I.J. Information Technology and Computer Science, **2024, 3, 36-51**
Published Online on June 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2024.03.04

A PRISMA-driven Review of Speech Recognition based on English, Mandarin Chinese, Hindi and Urdu Language

Muhammad Hazique Khatri*
Department of Computer Science, University of Karachi, Karachi, Pakistan
E-mail: haziqiqbal2101@gmail.com
ORCID iD: https://orcid.org/0000-0001-8683-8207
*Corresponding Author

Humera Tariq
Department of Computer Science, University of Karachi, Karachi, Pakistan
Email: humera@uok.edu.pk
ORCID iD: https://orcid.org/0000-0002-7684-9161

Maryam Feroze
Department of Computer Science, University of Karachi, Karachi, Pakistan
Email: maryam.feroze@uok.edu.pk
ORCID iD: https://orcid.org/0009-0004-9485-8469

Ebad Ali
ML-Labs, Dublin, Ireland
Email: ebadalie@gmail.com
ORCID iD: https://orcid.org/0000-0003-4813-5003

Zeeshan Anjum Junaidi
ML-Labs, Dublin, Ireland
Email: zeeshananjumjunaidi@gmail.com
ORCID iD: https://orcid.org/0000-0003-2207-1486

Received: 05 November 2023; Revised: 31 January 2024; Accepted: 08 March 2024; Published: 08 June 2024

Abstract: Urdu Language ranks ten and is continuously progressing. This unique PRISMA-Driven review deeply investigates Urdu speech recognition literature and adjoin it with English, Mandarin Chinese, and Hindi languages frame-works conceptualizing wider global perspective. The main objective is to unify progress on classical Artificially Intelligent (AI) and recent Deep Neural Networks (DNN) based speech recognition pipeline encompassing Dataset challenges, Feature extraction methods, Experimental design and the smooth integration with both Acoustic models (AM) and Language models (LM) using Transcriptions. A total of 176 articles were extracted from Google Scholar database for each language with custom query design. Inclusion criteria and quality assessment leads to end up with 5 review and 42 research articles. Comparative research questions have been addressed and findings were organized by four possible speech types: Isolated, connected, continuous and spontaneous. The finding shows that English, Mandarin, and Hindi languages used spontaneous speech size of 300, 200 and 1108 hours respectively which is quite remarkable as compared to Urdu spontaneous speech data size of only 9.5 hours. For the same data size reason, the Word Error Rate (WER) for English falls below 5% while for Mandarin Chinese the alternative metric Character Error Rate (CER) is mostly used that lies below 25%. The success of English and Chinese Speech recognition leads to incomparable accuracy due to wide use of DNNs like Conformer, Transformers, E2E-attention in comparison to conventional feature extraction and AI models LSTM, TDNN, RNN, HMM, GMM-HMM; used frequently by both Hindi and Urdu.

Index Terms: PRISMA, Speech-to-Text (STT), ASR, Transformer, Conformer, LSTM, Speech Recognition, HMM, Language Models.

1. Introduction

Speech-to-Text (STT) systems provide seamless communication between humans and machines. High-tech organizations have developed state-of-the-art (SOTA) technology based on Automatic Speech Recognition (ASR) systems [1-3], supporting many languages e.g., English, German, French, Spanish, Italian, and Japanese. Today's ASR technology is thus a commonly deployed Artificial Intelligent (AI) tools in many e-governments, e-business, and e-healthcare platforms [4-7]. Table 1 shows four possible speech types to begin with STT systems and are named as: Isolated speech, Connected word speech, Continuous Speech, and Spontaneous Speech. In isolated speech, speakers must stop between each spoken word. Connected words on the other hand require not to be silent between two or more words during speech recording. For Continuous and Spontaneous, speakers are allowed to speak naturally but spontaneous includes natural emotions as well e.g., singing, coughing, and laughing. Datasets for Speech recognition is widely available in recorded form and vary from each other based on the number of speakers, vocabulary size, sampling rate, and storage format of the recorded session e.g., mp3, wav, and flac.

Based on vocabulary size, any speech dataset is considered as small, medium, large, or very large. Small vocabulary consists of less than 50 words i.e. 10 or 20 words [8,9]. Medium vocabulary contains words in hundreds e.g., 115, 139, 250 [10-12]. A large vocabulary contains a few thousand words, such as 4000 or 5000 words [13,14] while a very large vocabulary has more than a thousand words i.e., 900,000 or 1,300,000 words [15,16]. According to Ethnologue [17], English, Mandarin Chinese, and Hindi are the world's topmost spoken languages with 1,452, 1,118, and 602 million speakers as shown in Fig. 1. Furthermore, it is notable that Urdu has a considerable number of speakers, making it the 10th most widely spoken language in the world. Despite its substantial speakers, limited research has been conducted on Urdu ASR systems, which are not considered sufficiently reliable or comprehensive for the ASR world. Hence, there is a need to develop an efficient ASR system for the Urdu language to position Urdu competitively in the realm of ASR technology. In this review, we include all four speech types of English, Mandarin Chinese, Hindi, and Urdu to investigate various aspects of ASR systems developed in these languages including datasets, feature extraction, experimental design, acoustic models and language models.

Table 1. Type of speech for english, chinese, hindi and urdu language

Type of Speech	Languages			
	English	**Chinese**	**Hindi**	**Urdu**
Isolated Word (simple)	Travel	旅行	यात्रा	سفر
Connected Word (simple)	Whatever	任何	जो भी	جو بھی
Continuous (challenging)	We have plenty of time for that	我们有足够的时间	हमारे पास इसके लिए काफी समय है	اس کے لیے ہمارے پاس بہت وقت ہے
Spontaneous (challenging)	<laugh>I laughed a lot at that	<笑>我对此笑了很多	<हंसना>मुझे उस पर बहुत हंसी आई	<ہنسنا>مجھے اس پر بہت ہنسی آئی

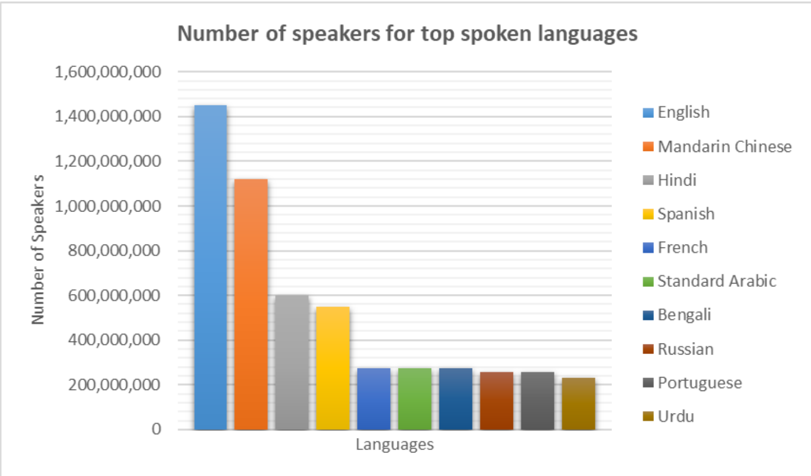

Fig.1. Top 10 languages of the world with total speakers

In this review, we include all four speech types of English, Mandarin Chinese, Hindi, and Urdu, and investigate various aspects of ASR systems developed in these languages including datasets, feature extraction, experimental design, acoustic and language models. Through a comparative analysis of ASR methodologies and findings, we aim to provide insights into the unique challenges and opportunities posed by each linguistic context, focusing on the development of Urdu ASR. We have taken inspiration from existing ASR studies [18,19] and applied the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) to this ASR review [20]. PRISMA is an evidence-based minimum set of items that helps researchers write high-quality Systematic Literature Reviews (SLRs) and meta-analyses. Most PRISMA literature review revolves around medical studies [21-23] but has recently become popular for NLP studies, e.g., speech recognition [19,18], Hate Speech Detection [24], Text-based Depression Detection [25], Phishing Email Detection [26], Alzheimer's disease detection from speech [27] and several others.

2. Speech Recognition Pipeline

ASR enables machines to recognize human speech and convert it into a sequence of text. The general Deep Neural Network (DNN) based high-level speech recognition pipeline comprised of a large speech database, pre-processing, feature extraction, experimental protocol, acoustic model, enormous transcription, and a language model. Fig. 2 shows common pre-processing techniques namely Noise Removal, Pre-emphasis, and Voice Activity Detection (VAD) [28,29]. Noise removal aims to remove unwanted background noise from speech signals [30]. Pre-emphasis improves the quality of the speech signals by boosting their high-frequency units [14,30,31]. Voice Activity Detection (VAD) detects speech segments and thus removes silence from the speech signals [11]. Next to pre-processing is feature extraction which aims to convert audio signals into pattern vectors. The Mel Frequency Cepstral Coefficient (MFCC) is the most common feature extraction technique as it mimics the human hearing system [8,32-38]. Deep Neural Network (DNN) experiment protocol requires splitting of the dataset into training and testing using cross-validation techniques like Hold-out [35,38-40] and K-folds [31].

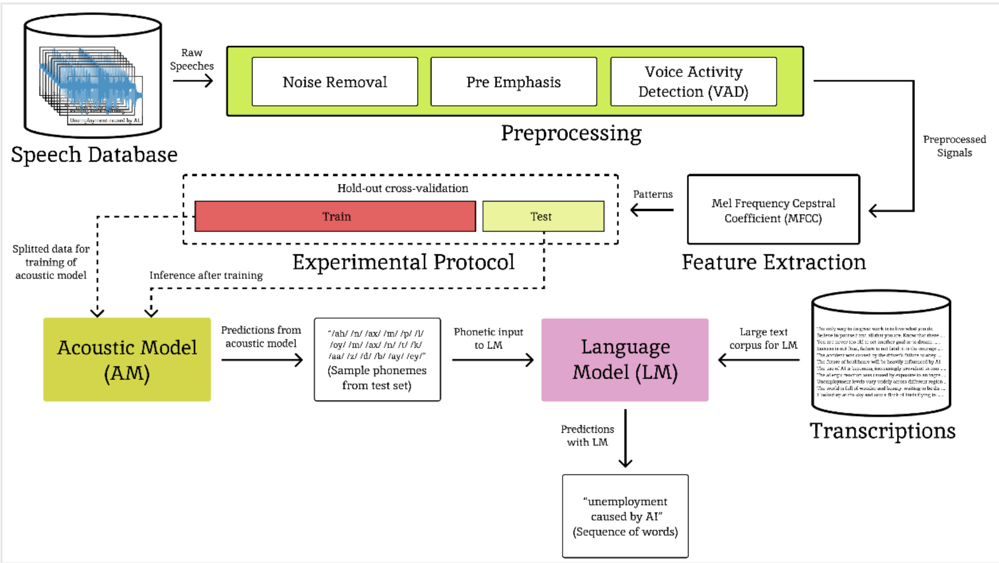

Fig.2. DNN based high level speech recognition pipeline

After data splitting, the training set is passed to the acoustic model, which generates a phonetic stream. The most common generative approaches for acoustic models are Gaussian Mixture with Hidden Markov Model (GMM-HMM) [8,34,36,37], Recurrent Neural Network (RNN) [33,38], Long Short-Term Memory (LSTM) [37,41,42], Convolution Neural Network (CNN) [9,33,36,43], Time Delay Neural Networks (TDNN) [34,37,38]. The most recent technique is transformer [44-46]. The acoustic model itself is insufficient to account for standard grammar rules, variations in speaking styles, and dialects. Therefore, a language model is trained using a large text corpus i.e., books or articles, which transforms the sequence of phonetic units from an acoustic model into the most likely word sequence based on the contextual information and probability of occurrence of different words in the transcription. For language modeling, the traditional n-gram model ('n' is the order of n words) has been frequently used [34,45,47,48]. However, the best outcomes have been achieved through neural networks such as Long Short-Term Memory (LSTM) [49], Recurrent Neural Networks (RNN) [50-52] and Hybrid Models [44,53]. The two most used evaluation metrics that measure the performance of an acoustic system are accuracy [10,30], word error rate (WER) [35,37,49,54], and character error rate (CER) [46,51,55]. For the evaluation of the language model, perplexity (PPL) [32,56] is used. It measures the word-level error from the predicted text. The formula of WER is available in Equation (1).

$$WER = \frac{S+D+I}{N} * 100 \quad (1)$$

Where S = number of substitutions in the predicted text, D = number of deletions in the output text, I = number of insertions, and N = actual words in the transcript.

To quantify errors at the character level, the character error rate (CER) is employed. The mathematical formula for

CER is identical to that of WER, but CER operates specifically on the character level.

To calculate accuracy, Equation (2) is used.

$$Accuracy = 100 - WER \tag{2}$$

$$PPL = \frac{1}{\prod_{i=1}^{n} (P(W_i \mid W_1, W_{2,\ldots,} W_{i-1}))^{\frac{1}{n}}} \tag{3}$$

The equation to calculate perplexity is available in Equation (3). Where W_i = i^{th} word in the sequence, n = total number of words in the sequence, and $P(W_i|W_1, W_2, \ldots, W_{i-1})$ = Conditional probability of the i^{th} word, given the previous (i - 1) words present in the sequence.

Once both the acoustic and language models have been trained, the test set from the hold-out cross-validation is utilized for inference. Let's pick an example from the test set. The acoustic model generates the phonetic sequence for our example as "/ah/ /n/ /ax/ /m/ /p/ /l/ /oy/ /m/ /ax/ /n/ /t/ /k/ /aa/ /z/ /d/ /b/ /ay/ /ey/", which is then processed by the language model to produce the meaningful sequence of words "unemployment caused by AI" as illustrated in DNN based high-level Speech Recognition Pipeline Fig. 2.

3. Artificial Intelligence (AI) Based Development for Urdu ASR

Languages evolve mainly from political, social, cultural, technological, and moral influences [57]. Urdu language development started back in the 12th century in North India, near Delhi [58]. Its vocabulary set is influenced by Persian, Arabic, Turkish, Hindi, and Punjabi. The Urdu language is famous for being adopted as the language of poets [59] and was trendy among poetry writers of the 17th to 19th century, including Khawaja Mir Dard, Mir Taqi Mir, Mirza Ghalib, and Dr. Allama Muhammad Iqbal. Urdu is a free word-order language [60]. It is very close to Hindi as they share phonological, morphological, and syntactic structures, but an Urdu NLP system can't be directly used for Hindi and vice versa because of script differences, vocabulary size, and missing diacritics problems [61].

Speech recognition for the Urdu language started in early 2000. In 2004, speech recognition using acoustic-phonetic modeling was developed for the Urdu language [62]. Another Recognition System for Urdu speech was proposed using Neural Networks in 2008 [63]. The first Large Vocabulary Continuous Speech Recognition (LVCSR) for Urdu was developed in 2010 [64]. Another system for speech recognition was developed for Urdu using the speaker-independent dataset [65]. The traditional approaches used to build Urdu classification models are the Hidden Markov Model (HMM), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), and Random Forest (RF) [30,11,31,66,67]. Though, the best results were achieved using Deep Learning techniques with a minimum Word Error Rate (WER) of 13.50%, 17%, 18.59%, 14%, and 4% available in [9,37,38,42,68] respectively.

Table 2 lists various research institutes along with their contributions and pipeline aspects, that are working to improve Urdu based ASR systems. The Centre of Language Engineering has made significant efforts for Urdu ASR. They collected isolated speech dataset for district names of Pakistan [11]. They also achieved the lowest WER for Urdu continuous speech recognition and designed a Large Vocabulary Continuous Speech Recognition (LVCSR) system using an extensive multi-genre Urdu broadcast speech corpus [37,38]. City, University of London examined the performance of two feature extraction techniques i.e., Discrete Wavelet Transform (DWT) and Mel Frequency Cepstral Coefficient (MFCC) for Urdu ASR [69]. Commission on Science and Technology for Sustainable Development in the South (COMSATS) University studied various architectures of Deep Neural Networks (DNN) namely Gated Rectified Units (GRUs) and Long Short-Term Memory (LSTMs) for Urdu ASR and achieved improvements over baseline [42]. Dr. Babasaheb Ambedkar Marathwada (B.A.M) university recorded Urdu digits and applied traditional statistical technique i.e. Hidden Markov Model (HMM) on them [67]. National University of Computer and Emerging Science (FAST) improved and compared statistical approaches for Urdu continuous speech recognition [14]. Information Technology University (ITU) compiled a large spontaneous speech dataset through telephonic community forums [47]. The National University of Science and Technology (NUST) contributed to in speaker independent Urdu ASR system with small size vocabulary and investigated two K-fold protocols using an isolated medium-sized speech corpus [31,65]. Shanghai Jiao Tong University developed a large Urdu digit dataset and SOTA digit recognition system [9]. United Arab Emirates (UAE) University was the first who implement a simple neural network based on 3 layers for Urdu digit recognition [63]. University of Engineering and Technology (UET) proposed the first Urdu end-to-end ASR system using Semi-Supervised Learning (SSL) for benchmarking [68]. University of Science and Technology Beijing developed an Urdu digit recognition system and provided a baseline using medium size vocabulary for Urdu ASR [30,66].

Table 2. AI-based development for Urdu speech recognition system

Institute	Work	Contribution(s)	ASR pipeline aspects
Center of Language Engineering, Pakistan	[11]	Build dataset for district names of Pakistan.	MFCC, HMM, Hold-out, K-folds, Accuracy, WER
	[37]	Develop Continuous ASR with lowest WER.	Text Norm, MFCC, Hold-out, TDNN-BLSTM, RNN LM, WER
	[38]	Designed LVCSR using extensive multi-genre Urdu broadcast dataset.	MFCC, i-vector, Hold-out, TDNN, LDA, MLLT, SAT, RNN LM, WER
City, University of London, UK	[69]	Examined the performance of feature extraction techniques for Urdu ASR.	Segmentation, Noise Removal, Pre-Emphasis, MFCC, Hold-out, LDA, Confusion-Matrix
COMSATS University, Pakistan	[42]	Examined DNN for Urdu ASR and achieved improvements.	Hold-out, MFCC, BLSTM, WER
Dr. B.A.M University, India	[67]	Recorded Urdu digits with the implementation of statistical techniques.	LPC, Hold-out, HMM, WER
FAST University, Pakistan	[14]	Improved and compared statistical methods for Urdu continuous ASR.	MFCC, Hold-out, SGMM, n-gram LM, WER
ITU, Pakistan	[47]	Compiled spontaneous dataset through the community forum.	SGMM, fMLLR, MMI, n-gram LM, WER, PPL
NUST University, Pakistan	[65]	Build speaker-independent Urdu ASR using small size vocabs.	Sphinx4, GMM/HMM, wordlist grammar, WER
	[31]	Investigated two experiments on K-fold protocols.	VAD, MFCC, K-folds, HMM, Accuracy
SJT University, China	[9]	Developed a large Urdu digit dataset and SOTA digit system.	Mel-spectrogram, CNN, Accuracy
UAE University, UAE	[63]	Designed the first Urdu digit recognition system.	MATLAB, NN, Accuracy
UET, Pakistan	[68]	Proposed first Urdu end-to-end ASR using SSL for benchmarking.	FBanks, MFCC, LLE, Hold-out, E2E, Maxout, WER
USTB, China	[30]	Trained an Urdu digit recognition system.	MFCC, Hold-out, SVM, Confusion-Matrix
	[66]	Provide a baseline for Urdu ASR.	MFCC, Hold-out, LDA, Confusion-Matrix

4. PRISMA Protocol

The four significant steps of PRISMA are: (1) Identification; In this step, articles are collected from various databases using search queries and duplications are removed. (2) Screening; The irrelevant or out-of-scope topics are removed by reading each article's title, abstract, and keywords. (3) Eligibility; The inclusion/exclusion criteria are prepared according to the scope of the literature and applied to each article, followed by a quality assessment. (4) Inclusion; the list of all collected papers is obtained in this final step. Gathering articles for this systematic study posed a challenge because on the one hand we were seeking to acquire Urdu ASR research work while on the other hand, we need to acquire ASR review papers on top languages, especially English, Mandarin Chinese, and Hindi to portray an enlighten ASR literature review.

We thus decided to study the ASR pipeline based on three main language categories criteria: top ten languages, top three languages, and finally Urdu language. The most general query words used to search and acquire literature are: “Speech recognition”, “Speech recognizer”, “ASR”, “Speech transformation”, “Speech to text”, “voice to text”, “Voice recognition”, "Deep Learning" and "Neural Network". To target the top 10 and top 3 language category articles, we set a time window from the year 2018 to 2022 but for Urdu, we start from 2015 and add additional keywords such as “Literature Review”, “Systematic Literature Review”, “SLR”, “Review” and “Study”. Further, we added keywords like "English", "Mandarin Chinese", "Hindi" and “Urdu”.

Fig. 3 demonstrates the PRISMA approach to finalize with 47 articles though initially, we started with total 176 Google Scholar database [70] articles. Among 176 total articles, 32 review papers talk about general ASR literature while the rest 144 were chosen to focus on a particular language. This way the 144 articles distribution for English, Mandarin Chinese, Hindi, and Urdu are 67, 41, 19, and 17 respectively. After careful screening of abstracts, keywords, and eligibility as shown in Tables 3 and 4, we finally select 47 open access papers among which 5 are review papers and the language-based article distribution after screening becomes 13, 8, 9, 12 respectively. Table 3 and Table 4 further explain the details of inclusion criteria for final selection of research articles. The finalized 5 review papers based upon discussion of the top 10 languages have significant citations, well organized, and have quality research questions. The 30 papers for the top three languages i.e., English, Mandarin Chinese, and Hindi are related to the dataset including benchmarking, state-of-the-art (SOTA) techniques, and standard research patterns. Finally, the 12 research papers for Urdu followed almost all the pipeline aspects and met our inclusion criteria.

5. Findings and Discussion

Findings in this section are under the ASR pipeline illustrated in Fig. 2. The latest ASR building blocks for any spoken language are based on datasets, pre-processing, feature extraction techniques, deep neural network-based (DNN) experimental design, acoustic model, and language models. The eight main raw attributes in each speech dataset include speech file format, sampling rate, speech duration, no. of speakers, vocabulary size, type of speech, channel, and gender.

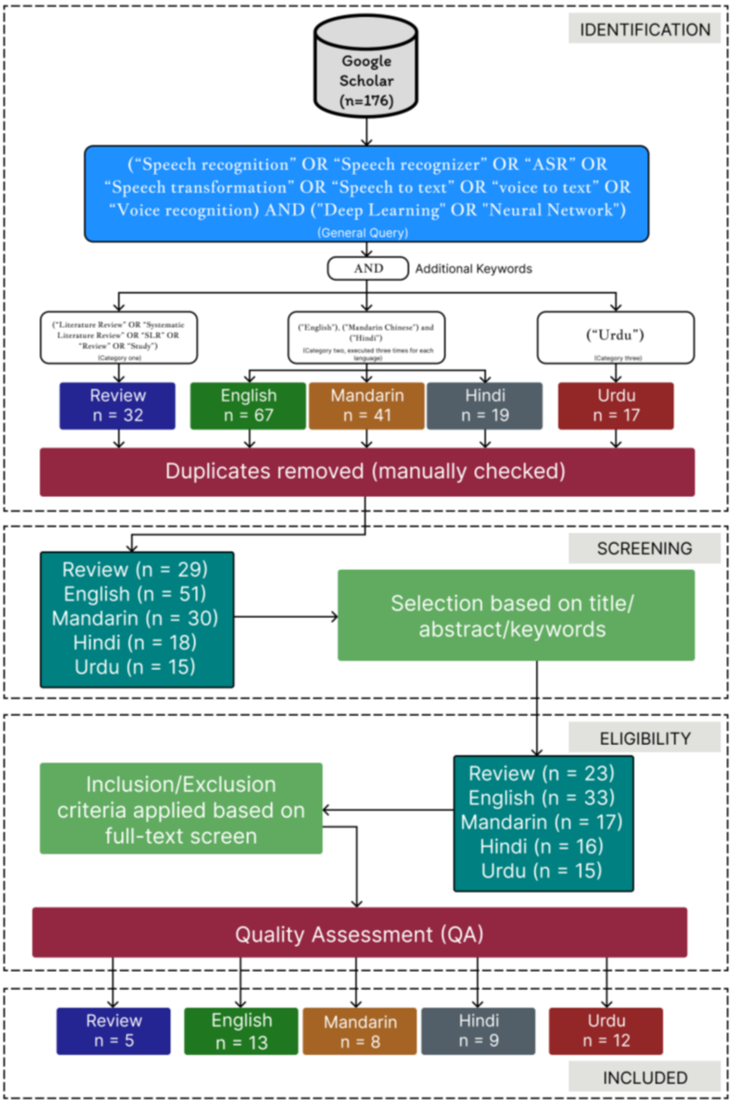

Fig.3. PRISMA protocol to study articles on ASR pipeline

Table 3. Inclusion and exclusion criteria for PRISMA

Papers	Inclusion	Exclusion
Review	Papers contain quality research questions and their answer.	Papers other than top languages.
	Papers with a core focus on ASR.	Papers other than open access.
	Review papers from 2018 to 2022.	Papers other than conferences and journals.
	Review papers on one or more languages for ASR.	Papers not written in the English language.
English, Mandarin Chinese, and Hindi	Papers with benchmark datasets.	Papers other than English, Mandarin Chinese and Hindi
	Research papers from 2018 to 2022.	Papers other than open access.
	Papers with state-of-the-art AI techniques.	Papers other than journal, conference, and research articles.
		Papers not written in the English language.
Urdu	Research papers from 2015 to 2022.	Papers other than open access.
		Papers other than journal, conference, and research articles.
	Research papers on Urdu ASR only.	Papers not written in the English language.
		Theoretical papers.

Table 4. Quality assessment questions

Papers	Quality Assessment (QA)
Review (Threshold = 2.5)	Is the type of review correctly identified and followed?
	Is the methodology used for the review discussed?
	Are research questions for the review correctly identified and answered?
	Are the review findings correctly reported?
	Is the limitation of review correctly mentioned?
English, Mandarin Chinese, Hindi, and Urdu (Threshold = 3)	Is the paper organized according to standards?
	Are the research objectives correctly discussed and answered?
	Are the methods used in the study reported clearly?
	Are the results correctly visualized i.e. through graphs, tables, etc.?
	Are the results obtained from the study contribute to the ASR domain?
	Are the limitations of the research discussed?

5.1. Benchmark Dataset and Attributes for English, Mandarin Chinese and Hindi

Librispeech, Switchboard (SWB), and Wall Street Journal (WSJ) are three commonly used benchmark English speech datasets developed by Johns Hopkins University (JHU), Texas Instruments (TI), Linguistic Data Consortium (LDC), and MIT Lincoln Laboratory, USA respectively as shown in Table 5.

Table 5. English speech datasets and their raw attributes

Dataset	Institute	Format, Sampling Rate	Duration (hrs), Speakers, Vocabulary Size, Type of Speech, [Channel]	Gender Distribution
Librispeech	JHU, USA	.flac, 16	1000, 2484, 0.9M, Continuous, [Mic, Headset, Hands-free]	52% Male, 48% Female
SWB	TI, USA	.wav, 8	300, 500, 3M, Spontaneous, [Telephone]	-
WSJ	MIT Lincoln Laboratory, USA	.wav, 16	~80, -, 47M Continuous, [Mic]	-

Librispeech corpus spans 1000 hours with a vocabulary size of 900,000 words in ".flac" format at a sampling rate of 16 KHz. Multiple recording channels like mic, headset, and hands-free were used to record speech by 2484 participants with 52% male and 48% female [15]. Switchboard (SWB) is a multiple-speaker collection of 300 hours of spontaneous telephonic conversations with a vocabulary size of 3,000,000 words. The database has a sampling rate of 8 KHz and is stored in ".wav" format [71]. The Wall Street Journal (WSJ) dataset is available in ".wav" format with a sampling rate of 16 KHz. This speech spans over 80 hours and has an extensive vocabulary of 47,000,000 words [72].

Table 6. Mandarin Chinese speech datasets and their raw attributes

Dataset	Institute	Format, Sampling Rate	Duration (hrs), Speakers, Vocabulary Size, Type of Speech, [Channel]	Gender Distribution
Aishell-I	BSST Co. Ltd, China	.wav, 16	~178,400, 1.3M, Continuous, [Mic, Mobile]	47% Male, 53% Female
HKUST	HLTC, HKUST	.wav, 8	200, 2412, -, Spontaneous, [Telephone]	51% Male, 49% Female

Beijing Shell Shell Technology (BSST) Co. Ltd developed a widely used open-source continuous rich benchmark dataset Aishell-I for the Mandarin Chinese speech recognition task available in Table 6. It is a continuous speech dataset available in ".wav" format and was recorded at 16 KHz sampling rate from 400 speakers having a vocabulary size of 1,300,000 words [16]. The second Chinese benchmark dataset in Table 6, HKUST is developed by the Human Language Technology Center (HLTC), Hong Kong University of Science and Technology (HKUST). The HLTC, HKUST speech data spans 200 hours and consists of telephonic conversations by 2412 participants on various topics with 51% male and 49% female. This dataset is recorded at 8 KHz sampling rate along with transcriptions and stored in ".wav" format [73].

The Tata Institute of Fundamental Research (TIFR) designed the only available Hindi benchmark speech corpus that spans over 2.3 hours with over 4000 words and involves 100 speakers. The speech was stored in “.wav” format at the sampling rate of 16 KHz. The duration of the training, testing, and validation speech set is 2.1, 0.1, and 0.1 hours respectively [13].

5.2. Urdu Speech Datasets

Table 7. Mandarin chinese speech datasets and their raw attributes

Institute	Work	Format, Sampling Rate	Duration (hrs), Speakers, Vocabulary Size, Type of Speech, [Channel]	Gender Distribution
SJTU, China	[9]	.opus, 48	-, 740, 10, Isolated, [Mobile]	49% Male, 51% Female
Dr. B.A.M University, India	[67]	.wav, 16	-, 50, 10, Isolated, [-]	30% Male, 70% Female
ITU, Pakistan	[47]	-	~1200, ~11K, 5K, Spontaneous, [Mobile]	-
FAST, Pakistan	[14]	.wav, 16	100, -, -, Continuous, [-]	-
CLE, Pakistan	[12]	.wav, 16	~0.3, 10, 250, Isolated, [Mic]	80% Male, 20% Female
	[11]	-, 8	~9, 300, 139, Isolated, [Mobile]	-
	[37]	.wav, 16	~309, 1733, 199K, Continuous, [Mic, Headset, Hands-free]	-
	[38]	.wav, 16	~102, 453, 200K, Continuous, [-]	73% Male, 27% Female

Table 7 shows that Shanghai Jiao Tong University (SJTU), China has introduced a large corpus specifically focused on Urdu digits, ranging from "صفر" (zero) to "نو" (nine) [9]. For data acquisition, a Mobile channel was employed to collect speech samples from 740 male and female participants at a sampling rate of 48,000Hz and was stored in the opus format. Dr. B.A.M University, India developed an isolated digit speech dataset from صفر (zero) to نو (nine) from 50 speakers, including 30% male and 70% female in .wav format with a sampling frequency of 16,000Hz [67]. Another spontaneous speech corpus from 11,017 speakers based on Urdu telephonic conversation was collected by Information Technology University (ITU), Pakistan. The speech spans ~1200 hours, vocabulary size of 5,000 words of Urdu speech using a mobile channel [47]. FAST University, Pakistan developed a 100 hours continuous speech dataset [14]. Table 7 further shows that the Center of Language Engineering (CLE), Pakistan designed four major Urdu speech datasets, and all the corpus types and relevant attributes are summarized in Table 7 on the same basis as described for other language corpus previously.

5.3. Feature Extraction Technique with Least Error Rate (WER and CER)

In this section, we address a critically important question of determining Which dataset and Which feature extraction technique yields the lowest word error rate (WER) or character error rate (CER) for English, Mandarin Chinese, Hindi, and Urdu benchmarks? Both WER and CER measure ASR prediction performance on a predicted word or character basis. Table 8 shows the feature extraction techniques that yield the least WER on benchmark databases. Fig. 4 demonstrates a rich Feature extraction pipeline to briefly explain Triangular features, Δ and ΔΔ features, Log Mel features, SpecAugment, and speed perturbation mentioned in Table 8.

The segmented frames of the speech signal are subjected to Fourier transform, to acquire the magnitude spectrum. For Triangular features, the magnitude spectrum is then passed to the filterbank to extract the energy distribution within different frequency bands. Delta (Δ) features are computed by subtracting the adjacent frames of the Triangular features. Similarly, double delta (ΔΔ) features are obtained by taking the difference between adjacent frames of the delta features. These calculated delta and double delta features are concatenated with the original Triangular features, creating a more comprehensive feature vector that captures both static and dynamic information. For the Log Mel feature, the magnitude spectrum obtained from the Fourier transform passed through a Mel filter, where triangular filters based on the Mel scale are applied. Finally, a logarithmic operation is performed on the resulting features, yielding the log mel spectrogram.

Data augmentation techniques are commonly utilized to increase the training data size without the need for acquiring additional speeches. These techniques include SpecAugment, which involves frequency and time masking to augment the features. Speed perturbation is another approach that alters the chronological behavior of the features. To extract i-vectors, the universal background model (UBM) is trained on the set of extracted acoustic features. It is then adapted to the specific speaker whose i-vector is to be extracted and statistical measures are computed to capture the distribution of the adapted features. The i-vector is extracted from computed statistics using techniques like factor analysis or total variability modeling.

Table 8 shows that, on the Librispeech continuous English dataset, the best WER of 1.9% and 3.9% was obtained on two test sets through 80-dimension Triangular features with the SpecAugment [49]. The SWB spontaneous benchmark of English achieved a WER of 4.3% by utilizing Log Mel features in combination with SpecAugment. Additionally, the i-vector extracted resulted in a total feature dimension of 260 [53]. The best WER of 1.42% was also achieved on the English WSJ continuous dataset through Triangular features with Δ and ΔΔ in combination with speed perturbation. The total dimension of the feature vector was 120 [50]. Table 8 also shows that the minimum CER of 6.4% is accomplished on the Mandarin Chinese Aishell continuous benchmark by utilizing 80-dimensional Mel features, combined with SpecAugment and speed perturbation [74]. In the case of the HKUST spontaneous dataset, the best CER of 23.09% was attained by extracting 40-dimensional Triangular features with Δ and ΔΔ. speed perturbation and

SpecAugment were also employed on the extracted features [40].

Previously, we highlighted that TIFR, Mumbai continuous, and the 250 isolated Urdu dataset serve as the sole benchmarks for Hindi and Urdu, respectively. Our investigation reveals that the best WER of 5.5% for the Hindi speech recognition task was obtained using the one-dimensional raw speech features [52]. In the case of Urdu, the lowest WER for 250 isolated word benchmark dataset is 2.47% by extracting the Log Mel spectrogram from the dataset [9].

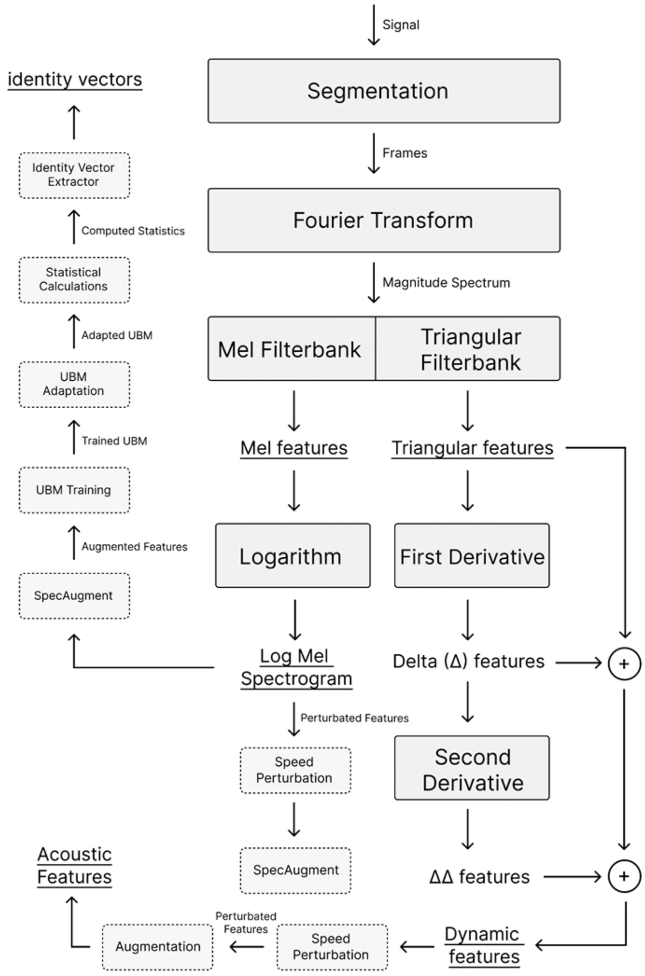

Fig.4. Rich feature extraction pipeline based on table 8

Table 8. Feature extraction techniques with least WER and CER

Benchmark Dataset	Acoustic Features	Augmentation	Dimension	Test Error Rates (%)
Librispeech 1000 hrs (Continuous)	Triangular features	SpecAugment	80	1.9, 3.9
SWB 300 hrs (Spontaneous)	Log Mel features	i-vectors SpecAugment	260	4.3
WSJ ~80 hrs (Continuous)	Δ, ΔΔ Triangular features	Speed Perturbation	120	1.42
Aishell ~178 hrs (Continuous)	Mel features	Speech Perturbation SpecAugment	80	6.4 (CER)
HKUST 200 hrs (Spontaneous)	Δ, ΔΔ Triangular features	Speech Perturbation SpecAugment	40	23.09 (CER)
TIFR, Mumbai 2.3 hrs (Continuous)	Raw speech features	N/A	1	5.5
Urdu 0.3 hrs (Isolated)	Log Mel Spectrogram	N/A	1024 pixels	2.47

5.4. Experimental Design

Deep learning experiments partition the dataset using cross-validation techniques, such as hold-out, k-folds, leave-one-out, leave-p-out, and nested k-folds to counter bias-variance tradeoff. In this section, we limit our discussion to the benchmark datasets and experiments that reported partitioning, error rates, and confidence intervals. Table 9 shows that speech datasets were partitioned based on their time durations. For instance, the English Librispeech continuous dataset consists of a rich training set of 980 hours and approximately 10 hours each for testing and validation, utilized by [32,44,45,49,75]. The authors that employed SWB and WSJ English datasets have offered limited details regarding the experimental design.

The 178 hours of Aishell continuous dataset partitioned approximately 150, 10, and 5 hours for training, validation, and testing, employed by [54,55,74,76]. Among these, [54] exclusively reported the Word Error Rate (WER) along with a confidence interval of 14.16±0.06. Another dataset of Mandarin Chinese, spontaneous HKUST that is used by [40,46,51,56] consists of 200 hours with ~173 hours of training data and ~5 hours each for training and validation.

Only the Hindi benchmark dataset of TIFR, Mumbai comprises approximately 2.1 hours of training data, while 1 hour each is allocated for validation and testing purposes [34,35,52,77,78]. The small-size benchmark corpus of Urdu consists of only ~23 minutes. The authors including [42,66,68] employed this dataset. Our study also found that only [31] applied a 10-fold cross-validation technique on this Urdu benchmark and reported 25.34±2.98 WER with a confidence interval.

Table 9. Experimental protocols employed to benchmark datasets

Benchmark Dataset	Total Time	Training (%)	Validation (%)	Testing (%)	Classes	Error Rates with Conf. Interval
Librispeech	~1000 hrs	98	1	1	52% Male, 48% Female	N/A
Aishell	~178 hrs	~85	~5	~3	47% Male, 53% Female	14.16±0.06
HKUST	200 hrs	~87	~3	~3	2412 speakers	N/A
TIFR, Mumbai	2.3 hrs	~92	~4	~4	100 speakers	N/A
Urdu	~0.3 hrs	10 folds cross-validation			80% Male, 20% Female	25.34±2.98

5.5. Acoustic and Language Models

To develop an ASR system, the acoustic model (AM) must be developed to generate the phonetic sequence i.e. /a/ /b/ /c/. An optional language model (LM) might convert phonetics into the most likely word sequence to apprehend grammar, speaking styles, and dialects.

A. Spontaneous Speech Recognition

Table 10 compares the experimental settings for English (SWB), Mandarin (HKUST), Hindi, and Urdu spontaneous datasets. Table 10 shows that English, Mandarin, and Hindi used >85 hours of training speech which is quite reasonable as compared to Urdu training size (8.5 hours). Notably, greater training sizes lead to the higher accuracy of the ASR system [78]. All studies used comparatively shorter (3 and 5 hours) speech for validation and testing.

Acoustic model WER is found as less than 10% in the case of English, Hindi, and Urdu using Log Mel and Δ, ΔΔ Log Mel with all DNNs while WER appears greater than 20% with MFCC and PLP. For the Mandarin dataset, Δ, ΔΔ Triangular features obtain CER ranging from 23 to 28 which is notably higher than WER. Another evaluation metric, perplexity (PPL) ranges from 1 to infinity and is used to measure the effectiveness of the LM. Higher PPL e.g. n-gram 243.66 indicates worse LM performance [79]; the smaller PPL e.g. 37.04 and 43.1 shows highly confident LM in predicting the word sequence [47,56]. The higher PPL i.e. 243.66 in Hindi is attributed to its larger training size (~100 hours) while Urdu with the same n-gram exhibits very low PPL because of the smaller training set (9.5 hours). The Mandarin PPL seems lower and nice i.e. 43.1 with SAN-LM and training size of (~87 hours) [56].

Table 10. Acoustic and language models for spontaneous datasets

Experiment Details	Acoustic Features	Augmentation	LM, PPL, AM, Test Error (%)
English SWB 300 hrs	Log Mel [53]	Identity vector, SpecAugment,	LSTM-Transformer,-, Conformer-LSTM,4.3
	Δ, ΔΔ Log Mel [48]	Identity vector	n-gram, -, LSTM, 7.6
	Δ, ΔΔ Log Mel, PLP [39]	Speaker vectors	n-gram, -, BLSTM-HMM, 7.6
Mandarin HKUST 200 hrs ~87, ~3, ~3 hrs	Δ, ΔΔ triangular [40] [56] [51]	Speed perturb, SpecAugment	SAN, -, CIF, 23.09(CER)
	Log Mel [46]	Speed perturb	SAN, 43.1, SAA, 24.1 (CER)
			RNN, -, E-RNA,27.7 (CER)
			-, -, Transformer, 26.64 (CER)
Hindi 1108 hrs ~100, 5, ~3 hrs	MFCC, PLP [79]	N/A	n-gram, 243.66, TDNN, 29.7
Urdu 9.5 hrs ~8.5, 1 hrs	MFCC, PLP [47]	N/A	n-gram, 37.04, SGMM, 24.19

Table 10 also shows that the least acoustic WER i.e. 4.3% has been achieved with Identity vector and SpecAugment on Conformer acoustic model on the English SWB dataset [53] as compared to LSTM and BLSTM-HMM acoustic models. Similarly, Transformer LM seems best as compared to n-gram LM [39,48]. The benchmark dataset of Mandarin Chinese, HKUST yielded the least CER of 23.09% along with Speed perturbation and SpecAugment on the CIF acoustic model [40] as compared to SAA (Self-Aligner Attention) [56], Transformer [46], and E-RNA (Extending Recurrent Neural Architecture) [51]. The WER on Hindi 29.7% and 24.19% Urdu spontaneous datasets reported a higher WER using TDNN and SGMM acoustic models.

B. Continuous Speech Recognition

Table 11. Acoustic and language models for continuous datasets

Experiment Details	Acoustic Features	Augmentation	LM, PPL, AM, Test Error (%)
Librispeech 1000 hrs 980, 10, 10 hrs	Triangular features [49]	SpecAugment	LSTM, 63.9, Conformer, (1.9, 3.9)
	Log Mel [44] [45]	Speed perturb, SpecAugment	NNLM+n-gram, Transformer, (2.3, 4.9)
			n-gram, -, Transformer, (2.5, 5.6)
	Raw Speech [75]	N/A	CNN, -, CNN, (3.26, 10.5)
	MFCC [32]	N/A	LSTM, 65.9, E2E-attention, (3.8, 13)
WSJ ~80 hrs ~75, 3, 2 hrs	Δ, ΔΔ Triangular [50]	Speed perturb	RNN, -, TDNN, 1.42
	Raw Speech [75] [80]	N/A	CNN, -, CNN, 3.5
			RNN+n-gram, -, E2E-SincNet, 4.7
Aishell-I 178 hrs ~85, 5, 3 hrs	Mel-filterbank [74]	Speed perturb, SpecAugment	-, -, LASO, 6.4(CER)
	Log Mel [55]	SpecAugment	-, -, SAN-M, 6.64(CER)
	Δ, ΔΔ Log Mel [54]	N/A	n-gram, -, CNN-BLSTM-CTC, 14
	Δ, ΔΔ MFCC [76]	N/A	-, -, CNN-BLSTM-CTC, 19
TIFR, Mumbai 2.3 hrs 2.1, 0.1, 0.1 hrs	Raw speech [52]	N/A	RNN, -, SincNet-CNN-LiGRU, 5.5
	Raw Speech [77]	Speed perturb	n-gram, -, SincNet-CNN-LiGRU, 8
	MFCC [34]	identity vector, Speed perturb, Tempo perturb	n-gram, -, TDNN, 10
	MFCC-GFCC,WERBC [78]	N/A	RNN, -, GMM, 12
	MFCC [35]	N/A	RNN, -, GMM-HMM,15.6
Urdu 100 hrs 100, - hrs	Δ, ΔΔ MFCC [14]	N/A	n-gram, -, SGMM, 9.6
Urdu 309.5 hrs 300, 9.5 hrs	MFCC [37]	Identity vectors	RNN,-, TDNN-BLSTM, 13.5
Urdu 102.5 hrs 98, 4.5 hrs	MFCC [38]	Identity vectors	RNN, -, TDNN, 18.6

In Table 11, we compare the experimental design of seven continuous speech datasets of our chosen four languages: (1) English Librispeech, WSJ, (2) Mandarin Aishell, (3) Hindi TIFR, Mumbai and (4) Urdu. The three languages: English, Mandarin, and Urdu have rich training sets of more than 75 hours as compared to Hindi i.e. 2.1 hours. The validation and testing continuous speech set size for Librispeech is far superior to the other six speech datasets i.e. 10 hours which is quite greater than 3 to 5 hours. The best acoustic error rates have been achieved in English, Mandarin, Hindi, and Urdu using Triangular features [49], Δ, ΔΔ Triangular features [50], Mel-filterbank [74], Raw speech features [52] and Δ, ΔΔ MFCC [14]. The effectiveness of the Librispeech LSTM language model in comparison to Librispeech NN-LM+n-gram, n-gram is evident by PPL lower values i.e. 63.9 and 65.9 [32,44,45,49].

Table 11 also shows that SpecAugment on Conformer [49] is comparable to Transformer [44,45] CNN [75] and E2E-attention [32] because its acoustic WER is 1.9% which is much less than 2.3%, 2.5%, 3.26% and 3.8% in case of Librispeech dataset. Similarly, the least WER i.e. 1.42% on the WSJ test set has been obtained with speed perturbation on TDNN acoustic in comparison with CNN WER 3.5% [75] and E2E-SincNet WER 4.7% [80]. Additionally, it is worth noting that the standalone WSJ RNN language model outperformed both CNN and RNN+n-gram models [75,80].

Table 11 further shows that the lowest CER i.e. 6.4% on Aishell corpus has been attained with Speed perturbation and SpecAugment on the LASO (Listen Attentively, and Spell Once) model [74] comparable to SAN-M (Memory Equipped Self-Attention) [55]. This least CER was achieved solely with the utilization of an acoustic model, without the incorporation of LM. Additionally, the best WER of 14.16% on this benchmark was obtained by utilizing the CNN-BLSTM-CTC acoustic model and the n-gram language model [54]. The same acoustic technique was used by [76] but resulted in a high WER of 19.2%, maybe because of employing an acoustic model only.

Table 11 also shows the best WER of 5.5% using SincNet-CNN-LiGRU [52] was obtained on TIFR, Mumbai benchmark which is superior to TDNN, GMM, and GMM-HMM [34,78,35]. Another study utilized the same SincNet-CNN-LiGRU acoustic model and achieved 8.0% WER [77]. The difference of 8.0% to 5.5% is might because of utilizing the RNN-LM approach Finally, for Urdu databases it can be seen that the best WER i.e. 9.6% was achieved by utilizing the Subspace Gaussian Mixture Model (SGMM) model with n-gram LM [14]. The TDNN-based acoustic models yielded high WERs i.e.13.5% and 18.6% with RNN-LM compared to 9.6% might be because of using only MFCCs [37,38].

C. Connected Words Speech Recognition

In Table 12 we studied the experimental protocols for connected word datasets. We found that very limited research work is available on connected word speech for that reason we included all publicly available research on this speech type. Table 12 shows that the words, sentence, and utterance-based training and testing have been performed on English and Hindi datasets. It can also be seen that all of the studies utilized the MFCC feature extraction approach [81] and PLP [82] with the HMM acoustic model. The least WER in available studies of English and Hindi are 13.33% and 6.67%.

Table 12. Acoustic and language models for connected words datasets

Experiment Details	Acoustic Features	AM, Test Error (%)
English 20 words, 5 sentences	MFCC [81]	HMM, 13.33
Hindi 20 words, 5 sentences	MFCC [81]	HMM, 6.67
Hindi 891 utterance 507, 384 utterance	PLP [82]	HMM, 11.46

D. Isolated Words Speech Recognition

Table 13. Acoustic and language models for isolated words datasets

<table>
<tr><th colspan="2">Experiment Details</th><th>Acoustic Features</th><th>AM, Test Error (%)</th></tr>
<tr><td colspan="2">English (1625 uttr)
75, 25 % (10-folds)</td><td>MFCC and RASTA-PLP [83]</td><td>Random Forest, 3.43</td></tr>
<tr><td colspan="2">Mandarin (600 uttr)
70, 30 % (7-folds)</td><td>Poisson Sub-Sampling [84]</td><td>Ensemble learning, 19.5</td></tr>
<tr><td colspan="2">Hindi (60 uttr)</td><td>MFCC [85]</td><td>KNN, 1.91</td></tr>
<tr><td colspan="2">Urdu (25518 uttr)</td><td>Log Mel Spectrogram [9]</td><td>CNN, 14</td></tr>
<tr><td colspan="2">Urdu (1500 uttr)
1000, 500 uttr</td><td>LPC [67]</td><td>HMM, 26</td></tr>
<tr><td colspan="2">Urdu (~9 hr)
80, 20 %</td><td>MFCC [11]</td><td>GMM, 7.13</td></tr>
<tr><td rowspan="6">Urdu Benchmark
~0.3 hours
[12]</td><td>-</td><td>Log Mel Spectrogram [9]</td><td>CNN, 2.47%</td></tr>
<tr><td>70, 30 %</td><td>MFCC [42]</td><td>BLSTM, 17</td></tr>
<tr><td>90, 10 %</td><td>MFCC + Δ + log filterbank + LLE [68]</td><td>E2E-Network, 22.08</td></tr>
<tr><td>10 folds</td><td>MFCC [31]</td><td>HMM, 25.34</td></tr>
<tr><td rowspan="2">70, 30 %</td><td>MFCC + Δ + ΔΔ [66]</td><td>SVM, 27</td></tr>
<tr><td>MFCC [30]</td><td>LDA, 29.33</td></tr>
</table>

Finally, in Table 13 we review the experimental design for isolated speech datasets. For Urdu, we spread our range from 2015 to 2022 to cover all major research. It can be seen that all the languages have applied hold-out (70%, 30% and 90%, 10%) and K-folds (10-fold) cross-validation techniques on their datasets. We found that MFCC was the most frequently used technique on isolated datasets [11,30,31,42,66,83,85].

Table 13 also shows that the conventional classification technique i.e. Random Forest (RF), Ensemble learning, and KNN have been applied in recent research of English, Mandarin Chinese, and Hindi having WER of 3.43%, 19.5%, and 1.91% respectively [83-85]. It can also be seen that the lowest WER of 2.47% was obtained on the Urdu benchmark dataset with the Log Mel spectrogram and CNN acoustic model [9,12].

6. Limitations

This comparative systematic literature review for Urdu ASR elucidates the AI building blocks used in its development. Our research compiled prominent ASR databases along with their raw attributes, SOTA feature extraction techniques, DNN experimental design for English, Mandarin Chinese, Hindi, and Urdu then proceeded to contrast their respective acoustic and language models based on the type of speech. Although our review thoroughly examines the comparative analysis of ASR systems, a more detailed discussion on comparing language characteristics, deep-learning models, and evaluation metrics is needed to provide a holistic understanding of the research landscape in this field. Our primary objective was to underscore the advancements made in Urdu speech recognition systems and ascertain the position of Urdu within the landscape of ASR advancements.

7. Conclusions

This PRISMA-based systematic review contributes to the novel comparative discussion of spontaneous, continuous, connected, and isolated speech datasets of the three most spoken languages English, Mandarin Chinese, and Hindi with Urdu. Huge speech datasets i.e. 1000 hours of English, 200 hours of Mandarin Chinese, and 1108 hours of Hindi are publicly available but the maximum known size for a private Urdu speech dataset is 1207 hours. We have had extensive discussions on language-oriented Automatic Speech Recognition (ASR) pipeline with the help of the year 2018-2022 literature. To deeply investigate Urdu ASR development from 2015-2017 is also analyzed. We conclude that both acoustic and language models turned into Deep Neural Networks (DNNs). The classical acoustic model HMM has been replaced by Conformer-LSTM and TDNN. Likewise, the n-gram language model evolves to LSTM and attention-based networks like Self-Attention (SAN). These models performed best on challenging spontaneous and continuous speech types as evidenced by respective low acoustic word error rate of less than 5%, character error rate of less than 25%, and perplexity ranges between 35% to 45%. The achievement of lower WER, CER and Perplexity requires Conformers, Transformers and Attention with particular focus on continuous and spontaneous speech recognition in comparison to conventional LSTM, TDNN, RNN, HMM, GMM-HMM developed with shorter speech size datasets.

Acknowledgment

This work has been done for the fulfilment of a 4-year undergrad Bachelor of Science in Computer Science (BSCS) at the Department of Computer Science, University of Karachi, Pakistan. We express our sincere gratitude to ML-Labs, Ireland, for their generous support and funding, which played a pivotal role in enabling us to accomplish the objectives of this review. Their assistance was instrumental in the successful completion of our research goals.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Google, "Language Support," [Online]. Available: https://developers.google.com/assistant/sdk/reference/rpc/languages. [Accessed 24 October 2023].

[2] Apple, "iOS and iPadOS 16 Feature Availability," [Online]. Available: https://www.apple.com/ios/feature-availability/#siri-on-device-speech. [Accessed 25 October 2023].

[3] Microsoft, "Cortana's regions and languages," [Online]. Available: https://support.microsoft.com/en-us/topic/cortana-s-regions-and-languages-ad09a301-ce3a-aee4-6364-0f0f0c2ca888. [Accessed 25 October 2023].

[4] Phonexia, "Phonexia Speech Platform for Government," [Online]. Available: https://www.phonexia.com/product/speech-platform-government. [Accessed 25 October 2023].

[5] Meity, "Ministry of Electronics & Information Technology," [Online]. Available: https://www.meity.gov.in/home. [Accessed 25 October 2023].

[6] Google, "Why Build," [Online]. Available: https://developers.google.com/assistant/why-build. [Accessed 25 October 2023].

[7] Voicebot.ai, "Bayer Launches AMI Voice Assistant for Doctors on Google Assistant," [Online]. Available: https://voicebot.ai/2021/04/19/bayer-launches-ami-voice-assistant-for-doctors-on-google-assistant/. [Accessed 25 October 2023].

[8] A. Ouisaadane and S. Safi, "A comparative study for Arabic speech recognition system in noisy environments," International Journal of Speech Technology, pp. 761-770, 2021.

[9] Aiman, Y. Shen, M. Bendechache, I. Inayat and T. Kumar, "AUDD: Audio Urdu Digits Dataset for Automatic Audio Urdu Digit Recognition," Applied Science, vol. 11, no. 19, 2021.

[10] M. M. H. Nahid, M. A. Islam and M. S. Islam, "A Noble Approach for Recognizing Bangla Real Number Automatically Using CMU Sphinx4," in 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016.

[11] M. Qasim, S. Nawaz, S. Hussain and T. Habib, "Urdu Speech Recognition System for District Names of Pakistan: Development, Challenges and Solutions," in 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA) , Bali, Indonesia, 2016.

[12] H. Ali, N. Ahmed, K. M. Yahya and O. Farooq, "A Medium Vocabulary Urdu Isolated Words Balanced Corpus for Automatic Speech Recognition," in 2012 International Conference on Electronics Computer Technology (ICECT 2012), 2012.

[13] K. Samudravijaya, P. V. S. Rao and S. S. Agrawal, "Hindi Speech Databases," in Sixth International Conference on Spoken Language Processing, Beijing, China, 2000.

[14] S. Naeem, M. Iqbal, M. Saqib, M. Saad, M. S. Raza, Z. Ali, N. Akhtar, M. O. Beg, W. Shahzad and M. U. Arshad, "Subspace Gaussian Mixture Model for Continuous Urdu Speech Recognition using Kaldi," in 14th International Conference on Open Source Systems and Technologies (ICOSST), 2020.

[15] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR Corpus Based Public Domain Audio Books," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.
[16] H. Bu, J. Du, X. Na, B. Wu and H. Zheng, "Aishell-I: An Open Source Mandarin Speech Corpus and A Speech Recognition Baseline," in 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), 2017.
[17] Ethnologue, "Ethnologue: Languages of the World," [Online]. Available: https://www.ethnologue.com/. [Accessed 25 October 2023].
[18] S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi, S. Alharbi, S. alturki, F. Alshehri and M. Almojil, "Automatic Speech Recognition: Systematic Literature Review," IEEE Access, vol. 9, pp. 131858-131876, 2021.
[19] A. Dhouib, A. Othman, O. E. Ghoul, M. K. Khribi and A. A. Sinani, "Arabic Automatic Speech Recognition: A Systematic Literature Review," Applied Science, vol. 12, no. 17, 2022.
[20] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman and f. t. P. Group, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," BMJ, 2009.
[21] A. Vanker, R. P. Gie and H. J. Zar, "The association between environmental tobacco smoke exposure and childhood respiratory disease: a review," Expert Review of Respiratory Medicine, vol. 11, no. 8, pp. 661-673, 2017.
[22] H. Zhang, D. Ren, X. Jin and H. Wu, "The prognostic value of modified Glasgow Prognostic Score in pancreatic cancer: a meta-analysis," Cancer Cell International, vol. 20, no. 462, 2020.
[23] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," Artificial Intelligence in Medicine, 2022.
[24] M. S. Jahan and M. Oussalah, "A systematic review of Hate Speech automatic detection using Natural Language Processing," Neurocomputing, vol. 546, 2023.
[25] D. William and D. Suhartono, "Text-based Depression Detection on Social Media Posts: A Systematic Literature Review," Procedia Computer Science, pp. 582-589, 2021.
[26] S. Salloum, T. Gaber, S. Vadera and K. Shaalan, "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques," IEEE Access, vol. 10, pp. 65703-65727, 2022.
[27] U. Petti, S. Baker and A. Korhonen, "A systematic literature review of automatic Alzheimer’s disease detection from speech and language," Journal of the American Medical Informatics Association, vol. 27, no. 11, pp. 1784-1797, 2020.
[28] Y. A. Ibrahim, J. C. Odiketa and T. S. Ibiyemi, "Preprocessing technique in automatic speech recognition for human computer interaction: an overview," Annals. Computer Science Series, vol. 15, no. 1, pp. 186-191, 2017.
[29] M. Labied, A. Belangour, M. Banane and M. Banane, "An overview of Automatic Speech Recognition Preprocessing Techniques," in International Conference on Decision Aid Sciences and Applications (DASA), 2022.
[30] H. Ali, N. Ahmad and X. Zhou, "Automatic speech recognition of Urdu words using linear discriminant analysis," Journal of Intelligent & Fuzzy Systems 28 (2015), vol. 8, pp. 2369-2375, 2015.
[31] Asadullah, A. Shaukat, H. Ali and U. Akram, "Automatic Urdu Speech Recognition using Hidden Markov Model," in 2016 International Conference on Image, Vision and Computing (ICIVC), 2016.
[32] A. Zeyer, A. Zeyer, R. Schluter and H. Ney, "Improved training of end-to-end attention models for speech recognition," 2018. [Online]. Available: https://arxiv.org/abs/1805.03294. [Accessed 20 October 2023].
[33] J. Islam, M. Mubassira, M. R. Islam and A. K. Das, "A Speech Recognition System for Bengali Language using Recurrent Neural Network," in 2019 IEEE 4th International Conference on Computer and Communication Systems, 2019.
[34] A. Kumar and R. K. Aggarwal, "Hindi speech recognition using time delay neural network acoustic modeling with i-vector adaptation," International Journal of Speech Technology, pp. 67-78, 2022.
[35] M. Dua, R. K. Aggarwal and M. Biswas, "Discriminatively trained continuous Hindi speech recognition system using interpolated recurrent neural network language modeling," Neural Computing and Applications, vol. 31, pp. 6747-6755, 2019.
[36] M. S. Yakoub, S.-a. Selouani, B.-F. Zaidi and A. Bouchair, "Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network," EURASIP Journal on Audio, Speech, and Music Processing, 2020.
[37] M. U. Farooq, F. Adeeba, S. Rauf and S. Hussain, "Improving Large Vocabulary Urdu Speech Recognition System using Deep Neural Networks," in Interspeech, 2019.
[38] E. Khan, S. Rauf, F. Adeeba and S. Hussain, "A Multi-Genre Urdu Broadcast Speech Recognition System," in 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), 2021.
[39] X. Cui, W. Zhang, U. Finkler, G. Saon, M. Picheny and D. Kung, "Distributed Training of Deep Neural Network Acoustic Models for Automatic Speech Recognition," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 39-49, 2020.
[40] L. Dong and B. Xu, "CIF: Continuous Integrate-And-Fire for End-To-End Speech Recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.
[41] S. Mussakhojayeva, Y. Khassanov and H. A. Varol, "A Study of Multilingual End-to-End Speech Recognition for Kazakh, Russian, and English," in Speech and Computer, 2021.
[42] T. Zia and U. Zahid, "Long short-term memory recurrent neural network architectures for Urdu acoustic modeling," International Journal of Speech Technology, vol. 22, pp. 21-30, 2019.
[43] R. Sharmin, S. K. Rahut and M. R. Huq, "Bengali Spoken Digit Classification: A Deep Learning Approach Using Convolutional Neural Network," Procedia Computer Science, vol. 171, pp. 1381-1388, 2020.
[44] Y. Wang, A. Mohamed, D. Le, C. Liu, A. Xiao, J. Mahadeokar, H. Huang, A. Tjandra, X. Zhang, F. Zhang, C. Fuegen, G. Zweig and M. L. Seltzer, "Transformer-Based Acoustic Modeling for Hybrid Speech Recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.
[45] Y. Shi, Y. Wang, C. Wu, C. Fuegen, F. Zhang, D. Le, C.-F. Yeh and M. L. Seltzer, "Weak-Attention Suppression For Transformer Based Speech Recognition," in Interspeech, 2020.
[46] S. Zhou, L. Dong, S. Xu and B. Xu, "A Comparison of Modeling Units in Sequence-to-Sequence Speech Recognition with the Transformer on Mandarin Chinese," in International Conference on Neural Information Processing, 2018.
[47] A. A. Raza, A. Athar, S. Randhawa, Z. Tariq, M. B. Saleem, H. B. Zia, U. Saif and R. Rosenfeld, "Rapid Collection of Spontaneous Speech Corpora using Telephonic Community Forums," in Interspeech, 2018.

[48] W. Zhang, X. Cui, U. Finkler, B. Kingsbury, G. Saon, D. Kung and M. Picheny, "Distributed Deep Learning Strategies For Automatic Speech Recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[49] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," 2020. [Online]. Available: https://arxiv.org/abs/2005.08100. [Accessed 27 October 2023].

[50] K. An, Y. Zhang and Z. Ou, "Deformable TDNN with adaptive receptive fields for speech recognition," in Interspeech, 2021.

[51] L. Dong, S. Zhou, W. Chen and B. Xu, "Extending Recurrent Neural Aligner for Streaming End-to-End Speech Recognition in Mandarin," in Interspeech, 2018.

[52] A. Kumer and R. K. Aggarwal, "An exploration of semi-supervised and language-adversarial transfer learning using hybrid acoustic model for hindi speech recognition," Journal of Reliable Intelligent Environments, vol. 8, pp. 117-132, 2021.

[53] Z. Tuske, G. Saon and B. Kingsbury, "On the limit of English conversational speech recognition," in Interspeech, 2021.

[54] Y. Wang, Z. LiMin, B. Zhang and Z. Li, "End-to-End Mandarin Recognition based on Convolution Input," in 2018 2nd International Conference on Information Processing and Control Engineering (ICIPCE 2018), 2018.

[55] Z. Gao, S. Zhang, M. Lei and I. McLoughlin, "SAN-M: Memory Equipped Self-Attention for End-to-End Speech Recognition," in Interspeech, 2020.

[56] L. Dong, F. Wang and B. Xu, "Self-attention Aligner: A Latency-control End-to-end Model for ASR Using Self-attention Network and Chunk-hopping," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[57] O. Mantiri, "Factors Affecting Language Change," SSRN Electronic Journal, 2010.

[58] ATLAS, "ATLAS - Urdu: Urdu Language," [Online]. Available: https://www.ucl.ac.uk/atlas/urdu/language.html. [Accessed 27 October 2023].

[59] F. Naz, W. Anwar, U. I. Bajwa and E. Munir, "Urdu Part of Speech Tagging Using Transformation Based Error Driven Learning," World Applied Sciences Journal, vol. 16, no. 3, pp. 437-448, 2012.

[60] K. Riaz, "Rule-based Named Entity Recognition in Urdu," in Proceedings of the 2010 Named Entities Workshop, 2010.

[61] A. Daud, W. Khan and D. Che, "Urdu language processing: a survey," Artificial Intelligence Review, vol. 47, pp. 279-311, 2017.

[62] M. U. Akram and M. Arif, "Design of an Urdu Speech Recognizer based upon acoustic phonetic modeling approach," in 8th International Multitopic Conference, 2004.

[63] A. Beg and S. K. Hasnain, "A Speech Recognition System for Urdu Language," in Wireless Networks, Information Processing and Systems, 2008.

[64] H. Sarfaraz, S. Hussain, R. Bokhari, A. A. Raza, I. Ullah, Z. Sarfaraz, S. Pervez, A. Mustafa, I. Javed and R. Parveen, "Large Vocabulary Continuous Speech Recognition for Urdu," in 8th International Conference on Frontiers of Information Technology, Islamabad, Pakistan, 2010.

[65] J. Ashraf, D. N. Iqbal, N. S. Khattak and A. M. Zaidi, "Speaker Independent Urdu Speech Recognition Using HMM," in 15th International Conference on Applications of Natural Language to Information Systems, 2010.

[66] H. Ali, A. Jianwei and K. Iqbal, "Automatic Speech Recognition of Urdu Digits with Optimal Classification Approach," International Journal of Computer Applications, vol. 118, no. 9, pp. 1-5, 2015.

[67] N. Shaikh and R. R. Deshmukh, "LPC and HMM Performance Analysis for Speech Recognition System for Urdu Digits," IOSR Journal of Computer Engineering (IOSR-JCE), vol. 19, no. 4, pp. 14-18, 2017.

[68] M. A. Humayun, I. A. Hameed, S. M. Shah, S. H. Khan, I. Zafar, S. B. Ahmed and J. Shuja, "Regularized Urdu Speech Recognition with Semi-Supervised Deep Learning," Applied Science, vol. 9, no. 9, 2019.

[69] H. Ali, N. Ahmed, X. Zhou, K. Iqbal and S. M. Ali, "DWT features performance analysis for automatic speech recognition of Urdu," SpringerPlus, vol. 3, 2014.

[70] Google, "Google Scholar," [Online]. Available: https://scholar.google.com/. [Accessed 27 October 2023].

[71] J. Godfrey, E. Holliman and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1992.

[72] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992, 1992.

[73] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang and D. Graff, "HKUST/MTS: A Very Large Scale Mandarin Telephone Speech Corpus," in 5th International Symposium on Chinese Spoken Language Processing, 2006.

[74] Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen and S. Zhang, "Listen Attentively, and Spell Once: Whole Sentence Generation via a Non-Autoregressive Architecture for Low-Latency Speech Recognition," 2020. [Online]. Available: https://arxiv.org/abs/2005.04862. [Accessed 27 October 2023].

[75] N. Zeghidour, Q. Xu, V. Liptchinsky, N. Usunier, G. Synnaeve and R. Collobert, "Fully Convolutional Speech Recognition," 2019. [Online]. Available: https://arxiv.org/abs/1812.06864. [Accessed 27 October 2023].

[76] D. Wang, X. Wang and S. Lv, "End-to-End Mandarin Speech Recognition Combining CNN and BLSTM," Symmetry, vol. 11, no. 5, 2019.

[77] A. Kumar and R. K. Aggarwal, "A hybrid CNN-LiGRU acoustic modeling using raw waveform sincnet for Hindi ASR," Computer Science, vol. 21, no. 4, pp. 397-417, 2020.

[78] Kumar and R. Aggarwal, "Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modeling," Journal of Intelligent Systems, pp. 165-179, 2021.

[79] Bhanushali, G. Bridgman, D. G, P. Ghosh, P. Kumar and e. al., "Gram Vaani ASR Challenge on spontaneous telephone speech recordings in regional variations of Hindi," in Interspeech, 2022.

[80] T. Parcollet, M. Morchid and G. Linares, "E2E-SINCNET: Toward Fully End-To-End Speech Recognition," in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

[81] S. Singhal and R. K. Dubey, "Automatic Speech Recognition for Connected Words using DTW /HMM for English/ Hindi Languages," in Communication, Control and Intelligent Systems (CCIS), 2015.

[82] S. Bhatt, A. Jain and A. Dev, "Monophone-based connected word Hindi speech recognition improvement," Sadhana, vol. 46, no. 99, pp. 1-17, 2021.

[83] D. E. B. Zeidan, A. Noun, M. Nassereddine, J. Charara and A. Chkeir, "Speech Recognition for Functional Decline assessment in older adults," in ICBRA '22: Proceedings of the 9th International Conference on Bioinformatics Research and Applications, 2022.
[84] C.-H. H. Yang, J. Qi, S. M. Siniscalchi and C.-H. Lee, "An Ensemble Teacher-Student Learning Approach with Poisson Sub-sampling to Differential Privacy Preserving Speech Recognition," [Online]. Available: https://arxiv.org/abs/2210.06382. [Accessed 27 October 2023].
[85] S. Bhable, A. Lahase and S. Maher, "Automatic Speech Recognition (ASR) of Isolated Words in Hindi low resource Language," Int`ernational Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 9, no. 2, pp. 260-265, 2021.

Authors' Profiles

Muhammad Hazique Khatri is a software engineer at Techwards. He graduated from the University of Karachi, Pakistan in 2023 with a Bachelor's degree, ranking second in his class with a CGPA of 3.7. His research interests span machine learning, deep neural networks, and natural language processing (NLP).

Dr. Humera Tariq received the B.E (Electrical) degree from NED University of Engineering and Technology in 1999 and then continue her studies at University of Karachi for Masters of Computer Science (MCS) in 2001. She stands First Class First amongst the Evening batch of MCS 2001-2003. She joined MS leading to PhD program in 2009 and completed MS course work with CGPA 4.0. She started her PhD work in the field of image processing in 2011 under the supervision of Meritorious Professor Dr. S.M. Aqil Burney.

Maryam Feroze is a lecturer at University of Karachi. She did BS in Computer Science from University of Karachi. Her research interest includes data mining and deep neural networks.

Ebad Ali is a Co-Founder and CTO of various ventures. Pursuing data science and audio processing with practical bent of software engineering. He was graduated from University of Karachi in 2014 and completed his master from National College of Ireland in 2019.

Zeeshan Anjum Junaidi is working as a senior software engineer at IDWise. He did his BS from University of Karachi in 2014.

How to cite this paper: Muhammad Hazique Khatri, Humera Tariq, Maryam Feroze, Ebad Ali, Zeeshan Anjum Junaidi, "A PRISMA-driven Review of Speech Recognition based on English, Mandarin Chinese, Hindi and Urdu Language", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.3, pp.36-51, 2024. DOI:10.5815/ijitcs.2024.03.04