2. Literature Review

2.1 Introduction

A complete analysis and synthesis of previous research and scholarly writings that are relevant to a certain subject or research issue is what is known as a literature review. Performing the function of an essential component in academic writing, including but not limited to research papers, theses, and dissertations. The fundamental objective of a literature review is to provide a comprehensive overview of the current state of knowledge on the subject that has been selected. This includes both studies and articles. An organized presentation of significant concepts, theories, approaches, and findings linked to the issue is something that is involved in this process. These presentations are typically organized thematically, chronologically, or methodologically. It is possible that the review could involve comparing, contrasting, or discovering patterns and trends across a number of different research. Additionally, the review will incorporate a critical examination of each source, evaluating their strengths and shortcomings simultaneously. It is essential to note that the purpose of a literature review is to locate gaps in the research that has already been conducted. This leads researchers to insert their work within a bigger framework contributing to knowledge development. A literature review goes beyond that of honouring the original authors by accurately citing all the works that are mentioned and allowing readers to get back to the sources to conduct a more in depth study on their subject. Doing literature reviews at the beginning of researchers' studies becomes an important part in the formulation of research topics, hypotheses and procedures. So it provides them with an appropriate foundation for placing new contributions into the academic discourse of their study field.

2.2 Previous Works in the fields of Speech Recognition System

Automatic speech recognition for under resourced languages, Besacier et al. (2014) do a thorough review of research. This survey addresses challenges commonly encountered, such as insufficient data, lack of linguistic proficiency, and standardized resources.

Through a diverse set of novel efforts in multilingual and cross-linguistic acoustic modeling, deep learning, and innovative data collection methods such as crowdsourcing and repurposing of existing audio archives, this study presents new approaches for acoustic modeling. It sketches the practical uses of ASR voice search in South Africa and the Avaaj Otalo project in India, demonstrating the importance of ASR voice search for documentation of endangered or unwritten languages. Despite these breakthroughs however, processes of adaptation are still inadequately addressed, the databases are still extensive, and the technical and ethical problems involved in implementing ASR technology in a minority language context are inadequately addressed. This equates to a crucial mile stone in directing ASR development towards closer bridging of linguistic disparities while preserving language variety[1].

This document summarizes the advancements, obstacles, and approaches of ASR systems for Indian regional languages as presented by More et al. in 2018. The authors indicate that although India recognizes 22 official languages, automatic speech recognition has been developed for only 14, presenting both linguistic and technical challenges. The authors elaborate extensively on these elements: speech kinds, encompassing isolated, connected, continuous, and spontaneous, and speaker models, which consist of speaker-dependent and speaker-independent categories. The texts delineate prevalent techniques, including Mel-Frequency Cepstral Coefficients for feature extraction and Hidden Markov Models for classification. Reported accuracies vary from 65.24% to 98.85%, predominantly concerning Hindi, Sanskrit, Tamil, Assamese, and Punjabi. Simultaneously, it highlights shortcomings in the existing review: there is an absence of discourse regarding under-represented languages, the compilation of various voice corpora presents significant challenges, and the advancement of scalable multilingual models requires additional effort. This analysis examines how ASR can address language disparities in India, advocating for the development of additional resources and innovative methodologies for speech recognition[2].

Sailor et al. (2018) detail the development of a Gujarati ASR system for the Low Resource Speech Recognition Challenge at INTERSPEECH 2018 and examine the limitations of speech recognition in Indian languages, which typically lack linguistic resources. The system achieved substantial reductions in Word Error Rate (WER) across test datasets by utilizing Amplitude Modulation (AM)-based features extracted through advanced auditory filterbanks, alongside Recurrent Neural Network Language Models (RNNLM) for language modeling and Time-Delay Neural Networks (TDNN) with Long Short-Term Memory (LSTM) for acoustic modeling. This led to a relative enhancement in perplexity, achieving 36.18% and 40.95% on the test and blind test sets, respectively, compared to a 3-gram LM baseline, while the incorporation of AM features and system combinations significantly diminished WER. This study introduces cutting-edge strategies to enhance the performance of automatic speech recognition in resource-limited languages through sophisticated modeling techniques and system integrations[3].

This research presents an End-to-End Automatic Speech Recognition system for the Gujarati language, utilizing a deep learning architecture of Convolutional Neural Networks, Bidirectional Long Short-Term Memory networks, and dense layers, optimized with a Connectionist Temporal Classification loss function. This study presents an advanced prefix decoding technique, using a 4-gram word-level language model and a bi-gram character-level language model, along with a BERT-based post-processing spell corrector. The created system, trained on the Microsoft Speech Corpus, achieves a 5.11% reduction in Word Error Rate (WER), improving from 70.65% to 65.54%. The majority of these discrepancies were due to consonant mismatches, diacritic inconsistencies, and independent vowel mismatches, illustrating the linguistic challenges present. The current study determines that enhancing ASR outcomes for resource-limited languages does not necessitate the expansion of their databases. The next sections aim to elucidate methods for optimization in low-resource environments[4].

The article by Parikh and Joshi examines various techniques for developing ASR systems in the Gujarati language, which encounters two primary challenges: linguistic diversity and limited resources. The primary methodologies encompass the statistical model—Hidden Markov Model, the hybrid model—HMM integrated with Artificial Neural Networks, and recent iterations of End-to-End models employing Recurrent Neural Networks with Connectionist Temporal Classification for continuous speech processing. Statistical models have achieved accuracies of up to 95.1% for isolated words, whereas hybrid HMM/ANN models demonstrate superior performance on more complex tasks. The paper highlights the challenges of diverse data gathering, the adaption of systems for real-world accents and dialects, and notes the potential of End-to-End systems, despite their substantial data requirements. This review emphasizes the necessity for new methods to improve the characteristics of Gujarati ASR systems in resource-constrained environments[5].

The review study by Kulkarni et al. (2016) examines ASR systems for Indian regional languages with HTK, analyzing 30 research based on characteristics such as Language, Utterance Type, Number of Speakers, Vocabulary Size, Recording Environment, and Performance Metrics. The paper summarizes the uses of HTK in other languages, including Hindi, Sanskrit, Tamil, and Gujarati, among others. It has seen success in both isolated and continuous speech recognition systems, with accuracy rates above 90%, and is often combined with feature extraction methods such as MFCC. The choice of HMM configurations is associated with enhancement in accuracy. In this research we demonstrate the strengths of the HTK toolkit and propose future investigation for improving strong ASR systems for underrepresented Indian languages[6].

Within a multilingual framework, Shetty et al. (2020) study the potential improvements of Transformer based speech recognition systems for low resource Indian languages. Datasets from Gujarati, Tamil, and Telugu are used in the study, of which the techniques have been innovated by us, whereby we integrate language identity at both the encoder and decoder levels using one hot vectors or acquired language embeddings.. The results were intriguing, demonstrating significant enhancements in both WER and CER for multilingual and monolingual contexts. Additional enhancements were achieved by the retraining of the multilingual model using target language data. Of the presented models, "Lang Embed Trans + LID" excelled by employing a learnt language embedding matrix in conjunction with retraining to enhance recognition accuracy. This study has significantly highlighted the potential of Transformer frameworks in addressing issues in voice recognition within multilingual and resource-limited contexts[7] .

Sailor and Hain (2020) proposed utilizing the MTL framework for multilingual speech recognition in Indian languages through language-specific phoneme recognition as an auxiliary task, which enhances multilingual senone classification by integrating language identity and phonetic information, thereby improving acoustic modeling.

Moreover, the aforementioned work enhances the MTL technique by integrating a structured output layer, thereby connecting the core and secondary activities and facilitating improved performance. The proposed MTL-SOL framework demonstrated a reduction in word error rates (WERs) by 3.1%-4.4% on the development sets and 2.9%-4.1% on the evaluation sets, relative to baseline systems, in experiments conducted across three datasets in Gujarati, Tamil, and Telugu from the Interspeech 2018 Low-Resource Speech Recognition Challenge. This study emphasizes the efficacy of integrating phoneme and linguistic identification data to enhance recognition in resource-limited multilingual context.[8].

Tailor and Shah (2015) has studied the latest developments in Indian Language ASR systems and their problems, techniques and applications. ASR applications for diverse phonetics and grammar in a large number of Indian languages suffer from resource scarcity and subtle pronunciation. In terms of breadth, examined research vary in breadth, with Gujarati ASR system using approaches like HMM and MFCC to yield performance from 72% to 96% based on the vocabulary size and speaker variability. Apparently other languages such as Hindi, Marathi and Tamil were addressed by other systems but they had been using approaches based on DTW, MLLR, or hybrid neural networks which brought inconsistent improvements at all. Therefore, ASR of Indo-Aryan languages has to be strengthened through frameworks, resources and adaptive techniques to support their advancement[9].

Mitra et al. (2017) show voice recognition under observed settings with noisy channels limiting the system’s robustness increased by unsupervised adaption strategies alone. They then looked at feature space maximum likelihood linear regression transformations and deep autoencoder bottleneck features that can help compensate for acoustic discrepancies. The authors demonstrated that features altered by fMLLR yield a 20% reduction in word error rates (WERs) using the DARPA RATS dataset, which comprises Levantine Arabic speech affected by multiple noisy channels. Like other traditional features such as MFCC, the DAE-BN achieves better performance than traditional features but with much stronger robustness against unknown conditions, and convolutional neural networks and time frequency CNN consistently result in leading performance with a small relative reduction of WER. In particular, entropic confidence measures applied to model selection for unsupervised data were optimal for minimizing

WER by an additional 2–4%. An exhaustive approach that emphasizes flexibility of a characteristic as much as a beneficial modeling in real world noisy situations[10].

Singh et al. (2019) presents research on automatic speech recognition systems for Indian languages from 2000–2018, an exhaustive, and their review examines research trends and challenges. The study also weaves together concerns for linguistic diversity, paucity of speech corpus and dialectal variations that make it difficult to develop ASR for India, except Hindi. This paper reviews such techniques as Mel Freqency Cepstral Coefficients, the use of Hidden Markov Models, and novel deep learning techniques like Deep Neural Networks and Recurrent Neural Networks. Moreover, the study however also accented on significant advancement of tool such as HTK and Kaldi as well as hybrid technique used to enhance performance. The importance of robust speech corpora, feature extraction, incorporate the dialectal and tonal variations have been emphasised by the authors. This work demonstrates how ASR can lay the ground for linguistic OAs while requiring high effort – in terms of resource building and advanced machine learning methods for Indian languages[11].

In a 2018 paper, Fathima et al. present the development of a TDNN based multilingual ASR system for Indian language at the Interspeech 2018 Low Resource Speech Recognition Challenge. The common phonetic characteristics of Indian languages were tackled using integrated acoustic modelling and language specific decoding for Gujarati, Tamil and Telugu, with word error rates of 16.07%, 17.14%, and 17.69%, respectively. These multilingual training datasets, hybrid and language specific lexicons, and cutting-edge Time Delay Neural Network (TDNN) using lattice free Maximum Mutual Information (LF-MMI) criteria represent significant innovations. Research indicates that multilingual data and language-specific decoding can significantly enhance the ASR performance of resource-limited languages, hence facilitating the integration of analogous phonetic systems into a unified framework for broader linguistic coverage[12].

In the publication by Messaoudi et al. (2021), they describe an end-to-end automatic speech recognition (ASR) system for the Tunisian dialect utilizing Mozilla’s Deep Speech framework and a newly developed paired text speech dataset, TunSpeech.The utilisation of TunSpeech—comprising 11 hours of speech—alongside MSA and other publically accessible datasets has mitigated the issue of data scarcity. This methodology utilised a recurrent neural network with a connectionist temporal classification loss function for training, resulting in a word error rate of 24.4% and a character error rate of 18.7% on the test set, comprising 15 hours of Tunisian dialect and 50 hours of Modern Standard Arabic material. The results indicate that the integration of MSA data diminishes perplexity and the out-of-vocabulary (OOV) rate, however the incorporation of synthetic dialectal data elevates the word error rate (WER) due to substandard text-to-speech quality. The book demonstrates the possibilities of integrating dialect and MSA data in strong ASR systems for under-resourced languages.[13].

Kanke et al. (2021) have introduced advancements in small-vocabulary automatic speech recognition systems pertaining to the Marathi language. The techniques to be examined about Marathi speech encompass MFCC, DTW, and HMM. This research focusses on ASR applications in HCI, particularly isolated word recognition, which is applicable in devices necessitating basic speech input, such as diallers or voice-activated devices. This highlights the limited advancement of Marathi ASR relative to other Indian languages, hence emphasising the necessity for more robust systems designed for regional languages. Among the employed strategies, both DTW and MFCC are noted for their simplicity, making them more appropriate for small datasets, whilst neural network-based methods are advantageous for large datasets with standardised corpora. The evaluation emphasises the necessity of building ASR interfaces in the Marathi language, which will be beneficial in various applications, particularly for individuals with disabilities and in rural regions.[14].

Adiga et al. (2021) provide the inaugural LV-D ASR system for Sanskrit, developed utilising an innovative 78-hour speech corpus named वाक् सञ्चयः (Vāksañcayah). The sample consists of 46,000 utterances from various domains and time periods, articulated by native speakers of six languages, presenting issues related to phonetic alterations due to Sandhi and the extensive lexicon of Sanskrit. Grapheme-based modelling adheres to the encoding of SLP1 and incorporates vowel segmentation for acoustic unit selection and language models. The testing findings demonstrated that SLP1 surpassed the native scripts in accuracy, achieving a minimum WER of 21.94% with BPE. Moreover, insights were offered for enhancing the ASR systems of Gujarati and Telugu languages, demonstrating constant improvement attributed to phoneme-grapheme correspondences.

It emphasises that for low-resourced and morphologically rich languages, such as Sanskrit, linguistically informed modelling decisions may be significant[15].

The lightweight ASR system for the Gujarati language proposed by Tailor and Shah (2018) was built on the HMM method. A voice corpus including 650 utterances was created to train and assess the developed system, sourced from speakers in South Gujarat. The authors subsequently conducted feature extraction employing linear prediction techniques. Subsequently, they employed Viterbi-based decoding for pattern recognition. The system's performance averaged 87.23% in Word Recognition (WR) and 12.7% in Word Error Rate (WER). The hurdles to system accuracy stemmed from gender differences, accents, and linguistic intricacies. The study continues by emphasising the potential benefits of augmenting the data set size and integrating hybrid models to enhance system performance[16].

Sailor and Patil (2018) developed a neural network-based implementation of an automatic speech recognition system focused on agricultural commodities in Gujarati. This language is categorised as a low-resource language. This study utilised a dataset collected from 1,005 farmers in Gujarat, accounting for dialectal and environmental noise variances. The acoustic models employed Time-Delay Neural Networks and Long Short-Term Memory. It employed a Convolutional Restricted Boltzmann Machine for auditory feature representation. Ultimately, it employed a Recurrent Neural Network Language Model for language modelling. The RNNLM-based rescoring achieved a 1.18% absolute decrease in Word Error Rate relative to bi-gram models. A subsequent improvement was noted for a system integrating elements from both ConvRBM and Mel filterbank, yielding a 5.4% relative decrease in WER. It also emphasises the significance of neural network methodologies in enhancing the performance of automatic voice recognition in low-resource languages, which could benefit speech-based agricultural information systems[17].

Mehra and Jain have introduced ERIL, an algorithm for emotion recognition from speech utilising eight Indian languages, including Hindi, Gujarati, Marathi, and Tamil, through the application of machine learning techniques. The proposed system utilises MFCC, LPC, and pitch characteristics for emotion extraction and subsequent classification via the CatBoost algorithm. The UTU Semi-Natural Emotion Speech

Corpus comprises 1,840 samples evenly distributed among six emotions: anger, sadness, happiness, fear, surprise, and neutrality. Findings demonstrate that ERIL attains an average accuracy of 95.05%, exceeding current state-of-the-art standards. This study demonstrates that ERIL is highly resilient to tiny datasets, multilingual data, and noisy settings in the context of human-computer interaction and emotion-aware intelligent systems. Subsequent enhancements will incorporate additional languages and potentially more sophisticated functionalities, including FFT and wavelets[18].

The study by Mehra and Verma (2022) introduces BERIS, an emotion identification system designed for multilingual Indian speech, with mBERT as its foundation. It fuses the audio features—MFCC, LPC, and pitch features—with the textual features obtained from mBERT to construct a most imposing dataset of nine emotions: neutral, fear, anger, sadness, happiness, surprise, excitement, disgust, and frustration. The training and exam recordings encompass eight Indian languages: Hindi, Gujarati, Tamil, Telugu, Oriya, Bangla, Punjabi, and Marathi. The dataset employed is the UTU Semi-Natural Emotion Speech Corpus. BERIS outperformed benchmarks in classification with CatBoost, achieving an average accuracy of 98.38%. The research discusses the challenges of emotion recognition in under-resourced Indian languages, highlighting the efficacy of BERIS in managing a varied array of datasets and suggesting new avenues for enhancing feature extraction through FFT and wavelets[19] .

Darekar and Dhande's research study discusses the multimodal feature fusion of speech metrics, including MFCC, pitch, and energy, to enhance emotion recognition performance in Marathi speech. The study attained enhanced accuracy in emotion classification for six emotional states—happy, angry, sad, startled, fear, and neutral—by integrating the results from the aforementioned individual feature evaluations via artificial neural networks (ANNs). The system, utilising a speech database of 1,200 audio recordings from professional Marathi actors, demonstrates that fusion methods greatly surpass single-parameter approaches, achieving accuracy levels over 95%. The study emphasises the effectiveness of including several speech factors for reliable emotion recognition and suggests the potential for utilising this extensive array of emotions in multilingual applications[20].

Vryzas et al. offer a web-based crowdsourcing approach that improves personalised search engine results using transfer learning. The authors propose a framework that initiates with training a CNN on a varied multi-user dataset to enhance the generalisation of emotion recognition. This is succeeded by the refinement of the model with minor user-specific datasets to improve personalisation. Through use of the VGGish model pre-trained on AudioSet, this paper establishes a benchmark to evaluate how SER is improved when the transfer learning from large domain is employed. The authors additionally notch up a dynamic AESDD emotional speech dataset with a web based infrastructure that could record and examine emotional conversations remotely. Experimental results also showed that the transfer learning strategy works well, reaching an accuracy up to 69.9% versus classic methods and underscored the value of incorporating pre-trained models with user specific data. The contribution of this study is to show that crowdsourcing and transfer learning are a promising future for speech emotion recognition technologies[21].

To cope with domain inconsistencies in cross corpus Speech Emotion Recognition (SER), Zheng et al. (2021) propose Multi scale discrepancy adversarial (MSDA) network. The MSDA framework uses three timescales domain discriminators, global, local and hybrid in order to align features of labeled source domain and unlabeled target domain while conserving the emotion specific traits. We show that MSDA significantly improves WA as well as UA on challenge datasets including IEMOCAP, CASIA and MSP-IMPROV relative to baseline models using Domain Adversarial Neural Networks (DANN). This framework proves to be effective in addressing the cross-domain challenge in SER and incorporated multi scale features at one side and adversarial training at the other have helped to be reliable in disparate datasets in real world applications[22].

2.3 Previous Works in the fields of Voice Recognition System

In this paper by Hanifa et al. (2021), the plant of speaker recognition technology is presented from improvements to applications to limitations. It sets speaker recognition apart from speech recognition in that we recognize individuals based on their own unique vocal traits (such as pitch and speaking style). Feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC), and classifiers like Hidden Markov Models

(HMM) and Neural networks, are examined. Obstacles such as variability in voice signals, limited training data and interference in background noise are examined, in turn, regarding applications in authentication, personalization, surveillance and forensic science. Moreover, they analyze the emerging dangers, in particular adverse attacks that corrupt machine learning models to cause incorrect results. It draws attention to the need for resilient, secure and efficient speaker recognition systems for current biometric applications[23].

A real time Automatic Speech-Speaker Recognition system capable of reliable performance across noisy environments is provided by Kakade and Salunke (2020). Feature extraction is done using Mel Frequency Cepstral Coefficient (MFCC), and recognition tasks are performed using Dynamic Time Warping (DTW) and Vector Quantization (VQ), which show usage in high security areas such as banking and forensic investigation. The research involved creation of a bespoke speech database, as well as utilising pre-processing to increase precision, for instance noise reduction and signal enhancement. Evaluation shows that the MFCC-DTW-VQ system achieves high recognition power at the expense of computational cost. This study stresses the important role of improving feature extraction and matching methods for reliable, real time recognition in noisy environments[24].

Mokgonyane et al. (2019) introduce a text independent speaker recognition system for Sepedi, a South African indigenous language, based on machine learning techniques. We evaluated four classifiers: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Multilayer Perceptrons (MLP), and Random Forest (RF), on a dataset of 5,000 audio samples from 50 speakers. The considered characteristics were key such as Mel Frequency Cepstral Coefficients (MFCC) and Auto-WEKA was used to fine the models' hyper parameters. We found that Random Forest (RF) achieved the best accuracy of 99.9% which is outperforming MLP (96.5%), SVM (96.4%) and KNN (85.1%). A graphical user interface (GUI) was also created for system testing and deployment, and the study. The document outlines the capability of speech recognition technology for verification and forensic use, especially in languages with limited resources[25].

He et al. (2020) detail the development of open-source, multi-speaker speech corpus for six Indian languages: Kannada, Malayalam, Marathi, Tamil, Telugu and Gujarati. The

2,000 high quality, US English based recordings as well as the 1500 Marathi inadequate recordings are grouped in 50 corpus, each labeled for multilingual text-to-speech (TTS) and speech synthesis systems, save for the Marathi corpus which contains only US female recordings. To ensure very high phoneme coverage and high audio quality, scalable techniques were used to process the crowd sourced recordings. All languages performed well in the MOS assessment, all with indices above 3.6. These datasets are made freely available for academic as well as commercial use, in order to address the resource deficiency in the area of speech technologies for Indian languages and in encouraging multilingual speech applications[26].

Liu et al. (2018) propose the use of a hybrid model that consists of a Convolutional Neural Networks (CNN) embedded in a Gaussi In order to work around acoustic properties of short utterances, the model utilizes a GMM based Universal Background Model (UBM) for initial feature alignment and CNN for further feature extraction based on spectrogram pictures. Then, both MFCC scores and CNN classifying results are used simultaneously using a dual judgement technique to further improve the decision accuracy. The experimental findings show that the integration of statistical and deep learning methods can reduce Equal Error Rate (EER) from 4.9% to 2.5% during short audio, indicating the efficacy of integrating to speech recognition methods using both statistics and deep learning[27].

In Tiwari et al. (2018), a multi-modal i vector speaker identification for voice interactive systems is presented with diverse speech lengths, including 0.25 second utterances. The work refines classic I-vector approaches using various enrollment models tailored to different speech lengths, and improves performance for short utterances, all in the context of the THUYG-20 Uyghur language speech database. Experimental results suggest a reduction of the EER from 4.01% to 3.21% using 10 second inputs, and subsequent improvements in the EER for shorter inputs using Gaussian Probabilistic Linear Discriminant Analysis (GPLDA). For the limited case of short speech, this approach outperforms traditional i-vector systems with demonstrating substantial accuracy gains. It is shown to be applicable in real contexts including speaker recognition in noisy environments and intelligent gadgets[28].

In their work, Maurya et al. (2018) use Mel Frequency Cepstral Coefficient (MFCC), along with Gaussian Mixture Model (GMM) and Vector Quantization (VQ) techniques for Speaker recognition for Hindi voice signals. The research is carried in text independent and text dependent settings, where we have 15 speakers (10 male, 5 female) who perform 17 trials each. Results show that MFCC-GMM performs significantly better than MFCC-VQ: 86.27% accuracy for text independent and 94.12% for text dependent vs. 77.64% and 85.49%, respectively, for MFCC-VQ. It also discusses such issues as brief utterances, ambient noise, and session unpredictability and need rigorous training so that the effect of emotional and dialectal variations can be alleviated. The emphasis of this study is on the additional accuracy and practical use of MFCC-GMM in real world speaker recognition tasks[29].

A large scale BiLSTM based end to end speaker recognition system optimized for constrained training cases is provided by Nammous et al. (2022). The system works across 4,000 speakers in the Fisher English Speech Corpus in segments of 1 to 10 second duration. For feature extraction, MFCC is used and for classification, BiLSTM networks are used. The model achieves individual segment accuracy of 76.9% and bundled segment accuracy of 99.5 in about 30 seconds of training data per speaker, the research shows. The model sets up a balance between computing efficiency and precision by relying on critical variables to the exclusion of preprocessing. Finally, the key point of this research is that BiLSTM models can be used for scalable and adaptive speaker identification in resource restricted environments[30].

The study of Kabir et al (2021) gives a full view into the speaker recognition system from basic ideas and methods to some interpretable trails for future development. In this context speaker identification, signature, and diarization are the respective principal subdomains considered for both text dependent and text independent recognition systems.. Their discourse encompasses the conventional feature extraction techniques of MFCC and LPC, in addition to contemporary modeling methodologies employing i-vector, x-vector, and deep learning approaches. The authors discuss the datasets involved in the work and the performance metrics such as EER and DET, revealing concerns in variability of data across the speakers, interference due to external noise, and constrained resources. It ultimately addresses future prospects that encompass more advanced multimodal systems, improved noise management, and expanded datasets, highlighting the increasing significance of speaker recognition in biometrics and human-computer interaction[31].

This work presents a limited-vocabulary voice recognition system for Gujarati, developed using the bootstrapping method, primarily derived from an existing English voice Recognition Engine. This methodology aligns Gujarati speech data with the English phoneme set by phonetic mapping, facilitating the development of preliminary acoustic models of Gujarati with minimal training data. The assessment was conducted on 31 speakers, with an overall recognition accuracy of 88.71%. Male speakers had a greater recognition rate than female speakers, with rates of 90.88% and 85.28%, respectively. This study demonstrates that bootstrapping approaches are effective for low-resource languages such as Gujarati, despite constraints like linguistic ambiguity and extensive phonetic rules, providing a solid platform for refinement and growth[32].

2.4 Previous Works in the field of Voice Recognition for Indian Languages

The study of Nawaz et al. introduces cross-modal verification and speaker recognition in multilingual situations with a newly constructed audio-visual dataset named MAV-Celeb in English, Hindi, and Urdu. This research seeks to answer two questions: whether face-voice relationship is language-independent and whether speakers can be reliably recognized across languages. Performance dropped dramatically when testing was conducted on languages unseen during training, demonstrating both tasks are language-dependent. A two-branch neural network was employed for cross-modal verification, resulting in a baseline equal error rate (EER) of 29.0% in multilingual contexts. Speaker recognition systems such as VGG-Vox and SincNet experienced a 40-60% decline in performance when evaluated on unfamiliar languages. These are the definitive points arising from the necessity to address domain shifts caused by linguistic variances, which would enhance performance in multilingual biometric systems[33].

Ghoniemah and Shaalan developed an Arabic text-independent speaker verification system utilizing a novel FHMM combined with WPDFDs for feature extraction. This approach will address the issues of Arabic speaker verification by incorporating fuzzy memberships using Kernel Fuzzy C-Means, hence enhancing HMM training and minimizing information loss. A database of 100 speakers recorded in noise-free situations was employed, yielding a recognition rate of up to 98.38%. The suggested system demonstrated resilience to noise, attaining state-of-the-art performance across a broad spectrum of signal-to-noise ratios. The proposed approach integrates wavelet-based feature extraction with fuzzy modeling to enhance speaker verification capabilities in Arabic and other low-resource languages[34].

The study by Patil and Basu (2008) discusses the method used to generate voice corpora which help the research of speaker recognition in Indian languages such as Marathi, Hindi, Urdu, and Oriya. This study delineated the particular issues associated with linguistic variety, dialectal variances, and recording conditions within the Indian context. In building the corpus, a total of 600 speakers from various geographical regions have been captured in phonetic and dialectical variance. The authors evaluated the efficacy of different feature extraction methods, including LPC, LPCC, and MFCC, in automatic speech recognition (ASR). The findings of this study demonstrate that the MFCC feature typically outperformed the others. Nevertheless, LPC exhibited superior performance compared to the others in the identification of mimics. This study highlights the necessity for meticulously crafted corpora with a bio-application foundation, encompassing biometric speaker identification and forensic applications using multilingual ASR systems[35].

Saleem et al. offer a forensic speaker recognition system that focuses on accent classification and language identification from brief utterances, with particular attention to Urdu and its regional variants in Pakistan. Traditional approaches like GMM-UBM and i-vectors are integrated with advanced techniques: CNN-VGGVox and DNN-x-vectors. The x-vector technique demonstrated superior performance, achieving 80.4% for forensic speaker recognition alone, 85.4% with AC, 90.2% with LI, and 95.1% with the combination of AC and LI. This study utilized speech corpora featuring accents such as Punjabi, Sindhi, Pashto, and Balochi, as well as several languages, to narrow the suspect search area to regional groupings, thereby enhancing forensic efficiency. This replaces the focus on integrating linguistic and auditory characteristics for effective forensic applications[36].

Kumar et al. (2009) created a multilingual speaker identification system employing an artificial neural network with backpropagation for five languages: Hindi, Punjabi, Telugu, Sanskrit, and English, utilizing features such as LPC and LPCC for system training and classification. A text-independent model was created with a dataset including 25 speakers, with each voice articulating the designated sentences in all five languages. The architecture had an input layer with 575 neurons, two hidden layers, and a single output layer with 25 neurons, resulting in an average recognition accuracy of 85.74%. This study's primary contribution is to identify the potential of artificial neural networks (ANN) in recognition for multilingual speakers while also delineating certain limits regarding the management of large-scale datasets and the variety of speaker accents and settings[37].

Sarkar et al. (2013) offer a study on multilingual speaker recognition performance utilizing the IITKGP Multilingual Indian Language Speech Corpus-MLILSC, which comprises 13 prevalent Indian languages. Closed-set speaker identification and verification are done using the GMM framework. The results indicate an exceptionally high average speaker recognition accuracy of 95.21%, demonstrating the resilience of GMM classifiers. Nonetheless, the speaker verification yields an average EER of merely 11.71%, indicating room for improvement. The language discrepancy has impacted verification more than identification. The issues related to language variety have been identified, and additional efforts are suggested to enhance multilingual speaker recognition in India using advanced models such as GMM-UBM and GMM-SVM[38].

Chojnacka et al. (2021) introduce SpeakerStew, a multilingual speaker verification system that generalizes across 46 languages by combining a TD component with a cutting-edge TI component. To improve the generalization capability of the aforementioned system, it aggregates multilingual training data and employs a triage mechanism that alternates between lightweight TD and larger TI models according to confidence scores. The configuration minimizes computational demands and delay while maintaining precision. In this method, SpeakerStew achieved considerable error rate and computation reduction in experiments, such as up to 73% fewer calls to the TI model and a 59% drop in reaction time for English data. Moreover, the performance for both familiar and unfamiliar languages is exceptionally commendable. The advanced multilingual models surpassed their monolingual equivalents. The approach provides scalable solutions for many languages across diverse resource configurations in a multilingual speaker verification task[39].

The research by Wang et al. presents a novel network model for text-independent speaker recognition, emphasizing short utterances and difficult situations associated with environmental noise. This model integrates unique feature extraction methods such as MC-spectrogram and MC-cube, developed from Mel-spectrogram and cochleagram, with multi-dimensional CNNs and asymmetrical Bi-directional LSTM (ABLSTM) layers for the learning of local and global voiceprint features. The study introduces three model variants: Considering the AISHELL-1 and VoxCeleb2 datasets, Audio-1DCNN-ABLSTM, MCS-2DCNN-ABLSTM and MCC-3DCNN-ABLSTM. The proposed techniques show improvements on accuracy as well and resilience over state of the art models and MCC-3DCNN-ABLSTM achieves the highest accuracy found in our experiments. The second part confirms that our proposed system reduces noise and reinforces voiceprint properties, being able of working in suboptimal situations for several speaker recognition applications[40].

Nayana et al. (2017) compare two methods for performance: An evaluation of text-independent speaker identification systems with Power Normalized Cepstral Coefficients and Relative Spectral Perceptual Linear Prediction (RSP) using GMM and i-vector strategies. We find that GMM is competitive with i-vector approaches to brief utterances of 2–3 seconds, achieving accuracy of 94.7% with PNCC, compared to 85% accuracy with the i-vector and PLDA.. While the performance of GMM was equivalent to that of i-vector with PLDA for longer utterances (6–9 sec.), appending pitch and formant information resulted in an improvement in accuracy for all the models. Here, the accuracy of the Gaussian Mixture Model (GMM) was 97.7% and the i vector was 90.7%. Robustness of the PNCC in noisy conditions and the potential benefits of incorporating pitch and formant information for speaker identification accuracy were the foci of this research[41].

Xu et al. introduced a deep multi-metric learning approach for text-independent speaker verification, integrating triplet loss, n-pair loss, and angular loss within a ResNet-based architecture enhanced by Squeeze-and-Excitation blocks for attention mechanisms. These three losses synergistically address the constraints of conventional single-loss metric learning to improve feature extraction. The extensive VoxCeleb2 dataset was utilized for trials, demonstrating that the suggested enhanced performance framework surpasses the state-of-the-art approaches with an EER of 3.48%. This research demonstrates that the integration of loss and attention techniques can enhance speaker verification systems, applicable in various domains like as authentication, forensic analysis, and intelligent interaction systems[42].

In their study, Shahnawazuddin et al. have offered various ways for increasing the speaker independence of the ASR systems, which are trained with limited data, particularly on the issues of pitch and speaking rate variations between adults and children. The primary authors propose two distinct FCSB-based methodologies: the alteration of children's speech to resemble adult acoustic profiles during testing and the creation of prosody-modified variants to enhance an adult speech training dataset. Experimental investigations on British English datasets indicated significant improvements for both techniques: 17% relative WER reductions for adult speech and 31% for children's speech with the augmented model. This work confirms the efficiency of FCSB-based approaches by bridging the acoustic mismatch and being more robust against noisy situations, advocating for deeper integration with noise-resistant front-end characteristics[43].

The work by Kinkiri et al. (2020) gives an insight into how one recognizes speakers with the typical qualities of a person's voice, and it states that both verbal and non-verbal cues are relevant in one way or another. It analyzes essential aspects including frequency, timbre, loudness, speech pace, accent, and pauses, which collectively contribute to the determination of their identification. The authors conducted an experiment utilizing a database of volunteers with varied linguistic backgrounds, performed spectral analysis to assess fundamental frequency range, and examined articulation and speech rates. The results indicated factors, including frequency peaks and speech pace, that would significantly narrow a speaker database to a limited range of potential matches. The study highlights the difficulties of using the human voice as a biometric identifier, while also identifying its applications in forensic and security systems, noting obstacles related to variations in accents, speech patterns, and environmental noise[44].

Bian et al. have introduced a novel framework utilizing self-attention for text-independent speaker recognition. This approach combines Residual Networks with a self-attention mechanism to successfully capture both local and global voiceprint characteristics. The suggested architecture employs a unique Cluster-Range Loss function that reduces intra-class fluctuations and enhances inter-class separations to achieve improved recognition accuracy. For the VoxCeleb dataset, the achievements contain EER of 5.5% in speaker verification and Top-1 accuracy of 89.1% in speaker identification, ensuring superior outcomes to other methods like i-vector/PLDA and x-vector models. Given these computational advantages, this method is expected to be effective: operating without fully connected layers, and making use of a lightweight attention mechanism, leading to scalable, high precision speaker recognition in practical applications[45].

In Chen et al. (2020), a novel SpeakerGAN model is introduced, a text-independent speaker recognition model using Conditional Generative Adversarial Networks. The SpeakerGAN consists of a generator and a discriminator from which both learn the speaker characteristics' distribution and identify speaker identities. To improve the learning efficiency and convergence, the model uses gated convolutional networks to learn in the generator, and a modified ResNet architecture in the discriminator. Specifically in low data scenarios, it achieves strong performance by jointly applying adversarial loss with classification loss and Huber loss. On LibriSpeech we evaluate SpeakerGAN and find it to outperform conventional systems like i-vector and x-vector by up to 87% reduction in error rate. On test clips of 1.6 seconds, the model had 98.2% accuracy identifying speakers. It demonstrates that this resource efficient design makes it very suitable for applying it to speaker recognition tasks[46].

2.5 Previous Works in the field of Voice Recognition for Gujarati Language

Patel and Nandurbarkar proposed a speaker recognition system for a Gujarati language using a weighted Mel Frequency Cepstral Coefficient and a Gaussian Mixture Model in 2015. To ensure quality the suggested speaker recognition system analyses 30 samples of voice from 20 male and 10 female individuals in a soundproof environment. Weighted MFCC is used as a feature extractor to enhance the system's ability to represent the speaker-specific vocal feature and GMM for speaker modeling.. Experimental findings indicate that the suggested weighted MFCC+GMM surpasses the traditional MFCC+GMM by approximately 1% in recognition accuracy. This paper presents insights into the feature extraction approaches essential for the enhancement of speaker recognition in low-resource languages like Gujarati[47].

Shah and Kavathiya have conducted a thorough assessment of the speech recognition system for Gujarati dialects, revealing significant research gaps regarding accuracy in various contexts and accents. The research covers feature extraction and classification methods such as MFCC and HMM as well as hybrid neural networks using successes on Gujarati Speaker recognition.. Despite the promising outcomes attained by several methods globally, including multilingual models like SPEAKERSTEW and neural network-based techniques, their use in Gujarati remains infrequently investigated. In the available literature, variable performance has been documented for Gujarati speech, attributed to problems such as accent variability, speaker variability, and background noise. The authors have emphasized the potential to develop a powerful, accent-sensitive speech recognition system for Gujarati, addressing the deficiencies of existing frameworks and the intricacies of implementation[48].

Ahuja and Vyas (2016) investigate the application of supra-segmental features, such as stress, intonation, tone, and vowel length, for text-independent identification of Gujarati speakers. Speech samples are collected from 1,400 speakers representing four principal dialects: Standard Gujarati, Kathiawari, Carotari, and Kutchi. This research identifies dialect-specific acoustic patterns through auditory and prosody analysis and evaluates their efficacy in forensic speaker profiling. These demonstrate considerable variations in vowel quality, intonation patterns, and lexical tones within these languages; speakers of Standard Gujarati had the fastest speaking rate, whereas Kathiawari speakers tended to lengthen their vowels longer. This study demonstrates the utility of supra-segmental features in differentiating dialectal accents and their applicability in speaker identification and forensic analysis[49].

Gupta et al. present a method, termed G-Cocktail, for addressing the "cocktail party problem" in the Gujarati language, which effectively separates and identifies voices from a composite audio stream. The model primarily utilizes MFCC, pitch, and the CatBoost algorithm for classification, while contrasting K-means, Naïve Bayes, and LightGBM. The dataset is compiled by Microsoft and LDC-IL, encompassing adult voices and expressive speech with systematically arranged recordings. The results generated using this G-Cocktail approach reached an accuracy of 98.33%. This surpassed current approaches such as HMM, TDNN, and RNN-CTC, particularly for small datasets susceptible to overfitting. It emphasizes its application in voice assistants and potential for advancement in multilingual and loud settings[50].

The study by Ambikairajah et al. (2011) serves as a tutorial on automatic language identification, detailing the system's evolution, methodology, and applications. It presents LID as a system that recognizes spoken languages from audio input by utilizing acoustic, phonotactic, and prosodic characteristics. The paper discusses front-end and back-end processes, including feature extraction techniques such as Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP), as well as modelling methods like Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), and Support Vector Machines (SVM). Advanced systems encompass those utilizing GMM-UBM and phonotactic models; they are also highlighted. It highlights that certain variables presenting challenges include brief utterances, noisy surroundings, and data paucity, while normalizing strategies encompass methods such as Cepstral Mean Subtraction to address variability. The authors underscore that the functions of LID are continually evolving in multilingual services and applications within biometrics, and additional research contributions are solicited to address certain limits exhibited by current systems[51].

Anusuya and Katti (2009) examine advancements in ASR systems, including around sixty years of progress. It addresses a range of topics, including acoustic-phonetic methodologies, pattern detection, and artificial intelligence techniques utilizing neural networks and support vector machines. Key difficulties addressed include speaker variability resulting from accent, noise circumstances, and computational limitations. Extraction techniques such as Mel Frequency Cepstral Coefficients and Linear Predictive Coding are significant components in Automatic Speech Recognition. Conventional uses of the system encompass telecommunications, education, healthcare, and assistive technology, wherein ASR may possess significant automation and accessibility capabilities. Significant progress has been made in handling spontaneous speech and multilingual contexts, but the research domain of managing spontaneous speech and multilingual contexts is still dynamic, spearheading innovation to make robustness and scalability a reality[52].

In Patil et al. (2023), the text was examined on topics of sociolinguistic and philological characteristics of Gujarati dialects to figure out speaker identification from a forensic speaker identification (FSI) perspective. To set up dialect-based framework for forensic profiling, the study analyses the differences in pronunciation, vocabulary and acoustic characteristic across different Gujarati regions with emphasis on major dialects like Kathiyawadi and Surati. To find such region-specific language factors, audio and phone recordings were gathered and both qualitative and quantitative approaches were used to data acquired from audio and phone recordings. Results show that even routine words can be extremely different from dialect to dialect, with measures like geography, education, and socioeconomic status easily separating a speaker by regional accent.. This research elucidates the function of forensic linguistics in criminal investigations, particularly its utility in identifying speakers in anonymous calls and deriving legal implications from linguistic data[53].

This study by Desai and Ramsay-Brijball delineates the historical and sociolinguistic development of the Gujarati language from its Sanskrit origins through the Indo-Aryan linguistic phases of Prakrit and Apabhramsa. The foundational philological framework of Gujarati's linguistic structure and its dialectical variations, including Kathiyawadi, Surati, and Charotari, is delineated. It also talks about the state of Gujarati in South Africa, particularly within the Indian communities there, understanding that much has happened along lines of migration, cultural assimilation, and generational shifts. The impact of colonialism, along with lexical borrowing from English, Afrikaans, and isiZulu, and regional sociolinguistic practices, has shaped this trajectory. The research emphasizes the necessity of documenting these developments to complement the socio-linguistic dynamics and safeguard the heritage of the language within a continually evolving linguistic environment[54].

Mesthrie's (2023) research examines a consonantal chain shift in Gujarati dialects, a neglected linguistic phenomenon characterized by systematic phonetic alterations. The transformations encompass /k/ and /kh/ changing to /c/ and /ch/, then growing into /s/ or

/ś/, which may further develop into /ḥ/ or /h/, and in certain instances, vanish completely (represented as /ø/). The study examines the variations in Surti, Kathiyawadi, and Charotari dialects, emphasizing regional differences: for example, Surti demonstrates changes such as /ch/ to /s/, but Charotari displays /k/ to /c/ or /kh/ to /s/. These modifications are congruent with historical phonetic developments of Indo-Aryan, but also show regional linguistic variation. These variations, he suggests, might reflect push chain mechanisms and possible Dravidian or Austro-Asiatic substrate influence, enriching our understanding of the development of Gujarati as a language and dialectal complexity[55].

According to Kurian's (2015) research, advancements in creation of speech corpora for Indian languages are evaluated with challenges as well as the initiatives undertaken in constructing linguistic resources in the midst of India’s varied linguistic landscape. It looks at important efforts around languages like Hindi, Marathi, Telugu, Punjabi, Kannada, and Garhwali and studies techniques such as mobile recordings, studio data collection, and phoneme specific. Examples include the building of many language mouth databases for massive vocabulary recognition, together with their implementation in areas like Mandi Information Systems and travel area automatic speech recognition systems.. Notwithstanding these gains, the research highlights significant deficiencies, including the paucity of open-access corpora, the lack of centralized repositories, and insufficient focus on dialectal variants. Kurian stresses the significance of systematic, national-level collaborations to build complete, publicly available voice databases for Indian languages, supporting advancement in speech technology research and applications[56].

Ardila et al. (2020) introduce Common Voice, a crowd-sourced multilingual speech corpus designed for automatic speech recognition and various speech technology applications. Featuring contributions from over 50,000 individuals, the corpus comprises validated recordings in 38 languages, equating to more than 2,500 hours of audio, making it the largest public-domain corpus for ASR. The data collecting employs a systematic crowdsourcing framework with integrated validation processes to ensure quality control. Experiments utilizing Mozilla’s DeepSpeech toolkit demonstrate the advantages of transfer learning, revealing a 5.99% average enhancement in Character Error Rate (CER) across 12 languages. Common Voice serves as an essential resource for the progression of ASR research, particularly for low-resource languages, providing community-driven development and accessibility under a Creative Commons CC0 license[57].

Tank and Hadia (2020) research conducted an emotional speech corpus in Gujarati and outlines the lack of resources for the development of speech emotion recognition (SER) in the common language. Six emotional states are articulated by nine people, each 20 to 25 years old, who are educated in acting; the corpus consists of sad, surprise, wrath, contempt, fear and happiness. They collected 1,296 samples using mobile devices containing 24 words. Analysis of metrics such as energy, pitch, and MFCC using MATLAB uncovered emotion-specific patterns: energy levels were elevated in joy and rage, but decreased in sadness and fear, with decreased pitch. The work demonstrates dataset's potential for use in SER tasks, and provides directions to fill gaps in it, namely having a larger set of speakers as well as natural speech itself, improving multimodal methodologies or changing the classifier for the sake of increasing identification accuracy[58].

Sztahó et al. (2019) analyze in which ways deep learning (DL) methods contribute to improving speaker recognition, they investigate the development of (speaker) identification and (speaker) verification. The text describes the transition from standard methods like i-vectors to deep learning based methods d-vectors, j-vectors, and x-vectors; the latter showing large improvement, having a 5.86% equal error rate (EER) on enriched datasets.The document also emphasizes developments such as end-to-end systems, SincNet, and corrective learning networks (CLNet), which improve accuracy in both text-dependent and text-independent tasks. Among other things, the paper points out still existing barriers to deep learning achieving more of a speaker recognition breakthrough, even though deep learning has greatly improved overall speaker recognition performance, namely dirty data, platform variability, and the effective use of unlabeled data. The research highlights the growing importance of deep learning in speaker recognition while suggesting generalizability advancements and motivated multi lingual datasets[59].

In this paper, in the context of language identification and despite a considerable absence of well established multilingual speech corpora for LI, Maity et al. (2012) introduce the

IITKGP-MLILSC — a rich multilingual speech corpus developed for LI, which includes 27 Indian languages. In speaker dependent configurations, spectral features like MFCCs and LPCCs with Gaussian Mixture Models (GMMs) used in the classification, the system attains a recognition rate of 96%. Nevertheless, performance was strongly affected by the lack of speaker specific data and dropped to 41% in the speaker independent case. A little accuracy improvement to 45% was achieved when speaker specific models were integrated. The study indicates areas including linguistic overlaps, and lack of use of prosodic cues, which suggests the need for sophisticated methods, such as neural networks in order to improve the generalizability and precision of LI systems[60].

In a linguistically varied country with over 1,652 dialects, Shrishrimal et al. (2012) provide an extensive study of evolution and use of voice databases for Indian languages and recognize their indispensably important role in promoting speech recognition and text to speech (TTS) systems. This document classifies these databases into general purpose, and extrapolates them in application specific segments including agriculture, travel and mobile communication. Institutes such as CDAC Noida, IIIT Hyderabad, TIFR are examined for their contributions. While the study reveals deficiencies in resources for underrepresented languages, and challenges such as managing noisy situations and continuous voice recognition, the study carries substantial advancements for Himalayan languages, including Hindi, Tamil, Telugu, and Bengali. The importance of the Linguistic Data Consortium for Indian Languages (LDC-IL) in developing databases is underlined, and we propose more inclusive and noise robust systems for practical usage[61].

The study attained notable outcomes, with SID accuracy at 94.49% and LID accuracy at 95.69% for 10-second utterances, however performance diminished for shorter durations. Identified challenges encompass linguistic overlaps that diminish classification accuracy and the necessity for augmented datasets featuring a greater diversity of speakers and languages. The research emphasizes the capability of sophisticated models to surmount these restrictions and enhance accuracy, especially for brief utterances and overlapping linguistic characteristics[62].

He et al. (2020) present open-source, high-quality, multi-speaker speech corpora for six Indian languages—Gujarati, Kannada, Malayalam, Marathi, Tamil, and Telugu—intended to facilitate text-to-speech (TTS) applications. Developed utilizing scalable and cost-effective approaches, these datasets include recordings from both male and female native speakers, with over 2,000 words per language emphasizing phonetic variation and variability. Stringent quality control measures guaranteed noise-free recordings, yielding Mean Opinion Scores (MOS) over 3.6, with numerous instances exceeding 4.0, indicating superior quality synthesized voices. The corpora tackle resource constraints for low-resource languages, while the study highlights the opportunity to extend to more languages and implement advanced modeling techniques to enhance TTS systems[63].

Kothari and Kumbharana (2015) describe the creation of a phoneme-based database for Gujarati text-to-speech (TTS) synthesis, utilizing the concatenative synthesis technique. The research aims to provide an extensive phoneme inventory comprising 872 phonemes, derived from diverse combinations of vowels, consonants, and diacritic modifications. The speech data, recorded by natural speakers at a 44,100 Hz sampling rate, underwent meticulous post-processing to remove noise and silence, guaranteeing superior quality output. The database is optimized for efficient phoneme matching and retrieval, utilizing the phonetic structure of Gujarati. The methodology exhibits adaptability to other Indian languages, such as Marathi and Hindi, through the modification of ASCII codes. This resource markedly improves the clarity, precision, and authenticity of Gujarati TTS systems, providing a solid basis for future developments[64].

Slavnov et al. (2020) present a voice corpus development system aimed at overcoming obstacles in speech recognition for various accents, dialects, and speech disorders. The system utilizes an acoustic-phonetic methodology and implements a graph database architecture to record phonemes, speaker characteristics (e.g., gender, age, native language), and metadata, hence assuring scalability and efficient data access. A web-based user interface, developed using NodeJS and React, enables annotation and data uploading, incorporating client-side segmentation to reduce server demand. The adaptability of graph databases, such as Neo4j, facilitates intricate queries, including the filtration of phonemes according to speaker attributes. The system seeks to enhance voice recognition precision for marginalized user demographics exhibiting speech variances or limitations. The authors suggest improvements in user collaboration and interface design to enhance the system's functionality[65].

Prahallad et al. (2012) discuss the establishment of the IIIT-H Indic Speech Databases, featuring speech corpora for seven Indian languages: Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil, and Telugu. The databases were created utilizing public domain texts, including Wikipedia articles, and recorded in regulated studio settings by native speakers. Phonemically balanced sentences were chosen utilizing Festvox scripts to guarantee linguistic diversity and thorough phonetic representation. The speech recordings were segmented and manually validated for accuracy, leading to the release of finished corpora for public usage. These databases facilitate applications such as text-to-speech systems, exemplified by prototype voices created with the Festvox framework. Identified challenges encompass managing phonotactic diversity, modeling prominence, and predicting phrase breaks in Indian languages. The authors highlight the databases' open accessibility to promote progress in speech technology research and applications[66].

2.6 Previous Works related to Feature Extraction Techniques

Hourri and Kharroubi (2019) propose a deep learning methodology for speaker recognition that transforms Mel-frequency cepstral coefficients (MFCCs) into deep speaker features (DeepSF) via deep neural networks (DNNs). The technique utilizes deep belief networks (DBNs) for weight initialization and deep neural networks (DNNs) for the supervised learning of feature distributions. A novel scoring technique for speaker verification, termed NearC, is based on cosine distance. Evaluations of the THUYG-20 SRE corpus demonstrated that this methodology surpassed conventional i-vector/PLDA and baseline MFCC-NearC systems, attaining an equal error rate (EER) as low as 0.43% in pristine conditions and exhibiting strong performance in noisy environments, with EERs not exceeding 3.19%. This technology considerably enhances noise resilience and accuracy, making it a strong option for secure voice biometrics and speaker verification applications in demanding circumstances[67].

Gade and Sumathi (2021) present a comprehensive analysis of automatic speaker recognition systems employing deep learning methodologies, concentrating on speaker verification and identification in difficult conditions such as noise and domain discrepancies. The review delineates the recognition process into preprocessing, feature extraction, feature selection, and classification stages, with sophisticated models including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and i-vector frameworks assuming pivotal roles. While deep learning methods have considerably increased recognition accuracy, persisting challenges include handling noisy settings, data shortages, and gradient instability during training. The authors emphasize promising advances such as x-vectors, generative adversarial networks (GANs), and multi-channel speech enhancement to address these difficulties. They recommend for more study into robust feature extraction and domain adaptation strategies to increase the reliability and scalability of speaker recognition systems for real-world applications[68].

The study by Zhang, Chen, and Wang (2023) analyzes progress in deep learning-driven speaker recognition, highlighting its use in biometric verification via distinct vocal characteristics. The paper explores three pivotal domains: domain adaptation, which tackles data distribution discrepancies through methods such as transfer learning and adversarial training; speech enhancement and de-reverberation, aimed at reducing noise and reverberation; and data augmentation, which improves model resilience by generating enriched datasets. Techniques like DNN-based embeddings (e.g., x-vectors) and GANs for noise reduction have significantly enhanced accuracy in real-world applications. Notwithstanding these advancements, difficulties such as inadequate data quality, speaker variability, and linguistic adaptability remain. The study offers a comparative review of current methodologies, identifying deficiencies and suggesting more research to enhance the efficacy and scalability of speaker identification systems[69].

Bansal et al. (2021) presented the Fused Features Hybrid Extraction Technique (FFHT) for speaker recognition, integrating features from the temporal, frequency, and cepstral domains, including MFCC, Zero Crossing Rate (ZCR), and RMSE, to enhance accuracy and efficiency. The approach utilizes a feed-forward back-propagation neural network optimized with Gradient Descent with Momentum, attaining a notable accuracy of

97.56% on the VoxForge dataset. This performance exceeds conventional methods, such as MFCC with MLP (94.44%) and clustering-based MFCC with ANN (93%). The technique notably decreases training duration and improves classification; however, the study's reliance on a restricted dataset of 34 speakers emphasizes the need for future investigations utilizing larger datasets and more sophisticated models to assess its generalizability and efficacy under diverse settings[70].

Li et al. (2020) introduced a speaker verification method that combines x-vector deep learning models with Probabilistic Linear Discriminant Analysis (PLDA) to tackle channel mismatch issues. The approach utilizes Time Delay Neural Networks (TDNN) for x-vector extraction and implements PLDA for channel compensation, thereby successfully mitigating noise and distortion in cross-channel settings. Upon evaluation using the AISHELL-2 dataset, the system exhibited a 35% decrease in Equal Error Rate (EER) relative to conventional i-vector techniques, indicating substantial enhancements in robustness and precision. The amalgamation of deep learning with PLDA demonstrates its capacity to enhance speaker verification efficacy in practical applications[71].

Hamidi et al. (2020) provide a comprehensive assessment of speaker recognition techniques, detailing the evolution from initial methods to contemporary innovations. The document classifies speaker recognition into two primary tasks: verification and identification, while also differentiating between text-dependent and text-independent systems. Essential processes such as feature extraction and modeling are examined, with prevalent approaches including MFCC, LPC, HMM, and neural networks emphasized. The review assesses the advantages and drawbacks of speaker identification systems, focusing on difficulties such mobility, unpredictability, and susceptibility to spoofing. Particular emphasis is placed on recognition systems for Arabic and Amazigh speakers, analyzing language-specific feature extraction and modeling methodologies. The authors conclude by highlighting ongoing challenges such as noise management and data variability, and they suggest future research avenues to improve the resilience and precision of speech recognition methods[72].

Prachi et al. (2022) created a speaker recognition system utilizing deep learning methodologies, particularly Convolutional Neural Networks (CNN) and Long Short-

Term Memory (LSTM) models. Using Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction, the study examined both open-set and closed-set implementations on the TIMIT and LibriSpeech datasets. The CNN model surpassed the LSTM model, with accuracies of 80.63% and 97.85% for closed-set recognition on TIMIT and LibriSpeech, respectively, in contrast to the LSTM model's 71.54% and 84.96%. Although open-set recognition accuracies were generally inferior, CNN consistently outperformed LSTM in all cases. The results highlight CNN's efficacy in audio categorization and recommend investigating advanced models such as SincNet to improve speaker recognition capabilities[73].

Ohi et al. (2021) provide an extensive assessment of developments in speaker recognition, emphasizing the transition from conventional methods such as GMM-UBM and i-vector systems to advanced deep learning frameworks. The research analyzes different architectures, encompassing stage-wise systems like x-vector and t-vector, as well as end-to-end methodologies such as Deep Speaker and SincNet, detailing their workflows and applications. Key issues, including domain adaptability, noise resilience, and generalization to real-world or "in-the-wild" scenarios, are thoroughly covered. The paper emphasizes unique techniques such as meta-learning for low-resource contexts and generative adversarial networks (GANs) for data augmentation and noise mitigation. The authors highlight advancements in speaker recognition accuracy through deep learning and call for additional study into explainability and improved adaptability to varied situations[74].

Hu et al. (2022) present a domain-robust deep embedding learning method for speaker verification, addressing the issues of domain shifts between training and testing datasets. This multi-task, end-to-end strategy mixes labeled source data with unlabeled target data to promote adaptability. The method uses Smoothed Knowledge Distillation (SKD) for self-supervised learning, successfully decreasing noise in pseudo-labels while collecting latent structural information. Domain resilience is further strengthened by Domain-Aware Batch Normalization (DABN), which eliminates cross-domain differences, and Domain-Agnostic Instance Normalization (DAIN), which mitigates within-domain variance. Assessed using the NIST-SRE16 dataset, the framework attained a 19% relative decrease in Equal Error Rate (EER) in comparison to baseline systems, indicating enhanced flexibility without need on adversarial training or data augmentation[75].

Sefara and Mokgonyane (2020) present a comparative comparison of machine learning and deep learning techniques for emotional speaker detection utilizing the RAVDESS dataset, which includes speech recordings spanning eight emotion categories. The research analyzes five machine learning models, namely Logistic Regression, Random Forest, and SVM, in conjunction with three deep learning models—MLP, CNN, and LSTM—utilizing information from the temporal, frequency, and spectral domains. Deep learning models exhibited exceptional performance, with the MLP attaining the greatest accuracy of 92%. The study emphasizes the efficacy of deep learning in addressing the complications arising from emotional variability in speaker recognition and reinforces the significance of feature engineering and normalization in enhancing model performance[76].

Costantini et al. (2023) evaluate CNN-based deep learning techniques against conventional machine learning methods for speech detection with the DEMoS dataset. A bespoke CNN architecture (CNN1) attained the maximum accuracy of 90.15% using grayscale spectrograms, surpassing AlexNet's 89.28% and a Naïve Bayes model's 87.09% accuracy. Although Naïve Bayes demonstrated marginally reduced accuracy, it achieved a commendable AUC of 0.985, coupled with expedited training durations and enhanced interpretability. The research highlights the importance of feature types, including MFCC and F0-related metrics, in vocal analysis. It also underscores the trade-offs among accuracy, training efficiency, and model complexity. The study highlights the efficacy of grayscale spectrogram-based CNNs for speaker recognition and advocates for the investigation of sophisticated architectures and augmentation techniques[77].

Gade and Sumathi (2023) presented a Hybrid Deep Convolutional Neural Network (HDCNN) model designed for speaker recognition in noisy settings, utilizing sophisticated data augmentation and feature extraction methods, including Mel-Frequency Spectral Coefficients (MFSC). The HDCNN design incorporates convolutional, pooling, and fully linked layers to efficiently capture speaker-specific spectral characteristics. Evaluated on datasets like ELSDSR and TIMIT, the model beat standard techniques, including ANN, CNN, SVM, and RF-SVM, obtaining an accuracy of 98.33%, precision of 95.19%, and recall of 95.17%. This illustrates its resilience in noisy environments. Despite its good performance, the study highlights problems linked to different noise situations and the model's computing intensity, proposing future research to better adaptability and efficiency for real-world applications[78].

Bousquet and Rouvier (2023) present an unsupervised subset selection approach designed to enhance training datasets for resource-limited speech identification systems. The method discerns the most informative speakers from extensive datasets, optimizing training efficiency and system precision. The method employs x-vector embeddings and agglomerative hierarchical clustering to condense the dataset to 30% of the speakers, while preserving performance similar to models trained on the complete dataset. The methodology was evaluated using datasets such as VoxCeleb, DeepMine, and TED-X Spanish, resulting in enhanced Equal Error Rate (EER) and detection cost function (DCF) scores, especially in contexts characterized by linguistic or environmental diversity. This study highlights the efficacy of subset selection in improving generalization and scalability in speaker identification systems within computational constraints[79].

Paramitha et al. (2022) evaluated the efficacy of CNN, LSTM, and GRU algorithms for speech emotion recognition utilizing the Berlin EMODB dataset. The research employed variables such as MFCC, ZCR, RMSE, Mel Spectrogram, and Chroma from the dataset, which classifies recordings into seven distinct emotions. Models were trained on 80% of the dataset and evaluated on the remaining 20%. CNN surpassed other models with an accuracy of 79.13%, while LSTM and GRU achieved 55.76% and 55.14%, respectively. The results illustrate CNN's proficiency in managing feature-dense datasets, while LSTM and GRU exhibited constraints in capturing intricate temporal connections. The authors propose enhancing feature extraction methods and investigating advanced model architectures to improve performance in subsequent research[80].

Choudhary et al. (2020) provide an outline of the Linguistic Data Consortium for Indian Languages (LDC-IL) effort, aimed at mitigating the deficiency of linguistic resources for Indian languages by the creation of extensive raw voice corpora. These datasets facilitate applications including Automatic Speech Recognition (ASR), Speech-to-Text (STT), and linguistic analysis. The corpora, developed with contributions from language specialists, encompass a diverse array of disciplines, including modern text, creative writing, phonetically balanced lexicon, and proper nouns. Data collection prioritizes variety by include many age groups, gender distributions, and regional dialects to ensure equitable representation. Sophisticated methods for segmentation, metadata mapping, and quality assurance guarantee data integrity. The program seeks to enhance research and technological applications within India's multilingual framework, while also tackling problems such as infrastructural limitations and ethical issues in data acquisition[81].

Bai and Zhang (2021) give an exhaustive assessment of deep learning strategies in speaker recognition, encompassing developments in subtasks such as speaker verification, identification, diarization, and robust recognition. The research examines how deep learning has transformed conventional systems like GMM-UBM and i-vector models via embedding techniques such as x-vectors and d-vectors, facilitating enhanced feature extraction. It analyzes technologies such as supervised and end-to-end frameworks, sophisticated pooling algorithms, and domain adaption strategies to mitigate noise and unpredictability. The research delineates primary problems such as domain discrepancies, the necessity for voice improvement, and the requirement for extensive datasets, while also citing publically accessible corpora. Bai and Zhang highlight the advantages of deep embeddings and neural network designs compared to conventional methods and recommend additional investigations into multimodal integration and real-time applications to enhance the discipline[82].

Shome et al. (2023) present a comprehensive evaluation of deep learning methodologies in speaker recognition, encompassing both speaker identification and verification. The paper analyzes essential elements of speech identification systems, encompassing preprocessing techniques, feature extraction methods such as MFCC and LPCC, and classification frameworks like GMM, ANN, and x-vector. It shows the advantages of deep learning models, particularly DNNs and TDNNs, in handling variability in speech signals induced by environmental noise, speaker-specific features, and linguistic diversity. The research addresses difficulties including data scarcity, domain discrepancies, and the substantial processing requirements of deep learning systems. It suggests future trajectories, encompassing the advancement of multimodal methodologies and real-time systems, to improve efficacy and relevance in practical contexts. The authors underline the crucial importance of advanced designs in overcoming these problems and increasing speaker recognition outcomes[83].

Premakanthan and Mikhael (2001) present a foundational assessment of speaker verification and recognition systems, underlining the necessity of selective feature extraction and normalization techniques. The research delineates a four-phase Automatic Speaker Verification (ASV) procedure: speech data acquisition, feature extraction and selection, clustering of feature vectors, and decision-making using pattern matching. The text examines multiple feature extraction techniques, such as Linear Prediction Coefficients (LPC), Cepstral Coefficients (CC), Mel Frequency Cepstral Coefficients (MFCC), and Discrete Wavelet Transform Coefficients (DWTC), highlighting their significance in capturing distinctive vocal characteristics. The significance of feature normalization is emphasized, especially for mitigating signal variability caused by noise or transmission discrepancies. Techniques such as Dynamic Time Warping (DTW), Vector Quantization (VQ), Hidden Markov Models (HMM), and Artificial Neural Networks (ANN) are examined for their efficacy in pattern matching. The research emphasizes persistent obstacles in attaining robust and precise systems and advocates for improvements in modeling and computing efficiency to address existing limits in speaker verification technology[84].

Bakshi and Kopparapu (2018) present an extensive overview of the features and databases utilized in spoken Indian language identification (SLID), highlighting the auditory, phonetic, and prosodic attributes that differentiate Indian languages. The research examines the difficulties of SLID within India's linguistically varied context, which encompasses more than 800 spoken languages. Essential characteristics such as MFCC, PLP, LPCC, and Delta Cepstrum are examined, along with their incorporation into deep learning frameworks, illustrating their efficacy in SLID tasks. Speech corpora are divided into general-purpose and application-specific databases, with the evaluation highlighting a substantial scarcity of high-quality, standardized datasets for Indian languages. The authors recognize data variability, phonotactic discrepancies, and the necessity for effective feature selection methods as crucial domains for more investigation. They highlight the significance of phonetic nuances and the creation of extensive corpora in enhancing SLID systems for Indian languages[85].

Shrawankar and Thakare (2009) conduct a comparative examination of diverse feature extraction strategies for voice recognition systems, highlighting their essential contribution to enhancing recognition performance. The research explores approaches such as Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), and Relative Spectral Filtering (RASTA), describing its mathematical underpinnings, advantages, and limitations. LPC is noted for its efficacy in low-bitrate encoding, whereas MFCC corresponds effectively with human auditory perception but is susceptible to noise interference. PLP integrates psychoacoustic models, enhancing noise resilience, while RASTA proficiently alleviates channel distortions. The authors advocate for hybrid methodologies that integrate various strategies to improve robustness and precision in real-world voice recognition applications[86].

Bouziane et al. (2021) present a systematic framework for the objective evaluation of feature extraction methodologies in automatic speaker recognition systems, tackling the absence of standardization in the modeling parameters of previous works. The research assesses characteristics such as MFCC, GFCC, and their dynamic versions through speaker modeling methodologies like GMM-UBM, GSV-SVM, and i-vector/CSS. The results demonstrate that MFCC variations routinely surpass GFCC features, with the HTK MFCC variant combined with GSV-SVM attaining the lowest Equal Error Rates (EER). The authors emphasize that the efficacy of feature extraction strategies is intricately linked to the chosen speaker modeling method, with extended test utterances typically improving accuracy. They recommend future study to optimize evaluation methodologies and generate robust features suitable for varied acoustic settings[87].

References

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey,” *Speech Commun*, vol. 56, no. 1, pp. 85–100, Jan. 2014, doi: 10.1016/j.specom.2013.07.008.

[2] S. S. More, D. Ambedkar, P. L. Borde, S. S. Nimbhore, and B. Ambedkar, "A Review on Automatic Speech Recognition System in Indian Regional Languages," 2018. [Online]. Available: https://www.researchgate.net/publication/333973981

[3] H. B. Sailor, M. Venkata Siva Krishna, D. Chhabra, A. T. Patil, M. Kamble, and H. Patil, "DA-IICT/IIITV System for Low Resource Speech Recognition Challenge 2018," in *Interspeech 2018*, ISCA: ISCA, Sep. 2018, pp. 3187–3191. doi: 10.21437/Interspeech.2018-1553.

[4] D. Raval, V. Pathak, M. Patel, and B. Bhatt, "End-to-End Automatic Speech Recognition for Gujarati," NLPAI, patna. [Online]. Available: https://github.com/apoorvnandan/speech-

[5] R. Parikh, H. Joshi, and R. B. Parikh, "Gujarati Speech Recognition-A Review", [Online]. Available: https://epgp.inflibnet.ac.in/view_f.php?category=1491

[6] D. S. Kulkarni, R. R. Deshmukh, P. P. Shrishrimal, and S. D. Waghmare, "HTK Based Speech Recognition Systems for Indian Regional languages: A Review IRJET Journal HTK Based Speech Recognition Systems for Indian Regional languages: A Review," *International Research Journal of Engineering and Technology*, 2016, [Online]. Available: www.irjet.net

[7] Vishwas M. Shetty, Metilda Sagaya Mary N J, and S. Umesh, IMPROVING THE PERFORMANCE OF TRANSFORMER BASED LOWRESOURCE SPEECH RECOGNITION FOR INDIAN LANGUAGES. IEEE, 2020.

[8] H. B. Sailor and T. Hain, "Multilingual Speech Recognition Using Language-Specific Phoneme Recognition as Auxiliary Task for Indian Languages," in *Interspeech 2020*, ISCA: ISCA, Oct. 2020, pp. 4756–4760. doi: 10.21437/Interspeech.2020-2739.

[9] J. H. Tailor, P. Scholar, D. B. Shah, and P. Post Graduate, "Review on Speech Recognition System for Indian Languages," 2015.

[10] Vikramjit Mitra, Horacio Franco, Chris Bartels, Julien van Hout, Martin Graciarena, and Dimitra Vergyri, "SPEECH RECOGNITION IN UNSEEN AND NOISY CHANNEL CONDITIONS," in

2017IEEEInternationalCOnferenceonAcoustics,Speechandsignalprocessing, IEEE, 2017.

[11] A. Singh, V. Kadyan, M. Kumar, and N. Bassan, “ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages,” *Artif Intell Rev*, vol. 53, no. 5, pp. 3673–3704, Jun. 2020, doi: 10.1007/s10462-019-09775-8.

[12] N. Fathima, T. Patel, C. Mahima, and A. Iyengar, “TDNN-based multilingual speech recognition system for low resource Indian languages,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2018, pp. 3197–3201. doi: 10.21437/Interspeech.2018-2117.

[13] A. Messaoudi, H. Haddad, C. Fourati, M. B. H. Hmida, A. Ben Elhaj Mabrouk, and M. Graiet, “Tunisian Dialectal End-to-end Speech Recognition based on DeepSpeech,” in *Procedia CIRP*, Elsevier B.V., 2021, pp. 183–190. doi: 10.1016/j.procs.2021.05.082.

[14] R. G. Kanke, M. A. Ambewadikar, and M. R. Baheti, “REVIEW ON SMALL VOCABULARY AUTOMATIC SPEECH RECOGNITION SYSTEM (ASR) FOR MARATHI,” *openaccessinternationaljournalofscience&engineering*, vol. 3297, no. 2, 2021, doi: 10.51397/OAIJSE02.2021.0001.

[15] D. Adiga, R. Kumar, A. Krishna, P. Jyothi, G. Ramakrishnan, and P. Goyal, “Automatic Speech Recognition in Sanskrit: A New Speech Corpus and Modelling Insights.” [Online]. Available: www.cse.iitb.ac.in/~asr

[16] J. H. Tailor and D. B. Shah, “HMM-Based Lightweight Speech Recognition System for Gujarati Language,” in *Lecture Notes in Networks and Systems*, vol. 10, Springer, 2018, pp. 451–461. doi: 10.1007/978-981-10-3920-1_46.

[17] H. Sailor and H. Patil, “Neural Networks-based Automatic Speech Recognition for Agricultural Commodity in Gujarati Language,” International Speech Communication Association, Oct. 2018, pp. 162–166. doi: 10.21437/sltu.2018-34.

[18] P. Mehra and P. Jain, “ERIL: An Algorithm for Emotion Recognition from Indian Languages Using Machine Learning,” *Wirel Pers Commun*, vol. 126, no. 3, pp. 2557–2577, 2022, doi: 10.1007/s11277-022-09829-1.

[19] P. Mehra and S. K. Verma, “BERIS: An mBERT-based Emotion Recognition Algorithm from Indian Speech,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 5, Apr. 2022, doi: 10.1145/3517195.

[20] R. V. Darekar and A. P. Dhande, Enhancing effectiveness of emotion detection by multimodal fusion of speech parameters. IEEE, 2016.

[21] N. Vryzas, L. Vrysis, R. Kotsakis, and C. Dimoulas, “A web crowdsourcing framework for transfer learning and personalized Speech Emotion Recognition,” *Machine Learning with Applications*, vol. 6, p. 100132, Dec. 2021, doi: 10.1016/j.mlwa.2021.100132.

[22] W. Zheng, W. Zheng, and Y. Zong, “Multi-scale discrepancy adversarial network for crosscorpus speech emotion recognition,” *Virtual Reality and Intelligent Hardware*, vol. 3, no. 1, pp. 65–75, Feb. 2021, doi: 10.1016/j.vrih.2020.11.006.

[23] R. Mohd Hanifa, K. Isa, and S. Mohamad, “A review on speaker recognition: Technology and challenges,” *Computers and Electrical Engineering*, vol. 90, Mar. 2021, doi: 10.1016/j.compeleceng.2021.107005.

[24] M. N. Kakade and D. B. Salunke, “An Automatic Real Time Speech-Speaker Recognition System: A Real Time Approach,” in *Lecture Notes in Electrical Engineering*, Springer Verlag, 2020, pp. 151–158. doi: 10.1007/978-981-13-8715-9_19.

[25] F. S. Central University of Technology, Institute of Electrical and Electronics Engineers, South African Institute of Electrical Engineers, Robotics Association of South Africa, S. A. Pattern Recognition Association of South Africa. Symposium (29th : 2019 : Bloemfontein, and S. A. Robotics and Mechatronics Conference (11th : 2019 : Bloemfontein,
Automatic Speaker Recognition System based on Machine Learning Algorithms.

[26] F. He *et al.*, “Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems,” 2020. [Online]. Available: http://www.openslr.org/78/

[27] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, “GMM and CNN Hybrid Method for Short Utterance Speaker Recognition,” *IEEE Trans Industr Inform*, vol. 14, no. 7, pp. 3244–3252, Jul. 2018, doi: 10.1109/TII.2018.2799928.

[28] V. Tiwari, M. F. Hashmi, A. Keskar, and N. C. Shivaprakash, “Speaker identification using multi-modal i-vector approach for varying length speech in voice interactive systems,” *Cogn Syst Res*, vol. 57, pp. 66–77, Oct. 2019, doi: 10.1016/j.cogsys.2018.09.028.

[29] A. Maurya, D. Kumar, and R. K. Agarwal, “Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach,” in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 880–887. doi: 10.1016/j.procs.2017.12.112.

[30] M. K. Nammous, K. Saeed, and P. Kobojek, “Using a small amount of text-independent speech data for a BiLSTM large-scale speaker identification approach,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 3, pp. 764–770, Mar. 2022, doi: 10.1016/j.jksuci.2020.03.011.

[31] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, “A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities,” *IEEE Access*, vol. 9, pp. 79236–79263, 2021, doi: 10.1109/ACCESS.2021.3084299.

[32] M. Himanshu, N. Patel, and P. V Virparia, “A Small Vocabulary Speech Recognition for Gujarati,” *International Journal of Advanced Research in Computer Science*, vol. 2, no. 1, [Online]. Available: www.ijarcs.info

[33] M. S. Saeed *et al.*, “Cross-modal Speaker Verification and Recognition: A Multilingual Perspective,” Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.13780

[34] R. M. Ghoniem and K. Shaalan, “A Novel Arabic Text-independent Speaker Verification System based on Fuzzy Hidden Markov Model,” in *Procedia Computer Science*, Elsevier B.V., 2017, pp. 274–286. doi: 10.1016/j.procs.2017.10.119.

[35] H. A. Patil and T. K. Basu, “Development of speech corpora for speaker recognition research and evaluation in Indian languages,” *Int J Speech Technol*, vol. 11, no. 1, pp. 17–32, Mar. 2008, doi: 10.1007/s10772-009-9029-5.

[36] S. Saleem, F. Subhan, N. Naseer, A. Bais, and A. Imtiaz, “Forensic speaker recognition: A new method based on extracting accent and language information from short utterances,” *Forensic Science International: Digital Investigation*, vol. 34, Sep. 2020, doi: 10.1016/j.fsidi.2020.300982.

[37] R. Ranjan, S. K. Singh, R. Kala, and R. Kumar, “Multilingual Speaker Recognition Using Neural Network Static hand gesture recognition using Deep Learning View project Expert System for Speaker Identification Using Lip Features with PCA View project MULTILINGUAL SPEAKER RECOGNITION USING NEURAL NETWORK,” 2009. [Online]. Available: https://www.researchgate.net/publication/272086352

[38] Sourjya Sarkar, K. Sreenivasa Rao, Dipanjan Nandi, and Sunil Kumar S. B., *Multilingual Speaker Recognition on Indian Languages*. IEEE, 2013.

[39] R. Chojnacka, J. Pelecanos, Q. Wang, and I. L. Moreno, “SpeakerStew: Scaling to Many Languages with a Triaged Multilingual Text-Dependent and Text-Independent Speaker Verification System,” Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.02125

[40] X. Wang, F. Xue, W. Wang, and A. Liu, “A network model of speaker identification with new feature extraction methods and asymmetric BLSTM,” *Neurocomputing*, vol. 403, pp. 167–181, Aug. 2020, doi: 10.1016/j.neucom.2020.04.041.

[41] P. K. Nayana, D. Mathew, and A. Thomas, “Comparison of Text Independent Speaker Identification Systems using GMM and i-Vector Methods,” in *Procedia Computer Science*, Elsevier B.V., 2017, pp. 47–54. doi: 10.1016/j.procs.2017.09.075.

[42] J. Xu, X. Wang, B. Feng, and W. Liu, “Deep multi-metric learning for text-independent speaker verification,” *Neurocomputing*, vol. 410, pp. 394–400, Oct. 2020, doi: 10.1016/j.neucom.2020.06.045.

[43] S. Shahnawazuddin, N. Adiga, B. T. Sai, W. Ahmad, and H. K. Kathania, “Developing speaker independent ASR system using limited data through prosody modification based on fuzzy classification of spectral bins,” *Digital Signal Processing: A Review Journal*, vol. 93, pp. 34–42, Oct. 2019, doi: 10.1016/j.dsp.2019.06.015.

[44] S. Kinkiri, B. Bakarat, and S. Keates, “Sensors & Transducers Identification of a Speaker from Characteristics of a Voice,” 2020. [Online]. Available: http://www.sensorsportal.com

[45] T. Bian, F. Chen, and L. Xu, “Self-attention based speaker recognition using Cluster-Range Loss,” *Neurocomputing*, vol. 368, pp. 59–68, Nov. 2019, doi: 10.1016/j.neucom.2019.08.046.

[46] L. Chen, Y. Liu, W. Xiao, Y. Wang, and H. Xie, “SpeakerGAN: Speaker identification with conditional generative adversarial network,” *Neurocomputing*, vol. 418, pp. 211–220, Dec. 2020, doi: 10.1016/j.neucom.2020.08.040.

[47] J. Patel and A. Nandurbarkar, “Development and Implementation of Algorithm for Speaker recognition for Gujarati Language,” *International Research Journal of Engineering and Technology*, 2015, [Online]. Available: www.irjet.net

[48] M. M. Shah and H. Kavathiya, “A Systematic survey on Voice Recognition for Gujarati Dialects.”

[49] P. Ahuja and J. M. Vyas, “Forensic speaker profiling: the study of supra-segmental features of Gujarati dialects for text–independent speaker identification,” *Australian Journal of Forensic Sciences*, vol. 50, no. 2, pp. 152–165, Mar. 2018, doi: 10.1080/00450618.2016.1237547.

[50] M. Gupta, R. K. Singh, and S. Singh, “G-Cocktail: An Algorithm to Address Cocktail Party Problem of Gujarati Language using CatBoost,” Mar. 17, 2021. doi: 10.21203/rs.3.rs-305722/v1.

[51] E. Ambikairajah, L. Wang, B. Yin, and V. Sethu, “Language Identification: A Tutorial.”

[52] M. A. Anusuya and S. K. Katti, “Speech Recognition by Machine: A Review,” 2009. [Online]. Available: http://sites.google.com/site/ijcsis/

[53] Sakshi A. Patil, Gaurav A. Varade, and Vikram Hankare, “Tracing Gujarati Dialects Philogically and Sociolinguistically,” *International Journal of Modern Developments in Engineering and Science*, vol. 2, no. 5, May 2023, [Online]. Available: https://www.ijmdes.com

[54] U. Desai and M. Ramsay-Brijball, “Tracing Gujarati Language Development Philologically and Sociolinguistic ally.”

[55] R. Mesthrie, “Uncovering a consonant chain shift in Gujarati,” Dec. 25, 2023, *Department of General Linguistics, Stellenbosch University*. doi: 10.5842/67-1-1010.

[56] C. kurian, “A Review on Speech Corpus Development for Automatic Speech Recognition in Indian Languages,” 2014.

[57] R. Ardila *et al.*, “Common Voice: A Massively-Multilingual Speech Corpus,” Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.06670

[58] V. P. Tank and S. K. Hadia, “Creation of speech corpus for emotion analysis in Gujarati language and its evaluation by various speech parameters,” *International Journal of Electrical and Computer Engineering*, vol. 10, no. 5, pp. 4752–4758, Oct. 2020, doi: 10.11591/ijece.v10i5.pp4752-4758.

[59] D. Sztahó, G. Szaszák, and A. Beke, “Deep learning methods in speaker recognition: a review.”

[60] Sudhamay Maity, Anil Kumar Vuppala, K. Sreenivasa Rao, and Dipanjan Nandi, *IITKGP-MLILSC Speech Database for Language Identification*. IEEE, 2012.

[61] P. P. Shrishrimal, R. R. Deshmukh, and V. B. Waghmare, “Indian Language Speech Database: A Review,” 2012.

[62] J. Basu, S. Khan, R. Roy, T. K. Basu, and S. Majumder, “Multilingual Speech Corpus in Low-Resource Eastern and Northeastern Indian Languages for Speaker and Language Identification,” *Circuits Syst Signal Process*, vol. 40, no. 10, pp. 4986–5013, Oct. 2021, doi: 10.1007/s00034-021-01704-x.

[63] F. He *et al.*, “Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems,” 2020. [Online]. Available: http://www.openslr.org/78/

[64] J. Kothari Associate Professor, S. M. P Shah, C. Kumbharana, and A. Professor, “A Phonetic Study for Constructing a Database of Gujarati Characters for Speech Synthesis of Gujarati Text,” 2015.

[65] Shaposhnikov, S..Nikolai V. Slavnov, Yurii V. Stroganov, and Alexander V. Kvasnikov, *System for Speech Corpus Development*. St. Petersburg Electrotechnical University “LETI,” 2020.

[66] K. Prahallad, E. Naresh Kumar, V. Keri, S. Rajendran, and A. W. Black, “The IIIT-H Indic Speech Databases.” [Online]. Available: http://www.isca-speech.org/archive

[67] S. Hourri and J. Kharroubi, “A deep learning approach for speaker recognition,” *Int J Speech Technol*, vol. 23, no. 1, pp. 123–131, Mar. 2020, doi: 10.1007/s10772-019-09665-y.

[68] V. S. Reddy Gade and M. Sumathi, “A Comprehensive Study on Automatic Speaker Recognition by using Deep Learning Techniques,” in *Proceedings of the 5th International Conference on Trends in Electronics and Informatics, ICOEI 2021*, Institute of Electrical and Electronics Engineers Inc., Jun. 2021, pp. 1591–1597. doi: 10.1109/ICOEI51242.2021.9452885.

[69] Y. Zhang, F. Chen, and X. Wang, “A Review of Robust Deep Learning-Based Speaker Recognition,” in *2023 5th International Conference on Artificial Intelligence and Computer Applications, ICAICA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 346–351. doi: 10.1109/ICAICA58456.2023.10405619.

[70] S. Bansal, R. K. Bansal, Y. Sharma, and G. Zail Singh, “ANN based efficient feature fusion technique for speaker recognition.”

[71] P. Li, G. Li, J. Han, T. Zhi, and D. Wang, “Channel Mismatch Speaker Verification Based on Deep Learning and PLDA,” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, 2020. doi: 10.1088/1742-6596/1682/1/012056.

[72] Mohamed Hamidi, Hassan Satori, Naouar Laaidi, and Khalid Satori, *Conception of Speaker Recognition Methods: A Review*. IEEE, 2020.

[73] N. N. Prachi, F. M. Nahiyan, M. Habibullah, and R. Khan, “Deep Learning Based Speaker Recognition System with CNN and LSTM Techniques,” in *2022 International Conference on Interdisciplinary Research in Technology and Management, IRTM 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/IRTM54583.2022.9791766.

[74] A. Q. Ohi, M. F. Mridha, M. A. Hamid, and M. M. Monowar, “Deep Speaker Recognition: Process, Progress, and Challenges,” 2021, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2021.3090109.

[75] H. R. Hu, Y. Song, Y. Liu, L. R. Dai, I. McLoughlin, and L. Liu, “DOMAIN ROBUST DEEP EMBEDDING LEARNING FOR SPEAKER RECOGNITION,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 7182–7186. doi: 10.1109/ICASSP43922.2022.9747364.

[76] T. J. Sefara and T. B. Mokgonyane, “Emotional Speaker Recognition based on Machine and Deep Learning,” in *2020 2nd International Multidisciplinary Information Technology and Engineering Conference, IMITEC 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/IMITEC50163.2020.9334138.

[77] G. Costantini, V. Cesarini, and E. Brenna, “High-Level CNN and Machine Learning Methods for Speaker Recognition,” *Sensors*, vol. 23, no. 7, Apr. 2023, doi: 10.3390/s23073461.

[78] V. S. Reddy Gade and M. Sumathi, “Hybrid Deep Convolutional Neural Network based Speaker Recognition for Noisy Speech Environments,” in *Proceedings of the 2nd International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 920–926. doi: 10.1109/ICAAIC56838.2023.10141080.

[79] P.-M. Bousquet and M. Rouvier, “Improving training datasets for resource-constrained speaker recognition neural networks.” [Online]. Available: https://hal.science/hal-04156025

[80] I. Gusti, B. Arya, P. Paramitha, H. B. Kusnawan, and M. Ernawati, “Performance Comparison of Deep Learning Algorithm for Speech Emotion Recognition.” [Online]. Available: http://jcosine.if.unram.ac.id/

[81] N. Choudhary and L. Ramamoorthy, “20 LDC-IL RAW SPEECH CORPORA: AN OVERVIEW.” [Online]. Available: https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf

[82] Z. Bai and X. L. Zhang, “Speaker recognition based on deep learning: An overview,” *Neural Networks*, vol. 140, pp. 65–99, Aug. 2021, doi: 10.1016/j.neunet.2021.03.004.

[83] N. Shome, A. Sarkar, A. K. Ghosh, R. H. Laskar, and R. Kashyap, “Speaker Recognition through Deep Learning Techniques: A Comprehensive Review and Research Challenges,” 2023, *Budapest University of Technology and Economics*. doi: 10.3311/PPee.20971.

[84] IEEE Circuits and Systems Society., *SPEAKER VERIFICATION/RECOGNITION AND THE IMPORTANCE OF SELECTIVE FEATURE EXTRACTION: REVIEW*. Institute of Electrical and Electronics Engineers, 2001.

[85] B. Aarti and S. K. Kopparapu, “Spoken Indian language identification: a review of features and databases,” *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 43, no. 4, Apr. 2018, doi: 10.1007/s12046-018-0841-y.

[86] U. Shrawankar, “TECHNIQUES FOR FEATURE EXTRACTION IN SPEECH RECOGNITION SYSTEM : A COMPARATIVE STUDY.”

[87] A. Bouziane, J. Kharroubi, and A. Zarghili, “Towards an objective comparison of feature extraction techniques for automatic speaker recognition systems,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 374–382, Feb. 2020, doi: 10.11591/eei.v10i1.1782.