IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 October 2023, accepted 10 November 2023, date of publication 20 November 2023, date of current version 28 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3334790

RESEARCH ARTICLE

Dementia Speech Dataset Creation and Analysis in Indic Languages–A Pilot Study

SUSMITHA VEKKOT[1], NAGULAPATI NAGA VENKATA SAI PRAKASH[2], THIRUPATI SAI ESWAR REDDY[2], SATWIK REDDY SRIPATHI[2], S. LALITHA[1], DEEPA GUPTA[2], MOHAMMED ZAKARIAH[3], (Member, IEEE), AND YOUSEF AJAMI ALOTAIBI[3], (Senior Member, IEEE)

[1]Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru 560035, India
[2]Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Bengaluru 560035, India
[3]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia

Corresponding author: Deepa Gupta (g_deepa@blr.amrita.edu)

This work was partially supported by the Researchers Supporting Project number (RSP2023R322), King Saud University, Riyadh, Saudi Arabia.

ABSTRACT The paper describes the creation, analysis and validation of a multilingual Dementia Speech dataset for Indic languages. Three popular Indian languages viz. Telugu, Tamil and Hindi are considered for the pilot study. Dementia and associated Alzheimers disease affect a large section of Asian population. Though there are promising studies in dementia detection focussed on Western ethnicity, the absence of a clinical dementia dataset for Indian languages forms the primary motivation for this study. This pilot study aims to overcome the challenges associated with data collection and validation in a clinical setting and deal with situations wherein clinical data is not readily available. The Indic dementia dataset is an enacted non-clinical dataset created from the manual translations of the benchmark clinical English DementiaBank dataset. The dataset created is validated using features extracted from the benchmark. The feature evaluation revealed a similarity of 92.6% for silences, 92% for mean pitch (Hz), 84.7% for jitter and 90.3% for shimmer. Subjective evaluation was also conducted based on clarity and similarity of utterances with DementiaBank data. An average MOS of 3.9 for clarity of speech and 3.76 for similarity with respect to DementiaBank was obtained across all three languages. A baseline classification using state-of-art deep network architecture gave a maximum of 78% accuracy in dementia detection using the Indic dementia dataset. The pilot experimentation in this work gives promising insights into the development of a multilingual dataset for analysis of clinical speech patterns in early dementia in the Indian population.

INDEX TERMS Dementia, Indic speech, pitch, silences, jitter, shimmer, Pearson correlation, cosine similarity.

I. INTRODUCTION

Dementia refers to various brain disorders that impact memory, thinking, demeanour and sentiment. Early symptoms of dementia can include memory loss, difficulty completing routine tasks, speech and vocabulary problems, and character changes. Alzheimer's disease (AD) is the most common cause of dementia. By 2030, it is estimated that the global outlay of dementia could rise to US$2 trillion, swamping health and social care strategies [1]. As far as Indian statistics is concerned, 11.44 million people in India are expected to be living with dementia, by 2050, according to the Global Burden of Disease study [2]. The most common types of dementia are Alzheimer's disease (AD), Vascular dementia, frontotemporal dementia, and Lewy's dementia [3]. Dementia is a condition which typically affects the elderly, and as the lifespan is longer, the early stage dementia can lead to advanced AD. Rarely can younger people acquire dementia, which is referred to as early-onset dementia [4]. Dementia can affect anyone, regardless of age,

The associate editor coordinating the review of this manuscript and approving it for publication was Stavros Ntalampiras.

sex, gender, or origin. However, statistics show that particular populations are more prone to develop it [5]. Studies have shown that women are more prone than males to develop dementia [6]. Furthermore, compared to other groups, black and minority ethnic people have a greater risk of early-onset dementia [7]. Several factors such as aging, medical history, genetics, and general lifestyle can increase the chances of developing the disease at an early stage. Nevertheless, we can delay or arrest the deterioration with proper medication if detected early.

Existing techniques of diagnosis often involve prolonged medical evaluations, involving lengthy questionnaires. The diagnosis of neurodegenerative syndromes, including AD, stay predominantly clinical and psychometric despite the remarkable advancements in information and communication technologies [3], [8]. Although no cure for this neurodegenerative disorder has been found, experts agree that early intervention is critical to delaying onset [9]. The early diagnosis of dementia is of paramount importance as it facilitates greater efficiency of pharmaceutical therapies for mitigating the side effects [10]. Additionally, early phases of cognitive decline could be stabilized or controlled in some cases. Consequently, there has been a need to substitute conventional pen-and-paper screening tests with adequate automated diagnostic methods [11]. There is a demand for cost-efficient and scalable methods to identify AD from the initial stage. Thus, researchers worldwide are trying to find non-invasive early-detection techniques and treatments for these disorders. Problems with inappropriate word usage, dysfluencies, differences in language and speech, etc characterize early symptoms of dementia. This makes current research methods in speech and language processing appropriate for the early detection of cognitive impairment due to dementia.

Dementia detection from spontaneous speech has been performed using speech input-based, language-based (text input-based), and multi-modal approaches. Investigations using speech and language for the detection of dementia or related disorders have been conducted in many languages, including English [12], French [13], German [14], Swedish [15] and Japanese [16], to name a few. Most studies are focused on textual feature representation and acknowledge that undersized datasets are a restriction and explain the challenges in gathering more data, including the difficulties in patient recruitment, the cost of running clinical investigations, and the manual effort needed for annotation. Among speech input-based strategies, prior work has focused on utilizing acoustic features [17] such as pitch, pause duration, speech rate, shimmer, or using standard feature banks [18]. These features capture diverse paralinguistic elements applicable to AD speech. Effectively combining these features from diverse speech segments to enhance the prediction of dementia is an ongoing research problem that our work addresses.

The main challenge in dementia detection is the non-availability of regional language datasets to analyze and detect dementia in local patients. Most existing studies utilized English datasets recorded in a clinical environment such as the DementiaBank dataset [19], [20] for dementia detection in English. The DementiaBank dataset depicts dementia and healthy controls describing a cookie theft picture to the instructor. The necessity of a balanced dataset for dementia and associated Alzheimer's disease detection led to the development of ADReSS dataset [21], [22]. The dataset aims to provide a standardized platform for researchers to evaluate and compare their AD recognition methods using spontaneous speech. This is important because previous studies have used different datasets, which can make it difficult to compare and reproduce results. It includes spontaneous speech from a diverse population, with variations in audio quality and imbalances of gender and age distribution.

The first challenge in the development of dementia detection systems is to identify a set of features that can be shared between languages and introduce a new set of features to model the hierarchy in which dementia patients and controls produce information units in the form of speech. The pathological pathway and the contributing elements behind this relationship have yet to be ascribed. Long-term studies are needed to estimate the importance of speech disability and dementia in local languages. The human voice is intricately merged with the psychological and physiological aspects of human behaviour and sentiment. Therefore, detailed investigations of the contributing factors to cognitive impairments due to disorders of the central nervous system require immediate attention. Research can benefit a large set of the local population, especially the elderly. This research discusses the creation and analysis of an enacted dementia dataset in Indic languages, viz. Telugu, Tamil and Hindi. The Indic dementia dataset consists of speech samples and transcripts for an Indian-translated version of the DementiaBank dataset, the details as discussed in section III. The main contributions of the research in this paper are:

- Creation and detailed feature analysis of an enacted multilingual Indic speech dataset for dementia detection covering three languages viz. Telugu, Tamil and Hindi
- Comparative Analysis of the features extracted from English DementiaBank dataset and the Indic dementia dataset
- Performance evaluation using standard metrics for a dementia detection task on the Indic language dataset developed in the work

The rest of the paper is organised as follows. Section II gives the related literature in the domain. Section. III describes in detail the methodology followed for the creation of the indic dementia dataset. Section. IV discusses the feature analysis performed and the performance evaluation of the results. The paper is concluded in Section. V with useful insights into future.

II. SURVEY OF EXISTING LITERATURE

There is an urgent requirement for cost-effective and customizable techniques to identify dementia, including its most subtle forms, especially from Subjective Memory Loss (SML) [23], to more serious conditions such as Mild Cognitive Impairment (MCI) [24] and Alzheimer's Dementia [25]. The need for dementia and Alzheimer's disease (AD) biomarkers that are inexpensive, safe, accurate, and non-invasive is the driving force behind many modern investigations [26]. Digital biomarkers have become increasingly common even though clinical evaluation is still the predominant analysis method, prognosis, and associated treatment method. The possibility of AI-enabled speech and language analysis as a biomarker for assessing the severity of the disease has shown promise. Problems with inappropriate word usage, impaired reasoning, differences in language and speech, etc. characterize early symptoms of dementia. This makes current research methods in speech and language processing appropriate for the early detection of cognitive impairment due to dementia. Researchers have resorted to machine learning and deep learning models as the main approaches to solving the problem.

Among dementia-based studies, the most prominent longitudinal observational study is the Framingham Heart Study (FHS) [27]. This study used 1264 voice recordings of neuropsychological tests given to individuals. The recordings had at least two speakers and lasted 73 minutes on average. Of the total number of voice recordings, 483 belonged to persons who had normal cognition (NC), 451 to those who had mild cognitive impairment (MCI), and 330 to those who had dementia (DE). Two deep learning models—a two-level long short-term memory (LSTM) network and a convolutional neural network (CNN) that used audio recordings were used to identify whether a participant had only NC or only DE, as well as to distinguish between recordings corresponding to DE from those corresponding to recordings corresponding to neither (i.e., NDE (NC+MCI)). Based on 5-fold cross-validation, the LSTM model distinguished between instances with DE and those with NC with a mean accuracy of 0.647±0.027, and F1 score of 0.596±0.047. The CNN model successfully distinguished between cases with DE and those with NC with a mean area under ROC (AUC) of 0.805±0.027, a mean balanced accuracy of 0.740±0.015, and a mean weighted F1 score of 0.742 ±0.033.

Despite loss of memory, there is also a growing interest in language impairments as a characteristic [28]. MCI is considered one of the first detectable indicators of the disease and plays a major role in the identification of the early stages of the disease [29]. Language deficiencies, among other cognitive impairments, deteriorate over the course of Alzheimer's disease. Language impairment in AD is primarily brought on by a deterioration in the semantic and pragmatic levels of language processing, which frequently accompany aphasia and dysarthria [30]. The inability to find the right words is one of the first indicators that communication is suffering. In addition to being affected by the disease, aging naturally changes the voice and speech over the course of a lifetime. The results of the initial investigations on older people's articulatory control during speaking motions showed that they performed less effectively than young people. They hypothesized that ageing causes a loss in movement amplitude accuracy, which affects temporal voice characteristics [31]. Literature [32] proposes a framework for identifying spoken languages using deep learning based on their frequency and timing. Auditory utterances are transformed into spectrograms when creating this framework. Subsequently, a convolutional neural network (CNN) is used to extract features from images for classification. The softmax activation function is finally used for the classification of many languages.

A. SPEECH ANALYSIS STUDIES

Studies on the use of speech and voice analysis in the detection of neurodegenerative diseases have increased [33], [34], [35], [36], [37]. The researchers in [28] used a systematic review to identify these traits and their diagnostic efficacy. Studies using innovative techniques for automatic speech analysis to gather behavioral evidence of linguistic deficits and their diagnostic efficacy for Alzheimer's disease and mild cognitive impairment [38], [39] were unravelled. The observed characteristics appear to represent signs of cognitive aging in elderly people. Research on the exact characteristics and cognitive changes involved may continue, leading to a more abstract representation of speech features and the establishment of a relationship between speech and cognitive decline in a targeted dementia population.

1) DEMENTIA DETECTION FOR DIVERSE POPULATIONS

An analysis of research on speech-based dementia detection over the past decade revealed that multiple nationalities have been analyzed for dementia detection studies based on speech characteristics. For the Spanish-speaking population, [40] examined the use of stop time and speech pace for Alzheimer's detection. Speech recordings from 68 participants, 37 of whom had dementia disease, and 31 healthy controls were analyzed for the study. The results showed a sensitivity of 84% and a specificity of 96%. Following the same lines, Orozco-Arroyave et al. [41] conducted a study to investigate the use of speech analysis for the detection of Alzheimer's disease in the Spanish population. 65 participants, including 35 people with Alzheimer's disease and 30 healthy controls, collected their spontaneous speech recordings. Characteristics such as articulation rate, speech rhythm, and voice quality are significant for dementia detection, resulting in an accuracy of 82.9%. Ahmed et al. [42] carried out a dementia study in the Urdu-speaking community. Speech recordings from 36 participants, including 18 people with dementia and 18 healthy controls, were analyzed. A sensitivity of 94% and a specificity of 78% were achieved with pitch and speech rate. In a study conducted, Satt et al. [43] examined the use of speech speed and pitch fluctuation to identify dementia. Speech recordings of 25 participants, including 12 people with dementia and 13 healthy controls, were analyzed. The outcomes demonstrated 100% sensitivity and 92.3% specificity.

Fraser et al. [44] recorded speech from 95 participants, including 45 with Alzheimer's disease and 50 with healthy controls. It was found that pause time and speech rate could accurately distinguish between individuals with dementia and healthy controls, with a sensitivity of 80% and a specificity of 84%. Using the same parameters, another study conducted by López-de-Ipiña et al. [45] on the Basque-speaking population on speech recordings from 82 participants, including 40 with dementia and 42 healthy controls that showed sensitivity of 82.5% and a specificity of 82.1%. Demirtas et al. [46] investigated detection of dementia in the Turkish population with speech collected from 30 participants, including 15 with dementia and 15 with healthy controls. It was found that speech features such as fundamental frequency and formant frequency could accurately distinguish between individuals with dementia resulting in a sensitivity of 93.3% and a specificity of 86.7%. The effect of voice quality parameters viz jitter and shimmer for dementia detection was examined by Yigit et al. [47] on a similar Turkish population with speech recordings collected from 40 participants, including 20 with dementia and 20 healthy controls. This resulted in a sensitivity of 80% and a specificity of 95%. The study in [5], [48] has been directed towards identifying how south Asian people perceive dementia. The purpose of the qualitative synthesis was to point out the gaps in the literature and offer fresh perspectives on our understanding of South Asians' attitudes, perceptions, and beliefs around dementia.

Parthasarathy et al. [49] conducted a study to investigate the use of speech-based biomarkers for the early detection of Alzheimer's disease. The study involved the analysis of speech from 47 participants, including 21 with dementia and 26 healthy controls. It was found that pause duration and articulation rate, could accurately distinguish between individuals with dementia with an accuracy of 91.5%. Quan et al. [50] investigated pitch and energy features for dementia detection. The study resulted in the highest accuracy of 87.3%. Machine learning (ML) algorithms were investigated in [51] to assess speech patterns for dementia identification. The support vector machine (SVM) technique was employed in the study to classify voice samples from participants with and without dementia. The findings demonstrated that the SVM algorithm has a dementia detection accuracy of 84.6%. In another similar study using ML, Wang et al. [52] used a random forest algorithm to analyze the data. The results showed that the random forest algorithm had an accuracy of 87.5% in detecting dementia. The efficacy of MFCC, GTCC, fundamental frequency, formants, signal energy, jitter, and shimmer speech characteristics were investigated in [53] for the purpose of dementia classification. For the dementia classification challenge, deep learning (DL) models were also investigated in addition to machine learning (ML) models like random forest, RepTree, and support vector machines [53], [54]. The best outcomes in the dementia recognition task utilising the Dementia Bank Pitts corpus dataset, employing the recommended collection of characteristics indicated above, are 87.6% with an ML model and 85% with a DL model.

B. DISCUSSION

Despite the fact that several studies have looked into language and speech features for the identification of Alzheimer's disease and mild cognitive impairment [37], [44] and have suggested various machine-learning models for this task [18], [55], the field still lacks balanced benchmark data against which different methods can be systematically compared. Although research in this domain has been promising, the datasets that have been used are often unbalanced and variable across different studies. Therefore, it is challenging to compare the effectiveness of the algorithm across different modalities.

Prosodic, voice quality and cepstral features make up the important speech feature set for the detection of dementia and associated Alzheimers from spontaneous speech [56], [57], [58], [59], [60]. Individual speech recordings made using digital technology are an appealing way to measure cognition, but not many systems might automatically interpret the data. The primary research problem is to investigate how speech characteristics can be used for dementia and AD recognition using balanced data and collaborative tasks, such as those offered by the standard benchmark datasets [9], [18], [61], [62]. In order to provide the community with benchmarks for the comparison of speech and language methods to cognitive evaluation, these tasks have brought together groups working on this dynamic area of study.

The findings of studies on the use of speech for dementia screening are encouraging. The analysis of speech patterns and attributes using machine learning and deep learning algorithms has produced highly accurate dementia detection rates. These techniques have the potential to be employed as a noninvasive, economical approach for dementia early detection. Further studies are required to confirm these results among a wider range of groups, address concerns about data privacy, and resolve ethical issues. The necessity of uniform evaluation procedures is one of the difficulties in employing speech analysis to identify dementia. Inconsistencies in the results may arise from the use of various speech parameters and analysis methods in various research. Further complicating the development of a universal method for speech analysis is the possibility that the characteristics of speech vary among dementia patients. The potential for false positives or false negatives is another restriction of speech analysis for dementia screening. It can be difficult to distinguish between people with dementia and healthy people, as speech changes can be caused by other conditions, such as aging or hearing loss. It might be difficult to recognise dementia in its early stages since some people with the disease may not develop speech problems until later stages of the disease.

According to recent studies [63], [64], a challenge in detecting cognitive decline through dementia is the lack of standardized data sets for analysis. Most of the work in speech-based dementia detection used the DementiaBank dataset [19], [20]. The ADDreSS challenge furnished a prospect for different strategies to be executed on a balanced dataset and eased the typical inclinations associated with other AD datasets. [21]. Both these datasets are widely used by the research community for the analysis and detection of AD-related dementia. Both datasets are recorded in English and, therefore, do not cater to a large number of multilingual Indian population. However, most research only targets a tiny subset of the monolingual population, and the assessment criteria are limited to particular aspects in benchmark datasets. There is a gap in research on a thorough investigation of dementia detection across a variety of demographics, with a focus on people with limited resources and those who speak several languages. To solve the problem, this study aims to create a multilingual data set that can be used to identify dementia in different Indian languages. The main characteristic of the dataset is that samples are produced from the benchmark clinical dataset by translating them into a particular language based on data analysis and visualization of numerous crucial elements for dementia studies. The created dataset is thoroughly analyzed to determine its viability for use.

III. METHODOLOGY

The main contribution of research in this work is the design of an Indic dementia speech dataset from the publicly available clinical DementiaBank dataset [19], [20]. The various modules in the creation of the data set are divided into three steps: designing the database, creating the database, and analyzing the created data set. Fig. 1 shows the various stages of database design followed in this paper. The methodology is divided into various blocks, the details of which are discussed in the following subsections.

A. DATABASE CREATION AND VISUALIZATION

English DementiaBank dataset [19], [20] is taken as a reference for the creation of the data set in this work. As shown in Fig. 1, audio files (dementia and control) and transcripts of these files were taken from the dementia bank. The translated transcripts are made in the three Indic languages viz. Telugu, Tamil and Hindi using English transcript files. The transcripts are designed by listening to the original English audio and noting down the variations in pauses and inundations that are evident in the dementia data samples. The transcripts are then translated for each of the native scripts by adding special characters and pauses, as discussed more in the database design section. After the transcripts are made ready in the native languages, the audio is recorded by maintaining the voice characteristics, pauses and word rate for a dementia patient, and the voice quality of the original recordings. This process is repeated until a suitable result is obtained, and continued until all of the audio files are generated. The speech samples were then subjected to various time and frequency domain feature analyses for features ranging from pitch, duration, short-time energy, formants, and cepstral domain. The detailed steps involved in the dataset design and feature analysis are provided in Fig. 2.

1) DATABASE DESIGN

The reference English DementiaBank dataset [42], [43] comprises dementia and control audio and transcript samples. There is an interviewer and an Alzheimer's patient in each dementia interview, and an interviewer and a normal person in each control interview. An image depicting the cookie theft situation was used to create the English dataset. An image is basically used to describe whether the dementia patient can easily identify and completely describe the details they can observe from the picture.

The Indic dataset generated by translating the English dementia dataset contains samples corresponding to dementia audio, control audio, and corresponding transcripts. Special symbols and characters are used in the creation of the transcript to describe different emotions that the person could speak of. The characters or symbols that we used and their purposes are listed below:

- ‘(.)‘: Sudden stop in the sentence spoken.
- ‘////////‘: There is a pause given by the speaker, but there would be a little unclear noise made by the dementia person, and the number of lines describes the total number of seconds that occur between two words spoken by the speaker.
- ‘……’: There is a pause without any noise and this continues for a time which would be the same as the total number of dots present in the transcripts (in seconds).
- laughs:Speaking the sentence while laughing
- *exc*: Exclaiming while speaking a sentence
- *confused*: Speaking a sentence in a confused tone
- ‘< has some text inside>‘: The text inside would be a word or sentence that they speak continuously at a very high speed that a normal person will not be able to understand as they are murmuring something.

Multiple speakers volunteered for creating the Indic dementia dataset. Three languages, viz. Telugu, Tamil, and Hindi were considered for recording. The decision was made based on the availability of native speakers in each of these languages. All speakers were final year undergraduate students in the age group of 20-25. All were native speakers with good command over the respective languages taken for the study. All voice samples except Telugu have a mix of male and female speakers. Sample transcripts in all three languages, along with reference English transcripts, are shown in Fig. 3.

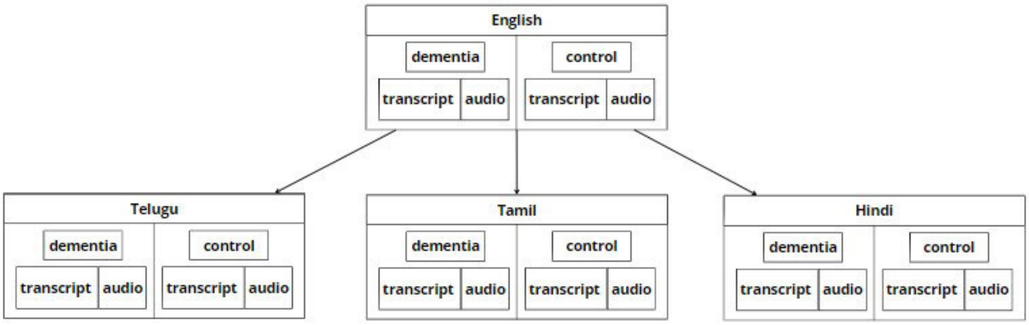


FIGURE 1. Schema of database design.

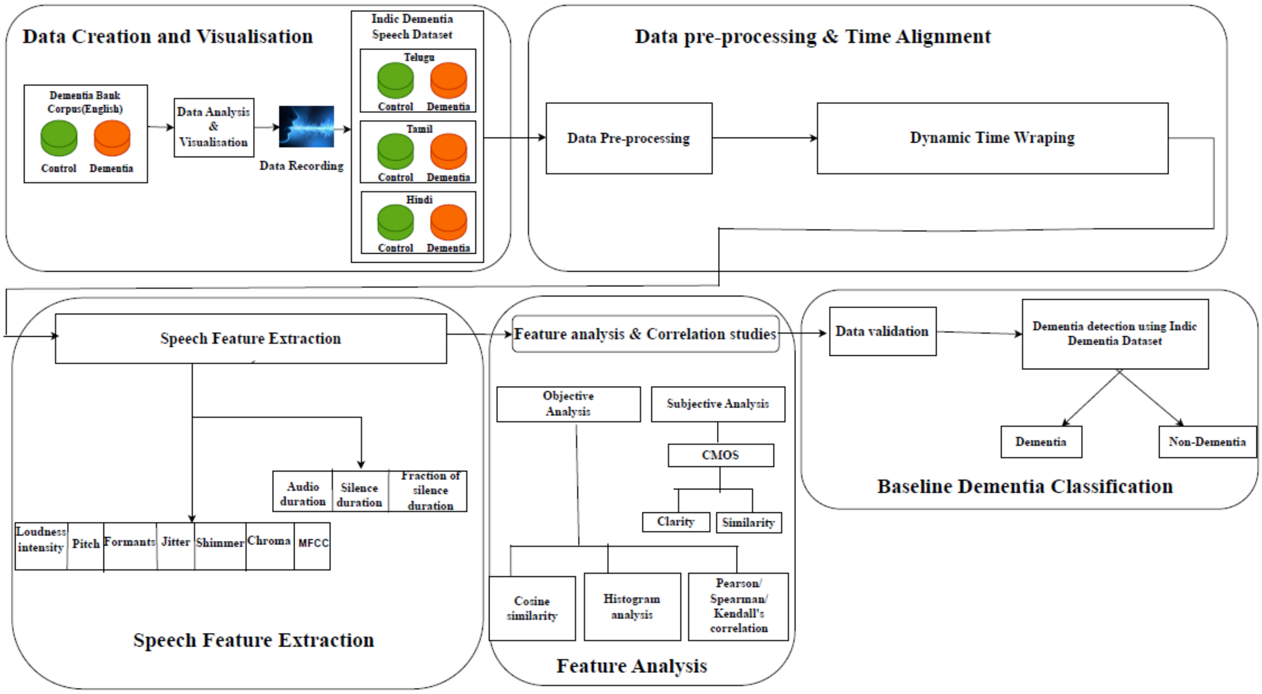
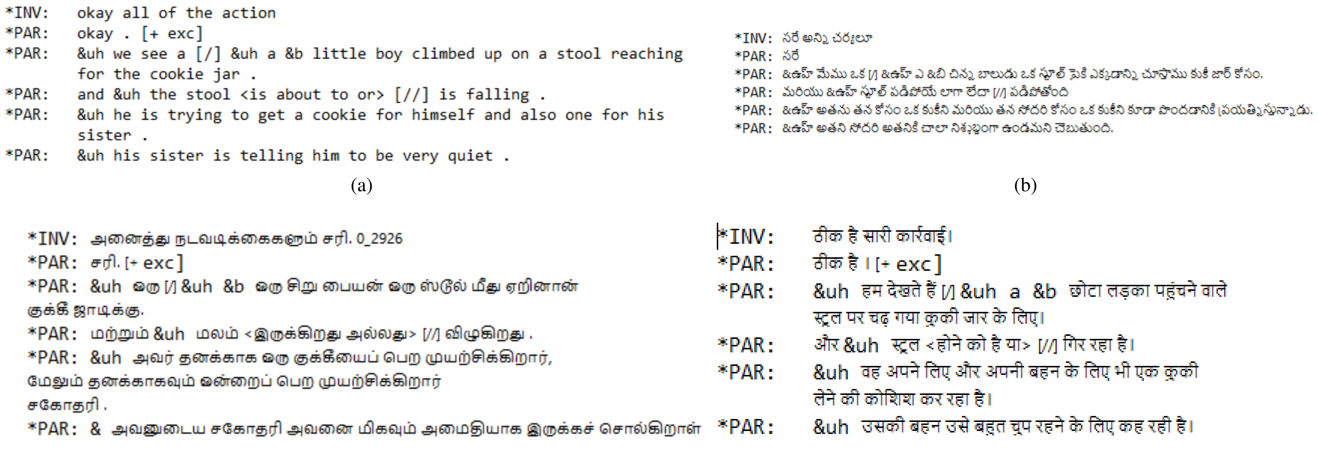

Data Creation and Visualisation
Dementia Bank Corpus(English)
Control
Dementia
Data Analysis & Visualisation
Data Recording
Indic Dementia Speech Dataset
Telugu
Tamil
Hindi
Data pre-processing & Time Alignment
Data Pre-processing
Dynamic Time Wraping
Speech Feature Extraction
Audio duration
Silence duration
Fraction of silence duration
Loudness intensity
Pitch
Formants
Jitter
Shimmer
Chroma
MFCC
Feature analysis & Correlation studies
Objective Analysis
Subjective Analysis
CMOS
Clarity
Similarity
Cosine similarity
Histogram analysis
Pearson/ Spearman/ Kendall's correlation
Feature Analysis
Data validation
Dementia detection using Indic Dementia Dataset
Dementia
Non-Dementia
Baseline Dementia Classification

FIGURE 2. Proposed architecture of creation and feature analysis of Indic dementia speech dataset.

2) RECORDING

According to findings in [65], unspecialized devices such as mobile phones can be reliable replacements for professional recording equipment, especially when conducting general prosodic analysis in real-time environments. A hand-held mobile device (Redmi Note 7 smartphone) has been used for recording, with specifications as provided in Table 2, with a good frequency range of 20Hz to 20kHz. The frequency range of the recorded samples, which is determined by the gender of the participants in each recording, spans from 100Hz to 220Hz. The Telugu audio samples were captured from 9 male participants, while the Tamil samples involved 3 males and 2 females. In the case of Hindi samples, there were 6 male and 3 female participants. A breakdown of the male and female distribution can be found in Table 1 of the manuscript. To summarize, Telugu recordings encompass frequencies between 100Hz and 140Hz, Tamil recordings range from 130Hz to 190Hz, and Hindi recordings cover a frequency spectrum of 120Hz to 220Hz.

A room with non-parallel walls, controlled reverberation atmosphere with reverberation time (RT60) was approximately 0.5 seconds and dimensions of 15ft x 11ft x 8ft was used for speech recording. The microphone was placed around 6-12 inches (15-30 cm) from the sound source. The rationale for keeping different distances is that when a person with dementia is talking, there are times when they may suddenly whisper or even speak very quickly in a very soft voice. The recordings underwent acoustic enhancement,



```
*INV:   okay all of the action
*PAR:   okay . [+ exc]
*PAR:   &uh we see a [/] &uh a &b little boy climbed up on a stool reaching
        for the cookie jar .
*PAR:   and &uh the stool <is about to or> [//] is falling .
*PAR:   &uh he is trying to get a cookie for himself and also one for his
        sister .
*PAR:   &uh his sister is telling him to be very quiet .
```

(a)

```
*INV:  సరే అన్ని చర్యలూ
*PAR:  సరే
*PAR:  &ఉహ్ మేము ఒక [/] &ఉహ్ ఎ &బి చిన్న బాలుడు ఒక స్టూల్ పైకి ఎక్కడాన్ని చూస్తాము కుకీ జార్ కోసం.
*PAR:  మరియు &ఉహ్ స్టూల్ పడిపోయే లాగా లేదా [//] పడిపోతోంది
*PAR:  &ఉహ్ అతను తన కోసం ఒక కుకీని మరియు తన సోదరి కోసం ఒక కుకీని కూడా పొందడానికి ప్రయత్నిస్తున్నాడు.
*PAR:  &ఉహ్ అతని సోదరి అతనికి చాలా నిశ్శబ్దంగా ఉండమని చెబుతుంది.
```

(b)

```
*INV:  அனைத்து நடவடிக்கைகளும் சரி. 0_2926
*PAR:  சரி. [+ exc]
*PAR:  &uh  ஒரு [/] &uh  &b  ஒரு சிறு பையன் ஒரு ஸ்டூல் மீது ஏறினான்
குக்கீ ஜாடிக்கு.
*PAR:  மற்றும் &uh  மலம் <இருக்கிறது அல்லது> [//] விழுகிறது .
*PAR:  &uh  அவர் தனக்காக ஒரு குக்கீயைப் பெற முயற்சிக்கிறார்,
மேலும் தனக்காகவும் ஒன்றைப் பெற முயற்சிக்கிறார்
சகோதரி .
*PAR:  &  அவனுடைய சகோதரி அவனை மிகவும் அமைதியாக இருக்கச் சொல்கிறாள்
```

(c)

```
*INV:    ठीक है सारी कार्रवाई।
*PAR:    ठीक है । [+ exc]
*PAR:    &uh  हम देखते हैं [/] &uh  a  &b  छोटा लड़का पहुंचने वाले
         स्टूल पर चढ़ गया कुकी जार के लिए।
*PAR:    और &uh  स्टूल <होने को है या> [//] गिर रहा है।
*PAR:    &uh  वह अपने लिए और अपनी बहन के लिए भी एक कुकी
         लेने की कोशिश कर रहा है।
*PAR:    &uh  उसकी बहन उसे बहुत चुप रहने के लिए कह रही है।
```

(d)

FIGURE 3. Sample transcripts from English DementiaBank and Indic dementia dataset: (a) English - DementiaBank (b) Indic - Telugu (c) Indic- Tamil (d) Indic- Hindi.

which included the elimination of stationary noise, and a uniform audio volume adjustment was applied to all speech segments to counterbalance any variations that may have been caused by the recording conditions, such as the positioning of the microphone. The angle of arrival (AoA) would essentially be zero degrees or very close to zero because the sound source is in the same location as the microphone. Moreover, AoA estimation techniques are generally used when we use multiple microphones in an array and need to estimate the direction of sound sources from a distance. The recording was carried out with a single microphone and placed very close to the sound source, and there is no meaningful AoA estimate in this case. The recording was carried out in a controlled studio environment with monochannel audio settings. The speech is re-sampled at 16kHz. The majority of the samples were captured at the standard conversational loudness level of 60dB, representative of human speech. Nonetheless, a subset of samples within the clinical dataset (English) involved patients whispering to the interviewer. As a result, the loudness covered a range of 40dB to 60dB. Three sessions were taken for recording a single file to maintain similar gaps between the same content in the untranslated and the translated audio. During a single session, the speakers read about fifteen lines of conversation while maintaining the quality of the speech signal. As the dementia clinical dataset (English) is recorded at the patients' comfort, the background noise varies in the clinical dataset. Due to these observations, the Indic dementia dataset was recorded without any noise cancellation. But it was made sure that no noises were made explicitly while recording to make this as authentic as possible. We tried to replicate the English audio environment as closely as possible by varying the type and intensity of background noise for each recording in Indic languages considered for the study. Speech data statistics for the indic dementia data set are represented in Table 1. The recording conditions that are considered are summarised and listed in Table 2.

B. DATA PREPROCESSING AND DYNAMIC TIME WARPING (DTW)

The data recorded from the above sessions are preprocessed for filtering and removal of high frequency noise. Furthermore, the data need to be time-aligned before further processing. Dynamic time warping (DTW) algorithm is used to compare two temporal sequences that may have different speeds and determine the similarity between the temporal sequences. In general, DTW is a technique that determines the best match between two sequences provided under specific constraints and rules. The various criteria that were taken into account for performing DTW are enumerated below.

- Each index from the first sequence should align with one or more indices from the second, and the converse is also true.
- First index from the first sequence needs to be aligned with the corresponding index from the second sequence.
- The final index from both sequences should be aligned
- The mapping of the indices should grow monotonically

The match that complies with all restrictions and has the lowest cost is taken as the optimal match. To calculate a measure of the sequences' similarity independent of nonlinear variations in time, the sequences are ''warped'' non-linearly. This method of sequence alignment is frequently employed in time series analysis. Although DTW measures a distance-like quantity between two given sequences, it does not always satisfy the triangle inequality. A sample of the corresponding aligned speech files in each language is depicted in Fig. 4 for visualisation.

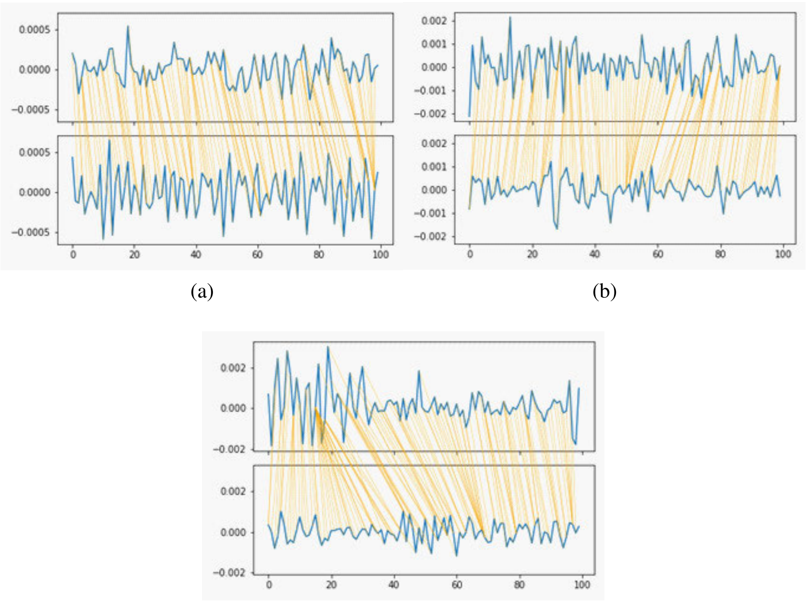
The DTW technique was employed on the recorded audio to achieve proper alignment of the audio. This step was needed to compensate for variations in ambient conditions during recording resulting from disparities in recording environments. DTW proved instrumental in enabling meaningful comparisons and drawing insightful conclusions. Moreover,

TABLE 1. Data Statistics for Indic Dementia dataset.

Language	Dementia	Control	Total samples	Males	Females	Average Duration (s)	Size on disc (MB)	
							Dementia	Control
Telugu	58	58	116	9	0	73.18	887	667
Tamil	41	36	77	3	2	72.70	390	525
Hindi	50	50	100	6	3	59.37	559	527

TABLE 2. The parameters used for recording with specifications of each.

Parameters	Specifications
Channel	Single/Mono
Sampling Rate	48kHz, resampled to 16kHz
Precision	32-bit
Bitrate	1.54Mbps
Encoding	integer PCM
Format	Raw *.wav
Speaker	Non-professional male, female
Recording environment	Normal room with controlled noise Ambient noise level 35 dBA
Recording equipment	Redmi Note 7 smartphone
Sensitivity	-42dB
Signal-to-noise ratio	65dB
Total samples	293
Total size on disc	3555MB
Sentences read per speaker	15
Average words per conversation	10



(a)

(b)

(c)

FIGURE 4. (a) Small chunk of the target language (Telugu) aligned with corresponding audio in English (b) Small chunk of the target language (Tamil) aligned with a small chunk of corresponding audio in English (c) Small chunk of the target language (Hindi) aligned with a small chunk of corresponding audio in English.

dealing with non-aligned audio posed increasing challenges in feature extraction and subsequent correlation analysis. We extended this alignment approach to features such as MFCC within the study as an additional experiment. Nonetheless, the DTW on features led to the exclusion of certain data points in both target and English languages. This, in turn, had an adverse impact on the accuracy of the audio features extracted for frame-by-frame analysis. Therefore, we focused on audio-wise alignment using DTW for the pilot study conducted.

C. SPEECH FEATURE EXTRACTION

The developed Indic language dataset needs to be correlated with the original English dataset to record accuracy and the level of details that are captured at the feature level. Since the recordings were performed manually in a controlled environment, multiple parameters are taken into account for correlating with the actual clinical data. In each of the analysis, the audio-based features are compared with parallel English data from DementiaBank dataset [19] which is taken as ground truth for all investigations. A description on the features extracted is given below:

- Loudness & intensity of utterance: Loudness is the manner in which intensity or amplitude of a sound is subjectively perceived as either "soft" or "loud" to our ears. Sound intensity is a physical sound property measured in decibels (dB). Loudness can be influenced by several factors like sound intensity, frequency, duration, and context. Acoustically, intensity is measured using the sound pressure level (SPL), measured in decibels (dB), with a reference sound pressure level of 20 micropascals (μPa). SPL is calculated using Eqn. 1:
$$SPL = 20log(p/p0) \tag{1}$$
where p is the measured sound pressure and p0 is the reference. In this case, the root-mean-square (RMS) amplitude is calculated as a measure of the average intensity over time.
- The pitch of the utterance $F0$: The pitch of an utterance refers to fundamental frequency component in speech. In this work, the short-time autocorrelation measure (RSS) is used to calculate the pitch of a sample of speech as given by Eqn. 2
$$RSS = \sum_{-\infty}^{\infty} s(m)w(n-m)s(k+m)w(n-k+m) \tag{2}$$
where $s(m)$ is the speech sample, k is the lag and $s(m)w(n-m)$ is the windowed version of the speech signal. After the computation of the autocorrelation function, the peaks are detected using a peak-picking algorithm. The location of the peak $'k'$ corresponds to the pitch period in samples. Using sampling rate of the signal, the pitch period is converted to pitch in Hz. Each audio recording in the target languages was juxtaposed with its corresponding English audio. Leveraging the openSMILE toolkit [66], pitch and formant features were extracted. The autocorrelation-based YIN algorithm facilitated pitch determination by fitting a sinusoidal function [67].
- Formants: Formants are the resonances of the vocal tract due to spectral shaping. The windowed speech is subjected to spectral shaping according to Eqn. 3
$$X(w,\tau) = \frac{1}{p}\Sigma H_{\omega}G_{\omega}W(\omega-\omega_k,\tau) \tag{3}$$
where a window $w(n,\tau)$ centered at τ is applied to the speech segment. The spectral shaping peaks are detected using a peak picking algorithm to yield the formants.
- Chroma: For perceptual discriminative studies, chroma features are commonly used in music applications. The entire spectrum is projected onto 12 bins, which represent the 12 distinct semitones (or chroma) of the musical octave in chroma features, which are highly efficient representations of music audio. Even without the absolute frequency (i.e., the original octave), knowing the distribution of chroma can provide useful musical information about the audio and video. The audio signals are divided into short frames of 20-30 ms, converted into frequency domain using STFT and subjected to a triangular filter banks and energy is determined. The spectrogram is then transformed into a chromagram, which represents the energy of each pitch class across the frequency spectrum. In speech, chroma features can be used to represent the distribution of pitch classes (or musical notes) over time. This can provide information about the tonality of speech and the variations in pitch that occur during speaking.
- Mel frequency cepstral coefficients (MFCC): MFCC gives the short-time power spectrum of speech signal. The mel coefficients mimic the perceptual capability of human ear. The speech signal is subjected to pre-emphasis, after which short-time fourier transform is computed. Mel warping is performed to transform the feature into logarithmic scale to simulate the perceptions in human hearing. Subsequently, the mel frequencies are subjected to discrete cosine transform to retrieve the 13 MFCC coefficients.
- Jitter & Shimmer: For voiced segment of speech. Jitter and shimmer are the cycle-to-cycle perturbations in $F0$ and amplitude respectively and are predominantly used to measure the characteristics of pathological voice. Stress can affect the tension and vibration patterns of vocal folds, subsequently impacting jitter and shimmer. Elevated stress levels have the potential to lead to heightened jitter or shimmer, which, in turn, indicate fluctuations in both pitch and intensity. These variations have the ability to shape the perceived quality of speech. Consequently, our study incorporates pitch, jitter, and shimmer as parameters rooted in features for the analysis of stress and tone. Jitter and shimmer are computed using Eqns. 4 and 5:
$$Jitter = \frac{1}{N-1}\frac{\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \tag{4}$$
$$Shimmer = \frac{1}{N-1}\frac{\sum_{i=1}^{N-1}|A_i - A_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \tag{5}$$
where T_i denotes fundamental period, A_i represents the peak-to-peak amplitudes and N is the count of pitch periods.
- Silences/Fraction of silences: Silences/pauses are observed to be the most statistically significant parameters for dementia detection and analysis [9]. Taking into account this observation from literature, we conducted a thorough analysis of variations in silences/pauses for dementia across all three languages taken for analysis.

TABLE 3. List of performance measures used for analysis of Indic dementia dataset.

Type	**Performance Measure**	**Description**	**Method of Computation**
Objective	Pearson Correlation (r_{XY})	Measures linear relationship between X & Y [68]	$r_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$
	Spearman rank correlation($R_{X\,i}$)	Measures the strength of association between ranked data	For $X = (X_1, X_2, ...X_n)$ and $Y = (Y_1, Y_2...Y_n)$ R_{X_i}= rank of X_i compared to others [69]
	Kendall tau Correlation (K_τ)	Degree of similarity between set of ranked objects	Normalize symmetric difference to between -1 to +1 [70]
	Cosine Similarity (C_s)	Maps the distance between symbolic descriptions into a numeric value [71]	Given features f_a and f_b, $C_s = \frac{f_a * f_b}{\lvert f_a\rvert * \lvert f_b\rvert}$
Subjective	Mean Opinion Score (MOS) [72]	Participants evaluate random speech samples from dataset for similarity	Ranges from 1-5, 1-No similarity, 5-Exactly similar [73], [74]

TABLE 4. Comparison of average pitch and short-time-energy across all the utterances in Indic dementia dataset with the corresponding parallel English DementiaBank dataset.

Language	**Pitch (Hz)**		**Short-Time Energy (STE)**	
	Dementia	Control	Dementia	Control
Telugu	148.87	132.78	13.54	15.25
Parallel English	139.75	120.33	15.22	18.71
Tamil	259.43	202.21	10.48	12.97
Parallel English	196.97	151.57	14.89	16.35
Hindi	110.41	113.07	21.35	20.70
Parallel English	105.97	115.45	24.24	20.50

In addition to the average silence across sentences, the fraction of silence in every audio was also taken as a parameter for analysis and is compared with that in the corresponding samples from the DementiaBank dataset [19], [20]. The fraction of silences is calculated using Eqn. 6:

$$sil_{frac} = \frac{Duration\ \ of\ \ silence\ \ in\ \ an\ \ Audio}{Total\ \ Audio\ \ duration} \tag{6}$$

- Average duration of audio: The average length of the audio in seconds is calculated across all the samples and the coefficients of correlation are plotted. The cosine similarity measures are also computed by comparison with English audios.

D. PERFORMANCE EVALUATION MEASURES

The features analysed in the preceding subsection are analysed using benchmark evaluation metrics for correlation and distance. Different objective and subjective performance measures are used for feature analysis, with details as provided in Table 3.

IV. FEATURE ANALYSIS AND DISCUSSION

The results section highlights the important feature analysis conducted as part of the dataset creation and validation pipeline. The audio signals from English and Indic languages are subjected to dynamic time warping before the feature extraction process. The details of the feature analysis conducted are described in the following subsections.

A. OBJECTIVE FEATURE ANALYSIS

The data created is objectively analysed for the similarity wih respect to time domain and frequency domain-based features. As discussed in literature, characteristics like articulation rate, speech rhythm, and voice quality, are statistically significant for dementia detection from speech. Therefore, to make a justified correlation between original clinical dataset and Indic dementia dataset, we have selected features corresponding to the above parameters. The most prominent features selected for analysis in this work are jitter, shimmer, pitch, formants, silence/ pause durations in seconds, fraction of silences (s), the first thirteen mel frequency cepstral coefficients (MFCC) and chroma-based spectral parameters. For correlation analysis, different metrics like cosine similarity, Pearson correlation, Kendall tau coefficient and Spearmans rank correlation are utilized.

Jitter and shimmer are important voice quality features relating to pitch and amplitude changes in dementia patients. Jitter and shimmer showcases the cycle-to-cycle variability of fundamental frequency and amplitude, respectively, that are extensively used to characterise pathological voice quality. Because they categorise some aspects of particular voices, it is normal to anticipate discrepancies in jitter and shimmer values between healthy and dementia patients. The average jitter and shimmer parameters for both control and dementia files from Indic dataset were computed and cosine similarity with respect to English files are plotted in Fig. 5.

From Fig. 5, it is evident that the dementia samples have a higher cosine similarity for jitter in audio. The similarity of jitter audios are higher than 80%, while shimmer cosine similarity touches above 90% for most cases. Dementia patients tend to speak with a perturbation in fundamental frequency. Also, the cycle-to-cycle amplitude variations are more pronounced due to uncertainties in verbal expression.

In order to account for the fluctuations in F0 (pitch), the average pitch across the utterances has also been compared across languages. A pitch detection algorithm with autocorrelation-based pitch tracking was utilized to find the segment-level pitch values. The loudness/intensity parameter is measured in terms of short-time energy (STE). Comparisons of average pitch and STE have been performed with the results as presented in Table 4.

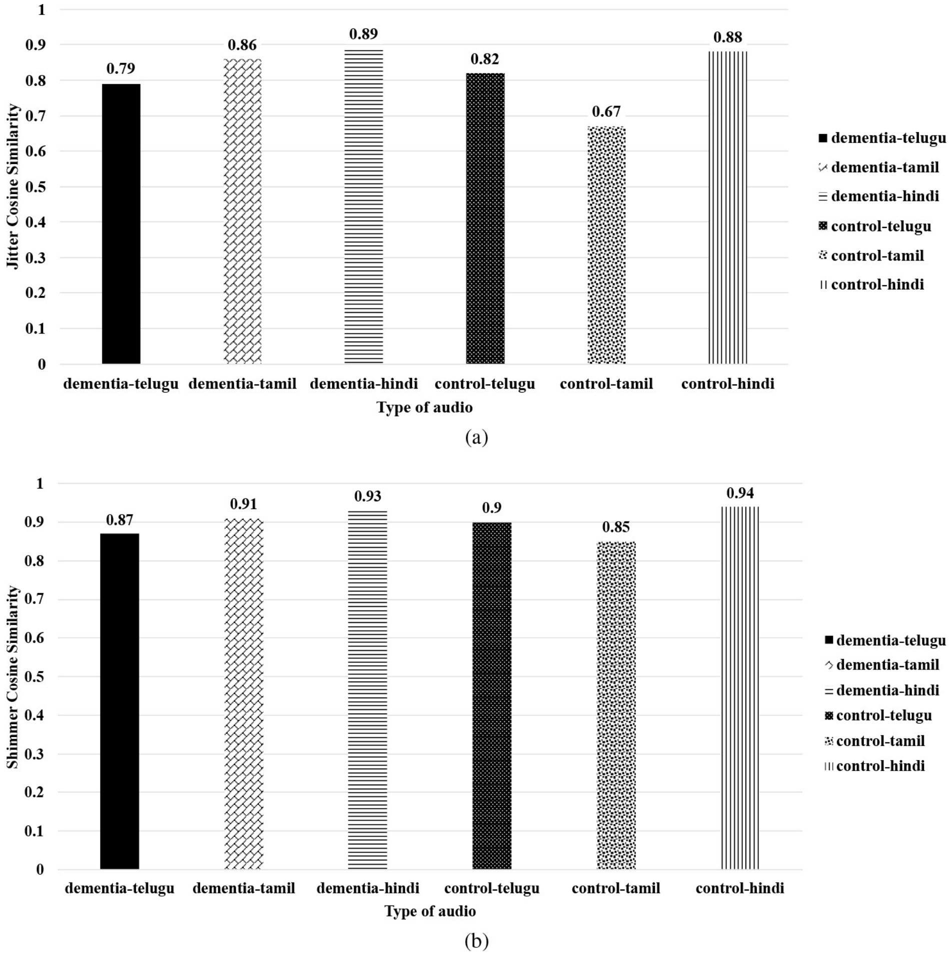


FIGURE 5. (a) Jitter cosine similarity analysis of Indic and English audios (b) Shimmer cosine similarity analysis of Indic and English audios.

From Table 4, it can be observed that the average pitch is closely similar to English utterances in Telugu and Hindi, closely followed by Tamil. The same pattern is observed in both dementia and control samples. In each case, the parallel English utterances from English dataset is taken for comparison. The south Indian languages depict the dementia samples as possessing a higher pitch than that of control subjects. The difference is not pronounced due to the fact that both dementia and control samples were recorded by individuals with same cognitive ability. Still, it is interesting to note that the values are closely matching with the actual clinical data in cookie theft scenario. This essentially points to the efficiency in data creation through enacting/simulation strategy. Similar trends are observed for STE values also.

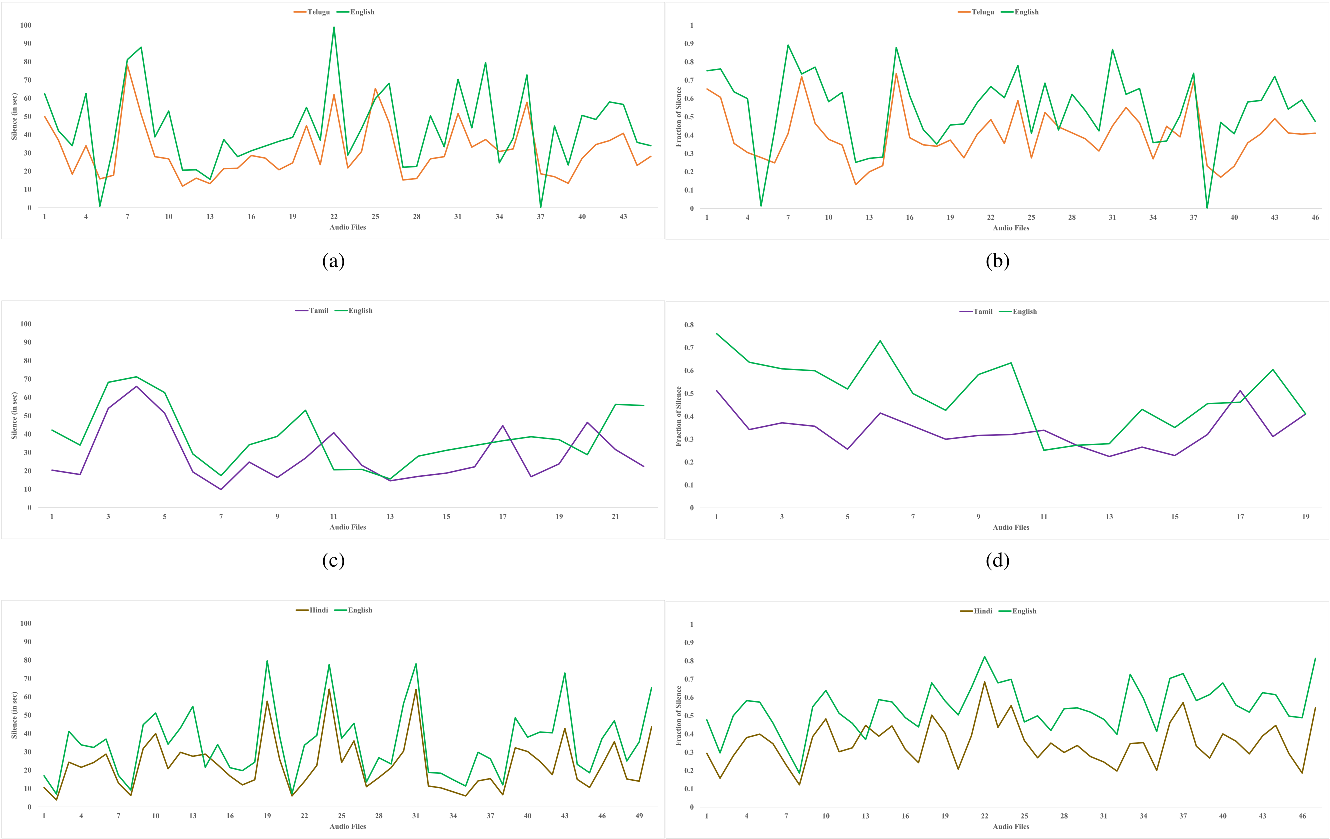
The formants are important parameters for determining vocal tract characteristics. Formant analysis between control and dementia subjects was conducted for all three languages for comparison with English. All results pertaining to the first three formant comparisons are listed in Table 5. From Table 5, it is observed that the F1 range for most samples is around 900-1000Hz and for F2 around 2020 Hz - 2080 Hz However, for F3 there is some variation in the languages, with range varying from 3000-3100 Hz. This can be attributed to the variations in speakers recording for different languages. Most of the time, the analysis is conducted based on overall trends rather than point-to-point values. In most cases, the formant values in English are slightly lesser than Indic languages due to the subtle variations in the verbal aspects of the language. However, the Indic dementia dataset still captures the overall trends.

In the reference dataset, the dementia patients are observed to be speaking with abrupt pauses in between continuous conversations. A thorough analysis was conducted on the silences/ fraction of silences and the audio duration. For sample-wise visualisation, the results are depicted in Fig. 6. The cosine similarity measure is also computed for silences across utterances, the results of which are reported in Table 6.

The cosine similarity between recorded audios in Indic Dementia dataset and the English DementiaBank dataset are as high as 0.94 for Telugu while the fraction of silences in audio is 0.98. The recordings were conducted to enact the nature and location of silences/pauses between utterances

TABLE 5. Comparison of average formant frequency 1 (F1), formant frequency2 (F2) and formant frequency 3 (F3) across all the utterances in Indic dementia dataset with the corresponding parallel English utterances from DementiaBank dataset.

Language	F1(Hz)		F2(Hz)		F3(Hz)	
	Dementia	Control	Dementia	Control	Dementia	Control
Telugu	985.05	1001.59	2013.12	2049.13	3046.11	3101.13
English	973.21	986.29	1987.18	2014.94	3005.47	3047.80
Tamil	956.90	983.18	1954.14	2016.49	2955.69	3054.09
English	922.65	961.37	1923.39	1960.01	2958.21	2962.00
Hindi	1014.91	1004.17	2069.67	2053.40	3128.60	3106.08
English	1012.56	1008.15	2059.29	2047.02	3110.92	3100.84



(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 6. (a) and (b) represents sample wise average pause duration and fraction of silence comparison between English and Telugu audios, (c), (d), (e) and (f) denotes the same for Tamil and Hindi respectively.

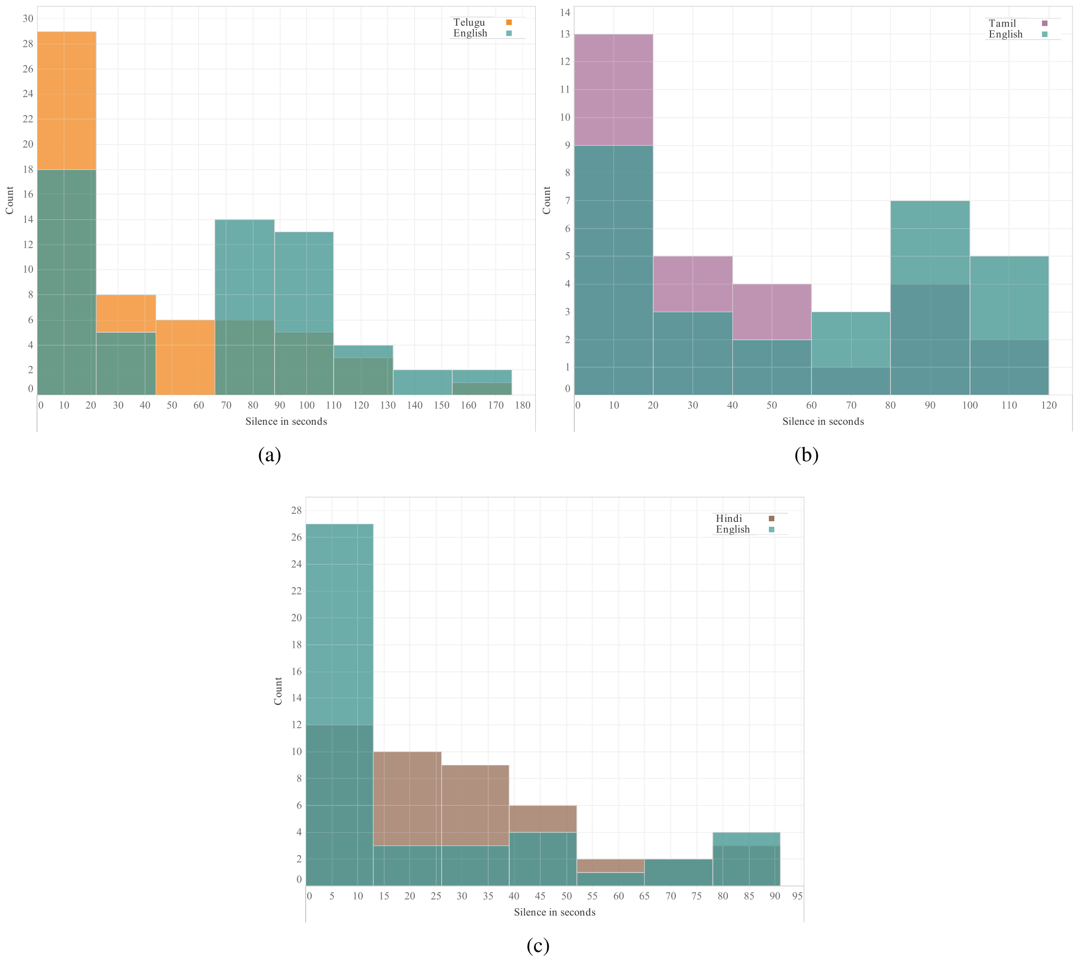
TABLE 6. Cosine similarity scores of mean pitch and pauses.

Language	Cosine Similarity Score:		
	Mean pitch	Silences	Fraction of silences
Telugu	0.91	0.94	0.94
Tamil	0.93	0.86	0.91
Hindi	0.92	0.98	0.98

accurately. In addition to the above, the histograms of silence duration comparisons were also conducted and analysed. The results of this comparison are shown in Fig. 7. From the histograms, it is observed that the sample-wise average silence ranges from 0-20 s for both Telugu and Tamil audios. The distribution of silence regions is almost similar for the three languages with respect to English. Samplewise comparison in Fig. 6 also provides the same observations.

The cosine similarity scores between the mean pitch, silence duration and fraction of silences of the target language and the corresponding English DementiaBank dataset [19], [20] are projected in Table 6 evidently shows that the similarity between pauses and silences is well-taken care of in the data creation. The high cosine similarity values validate that the data creation is performed after the visualization of statistically significant features for dementia.

Duration is one of the most prominent features for early dementia detection from speech. Hence, both dementia and control samples from the Indic dataset are compared with



(a)

(b)

(c)

FIGURE 7. Histogram comparisons of average silence duration between (a) Telugu and English audios (b) Tamil and English audios (c) Hindi and English audios.

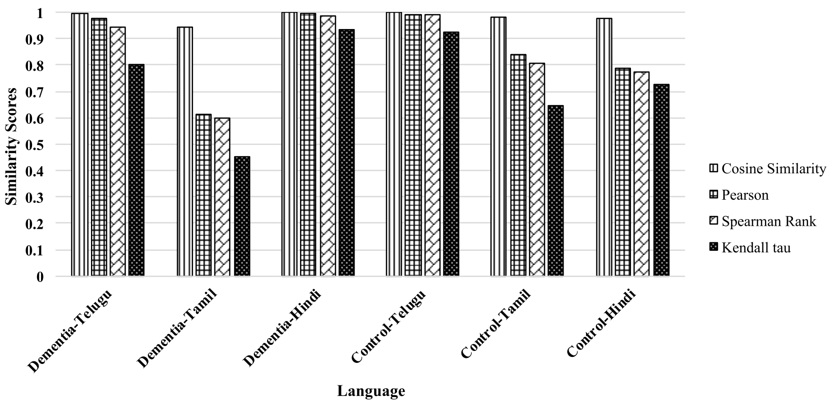
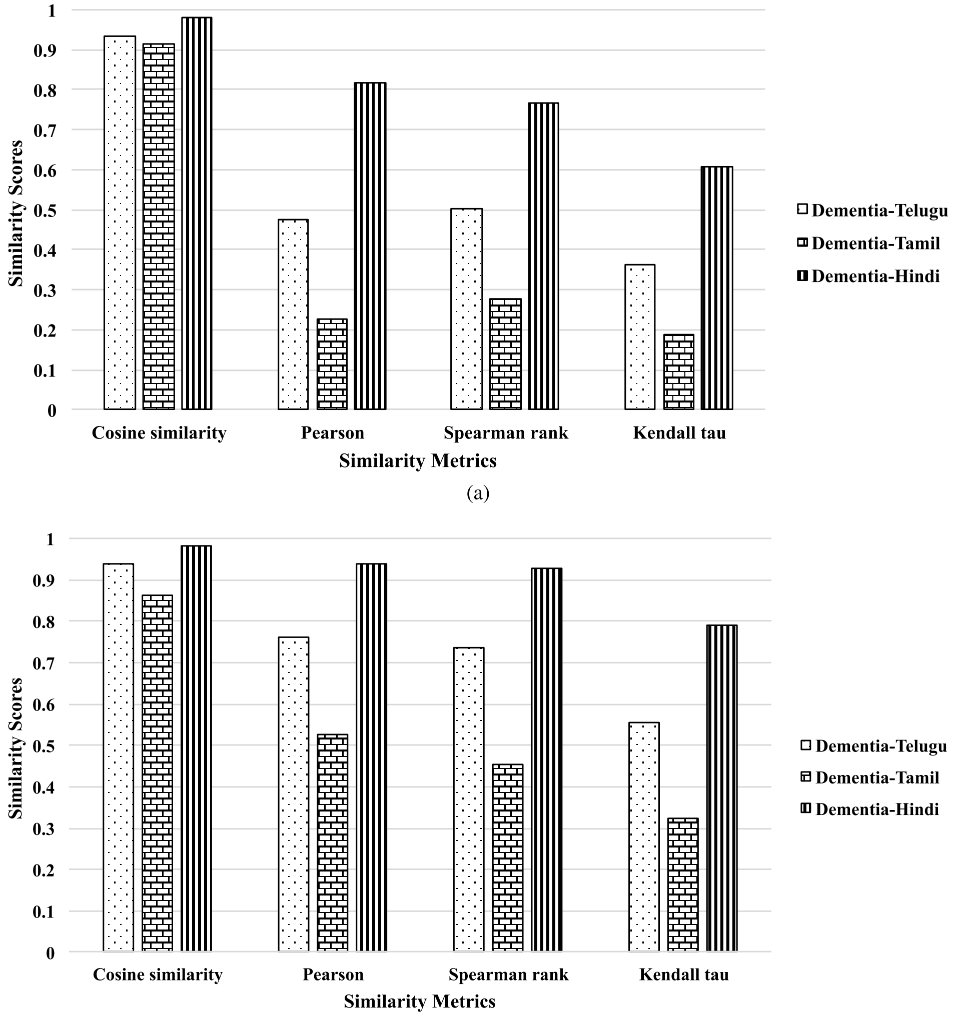


FIGURE 8. Comparison of average duration between the Indic dataset and the corresponding DementiaBank samples using various correlation metrics.

the English for duration correlation using various metrics as mentioned in Table 3 viz. cosine similarity, Pearson correlation, Spearman rank correlation and Kendall tau. The results are projected in Fig. 8. Cosine similarity is observed to be the highest in all cases, with values touching 0.9. Pearson's and Spearman rank correlation are also found to be higher for Telugu and Hindi as compared to Tamil. While recording, care was taken to ensure that the total duration of the utterance



(a)

(b)

FIGURE 9. **Correlation analysis of (a) Fraction of silence duration of Telugu, Tamil and Hindi with English audios from DementiaBank (b) Duration of silence (s) correlation analysis of Telugu, Tamil and Hindi with English audios from DementiaBank.**

is aligned to English utterances. This explains the reason for high cosine similarity and Pearson correlation for the Indic utterances with respect to English. Kendall tau is significant in time series data and computes the association between two variables in terms of relative position labels. Both Kendall tau and Spearman correlation coefficients assess the statistical dependence between two variables in terms of rank. The high correlations indicate that the relative position of each label in the sample is closely similar.

The same analysis was used on the fraction of silences and the duration of silences (s) for each Indic language sample created. The details of correlation analysis are depicted in Fig. 9.

From Fig. 9, it is observed that the silence correlations are better in Telugu and Hindi as compared to Tamil. Also, all four correlation metrics are highest in Hindi recordings. The recordings in Tamil were conducted by a smaller percentage of participants than that in Telugu and Hindi. The discrepancy can be attributed to the fact that the data visualization and understanding have been based on fewer native speakers of the language. On average, the Spearman and Pearson correlation of silences are touching 0.8 in most cases. In addition to the above mentioned time domain feature analysis, the frequency domain features like MFCC and chroma are also computed and compared. MFCCs represent the perceptual characteristics of human audio and is useful in representing the variations in response to different frequencies as perceived by the human ear. For computing the relationship between dementia samples from the two datasets, the cosine similarity metric is used. The cosine similarity between the first 13 MFCC coefficientsÂ of the Indic dataset (m1,m2…m13) and the original DementiaBank samples in English is calculated. The results are plotted in Fig. 10. From Fig. 10, the m1, m2, and m4 coefficients have the highest similarity to the original audio MFCCs. Because of the wide differences in recording language, the

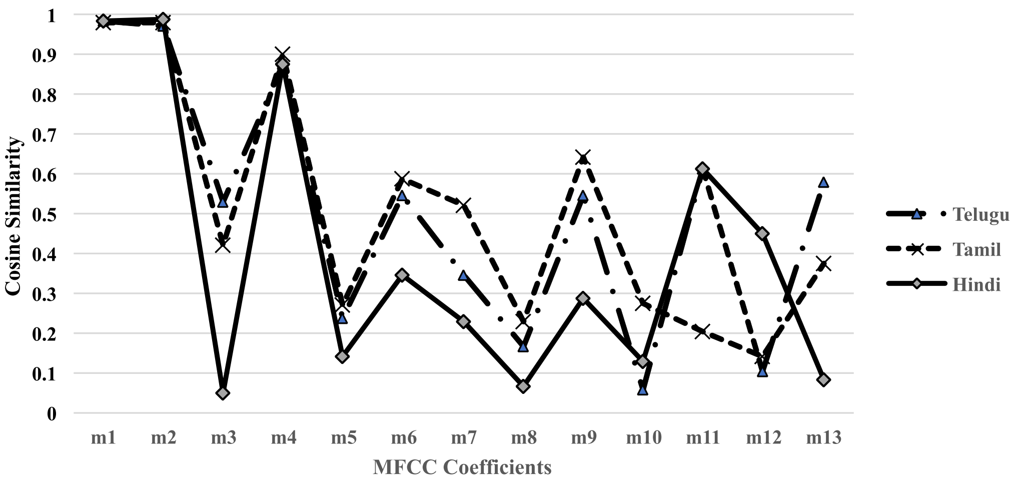


FIGURE 10. Cosine Similarity comparison of MFCC between the Indic dataset and the original DementiaBank samples.

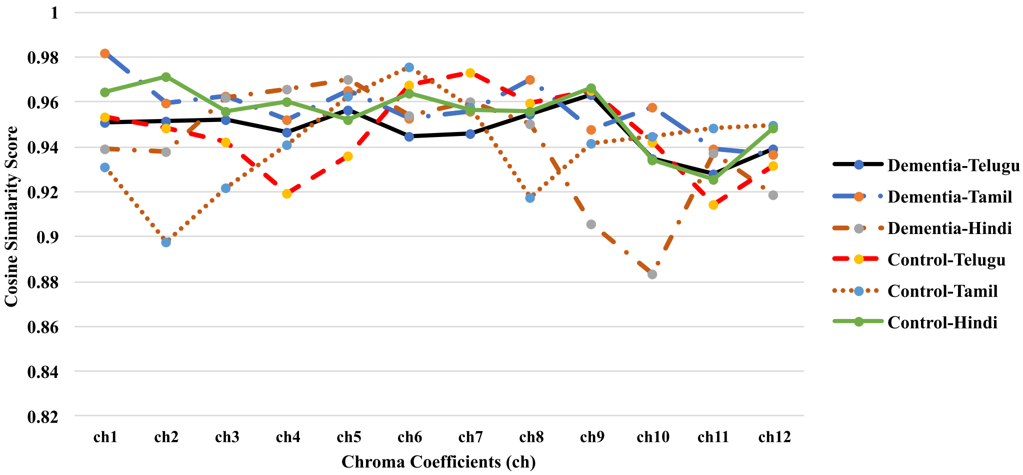


FIGURE 11. Comparison of chroma feature cosine similarity between the Indic dataset and the corresponding DementiaBank samples.

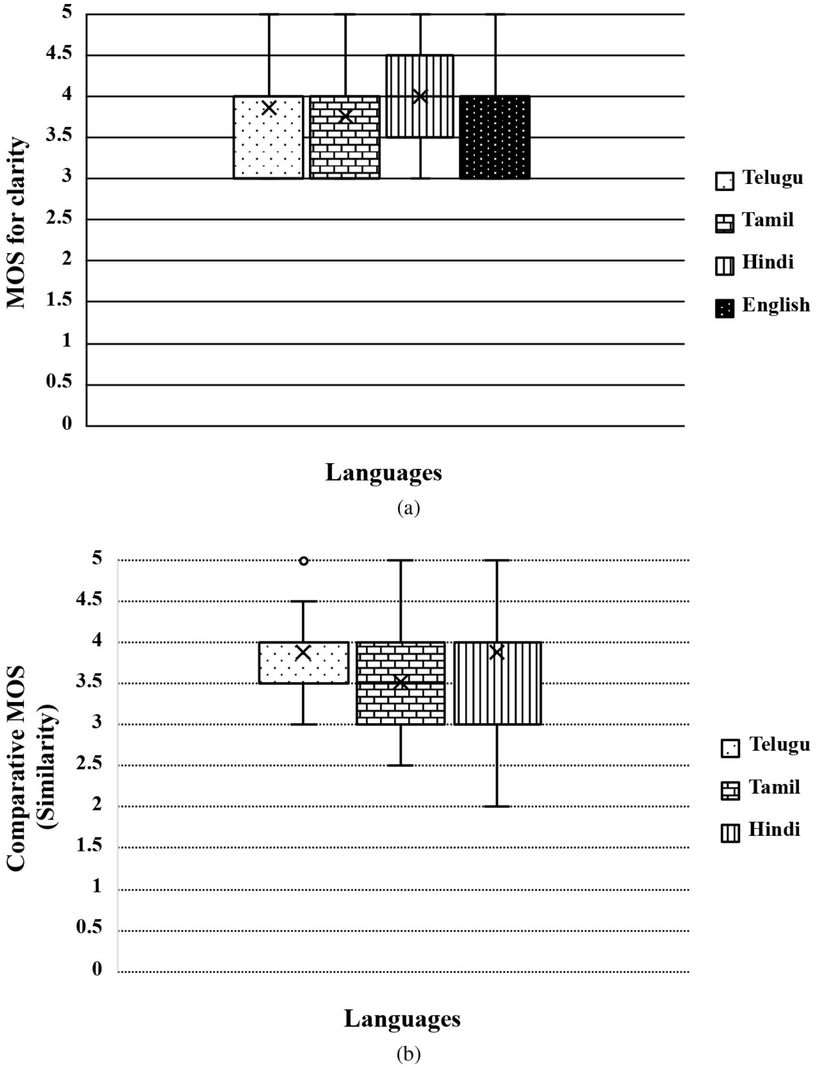
data is falling in the lower similarity regions for higher MFCC coefficients. The coefficients for MFCC vary depending on the language. We examined unnormalized MFCCs for dementia speech in this case.

Relying solely on MFCC-based analysis is inadequate for determining the spectral similarity of audio. As a result, the chroma features (ch1,ch2…ch12) for both the Indic dementia dataset and the English DementiaBank dataset were subjected to cosine similarity measures, the results as reported in Fig. 11. In this study, a chroma feature-based spectrogram is used to distinguish between tonal structure in dementia data from English and Indian languages. Comparing Figs. 10 and 11, it is evident that chroma-based features give a higher cosine similarity than MFCC, with most scores above 0.9 for all languages considered for the recording. Both control and dementia samples have been taken for comparison here. Since chroma features provide information on the distribution of pitch classes, the higher cosine similarity is understandable. Even with respect to pitch comparison in Table 4 and Table 6, higher values of correlation are obtained.

Therefore, from the objective evaluations conducted in this section, the overall cosine similarity with respect to pitch features touch 0.9, while silence and duration similarity scores are higher than 0.8 with respect to all languages considered for analysis.

B. SUBJECTIVE ANALYSIS

Subjective testing is very much essential to ascertain the clarity of speech and the similarity between the original English speech samples and the samples from the Indic dementia dataset. The prevailing method for assessing subjective speech quality is used in this research which is the Absolute Category Rating (ACR). In this method, participants listen to a set of stimuli processed under specific conditions and rate their perceived quality on a defined scale. Another established technique is the Comparison Category



(a)

(b)

FIGURE 12. MOS analysis of (a) Clarity of speech analysis of Telugu, Tamil and Hindi with English audios from DementiaBank (b) Similarity of pauses, silences and utterance semantics analysis of Telugu, Tamil and Hindi with English audios from DementiaBank.

Rating (CCR), where participants evaluate the quality of both reference and processed stimuli in relation to each other [72]. The International Telecommunication Union-Telecommunication (ITU-T) Recommendation P.800 introduced the CCR method for experimental use in laboratories. Drawing from the findings presented in [72], the research employed both ACR and CCR tests to subjectively assess the Indic dementia dataset. The testing involved playing audio samples from both the DementiaBank dataset [19], [20] and the Indic Dementia dataset to a group of 40 listeners, between the age-group of 20-25 and native speakers of one of the three languages selected, with reading and writing proficiency in the languages selected. The audio samples from DementiaBank dataset [19], [20] and the Indic dementia dataset were played one after the other. The listeners were asked to rate the samples on a scale of 1-5, with 5 being the highest quality score, based on several parameters including the clarity of speech, the similarity of translated transcripts, and the replication efficiency of pauses and silence durations in the dementia samples. It is worth noting that this is a comparative scoring strategy. The results obtained in both these cases were tabulated as in Fig. 12. Clarity scores in Fig. 12(a) are grounded in the ACR test, while similarity scores in Fig. 12(b) are derived from CCR.

Fig. 12 refers to the manner of articulation and clarity of recording the speech samples in the Indic dementia dataset. Among the languages, the average MOS was higher for Hindi in terms of clarity of recording. Both Telugu and Tamil showed the interquartile range between 3-4, with a tendency of sliding towards the upper quartile. With respect to similarity with original clinical data as recorded in Fig. 12, the average comparative MOS is ranging from 3.5-4, with a gradual sliding towards the upper quartile. Telugu audios showed a higher similarity to original audio. This is mainly

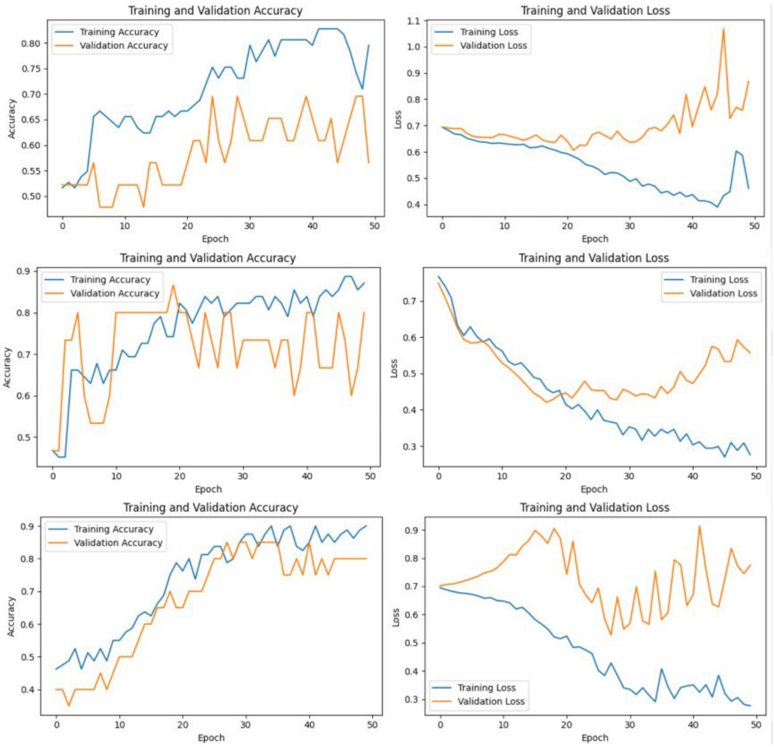


FIGURE 13. Best validation curves using MFCC feature modelling for Indic Dementia dataset for (a) Telugu (b) Tamil (c) Hindi.

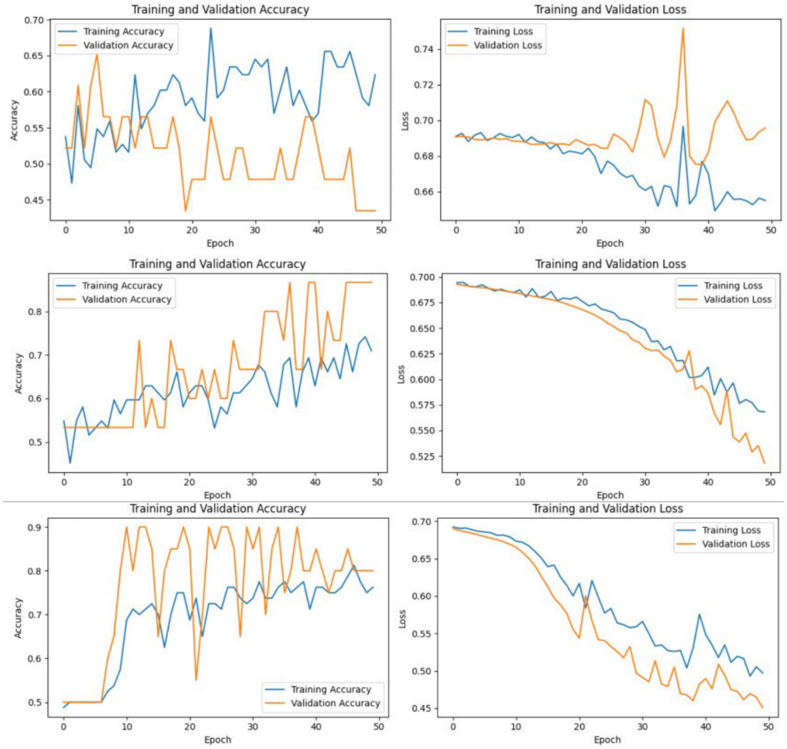


FIGURE 14. Best validation curves using Chroma feature modelling for Indic Dementia dataset for (a) Telugu (b) Tamil (c) Hindi.

attributed to the fact that the number of participants in Telugu data creation was higher than the other two languages in terms of male gender. However, the MOS scores reveal that a mutual agreement of above 3.5 is obtained in all the scores for clarity and similarity among all the listeners who participated in the perception testing experiments.

TABLE 7. Hyper-parameters for dementia detection model using various deep learning architectures.

Parameters	Values
Dataset split	80-20
Configuration	128-64-40
Drop-out	0.2
Activation Function-Dense layer	Relu
Optimizer	Adam
Learning rate	0.001
Loss Function	Binary Cross-entropy
Number of output nodes	1
Activation function-output layer	Sigmoid
Batch size	32
Epochs	50
k-fold validation	5

TABLE 8. Baseline classification of dementia from speech features using various deep learning architectures.

Baseline Classification using MFCC Features					
Language	Model used	Evaluation Metrics			
		Precision	Recall	F1 Score	Accuracy
Telugu	LSTM	0.66	0.66	0.65	**0.66**
	bi-LSTM	0.63	0.63	0.63	0.63
	GRU	0.57	0.57	0.56	0.57
Tamil	LSTM	0.66	0.66	0.66	0.66
	bi-LSTM	0.68	0.68	0.68	**0.68**
	GRU	0.62	0.62	0.62	0.62
Hindi	LSTM	0.74	0.74	0.74	0.74
	bi-LSTM	0.73	0.73	0.73	0.73
	GRU	0.80	0.78	0.78	**0.78**
Baseline Classification using Chroma Features					
Language	Model used	Evaluation Metrics			
		Precision	Recall	F1 Score	Accuracy
Telugu	LSTM	0.51	0.51	0.51	0.51
	bi-LSTM	0.58	0.58	0.57	**0.58**
	GRU	0.53	0.53	0.53	0.53
Tamil	LSTM	0.38	0.44	0.38	0.44
	bi-LSTM	0.56	0.55	0.54	0.55
	GRU	0.70	0.70	0.70	**0.70**
Hindi	LSTM	0.73	0.73	0.73	0.73
	bi-LSTM	0.78	0.78	0.78	**0.78**
	GRU	0.76	0.75	0.75	0.75

C. BASELINE DEMENTIA CLASSIFICATION USING INDIC DEMENTIA DATASET

The Indic dementia dataset created is validated using multiple parameters across temporal and spectral aspects. Further to the data analysis, it is required to test the effectiveness of the speech samples created for a dementia detection scenario. The last part of the pilot study in this paper deals with a case-study based validation of the dataset created by designing a classification system for dementia. MFCCs are widely used for audio analysis in deep learning models and have recently gained popularity for audio classification. As per the feature analysis in Section. IV, higher correlation with respect to DementiaBank [19], [20] were obtained from chroma features. In line with this understanding, the baseline classification was repeated using chroma features. In this work, sequential LSTM, BiLSTM and GRU models are used as baseline for dementia/non-dementia classification using the Indic dementia dataset created. Table 7 provides details of hyperparameters used for developing the baseline models.

The performance is analysed using benchmark evaluation metrics viz. precision, recall, accuracy and F1 score. The results are provided in Table 8. The best validation plots obtained for each language using MFCC features and chroma features are represented in Figs. 13 and 14 respectively. Bi-LSTM and GRU models with the hyper-parameter tuning as mentioned in Table 7 gave maximum performance using all the metrics considered for the study. A maximum F score of 78% is obtained for bi-LSTM model using Hindi audio files. The experiments were conducted on a baseline model and involved no feature selection. Only a single set of features viz MFCC or Chroma were used for classification in each case. Including more features which are statistically significant for dementia can yield better precision, accuracy and F-scores for dementia detection problems.

Three different models, specifically LSTM, bi-LSTM, and GRU, were trained using MFCC features on an Indic dataset. Additionally, all three models were trained using chroma features. Notably, the LSTM model achieved an accuracy of up to 66% for Telugu, while the bi-LSTM model achieved a higher accuracy of 68% for Tamil. The GRU model exhibited exceptional performance for Hindi, achieving an impressive accuracy of 78%. The resultant overall accuracy was computed by averaging the aforementioned values, yielded a total of 70.7%.

During the initial experimentation phase, the primary emphasis was placed on validating dementia speech data using the analysis of specific features that are shared among various Indian languages. While it is acknowledged that English and Indic languages exhibit distinct syntax and linguistic attributes, a comprehensive study of feature correlations indicated the existence of shared characteristics that hold significance in detecting dementia from speech. The collective analysis involves the consideration of pitch and formant values for comparison. Additionally, parameters like jitter and shimmer, which are crucial for feature-wise correlation, remain unaffected by linguistic variations within a multilingual context. Additional elements that underwent analysis include the pauses and the average duration of the audio. These aspects demonstrate variations between dementia and non-dementia subjects, regardless of the languages employed for analysis.

V. CONCLUSION AND FUTURE SCOPE

The Indic dementia dataset creation from speech is conducted as a pilot study for analysing the statistically significant features in dementia analysis from Indian languages. The recordings are performed in three different Indic languages varying in semantics and verbal structure. For validating the dataset created, all dementia-relevant speech features corresponding to time and frequency domain are analysed with various similarity measures and feature-wise correlation is performed. The audio files gave a similarity index of above 80% in all the analysis, particularly above 90%

in the representation of pauses between utterances. The feature evaluation revealed a similarity of 92.6% for silences, 92% for mean pitch (Hz), 84.7% for jitter and 90.3% for shimmer. Subjective evaluation was also conducted based on clarity and similarity of utterances with DementiaBank data. Average MOS of 3.9 for clarity of speech and 3.76 for similarity with respect to DementiaBank was obtained across all three languages. Finally, a baseline classification system is developed for dementia detection using the Indic dementia dataset. The models trained on Indic dementia dataset gave a detection with accuracy of 70.7% using MFCC features. The experiment yielded promising results for utilising the dataset for dementia detection in low-resource scenarios where clinical data is sparsely available.

Certain challenges were associated with the recording and data visualisation of the Indic dementia dataset. The pilot experimentation for acted dementia dataset was conducted using amateur recordists who are undergraduate students in the 20-25 age-range. Also, the number of participants for each recording was limited to 10. The number of speakers recorded for each language adheres to established standards for dataset creation. Since the Indic Dementia dataset is simulated, great care has been taken to emulate the process of generating speech data for individuals with dementia. The selected speakers fell within the narrow age range of 20-25 years, and thus may not provide a comprehensive representation of the entire study population. However, the initial pilot study was undertaken as a foundational phase of experimentation. Based on the insights gained from this study, there is potential to expand the dataset by considering the demographic characteristics of the broader population. Given the dearth of investigations into the speech traits of the dementia-affected Indian population using speech analysis, this study can be seen as an initial effort to establish a model in this respect. Furthermore, the intricate process of identifying a suitable group of volunteers encompassing a diverse range of linguistic backgrounds for recording presented a notable challenge during the execution of the pilot study. The dataset can be further extended to include professional artists to emulate the nuances in dementia speech in a better manner and quality. Also, since dementia impacts the elderly more significantly, promising results can be obtained if the professional recordists' age group falls between 50-70 in the subsequent recordings. As India is a multilingual economy with a large variety of dialects and languages, the dataset can be further extended to include other prominent Indian languages and multiple accents for a wider representation of the sample population under study.

ACKNOWLEDGMENT

The authors extend gratitude to King Saud University for partially funding this research through the Researchers Supporting Project number (RSP2023R322), King Saud University, Riyadh, Saudi Arabia. The authors would like to thank the undergraduate students who took part in the recordings and the listening tests.

REFERENCES

[1] Alzheimer's Association, "2021 Alzheimer's disease facts and figures," *Alzheimer's Dementia*, vol. 17, no. 3, 2021.

[2] S. Bhattacharya, P. Heidler, and S. Varshney, "Incorporating neglected non-communicable diseases into the national health program—A review," *Frontiers Public Health*, vol. 10, Jan. 2023, Art. no. 1093170.

[3] C. R. Jack, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, and N. Silverberg, "NIA-AA research framework: Toward a biological definition of Alzheimer's disease," *Alzheimer's Dementia*, vol. 14, no. 4, pp. 535–562, 2018.

[4] E. Kennedy, S. Panahi, I. J. Stewart, D. F. Tate, E. A. Wilde, K. Kenney, J. K. Werner, J. Gill, R. Diaz-Arrastia, M. Amuan, A. C. Van Cott, and M. J. Pugh, "Traumatic brain injury and early onset dementia in post 9–11 veterans," *Brain Injury*, vol. 36, no. 5, pp. 620–627, Apr. 2022.

[5] M. Hossain, J. Crossland, R. Stores, A. Dewey, and Y. Hakak, "Awareness and understanding of dementia in south asians: A synthesis of qualitative evidence," *Dementia*, vol. 19, no. 5, pp. 1441–1473, Jul. 2020.

[6] Y. Tao, M. E. Peters, L. T. Drye, D. P. Devanand, J. E. Mintzer, B. G. Pollock, and C. A. Munro, "Sex differences in the neuropsychiatric symptoms of patients with Alzheimer's disease," *Amer. J. Alzheimer's Disease Dementias*, vol. 33, no. 7, pp. 450–457, 2018.

[7] P. L. K. Bothongo, M. Jitlal, E. Parry, S. Waters, I. F. Foote, C. J. Watson, J. Cuzick, G. Giovannoni, R. Dobson, A. J. Noyce, N. Mukadam, J. P. Bestwick, and C. R. Marshall, "Dementia risk in a diverse population: A single-region nested case-control study in the East End of London," *Lancet Regional Health-Eur.*, vol. 15, Apr. 2022, Art. no. 100321.

[8] A. H. Alkenani, Y. Li, Y. Xu, and Q. Zhang, "Predicting Alzheimer's disease from spoken and written language using fusion-based stacked generalization," *J. Biomed. Informat.*, vol. 118, Jun. 2021, Art. no. 103803.

[9] P. Mahajan and V. Baths, "Acoustic and language based deep learning approaches for Alzheimer's dementia detection from spontaneous speech," *Frontiers Aging Neurosci.*, vol. 13, Feb. 2021, Art. no. 623607.

[10] M. A. Myszczynska, P. N. Ojamies, A. M. B. Lacoste, D. Neil, A. Saffari, R. Mead, G. M. Hautbergue, J. D. Holbrook, and L. Ferraiuolo, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nature Rev. Neurol.*, vol. 16, no. 8, pp. 440–456, Aug. 2020.

[11] E. Eyigoz, S. Mathur, M. Santamaria, G. Cecchi, and M. Naylor, "Linguistic markers predict onset of Alzheimer's disease," *EClinicalMedicine*, vol. 28, Nov. 2020, Art. no. 100583.

[12] M. Asgari, J. Kaye, and H. Dodge, "Predicting mild cognitive impairment from spontaneous spoken utterances," *Alzheimer's Dementia, Transl. Res. Clin. Intervent.*, vol. 3, no. 2, pp. 219–228, Jun. 2017.

[13] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's Dementia, Diagnosis, Assessment Disease Monitor.*, vol. 1, no. 1, pp. 112–124, Mar. 2015.

[14] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of Alzheimer's disease in conversational German," in *Proc. Interspeech*, Sep. 2016, pp. 1938–1942.

[15] L. Kristina, C. F. Kathleen, and K. Dimitrios, "Automated syntactic analysis of language abilities in persons with mild and subjective cognitive impairment," in *Proc. Med. Informat. Eur. (MIE) Conf.*, 2018, pp. 705–709.

[16] S. Daisaku, W. Shoko, K. Ayae, and A. Eiji, "Detecting Japanese patients with Alzheimer's disease based on word category frequencies," in *Proc. ClinicalNLP*, 2016, pp. 78–85.

[17] E. Ambrosini, M. Caielli, M. Milis, C. Loizou, D. Azzolino, S. Damanti, L. Bertagnoli, M. Cesari, S. Moccia, M. Cid, and C. G. De Isla, "Automatic speech analysis to early detect functional cognitive decline in elderly population," in *Proc. EMBC*, 2019, pp. 212–216.

[18] S. Luz, F. Haider, S. L. F. Garcia, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech," *Frontiers Comput. Sci.*, vol. 3, Oct. 2021, Art. no. 780169.

[19] F. Boller and J. Becker, "Dementiabank database guide," Univ. Pittsburgh, Tech. Rep., 2005.

[20] DementiaBank. *TalkBank*. Accessed: Jun. 21, 2022. [Online]. Available: https://dementia.talkbank.org/

[21] S. Luz, F. Haider, S. D. Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 1–5.

[22] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSS challenge," 2021, *arXiv:2104.09356*.

[23] F. Webster-Cordero and L. Giménez-Llort, "The challenge of subjective cognitive complaints and executive functions in middle-aged adults as a preclinical stage of dementia: A systematic review," *Geriatrics*, vol. 7, no. 2, p. 30, Mar. 2022.

[24] A. Breton, D. Casey, and N. A. Arnaoutoglou, "Cognitive tests for the detection of mild cognitive impairment (MCI), the prodromal stage of dementia: Meta-analysis of diagnostic accuracy studies," *Int. J. Geriatric Psychiatry*, vol. 34, no. 2, pp. 233–242, Feb. 2019.

[25] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated screening for Alzheimer's dementia through spontaneous speech," in *Proc. Interspeech*, 2020, p. 2222.

[26] A. Mandell and R. Green, "Alzheimer's disease," in *Handbook of Alzheimer's Disease*, A. E. Budson and N. W. Kowall, Eds. Malden, MA, USA: Wiley, 2021. ch. 1.

[27] C. Xue, C. Karjadi, I. C. Paschalidis, R. Au, and V. B. Kolachalama, "Detection of dementia on voice recordings using deep learning: A Framingham heart study," *Alzheimer's Res. Therapy*, vol. 13, no. 1, pp. 1–15, Dec. 2021.

[28] I. Martínez-Nicolás, T. E. Llorente, F. Martínez-Sánchez, and J. J. G. Meilán, "Ten years of research on automatic voice and speech analysis of people with Alzheimer's disease and mild cognitive impairment: A systematic review article," *Frontiers Psychol.*, vol. 12, Mar. 2021, Art. no. 620251.

[29] I. Vigo, L. Coelho, and S. Reis, "Speech- and language-based classification of Alzheimer's disease: A systematic review," *Bioengineering*, vol. 9, no. 1, p. 27, 2022.

[30] S. H. Ferris and M. Farlow, "Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors," *Clin. Interv. Aging*, vol. 8, pp. 1007–1014, Aug. 2013.

[31] K. J. Ballard, D. A. Robin, G. Woodworth, and L. D. Zimba, "Age-related changes in motor control during articulator visuomotor tracking," *J. Speech, Lang., Hearing Res.*, vol. 44, no. 4, pp. 763–777, Aug. 2001.

[32] G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, and M. Masud, "Spoken language identification using deep learning," *Comput. Intell. Neurosci.*, vol. 2021, Sep. 2021, Art. no. 5123671.

[33] R. Negrón, "Audio recording everyday talk," *Field Methods*, vol. 24, no. 3, pp. 292–309, Aug. 2012.

[34] W. Schmitt and A. Eiling, "Theoretical description of FM audio recording in Hi-Fi-VHS," *IEEE Trans. Magn.*, vol. 26, no. 5, pp. 2131–2133, Sep. 1990.

[35] M. G. Croll, "Broadcasters' use of optical discs for audio recording," in *Proc. 8th Int. Conf. Video, Audio Data Recording*, 1990, pp. 140–143.

[36] S. Qi, Z. Huang, Y. Li, and S. Shi, "Audio recording device identification based on deep learning," in *Proc. IEEE Int. Conf. Signal Image Process. (ICSIP)*, Aug. 2016, pp. 426–431.

[37] M. R. Kumar, S. Vekkot, S. Lalitha, D. Gupta, V. J. Govindraj, K. Shaukat, Y. A. Alotaibi, and M. Zakariah, "Dementia detection from speech using machine learning and deep learning architectures," *Sensors*, vol. 22, no. 23, p. 9311, Nov. 2022.

[38] F. Bertini, D. Allevi, G. Lutero, D. Montesi, and L. Calza, "Automatic speech classifier for mild cognitive impairment and early dementia," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, pp. 1–11, Jan. 2022.

[39] A. Shimoda, Y. Li, H. Hayashi, and N. Kondo, "Dementia risks identified by vocal features via telephone conversations: A novel machine learning prediction model," *PLoS ONE*, vol. 16, no. 7, Jul. 2021, Art. no. e0253988.

[40] J. Ramirez, J. M. Gorriz, F. Segovia, A. Ortiz, and R. Chaves, "Automatic diagnosis of Alzheimer's disease using support vector machine and speech signal," *J. Alzheimer's Disease*, vol. 29, no. 2, pp. 315–327, 2012.

[41] J. R. Orozco-Arroyave, F. Honig, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, and J. Rusz, "Automatic detection of Alzheimer's disease using spontaneous speech analysis and robust feature selection," *Biomed. Eng.*, vol. 12, no. 1, p. 57, 2013.

[42] M. Ahmed, U. Ghafoor, and S. Lee, "Detection of dementia in Urdu speaking population using speech analysis," in *Proc. IEEE 15th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Lisbon, Portugal, Mar. 2013, pp. 398–401.

[43] A. Satt, S. Hoermann, K. Anneken, U. Wehrmann, and J. Schroder, "Speech analysis for dementia detection—An exploratory study," *Current Alzheimer Res.*, vol. 10, no. 3, pp. 247–253, 2013.

[44] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *J. Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, Nov. 2015.

[45] C. Laske, H. R. Sohrabi, S. M. Frost, K. Lopez-de-Ipina, P. Garrard, M. Buscema, and S. E. O'bryant, "Innovative diagnostic tools for early detection of Alzheimer's disease," *Alzheimer's Dementia*, vol. 11, no. 5, pp. 561–578, May 2015.

[46] B. Demirtas, H. Ozturk, E. Kurt, and S. Yildirim, "Speech analysis for detection of dementia in the Turkish population," *J. Med. Syst.*, vol. 41, no. 9, p. 139, 2017.

[47] E. Yigit and K. Knel, "Speech analysis for dementia screening: A comparative study," *Appl. Acoust.*, vol. 119, pp. 23–29, Jan. 2017.

[48] C. M. Giebel, A. Worden, D. Challis, D. Jolley, K. S. Bhui, A. Lambat, E. Kampanellou, and N. Purandare, "Age, memory loss and perceptions of dementia in south Asian ethnic minorities," *Aging Mental Health*, vol. 23, no. 2, pp. 173–182, Feb. 2019.

[49] R. Parthasarathy, L. Almasy, R. C. Gur, J. A. Turner, V. L. Nimgaonkar, and N. Tandon, "Speech-based biomarkers for early detection of Alzheimer's disease: The role of machine learning," *Alzheimers Dementia, Diagnosis, Assessment Disease Monitor.*, vol. 12, no. 1, 2020, Art. no. e12032.

[50] X. Quan, L. Li, Q. Liu, X. Li, and Y. Li, "Investigating the use of speech features for dementia detection," *Appl. Sci.*, vol. 12, no. 2, p. 490, 2022.

[51] K. Yu, Y. Zhu, J. Lu, Y. Zhao, and S. Li, "Speech pattern analysis with machine learning for dementia detection," *J. Healthcare Eng.*, 2021, Art. no. 6623578.

[52] H. Wang, S. Fang, J. Liu, X. Guan, W. Wang, and H. Xue, "A speech-based dementia detection system using a random forest algorithm," *Frontiers Neurosci.*, vol. 15, Feb. 2021, Art. no. 625803.

[53] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 272–281, Feb. 2020.

[54] J. Zhang, Y. Liu, X. Zhang, Z. Wu, X. Wang, and W. Wang, "A speech-based dementia detection system using a convolutional neural network," *Appl. Sci.*, vol. 11, no. 17, p. 7879, 2021.

[55] U. Petti, S. Baker, and A. Korhonen, "A systematic literature review of automatic Alzheimer's disease detection from speech and language," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 11, pp. 1784–1797, Nov. 2020.

[56] M. Zolnoori, A. Zolnour, and M. Topaz, "ADscreen: A speech processing-based screening system for automatic identification of patients with Alzheimer's disease and related dementia," *Artif. Intell. Med.*, vol. 143, Sep. 2023, Art. no. 102624.

[57] S. Gregory, N. Linz, A. König, K. Langel, H. Pullen, S. Luz, J. Harrison, and C. W. Ritchie, "Remote data collection speech analysis and prediction of the identification of Alzheimer's disease biomarkers in people at risk for Alzheimer's disease dementia: The speech on the phone assessment (SPeAk) prospective observational study protocol," *BMJ Open*, vol. 12, no. 3, Mar. 2022, Art. no. e052250.

[58] M. Parsapoor, "AI-based assessments of speech and language impairments in dementia," *Alzheimer's Dementia*, vol. 19, no. 10, pp. 4675–4687, Oct. 2023.

[59] L. Ilias, D. Askounis, and J. Psarras, "Detecting dementia from speech and transcripts using transformers," *Comput. Speech Lang.*, vol. 79, Apr. 2023, Art. no. 101485.

[60] P. Priyadarshinee, C. J. Clarke, J. Melechovsky, C. M. Y. Lin, and J.-M. Chen, "Alzheimer's dementia speech (audio vs. text): Multi-modal machine learning at high vs. low resolution," *Appl. Sci.*, vol. 13, no. 7, p. 4244, Mar. 2023.

[61] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, ''Multimodal inductive transfer learning for detection of Alzheimer's dementia and its severity,'' 2020, *arXiv:2009.00700*.

[62] M. Rohanian, J. Hough, and M. Purver, ''Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech,'' 2021, *arXiv:2106.09668*.

[63] R. Voleti, J. M. Liss, and V. Berisha, ''A review of automated speech and language features for assessment of cognitive and thought disorders,'' *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 282–298, Feb. 2020.

[64] S. D. L. F. Garcia, C. W. Ritchie, and S. Luz, ''Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: A systematic review,'' *J. Alzheimer's Disease*, vol. 78, no. 4, pp. 1547–1574, Dec. 2020.

[65] Y. Guan and B. Li, ''Usability and practicality of speech recording by mobile phones for phonetic analysis,'' in *Proc. 12th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Jan. 2021, pp. 1–5.

[66] F. Eyben, M. Wöllmer, and B. Schuller, ''Opensmile: The Munich versatile and fast open-source audio feature extractor,'' in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 1459–1462.

[67] A. de Cheveigné and H. Kawahara, ''YIN, a fundamental frequency estimator for speech and music,'' *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[68] V. A. Profillidis and G. N. Botzoris, ''Statistical methods for transport demand modeling,'' *Model. Transp. Demand*, pp. 163–224, 2019.

[69] G. J. Glasser and R. F. Winter, ''Critical values of the coefficient of rank correlation for testing the hypothesis of independence,'' *Biometrika*, vol. 48, pp. 444–448, Dec. 1961.

[70] A. Herve, ''The Kendall rank correlation coefficient,'' in *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA, USA: Sage, 2007, pp. 508–510.

[71] A. Huang, ''Similarity measures for text document clustering,'' in *Proc. 6th New Zealand Comput. Sci. Res. Student Conf.*, Christchurch, New Zealand, vol. 4, 2008, pp. 9–56.

[72] B. Naderi, S. Möller, and R. Cutler, ''Speech quality assessment in crowdsourcing: Comparison category rating method,'' in *Proc. 13th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2021, pp. 31–36.

[73] S. Vekkot, D. Gupta, M. Zakariah, and Y. A. Alotaibi, ''Emotional voice conversion using a hybrid framework with speaker-adaptive DNN and particle-swarm-optimized neural network,'' *IEEE Access*, vol. 8, pp. 74627–74647, 2020.

[74] S. Vekkot and D. Gupta, ''Fusion of spectral and prosody modelling for multilingual speech emotion conversion,'' *Knowl.-Based Syst.*, vol. 242, Apr. 2022, Art. no. 108360.

SUSMITHA VEKKOT received the Ph.D. degree in speech processing from Amrita Vishwa Vidyapeetham, India, in 2021. She is currently an Assistant Professor (Senior Grade) with the Department of ECE, Amrita Vishwa Vidyapeetham, Bengaluru Campus, India. She has been an active Researcher. She has published articles in several international conferences and journals of repute. Her research work is published in reputed international journals like IEEE Access, *Knowledge-Based Systems*, *International Journal of Speech Technology*, and *Sensors*. She has conducted and participated in a number of short-term courses, seminars, and conferences conducted at the national/international level. She has more than a decade of experience in teaching and research space. She has around 25 publications from reputed Scopus-indexed journals, conferences, and book chapters. Her primary research interests include signal, speech and image processing, and machine learning applications of speech and image processing in health care and human–computer interaction. She was a recipient of Government of India's Visvesvaraya Full-Time Scholarship for pursuing the Ph.D. degree.

NAGULAPATI NAGA VENKATA SAI PRAKASH is currently pursuing the B.Tech. degree in artificial intelligence engineering with the Amrita School of Computing. His research interests include data science, machine learning, and deep learning.

THIRUPATI SAI ESWAR REDDY is currently pursuing the B.Tech. degree in artificial intelligence engineering with the Amrita School of Computing. His research interests include data science, machine learning, and deep learning.

SATWIK REDDY SRIPATHI is currently pursuing the B.Tech. degree in artificial intelligence engineering with the Amrita School of Computing. His research interests include NLP, machine learning, data science, and deep learning.

S. LALITHA received the B.Tech. degree from the Vijayanagar Engineering College, Bellary, Gulbarga University, in 1998, the M.Tech. degree from the M. S. Ramiah Institute of Technology, Bengaluru, VTU, in 2008, and the Ph.D. degree in speech signal processing from the Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru. She is also an Assistant Professor (Senior Grade) with the Department of ECE, Amrita Vishwa Vidyapeetham. Having a teaching experience of more than 17 years. Her major research interests include speech and audio signal processing, image signal processing, natural language processing, and artificial intelligence. She has around 27 publications from reputed Scopus indexed journals, conferences, and book chapters.

DEEPA GUPTA received the Ph.D. degree in natural language processing in example-based machine translation from the Department of Mathematics and Computer Application, Indian Institute of Technology Delhi, in 2005. She is currently a Professor and the Research Head of the Amrita School of Computing, Bengaluru Campus, India. She joined Amrita University, Bengaluru, in 2009, and has been heading the Human Language Technology Laboratory on campus, since 2015. Since then, she has guided Ph.D. and graduate/postgraduate students. She was a Postdoctoral Researcher with FBKIRST (Centre for Scientific and Technological Research), Trento, Italy, for two years on the EU-funded TC-star project. She was an Assistant Professor, IIIT-Bangalore, before joining the Amrita School of Computing. A total of six Ph.D. students are awarded under her guidance. She had completed three government-funded projects and a couple of company consultancy projects. Her research interests include text analytics, clinical data mining, speech processing, and other areas in natural language processing. Her research has been published in international journals like IEEE Access, *Information Processing and Management*, *Knowledge-Based Systems*, and *Expert Systems with Applications*, along with several peer-reviewed international conferences.

MOHAMMED ZAKARIAH (Member, IEEE) received the B.Sc. degree in computer science and engineering from Visvesvaraya Technological University, India, in 2005, the master's degree in computer engineering from Jawaharlal Nehru Technological University, India, in 2007, and the Ph.D. degree in informatics. As a researcher, he has published more than 45 articles in reputed journals indexed in ISI Thomson Reuters in various topics ranging from bioinformatics, image processing, speech processing, and audio forensics in reputed ISI indexed journals like *Molecules*, *Sensors*, *Applied Sciences*, *Electronics*, *Multimedia Tools and Applications*, IEEE Access, and *Applied Soft Computing*. He is experienced in machine learning, artificial intelligence, and image and speech Processing. He has worked on five government-funded (KACST) projects and has experience in writing research grant proposals.

YOUSEF AJAMI ALOTAIBI (Senior Member, IEEE) received the B.Sc. degree from King Saud University, Riyadh, Saudi Arabia, in 1988, and the M.Sc. and Ph.D. degrees in computer engineering from the Florida Institute of Technology, Melbourne, FL, USA, in 1994 and 1997, respectively. From 1988 to 1992 and from 1998 to 1999, he was a Research Engineer with Al-ELM Research and Development Corporation, Riyadh. He was an Assistant Professor and an Associate Professor with the College of Computer and Information Sciences, King Saud University, from 1999 to 2008 and from 2008 to 2012, respectively, where he has also been a Professor, since 2012. His research interests include digital speech processing, specifically speech recognition and Arabic language and speech processing.