classify different disfluency types. This will support future machine learning and speech recognition experiments.

5 Artificial Data Collection

After collection of data samples from movie all samples are preprocessed and denoised. Then acted stuttered speech samples from '50' males and '50' females' same

samples as per movie is collected and preprocessed and denoised. The collected database consist of total '3000' samples in that there are '2000' utterance of words, '1000' utterance of compound words sentences. All acted speech samples were recorded in a controlled environment using a Maono AU-A04 condenser microphone (mono) this microphone is connected to a computer system in the college laboratory. For the minimization of background disturbances and to maintain uniform recording conditions a device Marpac All-Natural White Noise Sound Machine was used during the overall recording process. The microphone was set to positioned at an approximate distance of 10 centimeters from the speaker for ensuring clear speech sample capture and also it reduces breathing and plosive noise. The recordings were collected in a quiet indoor setup and after preprocessed including denoising and normalization techniques for maintaining consistent audio quality across all speakers. The details of recording setup and specification of equipment used during development of corpus provided in Table 5.

Table 5. Recording Setup and Equipment Specifications

Parameter	Description
Recording Microphone	Maono AU-A04 Condenser Microphone Kit (Black)
Microphone Type	Condenser Microphone
Channel Mode	Mono
Recording Location	College Computer Laboratory
Noise Control Device	Marpac All-Natural White Noise Sound Machine (Home and Away Bundle, Tan)
Recording Environment	Controlled indoor setup with reduced external noise
Microphone Distance	~15–20 cm from speaker
Recording Format	WAV
Post-processing	Denoising, normalization, silence trimming

The selected configuration ensured consistency in recording conditions and minimized environmental noise across all speakers.

The distribution of speakers and utterance statistics of the developed corpus are presented in Table 6.

Table 6. Speaker Distribution and Utterance Statistics

Category	Count
Total speakers	100
Male speakers	50
Female speakers	50
Total utterances	3000
Word utterances	2000
Sentence utterances	1000
Avg utterances per speaker	30
Avg words per speaker	20
Avg sentences per speaker	10

The dataset follows balanced distribution across gender categories, resulting in nearly 1500 utterances from male speakers and 1500 utterances from female speakers. For dataset each speaker contributed for 30 utterances consist approximately 20 isolated words and 10 compound/sentence samples. All recordings were preprocessed and denoised to ensure uniformity and improved speech quality. Table 7 provides disfluency labels and representative examples used during annotation.

Table 7. Disfluency Label Descriptions and Examples

Label	Description	Example
R	Sound/word repetition	"ma-ma-ma"
P	Sound prolongation	"sssspeech"
B	Block	pause before speech
I	Interjection	"uh… speech"

These labels were continuously applied across the dataset to facilitate supervised learning and disfluency classification tasks.

Table 8. Summary of Artificial Stuttered Speech Dataset

Parameter	Value
Total Samples	3000
Total Duration	~2 hours
Sampling Rate	16 kHz
Bit Depth	16-bit PCM
Audio Format	WAV
Channels	Mono (1 Channel)
Average Duration per Sample	~2.4 seconds
Minimum Duration	~1 sec
Maximum Duration	~5 sec

Table 8 highlights characteristics of the corpus, including dataset size, sampling rate, audio format, and duration statistics which demonstrate the suitability of the dataset for speech processing applications.

5.1 Speaker selection

Firstly, speech sample trials from many speakers taken then on the basis of speaker stuttering act intensity the speakers were selected for taking samples also some speakers are selected on the basis of availability and other factors.

5.2 Data collection

For the collection of acted data recording audio samples, high quality Maono microphone device was used and for white noise, device called Morpoc was used. For analysis and listening of recorded speech samples RedGear 7.1 headphone was used.

5.3 Data Collection Statistics

collected speech samples from ‘100’ speakers. All speakers were categorized on the basis of gender. The collected the data from ‘50’ males and ‘50’ females’ speakers from the age group of ‘19’ to ‘50’. The developed database consists of ‘2000’ utterances of word, ‘1000’ utterances compound words and sentences.

6 Recording Procedure

The selected words from movie were recorded from speakers using high quality Maono microphone device. The Maono microphone have different specifications in terms of technical specifications. The selected microphone is having noise cancelling facility and other technical features. The distance of the both the microphone from speaker was same. For the recording in normal environment the headset from circle used also the circle headset having noise cancellation facility and other technical specifications (Rusz et al., 2021). The data was recorded in lab environment and also recorded in normal environment. The PRAAT software used for recording the speech samples. The PRAAT is having graphical user interface, and the functionalities like spectral analysis, pitch analysis, formant analysis, intensity analysis, other functionalities for drawing the Cochlea gram, Spectrogram, speech signal plots features and most important that it is open-source software. The speakers were asked to pronounce each word and keeping in mind as stutter of word is categorized in list. The recording was continuously carried out until the speaker does not pronounce the word as same as stuttered person is pronounced. The recording sessions lasted for ‘350’ minutes as mimicking the stutter correctly was difficult.

6.1 Artificial Database Details

The developed artificial dataset contain total ‘3000’ stuttered speech samples. There are total ‘2000’ utterances of isolated words, ‘1000’ utterances of compound words and continuous spoken sentences included in the dataset. All collected stuttered speech samples were denoised and filtered for ensuring uniform speech samples quality.

7 Spectrograms Analysis

For the visual analysis of the acoustic characteristics of the collected stuttered speech samples the spectrogram representations were generated by using the Short-Time Fourier Transform (STFT). Spectrograms provides time–frequency visualization of

speech signals gives the identification of disfluency patterns such as repetitions, prolongations, and blocks.

Fig. 1. Spectrogram of sample collected from movie of word "GHABRAT" before and after denoising

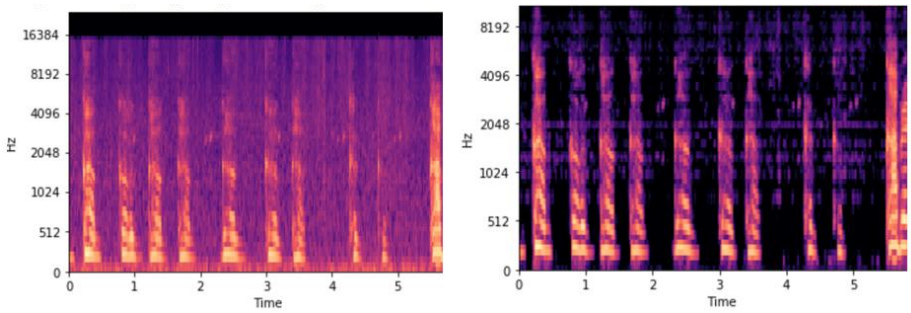
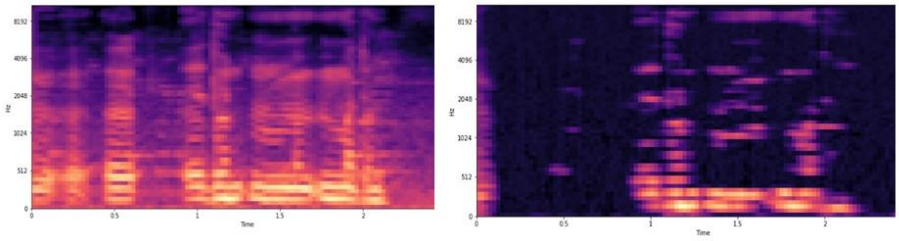


Figure 1 left side shows the spectrogram of the word "GHABRAT" which extracted directly from the movie. In the figure it can be observed that the speech signal contains background noise components particularly in higher frequency bands. This due to music and environmental sounds present in the original movie recording. And the right spectrogram of the same word after the process of denoising. A noticeable reduction in background noise and improvement clarity in frequency distribution is observed. The energy concentration in speech-dominant frequency regions becomes more distinct which indicates effective preprocessing and signal enhancement.

Fig. 2. Spectrogram of sample "ADHUN-MADHUN" collected from speakers before and after denoising



Similarly, Figure 2 left side represents the spectrogram of the word "ADHUN-MADHUN" recorded artificially from speaker. The repetition and prolongation characteristics identified through repeated spectral patterns and elongated energy bands across the time. After denoising process as shown right side the spectral structure becomes cleaner, with improvement in separation between speech components and noise artifacts. The spectrogram analysis confirms that the preprocessing steps can effectively reduced unwanted background noise with preserving essential characteristics of speech. These clean spectral representations can enhance the suitability of the

dataset for feature extraction and subsequent speech recognition or disfluency classification tasks.

8 Acknowledgement

The authors would like to thank Vivekanand College and Shivaji University for providing laboratory facilities and recording equipment support.

9 Conclusion

The paper present the development process of stuttering Marathi speech database using Marathi movie having a stuttering character acting like a stuttering person in the movie. Also, the similar speech samples were also collected from Marathi speakers by mimicking the words and sentences. The collected speech samples were denoised using STFT and made ready after annotations for experimentation. The developed database will be useful for further research by considering many speech denoising or enhancement techniques. Despite previous Marathi stuttered speech datasets the proposed corpus provides balanced gender distribution, structured disfluency annotations, and standardized recording conditions, thereby improving its suitability for machine learning applications. Future work includes applying deep learning models such as CNN and Transformer-based architectures for automatic stutter classification using the developed corpus.
The developed Marathi artificial stuttered speech corpus initially will be used by the authors for the ongoing research work. After completion of the current research work the dataset is planned to be made publicly available for the research community.

References

Barry, W.J., Dalsgaard, P.: Speech database annotation: The importance of a multi-lingual approach. In: EUROSPEECH (1993).

Goldman, J.P., Honnet, P.E., Clark, R., Garner, P.N., Ivanova, M., Lazaridis, A., et al.: The SIWIS database: A multilingual speech database with acted emphasis. In: Interspeech 2016, pp. 1532–1535 (2016).

Karray, F., Alemzadeh, M., Saleh, J.A., Arab, M.N.: Human-computer interaction: Overview on state of the art. International Journal on Smart Sensing and Intelligent Systems 1(1), 137 (2008).

Rusz, J., Tykalova, T., Ramig, L.O., Tripoliti, E.: Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. Movement Disorders 36(4), 803–814 (2021).

Shrishirmal, P.P., et al.: Development of Marathi Language Speech Database from Marathwada Region. In: 18th Oriental COCOSDA, Shanghai, China (2015).

Sutton, J., Austin, Z.: Qualitative research: Data collection, analysis, and management. The Canadian Journal of Hospital Pharmacy 68(3), 226 (2015).

Waghmare, S., Deshmukh, R., Shrishrimal, P., Waghmare, V., Janvale, G.: Stuttered Isolated Spoken Marathi Speech Recognition by using MFCC and LPC. International Journal of Innovations in Engineering and Technology (IJIET) 8, 122 (2017).

Waghmare, V.B., et al.: Emotion recognition system from artificial Marathi speech using MFCC and LDA techniques. In: Fifth International Conference on Advances in Communication, Network, and Computing (CNC) (2014).

Yu, G., Mallat, S., Bacry, E.: Audio denoising by time-frequency block thresholding. IEEE Transactions on Signal Processing 56(5), 1830–1839 (2008).