Journal of Multi Disciplinary Engineering Technologies
Volume 18, Issue 01

Hindi Analyser: A Context-Aware Approach for Semantic Clustering of Hindi Words

Shefali Arora[1], Ruchi Mittal[2*], Shikha Gupta[3]

[1]Department of Computer Science, NIT Jalandhar, Jalandhar, Punjab, India.
[2]Iconic Data, Tokyo, Japan.
[3] Department of Information Technology, MAIT, Delhi, India.

*Corresponding author(s). E-mail(s): ruchimittal138@gmail.com;
Contributing authors: arorashef@gmail.com; shikha.gpt1@gmail.com;

Abstract

Over the years, automatic speech recognition has become an important area of research due to its applications in the domain of speaker recognition, topic modelling context retrieval. However, a lot of work is needed in the domain of audio signal processing of words in different languages. Hindi is India's national language and there is a need to build speech recognition and classification systems so as to improve its context understanding in different application areas such as news retrieval and topic modelling. In this paper, we propose a framework HindiAnalyzer which recognizes various speech samples in Hindi and analyses the significant keywords spoken by individuals in audio files using an integrated BERT and K-means clustering approach. The framework makes use of cosine similarity between keywords extracted using a pre-trained multilingual BERT model. This is followed by K-means clustering of the extracted embedding. To reduce the dimensionality, Principal Component Analysis (PCA) is used for extraction and visualization of clusters. A large dataset of 2000 samples are used in Hindi language to validate the proposed framework. The results of the studies demonstrate that the suggested BERT and cosine-similarity based method for completing the task has a higher level of accuracy when it comes to modeling the keywords as clusters. Furthermore, performance indicators such as the Silhouette score are used to gauge the relationship between the terms in the cluster.

Keywords: BERT, K-means clustering, Cosine similarity, Topic modeling

1 Introduction

It is getting harder and harder to extract insightful information from the deluge of sound, which includes anything from speeches and recorded phone conversations to music and podcasts. Since MP3s and other audio sources are so common, it is essential to have effective methods for cataloguing, searching, and indexing this data. Among the main strategies used in this endeavor are topic modeling and context retrieval, which are powerful tools for examining themes and patterns and extracting comprehensive insights from vast amounts of audio information. Leveraging recent

Published on: November, 2024

advances in machine learning, particularly deep learning models and transformer technologies, has the potential to unlock new levels of comprehension and value in audio processing tasks. Over the past decade, artificial intelligence has advanced significantly, revolutionizing how humans interact with technology. A key player in this paradigm shift is speech synthesis, a subset of natural language processing that enables machines to convert written text into audible speech, facilitating dynamic user-device communication. The evolution from artificial and monotonous voices to more organic and expressive tones has made speech synthesis a crucial element of user-centered design [1]. In the digital realm, seamless human-machine interaction is now considered the ideal user experience. When it comes to chatbots, voice assistants, and smart home applications, user happiness is directly impacted by both factors: naturalness of interaction and convenience of use. Speaking is important because it involves not just mimicking the human voice but also reconstructing the meaning of human speech from context to emotion and intent [2]. Furthermore, there are still significant and novel issues with the speech synthesis paradigms of today. Issues like the uncanny valley arise when the realistic audio of artificial voices is unable to replicate the natural intonation and rhythm of human voices. The last synthesis-related concern is the need for real-time operations combined with the capacity to take linguistic nuances into account. This research analyzes these difficulties by elucidating current shortcomings and proposing novel solutions [3].

In this study, an exploratory approach is taken to the general examination of voice synthesis and its notion. We examine the specifics of a few synthesis models that rely on formant, concatenation, and the more modern deep learning methods. Furthermore, in the larger context of growing affect creation and detection in speech replication, our study goes beyond the fundamentals of text-to-voice interpretation. We want to provide a thorough overview of the state of speech synthesis development to date, as well as the emerging implications for HM interaction.

Beyond PDA, advancements in speech synthesis are applicable in a wide range of fields, such as medicine, education, call center services, and entertainment. Natural and free-speaking materials have the potential to drastically alter how humans interact with machines or equipment in these domains. Like any device that records human speech, there are countless applications for this technology that could transform a variety of contexts, such as helping those who struggle with speech or developing language learning settings [4]. In the contemporary context of a human-robot interaction, bridging the gap between artificial and human intelligence is particularly critical. Speech synthesis is a medium for natural language communication that transcends traditional user interaction surfaces. It combines context sensitivity, emotional appeal, and language correctness to produce novel and disruptive applications [5].

2 Related Work

In changed landscape of computer science, the key activities are modeling, analysis and synthesis of human behavior. This exploration is not merely about adopting technological methodologies for machine-readable data; it delves deeper into understanding human behavior from multiple dimensions—cognitive, social, and psychological—that are not always directly observable. Such comprehensive approaches necessitate advanced technology and have historically centered around the automatic analysis and synthesis of facial expressions and human motion through methods used in robotics and computer graphics to replicate visible aspects of human behavior [6].

As we progress into more refined interactions between humans and machines, the role of affective computing, initiated by Picard in 2000, becomes increasingly significant. This domain, alongside social signal processing (SSP) [7], social robotics [8], and intelligent virtual agents [**?**], has matured into a well-established field within the computing community. These disciplines are crucial in establishing our current studies focused on human- machine interfaces that afford the machines to be more than just physically imitating users but also emotionally and contextually responding to them.

The variety and diversity of human-machine interaction (HMI) are further illustrated by subsequent advances. For instance, social robots have been employed to assist autistic children [9], and several techniques have been developed to recognize human personality traits in various contexts [10]. These developments highlight the flexibility of HMI, where it's crucial to engage and manage people's behavior in addition to other aspects. The characteristics of mimicry that emerged from [11] and multimodal affective interaction that emerged from [12] demonstrate that human behavior is a cocktail that necessitates the identification of numerous behaviors in order for the system to learn and interact with humans.

The study by Vinciarelli and his associates (2015) [3] draws attention to the challenges that come with voice synthesis, a crucial component of HMI that must overcome issues like the uncanny valley. Natural human contact with technology is facilitated by the shift from text-to-speech technology to speech synthesis with emotion and an awareness of the speech environment in which specific emotions are suitable. In order to replace the current static user interfaces with active or vibrant ones, the synthesized voices must not only sound natural but also understand the context and emotions of the conversation.

Thus, the body of work in HMI and related fields illustrates a trajectory from foundational theories of signal processing and behavior analysis to sophisticated applications in real-world settings, underpinning the evolution of our interaction paradigms with machines. This historical context sets the stage for the research presented in this paper, where we propose HindiAnalyzer, a tool designed to leverage these multidisciplinary insights to improve speech recognition and semantic analysis in Hindi, thereby contributing to the broader domain of audio processing and topic modeling in digital communications.

3 Methodology

The framework for Hindi Analyzer for context-understanding of Hindi language is presented in Figure 1. The collected audio dataset is pre-processed in order to convert it into text transcriptions. This is followed by the tokenization of words using the bert-base-multilingual-cased model in order extract keyword embeddings. These embeddings are clustered with the help of the K-means clustering algorithm which makes use of cosine similarity to group words into different categories. Followed by dimensionality reduction and visualization, the performance of the proposed framework is evaluated with the help of performance metrics such as Silhouette score.

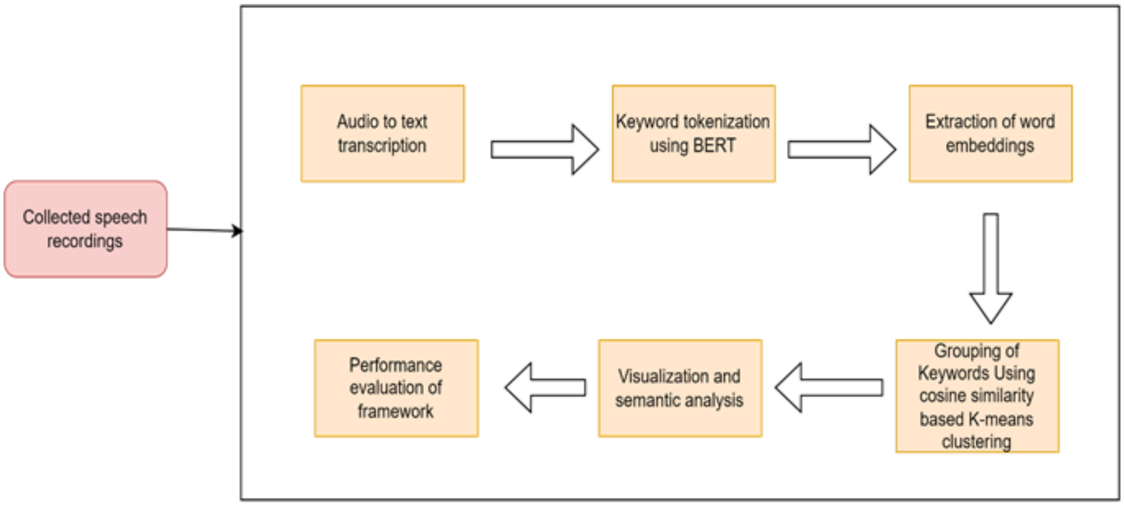
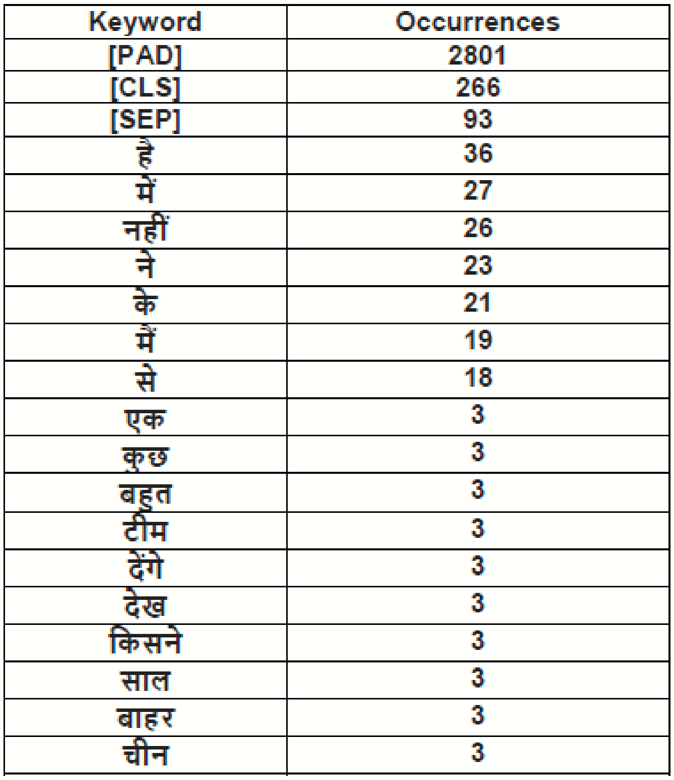

Collected speech recordings
Audio to text transcription
Keyword tokenization using BERT
Extraction of word embeddings
Grouping of Keywords Using cosine similarity based K-means clustering
Visualization and semantic analysis
Performance evaluation of framework

Fig. 1 Proposed Framework



Keyword	Occurrences
[PAD]	2801
[CLS]	266
[SEP]	93
है	36
में	27
नहीं	26
ने	23
के	21
मैं	19
से	18
एक	3
कुछ	3
बहुत	3
टीम	3
देंगे	3
देख	3
किसने	3
साल	3
बाहर	3
चीन	3

Fig. 2 Sample of Hindi keywords and their occurrences

The audio signals from different speakers are processed and converted into text in order to extract keywords from the speech. The process of keyword extraction from the text is done using a pre-trained multilingual BERT model which extracts embeddings from the samples. After the training of an autoencoder which converts these embeddings into a lower dimensional space, K-means clustering is carried out to group these keywords into their relevant groups. Further, the commonly related words in a cluster are extracted with the help of cosine similarity. This unsupervised problem is made up of multiple steps i.e., acquisition of input speech signals, pre-processing of signals, augmentation of captured speech signals and conversion to text, designing of autoencoder architecture followed by clustering and evaluation of the model using several performance metrics.

3.1 Data Collection

As there is a scarcity of Hindi language public speech dataset, an open source audio speech dataset of Hindi language is taken from Kaggle which comprises 1998 recordings of male and female speakers. The audio recordings consist of discussion on various news agendas by male and female adults. The recordings consist of around 1600 audios by males and around 350 audios by females. Figure 2 shows the examples of keywords extracted from these audios using BERT embeddings.

3.2 Keyword Extraction using BERT embeddings

Feature extraction and their classification is challenging in the case of audio clips particularly for languages other English. In order to extract keywords from audio files, the audio recordings are transcribed into text using GoogleSpeechRecognition. Further, bert-base-multilingual-cased model is used for tokenization of Hindi sentences retrieved from audio samples. The pre-trained BERT model makes use of WordPiece tokenization to split words into smaller units known as tokens. Further, each word or subword is looked up in the model's vocabulary and words are segmented

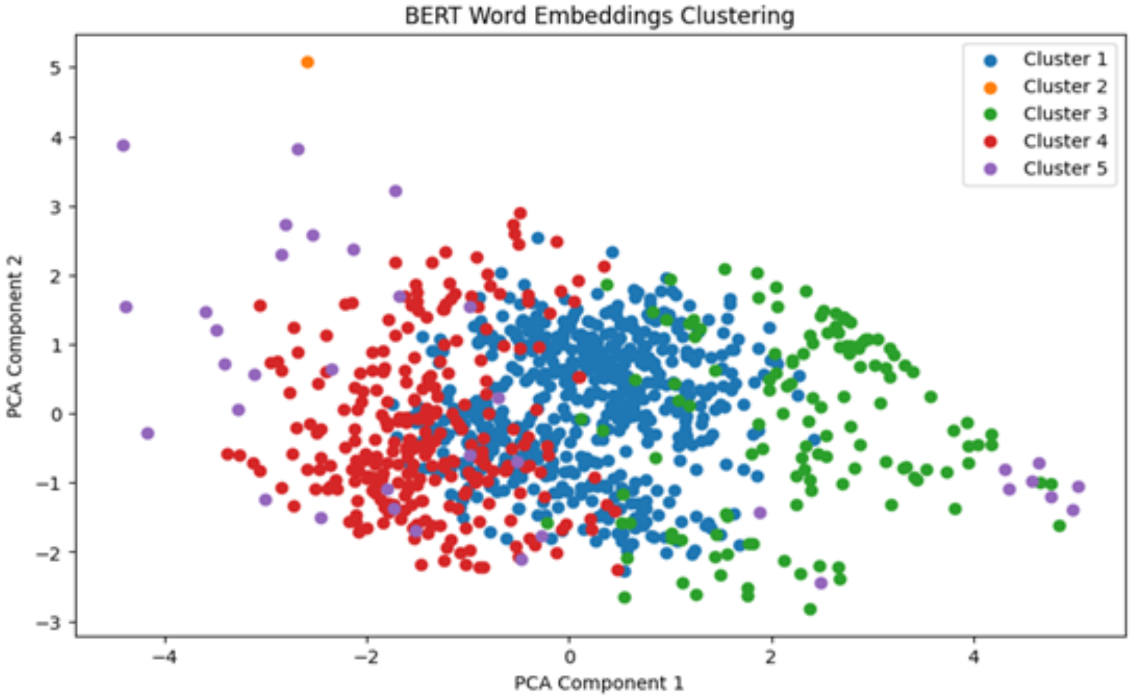


Fig. 3 Visualization of clusters for k=5

further if not found. Each word is mapped to a token id in the vocabulary. Special tokens such as [CLS] and [SEP] are added by the model to mark the beginning and end of sequences. To ensure uniform sequences, the token [PAD] is added to the sentences or if input size limit is exceeded, then length is truncated. The encoded sequences obtained from the keywords are fed into BERT model to obtain contextualized embeddings. This will ensure that the semantic meaning of input sequences is understood by the model. Table 1 depicts the occurrences of common keywords extracted by the BERT model from the audio samples. Further, these embeddings are input to a K-means clustering model in order to group similar words in one cluster.

3.3 K-means clustering on BERT embeddings

The extracted embeddings using the multilingual BERT model are further fed to as an input to the K-means clustering algorithm in order to cluster the keywords.

The algorithm used for integrated BERT-K means clustering shown in Algorithm 1:

X¡ - keywords represented as BERT embeddings. K¡- Number of clusters. C¡-Set of clusters which each cluster ci with a set of keywords

Step 1: Initialize k centroids µ1, µ2. . . . µk as initial clusters

Step 2: For each keyword Xj in X, calculate cosine similarity with each centroid. Assign each keyword to Xj to the cluster based on cosine similarity.

Step 3:For each cluster Ci, update the new centroid as the mean of all keywords assigned to that cluster.

Step 4:Repeat steps 2, 3, and 4 until convergence criteria are met, such as a maximum number of iterations or minimal change in cluster assignments.

Step 5:Apply Principal Component Analysis for reducing dimensionality of BERT embeddings.

Step 6:Analyze the cosine similarity and other metrics with respect to words in each cluster to understand the characteristics of each cluster.

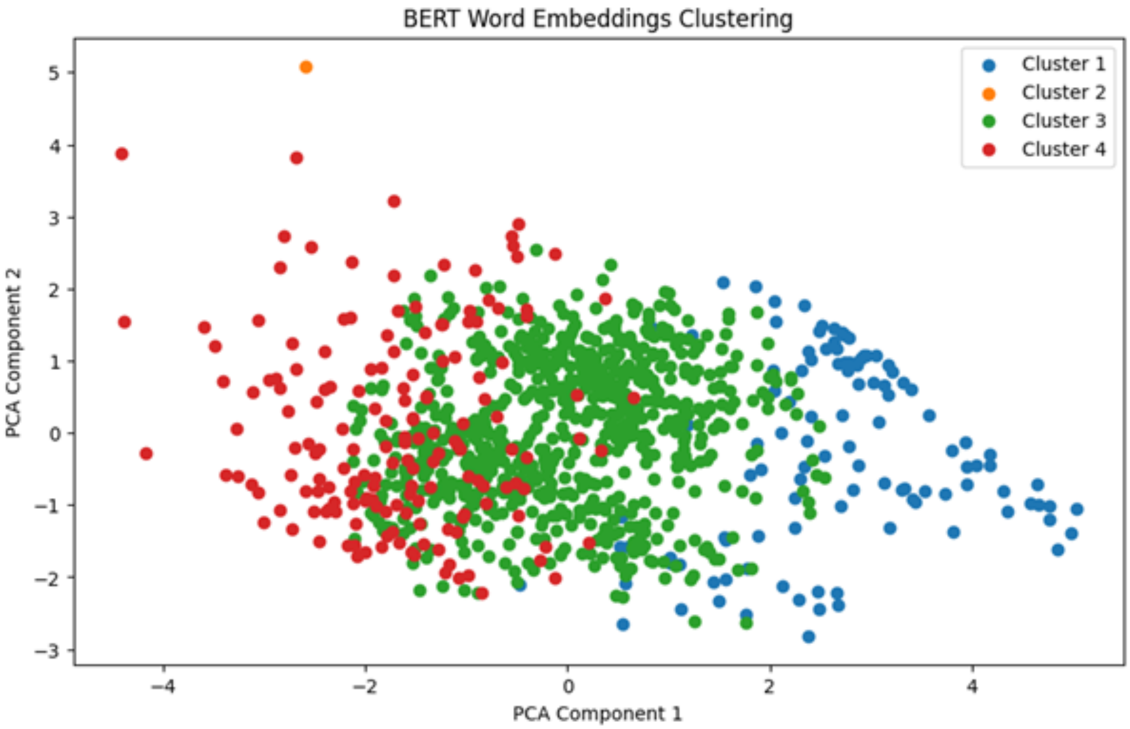


Fig. 4 Visualization of clusters for k=4

Table 1 Silhouette score for different cluster groups

No. of clusters (k)	Silhouette Score
3	0.4625403881072998
4	0.4089522659778595
5	0.31622588634490967

Step 7:Compute Silhouette Score to evaluate the performance of the proposed model.

This algorithm describes the steps involved in K-means clustering with BERT embeddings, which include initialization, assigning data points to clusters based on cosine similarity, updating cluster centres, calculating cosine similarity within each cluster, convergence check, evaluation metrics, visualisation, and cluster analysis.

3.4 Dimensionality Reduction

Text mining relies heavily on dimensionality reduction. Reducing dimensionality improves clustering performance by allowing for easier text analysis. Mining approaches process data with fewer terms. This study uses dimensionality reduction approaches like as PCA for effective clustering of keywords. The PCA (Principal component analysis) reduces the dimensionality of huge data sets by condensing variables into a smaller set while retaining information.

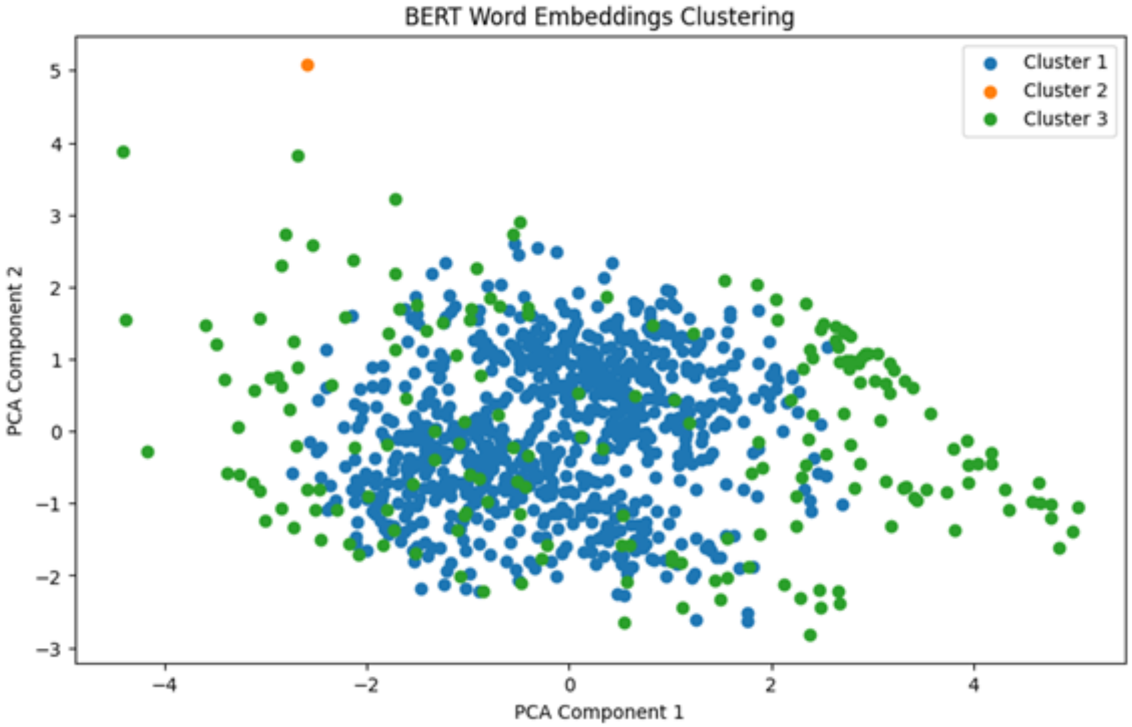


Fig. 5 Visualization of clusters for k=4

Table 2 Average cosine similarity between clusters

No. of clusters (k)	Average cosine similarity score
1	0.9143999814987183
2	0.9304097890853882
3	0.949791669845581
4	0.9459192156791687
5	0.8954121470451355

4 Experimental Results

4.1 Results

The outcome of the results is evaluated based on Silhouette Score Coefficient. The Silhouette Coefficient is a metric applied to determine the goodness of a clustering method. The value ranges from -1 to 1. This indicates the separation of clusters from each other as strong, reasonable or weak. The Silhouette score coefficient is calculated as follows:

$$Sj = \frac{B(j) - A(j)}{\max(B(j), A(i))}$$

where, A(j) is the average separation between that point and all other points belonging to the same cluster. B(j) is the average distance from that point to its cluster of all the points in the nearest cluster.

After computing the silhouette coefficient for each point, the average score is calculated in order to model the topics using PCA dimensionality reduction. Figures 3,4 ad 5 depict the visualization of clusters based on the number of PCA components. Table 2 depicts the Silhouette score based on the number of clusters.

Table 2 depicts the average Silhouette score for various number of clusters. It is observed that a reasonable score is achieved between clusters which proves that the use of dimensionality reduction followed by K-means clustering on BERT embeddings helps to form meaningful clusters which give contextual information about different topics. In case of the given audio dataset, it is observed that the clusters formed for the extracted keyword embeddings correspond to the following broad topics: politics, entertainment, nations, general keywords. Table 3 depicts the average cosine similarity between words in each cluster, which proves that the extracted keywords have a strong semantic relation with each other.

5 Conclusion

The study extends its exploration to the field of automatic speech recognition, particularly highlighting the importance in contexts like news retrieval and speaker recognition. It introduces "HindiAnalyzer," a framework designed for recognizing Hindi speech samples and analyzing key spoken words using an integrated approach of BERT and K-means clustering. This innovative combination, along with PCA for dimensionality reduction, demonstrated superior performance in extracting meaningful insights from speech data. The benefit of this system in enhancing context in numerous applications, including the availability of a bulk 2000-sample Hindi collection, has been highlighted by the incorporation of multiple sources, including huge datasets. A high degree of cluster separability is shown by the obtained Silhouette score of 0.46. Additionally, this method offers guidance for complex SR &C systems in a variety of languages, with a focus on Hindi because of its contextual sensitivity, and validates the results of the HMM-based model. HMM structure optimization and fine-tuning in relation to the HCI domain should be considered in the future. Models could be modified to better reflect the dimensions of the various interactive systems, which could improve user happiness and performance. Additionally, exploring real-time adaptation mechanisms for HMM-based voice synthesis in dynamic HCI scenarios could enhance the model's responsiveness.

Declarations

- The authors received no specific funding for this study.
- The authors declare that they have no conflicts of interest to report regarding the present study.
- No Human subject or animals are involved in the research.
- All authors have mutually consented to participate.
- All the authors have consented the Journal to publish this paper.
- Authors declare that all the data being used in the design and production cum layout of the manuscript is declared in the manuscript.

References

[1] Naik, A.: Hmm-based phoneme speech recognition system for the control and command of industrial robots. Technical Transactions **118**(1) (2021)

[2] Yu, D., Deng, L.: Automatic Speech Recognition vol. 1, (2016). Springer

[3] Vinciarelli, A., Chatziioannou, P., Esposito, A.: When the words are not everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls. Frontiers in ICT **2**, 4 (2015)

[4] Padoy, N., Hager, G.D.: Human-machine collaborative surgery using learned models. In: 2011 IEEE International Conference on Robotics and Automation, pp. 5285–5292 (2011). IEEE

[5] Raghudathesh, G., Chandrakala, C., Rao, B.D.: Review of toolkit to build automatic speech recognition models. In: Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2022, Volume 2, pp. 449–459 (2022). Springer

[6] Ekman, P., Davidson, R.J.: Voluntary smiling changes regional brain activity. Psychological Science **4**(5), 342–345 (1993)

[7] Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and vision computing **27**(12), 1743–1759 (2009)

[8] Breazeal, C.: Social interactions in hri: the robot view. IEEE transactions on systems, man, and cybernetics, part C (applications and reviews) **34**(2), 181–186 (2004)

[9] Scassellati, B., Admoni, H., Matarić, M.: Robots for Use in Autism Research

[10] Vinciarelli, A., Mohammadi, G.: A survey of personality computing. IEEE Transactions on Affective Computing **5**(3), 273–291 (2014)

[11] Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: A survey of evaluation methods across disciplines. IEEE Transactions on Affective Computing **3**(3), 349–365 (2012)

[12] Hussain, M.S., Calvo, R.A., Chen, F.: Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. Interacting with computers **26**(3), 256–268 (2014)