International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Automatic Speech Recognition System for Low Resource Punjabi Language using Deep Neural Network-Hidden Markov Model (DNN-HMM)

Rajni Sobti[1,2], Kalpna Guleria[1*], Virender Kadyan[3]

Submitted: 10/01/2024 **Revised**: 16/02/2024 **Accepted**: 24/02/2024

Abstract: In recent years, speech recognition technology has advanced significantly, enabling seamless human-machine interaction. The majority of these advances, however, have focused on major languages with abundant data and resources, neglecting the rich linguistic diversity inherent in low resource languages. There are unique challenges associated with speech recognition in low resource languages, because there is a lack of comprehensive linguistic resources as well as data. To ensure inclusivity and promote global accessibility, researchers recognize the need to bridge this gap. This article focuses on the development of children ASR in the Punjabi languages, along with the potential benefits that can be gained from addressing this understudied field. For the purpose, the speech data from children have been collected (who speak Punjabi) and collected audio data were then segmented using PRAAT (open source software) followed by transcription of segmented audio files. The feature extraction has been implemented using the MFCC algorithm. The process of acoustic modelling has been implemented using various models which include MONO, Tri1, Tri2 and Tri3. The acoustic model then was trained with DNN-HMM to increase the accuracy of the children's ASR in Punjabi language. The results reveal 83.9% accuracy of children ASR in the Punjabi language. Further, the comparison with the existing models shows that the proposed DNN-HMM model gives better results.

Keywords*: Children Automatic Speech Recognition; Low Resource Language; Punjabi Speech; Data Collection; Deep Neural Networks; DNN-HMM.*

1. Introduction

Automatic speech recognition (ASR) converts spoken language into text. It is a powerful tool that can be used in a variety of applications. These include transcribing audio recordings, generating subtitles for videos, and controlling devices with voice commands (Kadyan 2018). Introducing speech recognition systems into smart devices such as Alexa and Siri are the best examples of successful ASR (Hasija et al., 2022). ASR is not only limited to Alexa and Siri, it has other numerous applications in the fields of robotics, call classification, medical healthcare /reports dictation systems. Helping differently abled people is one of the most promising applications of speech recognition. Disabled individuals are likely to benefit from such technological advancements (Noyes et al.,1989). Speech recognition systems offer numerous benefits for adult users, however, research focuses on children's speech started quite late. Since 1990, speech recognition technology has been applied to applications for young children. Education, entertainment, assessment, and health care are some of the areas where speech recognition is used for children nowadays (Sobti et al., 2022). According to the studies, ASR provides high accuracy for adult speech, whereas children's ASR lags behind in accuracy due to their vocal variations (Bhardwaj et al., 2021). To address this, ASR systems must be trained with more data from children and focus should be on improving the accuracy of recognizing children's speech (Bhardwaj et al., 2021; Bhardwaj and Kukreja, 2021; Shivkumar et al., 2014; Nagano et al., 2019; Shahnawazuddin et al.,2020). Additionally, very little research is being conducted on the ASR system for children, particularly in regional Indian languages (Bhardwaj et al., 2021). In recent years, speech recognition technology has advanced substantially, allowing seamless interactions between humans and machines. The majority of these advances have concentrated on major languages with abundant resources and data, leaving linguistic diversity in low resource languages behind. Creating speech recognition solutions for low-resource languages poses unique challenges due to the lack of comprehensive linguistic resources and data. The low-resource languages are typically spoken by fewer people than major languages. They are often called minority languages or under-resourced languages. For these languages, there may not be sufficient linguistic resources such as speech corpora, lexicons, and language models for training and development of robust

[1]*Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*
sobtirajni15@gmail.com
[2]*University Institute of Engineering & Technology, Panjab University, Chandigarh, India-160014*
guleria.kalpna@gmail.com
[3]*Speech and Language Research Centre, School of Computer Sciences, University of Petroleum and Energy Studies, Dehradun, Uttrakhand, India-248007*
ervirenderkadyan@gmail.com
[*]*Corresponding Author: guleria.kalpna@gmail.com* (Kalpna Guleria)

speech recognition systems. The lack of available data makes low resource languages challenging in multiple areas of speech recognition, such as acoustic modeling, language modeling, and pronunciation modelling.

Punjabi is a regional language predominantly spoken in the Indian state of Punjab. Apart from India, Punjabi is also recognized worldwide, but still it is consider as under resource language due to lack of its digital resources. The Punjabi language possesses tonal characteristics (hasija et. al 2021A), despite a tonal language; it is recognized as one of the languages with limited available resources or low resource language (hasija et. al 2021A). Among 22 major languages in India, the development of ASR system primarily focused on English, Hindi and Marathi, leaving other languages, including Punjabi with limited ASR support. Not much research has been conducted for development of ASR in Punjabi language when compared to other languages, like English and Italian (Kadyan, 2018)**.** Research on children's speech in Punjabi is lacking, while most research has been conducted on adults' speech. Punjabi ASR performance is adversely affected by data scarcity which, when applied to various acoustic model algorithms, results in overfitting due to a lack of resources (Bawa and Kadyan, 2021)**.** Due to the fact that DNNs are data-driven, the recognition rate of ASR is better with large amounts of speech data. Using a large amount of children's speech corpora for ASR training is one way of improving recognition rates of the children's ASR system. However, the ASR system for Punjabi language faces a significant challenge due to the insufficient availability of a suitable children`s speech corpus. In addition, collecting speech data for children is more challenging than for adults (Bhardwaj et al., 2021)**.** The process of recording and transcribing input data for a low-resource language demands substantial human efforts and resources. Due to these limitations, real-time speech recognition engines cannot be built efficiently without sufficient training data (Chohan and Garcia, 2019; Lata and Arora, 2013)**.**

Motivation

Studying native languages for the design of native ASR systems is an active area of research which could recognize the importance of enabling people to interact comfortably in their native languages. Efforts are underway to design ASR systems specifically tailored to support such interactions. It is essential to have a native ASR application in India because the majority of the population doesn't speak English. It is important that ASR systems are designed which can handle customer inquiries via a voice-based interface at railway stations and in other institutions (Kadyan 2018, Lata and Arora, 2013). Languages that are native are also called unreserved languages (UR) because there are fewer electronic resources available for these languages. It is challenging to work for UR ASR due to the lack of resources and corpora (Kadyan 2018). The available research on Punjabi ASR has been carried out using adult speech, however, ASRs developed using children speech are lacking. Since, Punjabi language (in spite of its native character) is recognized worldwide, which encourages us for the design of Punjabi ASR. ASR system for children is as important as those for adults, as children are more dependent on these systems nowadays, such as computer games, reading tutors, foreign language learning tools, etc. The contribution of this article is :

1. This article presents the development of ASR system on children speech in the Punjabi language using DNN-HMM acoustic model.
2. Further, the article describes the process of feature extraction using MFCC algorithm. The process of acoustic modelling has been implemented using various models which include MONO, Tri1, Tri2 and Tri3.
3. The proposed system is compared with the existing ASR models which show that the proposed DNN-HMM model gives better results exhibiting 83.9% accuracy and 16.1% WER for Children's ASR developed in Punjabi language.
4. The article also presents in depth details of low resource language and various challenges encountered for developing Punjabi-ASR. It also spotlights the basics of the Punjabi language, a low resource language.

The remaining article is divided as: the theoretical background of the ASR systems has been discussed in section 2. Section 3 focused on low resource languages. Section 4 elaborates on the Punjabi language. Section 5 describes materials and methods for development of Punjabi ASR. Experimental results and discussion is present in Section 6. Finally the article is concluded in Section 7.

2. Theoretical Background

2.1 MFCC (Mel Frequency Cepstral Coefficient)**:** A raw speech signal has very complex representation and for any ASR system it cannot act as effective input. The important information is extracted from the raw speech signal and fed as input to any ASR system. ASR systems initiate by extracting features from speech signals to discern and identify the underlying components of the signals. Feature extraction encompasses three distinct processing stages: *i)* static feature extraction; *ii)* normalization, and *iii)* incorporation of temporal information. At each stage, a comprehensive evaluation of techniques is conducted, considering both theoretical aspects and their relative performance. These features capture the essential information while filtering out irrelevant or anomalous details from the signal. Useful information that is extracted from audio signals is called features and the process of extracting this information is called feature extraction. In

order to develop a successful speech recognition system, feature extraction plays an integral part. This process eliminates redundant and unwanted information from speech signals. Number of features extraction methods is available in the literature (Kim and Stern, 2012; Hermansky, 1990; Alim and Rashid, 2018) but most commonly used technique is MFCC (Kherdekar and Naik, 2019). MFCC has become de facto for extraction of features from speech signals. It leverages key aspects of speech perception and production to capture relevant characteristics. By utilizing these principles, MFCC effectively encapsulates comprehensive details of the speech signal within the feature vector (Dua et al., 2018). Following steps have been used by MFCC for feature extraction.

Pre-Processing: As audio clips are sampled and converted into discrete spaces, A/D conversion digitizes them. In most cases, sampling frequencies of 8 or 16 kHz are used.

Pre-emphasis: The pre-emphasis step improves the signal-to-noise ratio by increasing energy at high. Pre-emphasis boosts higher frequencies using a filter. Filters maximize higher frequencies over lower frequencies when passing data through them. Eq. 1(Shi et al., 2018; Bawa and Kadyan, 2021) represents high pass filter.

$$y(n)=x(n)- \alpha x(n-1) \quad (1)$$

where, y(n) is an output signal; x(n) is an input signal; n is number of frame and α is the filter coefficient.

Windowing: A windowing technique involves slicing the audio waveform into sliding frames. Here, frames of 20–40 ms are framed for the pre-emphasis signal. As a standard, 25 milliseconds are considered adequate. On adjacent frames, 10 ms of overlap is performed (Liu et al., 2018). However, chopping it off at the edge of the frame will not work. It is likely that the sudden drop in amplitude will create a lot of noise in the high frequencies as a result of the sudden drop in amplitude. It is best to slice the audio by gradually decreasing its amplitude near the frame edges. A time domain window (w) is applied to the original audio clip. Eq. 2 represents relation between sliced frame i.e X[n] and original audio clip s[n].

$$X[n] = w[n]s[n] \quad (2)$$

The corresponding equation for (*w)*, Eq.3 (Liu et al., 2018; Shi et al., 2018; Bawa and Kadyan, 2021), is given as follows:

$$w[n] = (1 - \alpha) - \alpha cos\left(\frac{2\pi n}{L-1}\right) \quad (3)$$

$\alpha = 0.4614$ for *hamming window* and L is the window width.

Fast Fourier Transformation (FFT): After windowing, Fourier transformation is applied on the signal, which is a filtering operation and used to convert the signal into the frequency domain from time domain (Gupta et al., 2013).The set of N samples shown below utilizes the Fast Fourier Transforms (FFT) algorithms to efficiently compute the Discrete Fourier Transform (DFT) (Gupta et al., 2013; Speech Recognition — Feature Extraction MFCC & PLP | by Jonathan Hui | Medium.)

$$X[k] = \sum_{n=0}^{N-1} x[n]\, exp\left(-j\frac{2\pi}{N}kn\right) \quad \text{for} \quad 0 \le k \le N-1 \quad (4)$$

Mel Filter Bank (MFB): Mel spectrums are obtained by applying a **MFB**, which consists of a series of bandpass filters, to the signals that have undergone Fourier transformation. Human auditory systems are simulated using the **MFB**. Human hearing is capable of processing speech signals with varying frequency distributions. This makes it easier for human hearing to process and recognize speech. The information that is extracted is then used to classify the signal and determine what it is. Mels are units of measurement based on the human ear's perceived frequency. In humans, pitch does not relate linearly to frequency, as pitch is not perceived linearly. Under 1 kHz, Mel scales usually have linear spacing, and above 1 kHz, logarithmic spacing (Gupta et al., 2013; Bawa and Kadyan, 2021). Mel's approximation from physical frequency is expressed in Eq. (5)

$$f_{mel} = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \quad (5)$$

where, f_{mel} expressed the perceived; frequency f signifies the physical frequency (Hz),

(Gupta et al., 2013; Liu et al., 2018;Bawa and Kadyan, 2021).

Log: Power spectrums are produced by the **MFB**. A logarithmic distribution is used here. Following that, the log will be retrieved from the Mel filter bank output. Moreover, this minimizes non-relevant acoustic variants.

Discrete Transformation (DCT): By applying DCT to transformed Mel frequency coefficients, cepstral coefficients are obtained. It is possible to make this system robust by extracting only the first few MFCC coefficients. The higher order DCT components are either truncated or and ignored. Therefore, it resulted in 2–13 MFCC coefficients which are retained and rest are represented through fast changes which are further discarded. In Eq.(6) (Shi et al.,2018; Gupta et al.,2013) *c(n)* expressed cepstral coefficients and C is the no. of MFCCs. MFCC is calculated using the procedure illustrated in Fig. **1.**

$c(n) = \sum_{m=0}^{M-1}(log\ s(m))cos\left(\frac{\pi n(m-0.5}{M}\right)$ for n=0,1,2…,C-1 (6)

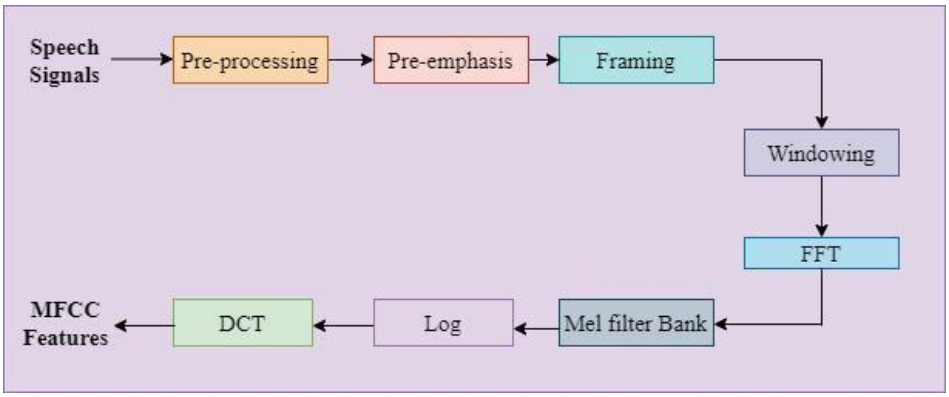

Fig 1: Complete procedure to compute MFCC coefficients.

2.2 Statistical Framework of ASR: Speech-to-text mapping is a task that maps spoken audio to written text. Under statistical framework maximum posterior probability can be calculated by observing a word sequence as an acoustic sequence. (Lu et. al, 2019)

$$W^* = arg_w max\ P(W|X) \quad (7)$$

In Eq.(7), W= [w_1, w_2, ---------- w_n] and X=[x_1,x_2,----------x_n] are the given predicted word sequence and given observation acoustic sequence, respectively.

Applying Bayesian theory, Eq. 7 reduces to:

$$W^* = arg_w max \frac{P(X|W)(PW)}{P(X)} \propto arg_w max P(X|W)P(W) \quad (8)$$

In Eq. 8 P(X|W) is considered as acoustic model; P(W) is the probability of observing the sequence of words. Training of acoustic and language models is done separately. Since optimization of Eq. 7 is a tedious work, Eq. 8 can be rewritten as Eq. 9 (Lu et. al, 2019) using divide and conquer strategy:

$$W^* \approx \arg_w \max \sum_L P(X \mid L)P(L \mid W)P(W) \quad (9)$$

In Eq. 9, P(L|W) and P (X|L) are pronunciation model and generative model, respectively.

Further, this generative model could be formulated by summation of hidden state variable:

$$W^* \approx \arg_w \max \sum_L P(X \mid S)P(S \mid L)P(L \mid W)P(W) \quad (10)$$

In Eq. 10, S=[s_1,s_2,----------s_T]: hidden state sequence; P (X|S): likelihood of acoustic observation sequence.

Generally, the conventional algorithms are based on Eq. 10 to estimate sequential probability which involves complex computation process. HMM and Gaussian Mixture Model (GMM) are popular for approximating hidden states, where a markov chain assumption can approximate the sequence of probability of hidden states, while a GMM model can approximate the output probability of each hidden state. Typically, a conventional ASR consists of several components as shown in Fig. 2.

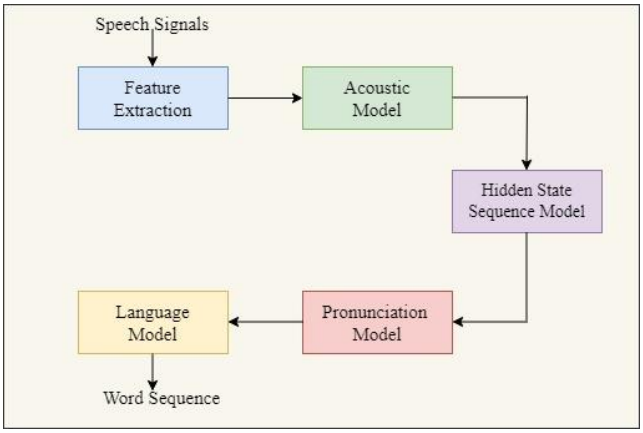

Fig 2: Conventional ASR and its components.

2.3 Deep Neural Networks (DNN): A number of studies have found that deep learning can be a powerful method for building complex and dedicated analysis system using large scale training data (Zhang et. Al 2017). Along with gaming (Mnih et al. 2015) deep learning has also been successfully applied to other fields, like language translation (Wu et al. 2016), visual recognition, music information retrieval and of course in automatic speech recognition systems (Liu et al. 2017; Schedl et al. 2016 ;Zhang et al.,2018; Russakovsky et al. 2015). DNN is used to train acoustic model, and they are among the most effective and out performing methods available. DNNs are feed forward neural networks with neurons arranged in layers that are fully connected (Serizel and Giuliani, 2017). In DNN, inputs and outputs are hidden in more than one layer of hidden units (Hinton et al., 2012). A DNN is called a deep neural network because it is composed of many layers. An input and output layer are separated by a hidden layer. When DNN is used in ASR, the posterior probability of (sub) phonetic units provided by output layer and feature vectors (augmented with context) are processed by input layers (Serizel & Giuliani, 2017).

There are many units between outputs and inputs in a DNN. Every hidden unit, *p,* uses logistic function input from layer below, x_p, to the scalar state, y_p which it sends to above layer (Hinton et al., 2012).

$$y_p = logistic(x_p) = \frac{1}{1+e^{-x_p}} \; ; \quad x_p = b_p + \sum_i y_i \, . \, w_{ip} \tag{11}$$

In Eq. (11), b_p *is* bias of hidden unit *p* ;

i represents index over units in layer below.

w_{ip} is the weight on a connection to unit *p* from unit *i* in layer below.

With multiclass classification, output unit *p* converts its total input, x_p into a class probability P_p , using *"softmax"* non-linearity as:

$$p_p = \frac{exp\,(x_p)}{\sum_k \exp\,(x_k)} \tag{12}$$

In Eq. 12, k is an index over all classes. Further, Eq. (13) represents natural cost function '*C'* which represents the cross-entropy between softmax output *p* and target probabilities *d*. DNN classifier is trained by providing supervised information through target probabilities having values of 1 or 0.

$$C = -\sum_p d_p \; log \; p_p \tag{13}$$

The derivatives are computed on a mini-batch 't' as it is always advisable for more efficiency. The improvement in this stochastic gradient descent method could be done using "momentum" coefficient, 0<α<1 (Hinton et al., 2012):

$$\Delta W_{ip}\,(t) = \alpha \Delta w_{ip}(t-1) - \epsilon \frac{\partial C}{\partial w_{ip(t)}} \tag{14}$$

For understanding details regarding training of DNN the readers may refer to Hinton et al. (2012).

3. Low Resource Languages

In recent years, speech recognition technology has advanced significantly, enabling seamless human-machine interaction. The majority of these advances, however, have focused on major languages with abundant data and resources, neglecting the rich linguistic diversity inherent in low resource languages. There are unique challenges associated with speech recognition in low resource languages due to lack of comprehensive linguistic resources as well as data (Hasegawa et al., 2016; Besacier et al., 2014). To ensure inclusivity and promote global accessibility, researchers recognize the need to bridge this gap.

Understanding Low Resource Languages*:* Low resource language referred to as under resource languages. These languages are spoken by smaller population in comparison to major languages. These languages often face challenges such as lack of comprehensive linguistic resource, absence of a unique writing system, limited linguistic expertise, scarcity of resources on web, inadequate electronic resources such as speech corpora, lexicons, and language models essential for training & development of robust speech recognition systems (Deka et al.,2018; Berment et al.,2004; Besacier et al., 2014).

Challenges in Low resource languages: for designing ASR in native language, a significant hurdle is the difficulty in locating native speakers possessing the necessary technical skills. Technology experts (the system developers) and language experts (the speakers) must bridge the gap. The *second* challenge is availability of high quality speech corpora. *Third*, low resource language has poorly addressed linguistic literature, many dialects and phonetic intricacies. *Fourth*, pronunciation variability is another challenge i.e. low resource language may lack with standardized pronunciation guidelines (Besacier et al., 2014).

All the above mentioned challenges pose difficulties in developing a robust ASR system for low resource languages. Since the researchers have started working on low resource languages, therefore, there exist remedies for the above mentioned challenges. The quantity of speech data can be increased through diverse data augmentation techniques, including speech, pitch, tempo, and volume perturbation; reverberation and spectral augmentation.

These techniques enable the augmentation of speech data by introducing variations in different aspects, expanding the available dataset. However, originally recorded small corpus specific to the concerned language must be available. Further, Advanced machine learning algorithms and DNNs can also be used to increase the accuracy of ASR system developed for low resource languages.

4. Punjabi language

India is a diverse country in terms of culture, religion and languages. Here, variety of languages is spoken. Out of these different languages there are only 22 official languages (Dhanjal 2014). Most of these languages belong to Indo Aryan family of languages. Besides being one of the official languages, Punjabi is also an Indo-Aryan language. Punjabi is world`s 10th most influential language (Guglani, 2022; Dhanjal 2014). Punjabi is vibrant, very easy to learn and widely spoken language. Punjabi language is not only limited to one state but also spoken in other states of India. Punjabi literature, music, and poetry emphasize its rich cultural heritage. Punjabi poetry reflects its vivid imagery and emotional expression and Punjabi music is popular for its upbeat rhythms. Though Punjabi is widely spoken outside Punjab and internationally, still it get limited international recognition. Due to this, Punjabi speakers in other parts of the world may have difficulties accessing resources or communicating. Dialect of any region represents the characteristics of that area and in Punjabi language, there are many dialects. A list of 31 dialects has been published by Panjabi University, Patiala, India (Dhanjal 2014). Further, Malwai is the dominating dialect in the state amongst presently spoken dialects viz *Malwai, Majhi, Doadi and Puadhi* (Dhanjal 2014). Due to many dialects of Punjabi, there is lack of standardization form of the language. It is challenging for speakers of different dialects to understand each other.

It is the Gurmukhi script that is used to write Punjabi. Since Gurmukhi is written with 35 letters, it is referred to as "Painti". A total of 35 characters of the Gurmukhi script are divided into seven rows of five characters each. By placing a dot underneath already existing letters, five more letters have been added to the language. This is called the “New Group” (navīn ṭoli), which constitutes an eighth line in the script, increasing the number of letters to 40. In addition to that, there is a new letter ਲ਼ /L/ which has been added recently to this new group. As a result of these additional sounds, the script now contains 41 letters shown in Fig 3.

Fig 3: 41 letters of Gurmukhi Script

The Gurmukhi script consists of three vowels characters ੳ, ਅ, and ੲ. Ten vowel accents are used in Punjabi to make 10 vowel sounds. These 10 vowel sounds are used to distinguish words in the language and create an auditory distinction between similar words. This distinction helps to ensure that the correct meaning is conveyed in conversations and other forms of communication.

Vowel accents consist of the three vowel characters ੳ, ਅ, and ੲ. ਅ is the only vowel that requires no modification of vowel accent. Additionally, both other vowels must be paired with their corresponding accents (Dhanjal, 2014;

https://punjabi.lrc.columbia.edu). Ten vowel symbols are there, known as a *ləga/matra*.

5. Material and Methods

To develop a children ASR system in the Punjabi language, the speech data were collected from children who speak Punjabi. The collected audio files were then segmented using PRAAT (open source software) followed by transcription of segmented audio files. Subsequently, the features were extracted from the acquired dataset. A description of the proposed method in term of block diagram is given in Fig. 4. The following subsections cover the research materials and the methods used in this work

5.1 Data Collection:

This article presents the speech data collection of children in the Punjabi language which is considered as low-resource language (Bawa and Kadyan, 2021). Due to the severe shortage of children`s speech corpora in the Punjabi language, this initiative was set up in order to address the issue. In the present study, corpus has been collected from children aged between 7-13 years. Description of collected child corpora is given below in table 1.

Table 1: Description of collected child corpora in the Punjabi language

Term	**Value**
Age Group	7 - 13 years
Number of Speakers	45
Number of Unique Sentences	2370
Type of Speech	Connected data
No. of words	27,381

5.2 Experimental setup: The baseline system is designed/built using collected children speech corpus. This corpus is composed of connected sentences. In this corpus, the data have been collected from children aged between 7-13 years of male and female both. The whole data have been collected in real environment and saved in *.wav* format. The data were recorded using mobile microphone in a separate classroom at sampling frequency of 16,000 kHz. Collected data were segmented using PRAAT software. After segmentation and transcription of audios, there are total 2370 utterances out of which 1885 were kept for training and 485 for testing. The baseline system was designed using Kaldi toolkit (open source software) on Ubuntu version 18.0. Further, experiments were conducted on GPU (graphical processing unit). Front end acoustic features were extracted using MFCC. In MFCC, overlapping of hamming window of duration 25ms with frame shift of 10 ms is used for short-time signal analysis. Finally, 39 features were extracted from frame. Out of these 39 features only first 13 features are used which are considered as energy parameters of each frame. Further, linear discriminative analysis (LDA) was found to reconstruct these extracted Y features and improve training acoustic of datasets significantly. The extracted first 13 features are combined with frames which resulted into 117 dimensions and these were further reduced to 40 dimensions using LDA approach. In addition, the triphone model was also used to apply these features to HMM state alignments. Furthermore, feature space was transformed using MLLT, through the state conditional covariance of combined features spaces. Finally, for the baseline system, extracted MFCC features were computed for delta, double delta, and LDA. DNN-HMM classifier is used to process MLLT. The performance of ASR system can be computed using word error rate (WER) as (Hasija et al., 2021B):

$$WER = \frac{(Insertions\ (I) + Deletions\ (D) + Substitutions\ (S)}{Total\ Words\ in\ Correct\ Transcription\ (N)} *100 \quad (15)$$

In Eq. (15), the word error could be in terms of substitution (S), deletion (D), and insertion (I). Block diagram of baseline system for children is given in Fig 4**.**

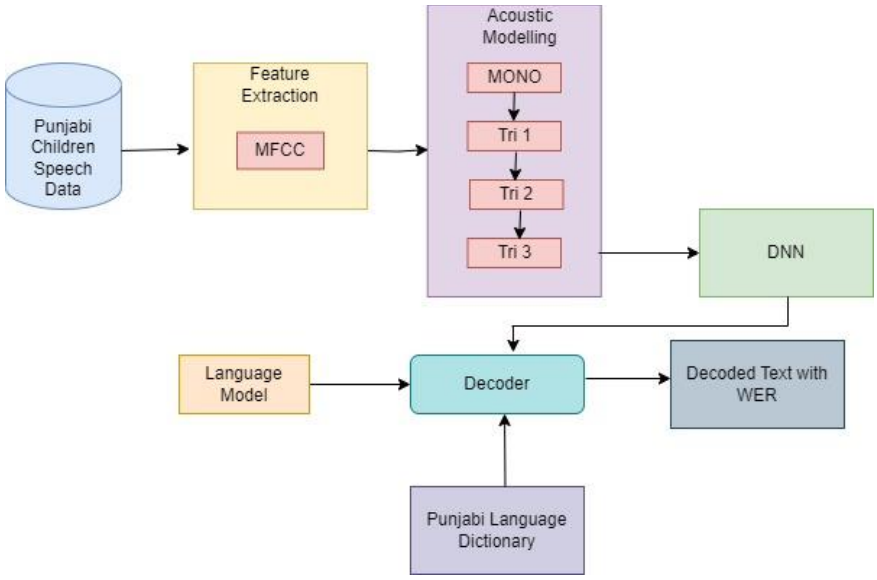

Fig 4: Block diagram for children ASR

The flowchart of the procedure to implement children ASR system is given below in Fig 5.

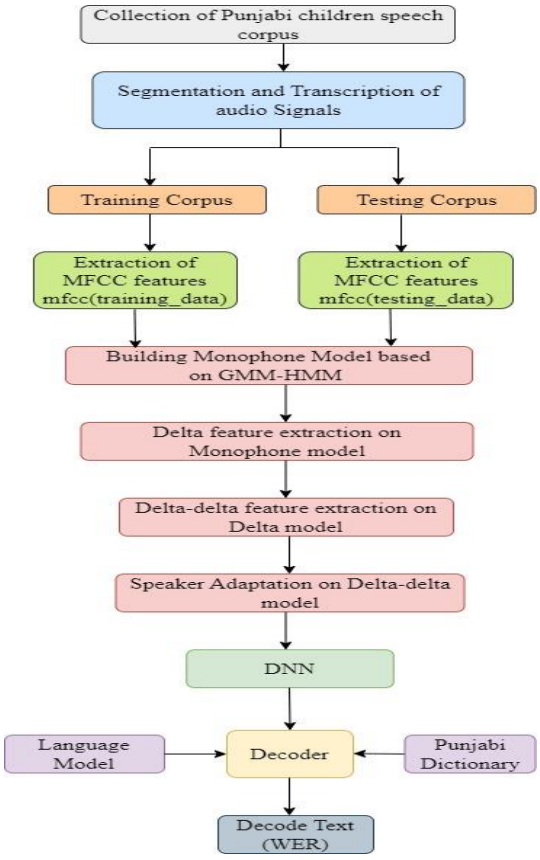

Fig 5: Flow chart of procedure to implement children ASR system

6. Experimental Results and Discussion

6.1 WER/ Accuracy of the developed Punjabi ASR for Children

Experiment was conducted on collected children dataset for Punjabi language. The word error rate (%) of system using MFCC on features extraction and DNN-HMM as classifier is shown in table 2.

Table 2: WER (%) of children ASR system

Acoustic Models	Mel Frequency Cepstral Coefficients (MFCC)
Mono	21.20
Tri1	29.50
Tri2	28.74
Tri3	18.41
DNN	16.10

In Kaldi, MONO, TRI 1, TRI 2, and TRI 3 refer to different stages of acoustic modeling in the context of speech recognition systems. Each of these stages involves progressively more complex models, TRI 3 being the most advanced and typically providing the best results. The development of DNN has led to significant advancements in many fields, including speech recognition. DNN can model complex relationships and capture fine-grained acoustic variations. DNN has shown superior performance in various speech recognition benchmarks and real-world applications. They often achieve lower word error rates and higher recognition accuracy compared to traditional models.

As part of this system, MFCC features were extracted from the speech corpus. Then HMM-DNN is applied as acoustic model. The comparison of WER (%) of Mono, Tri 1, Tri2, Tri3 and DNN is shown in Fig 6 and the accuracy is given in Fig 7.

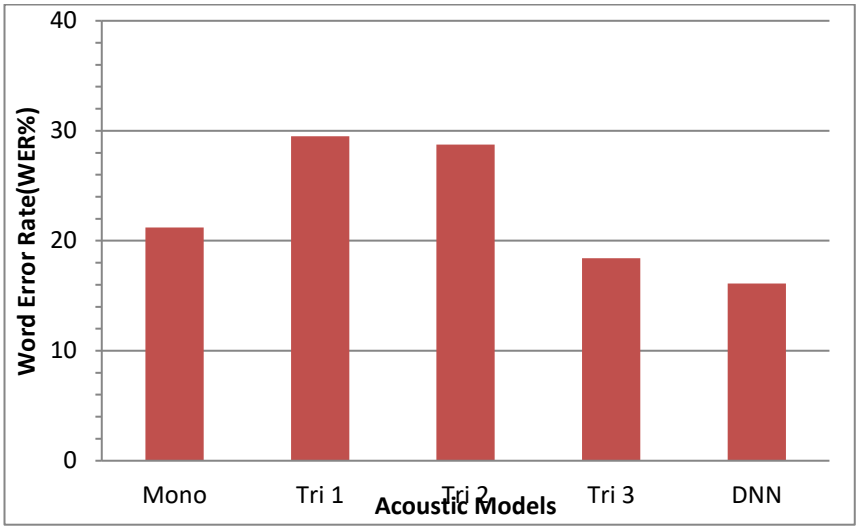

Fig 6: WER of children ASR system

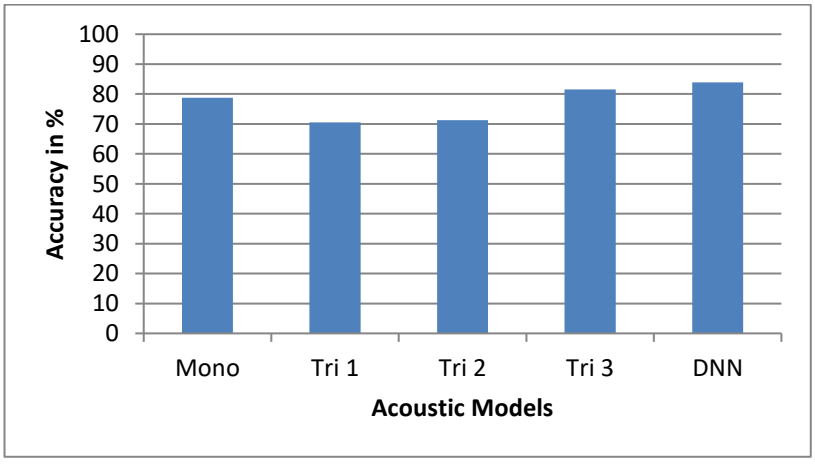

Fig 7: Accuracy of children ASR system

It is observed from Fig. 6 that this Punjabi ASR system has shown word error rate of 21.20% for Mono which is better than Tri 1 and Tri 2 system, i.e 29.50% and 28.74% respectively. However, Tri3 gave better performance (18.41%WER) compared to Mono, Tri1, and Tri2. DNN outperforms all other techniques applied here, showing WER of 16.10%. Consequently, Fig. 7 depicts the highest accuracy for the applied DNN approach compared with others i.e. Mono, Tri1, Tri2, and Tri3.

6.2 Comparison with earlier Punjabi Speech Recognition research:

This section represents comparison with the existing Punjabi ASR with the one developed in this study in terms of WER/accuracy. Table 3 shows comparative results of the existing Punjabi ASR with the present results. The comparison in terms of accuracy and WER with the proposed work is shown in Fig. 8 and Fig. 9, respectively.

Table 3: Comparison with existing Punjabi ASRs

S.no	Ref.	Data Set	Feature Extraction Technique	Acoustic Modelling Technique	Performance
1.	(Kadyan et al.,2017)	Total utterances: 45,000. Male participants: 15 Female participants 10. Isolated data set was used.	MFCC, PLP, RASTA-PLP	HMM + GA (Genetic Algorithm), HMM + DE (Differential Evolution)	Accuracy for 1) MFCC: 67.38%. 2) PLP: 61.17%, 3) RATA-PLP: 58.67%
2.	(Kadyan et al.,2018)	Total utterances: 4033. For training 3611 utterances were used and for testing 422 utterances were used. Used dataset was phonetically rich.	MFCC, GFCC	GMM + HMM and DNN + HMM	WER(MFCC) for 1)DNN+HMM:5.22% 2)GMM+HMM: 7.01% WER(GFCC) for 1)DNN+HMM: 24.67% 2)GMM+HMM: 34.4%
3.	Guglani and Mishra (2020)	Total utterances: 1500.	Pitch Extraction Technique: SAcC	----------	69.4% WER on SAcC Pitch
4.	Proposed Work	Total utterances: 2370 For training 1885 utterances were used and for testing 485 utterances were used.	MFCC	DNN-HMM	MFCC was used as feature extraction and Acoustic model was trained using DNN-HMM, the WER was 16.1% (84% accuracy)

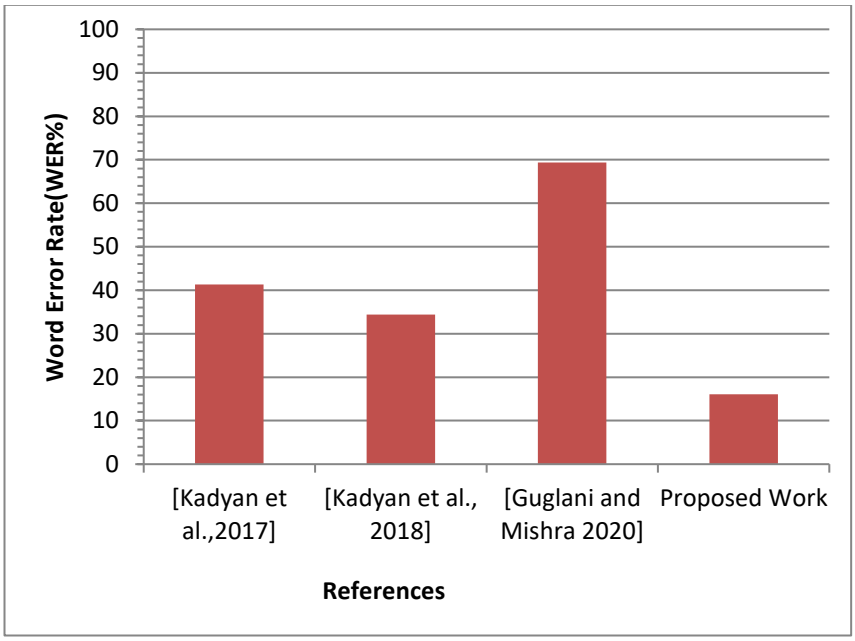

Fig 8: Comparison of word error rate.

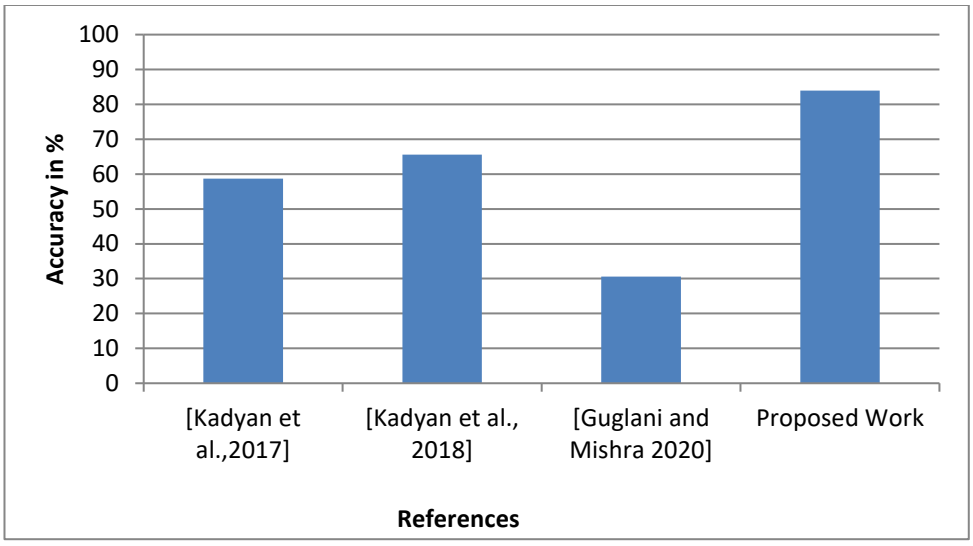

Fig 9: Comparison in terms of accuracy.

From Fig. 8, it is observed that the proposed work shows better WER (16.1%) compared with WER of existing Punjabi ASRs. Consequently, the proposed work shows better accuracy of 84% compared with accuracy of existing Punjabi ASRs, represented in Fig 9.

7. Conclusion

Traditionally, there has been a strong focus on international languages in ASR research. Developing ASR systems for languages with limited resources can be challenging. Little attention has been paid to the Punjabi ASR system. Here, ASR system has been developed with collected children speech data in Punjabi language. The developed Punjabi ASR system for children has shown word error rate of 21.20% for Mono which is better than Tri 1 and Tri 2 system, i.e 29.50% and 28.74% respectively. However, Tri3 gave better performance (18.41%WER) compared to Mono, Tri1, and Tri2. DNN outperforms all other techniques applied here, showing WER of 16.10%, consequently, showing better accuracy for the applied DNN approach compared with others i.e. Mono, Tri1, and Tri2. The comparison with the existing ASRs shows that the proposed work outperforms all in terms of WER/accuracy. Therefore, the development of the proposed Punjabi ASR for children using DNN-HMM could lead to significant advancements in many fields, including speech recognition.

References:

[1] Alim, S. A., & Rashid, N. K. A. (2018). Some commonly used speech feature extraction algorithms (pp. 2-19). London, UK: IntechOpen. http://dx.doi.org/10.5772/intechopen.80419

[2] Bawa, P., & Kadyan, V. (2021). Noise robust in-domain children speech enhancement for automatic Punjabi recognition system under mismatched conditions. Applied Acoustics, 175, 107810. https://doi.org/10.1016/j.apacoust.2020.107810

[3] Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. Speech communication, 56, 85-100. https://doi.org/10.1016/j.specom.2013.07.008

[4] Bhardwaj, V., Kukreja, V., & Singh, A. (2021). Usage of Prosody Modification and Acoustic Adaptation for Robust Automatic Speech Recognition (ASR) System. Rev. d'Intelligence Artif., 35(3), 235-242. https://doi.org/10.18280/ria.350307 |

[5] Bhardwaj, V., & Kukreja, V. (2021). Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions. Applied Acoustics, 177, 107918. https://doi.org/10.1016/j.apacoust.2021.107918

[6] Chohan, M. N., & García, M. I. M. (2019). Phonemic comparison of English and punjabi. International Journal of English Linguistics, 9(4), 347-357. https://doi.org/10.5539/ijel.v9n4p347

[7] Rumelhart, D. E., Hinton, G. E., and Williams, R. J.(1986) "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536.

[8] Deka, B., Chakraborty, J., Dey, A., Nath, S., Sarmah, P., Nirmala, S. R., & Vijaya, S. (2018). Speech corpora of under resourced languages of north-east India. In 2018 Oriental COCOSDA-International Conference on Speech Database and Assessments (pp. 72-77). IEEE. https://doi.org/10.1109/ICSDA.2018.8693038

[9] Dua, M., Aggarwal, R. K., & Biswas, M. (2018). Optimizing integrated features for Hindi automatic speech recognition system. Journal of Intelligent Systems, 29(1), 959-976. https://doi.org/10.1515/jisys-2018-0057

[10] Guglani, J., & Mishra, A. N. (2020). Automatic speech recognition system with pitch dependent features for Punjabi language on KALDI toolkit. Applied Acoustics, 167, 107386. https://doi.org/10.1016/j.apacoust.2020.107386

[11] Gupta, S., Jaafar, J., Ahmad, W. W., & Bansal, A. (2013). Feature extraction using MFCC. Signal & Image Processing: An International Journal, 4(4),

[12] https://doi.org/101-108. 10.5121/sipij.2013.4408

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," Journal of the Acoustical Society of America, Vol-4, Aug, 1990. https://doi.org/10.1121/1.399423

Hasegawa-Johnson, M. A., Jyothi, P., McCloy, D., Mirbagheri, M., Di Liberto, G. M., Das, A., & Lee, A. K. C. (2016). ASR for under-resourced languages from probabilistic transcription. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(1), 50-63. https://doi.org/10.1109/TASLP.2016.2621659

[14] Hasija, T., Kadyan, V., & Guleria, K. (2021, A March). Recognition of Children Punjabi Speech using Tonal Non-Tonal Classifier. In 2021 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 702-706). IEEE.

[15] Hasija, T., Kadyan, V., & Guleria, K. (2021, B August). Out Domain Data Augmentation on Punjabi Children Speech Recognition using Tacotron. In Journal of Physics: Conference Series (Vol. 1950, No. 1, p. 012044). IOP Publishing.

[16] Hasija, T., Kadyan, V., Guleria, K., Alharbi, A., Alyami, H., & Goyal, N. (2022). Prosodic feature-based discriminatively trained low resource speech recognition system. Sustainability, 14(2), 614.

[17] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine, 29(6), 82-97. https://doi.org/10.1109/MSP.2012.2205597

[18] https://punjabi.lrc.columbia.edu/?page_id=11
Jyoti Guglani," Continuous Speech Recognition Of Punjabi Language", Ph.D Dissertation. Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, 2022.

[19] Kadyan, V. Acoustic Features Optimization for Punjabi Automatic Speech Recognition System. Ph.D. Dissertation, Chitkara University, Rajpura, India, 2018.

[20] Kadyan, V.; Mantri, A.; Aggarwal, R.K. A heterogeneous speech feature vectors generation approach with hybrid hmm classifiers. Int. J. Speech Technol. 2017, 20, 761–769. https://doi.org/10.1007/s10772-017-9446-9

[21] Kherdekar, V. A., & Naik, S. A. Speech Recognition System Approaches, Techniques And Tools For Mathematical Expressions: A Review. International Journal Of Scientific & Technology Research Volume 8, Issue 08, August 2019, ISSN 2277-8616,pp 1255-1263.

[22] https://api.semanticscholar.org/CorpusID:202890136

[23] Kim, C. and Stern, R. M., "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in Proc. of ICASSP, Vol-1, May, 2012. https://doi.org/10.1109/TASLP.2016.2545928

[24] Lata, S., & Arora, S. (2013, August). Laryngeal tonal characteristics of Punjabi—an experimental study. In 2013 International Conference on Human Computer Interactions (ICHCI) (pp. 1-6). https://doi.org/10.1109/ICHCI-IEEE.2013.6887793

[25] Liu, Z., Wu, Z., Li, T., Li, J., & Shen, C. (2018). GMM and CNN hybrid method for short utterance speaker recognition. IEEE Transactions on Industrial informatics, 14(7), 3244-3252. https://doi.org /10.1109/TII.2018.2799928

[26] Lu X., Li S., and Fujimoto M.(2019), Automatic Speech Recognition. Book chapter, Speech-to-Speech Translation, SpringerBriefs in Computer Science, https://doi.org/10.1007/978-981-15-0595-9_2

Schedl M., Yi-Hsuan Yang, and Perfecto Herrera-Boyer. 2016. Introduction to intelligent music systems and applications. ACM Trans. Intell. Syst. Technol. 8, 2 (Oct. 2016), 17:1–17:8. https://doi.org/10.1145/2991468

[27] Nagano, T., Fukuda, T., Suzuki, M., Kurata, G. (2019). Data augmentation based on vowel stretch for improving children's speech recognition. In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 502-508. https://doi.org/10.1109/ASRU46091.2019.9003741

[28] Noyes, J. M., Haigh, R., & Starr, A. F. (1989). Automatic speech recognition for disabled people. Applied Ergonomics, 20(4), 293-298.

[29] https://doi.org/10.1016/0003-6870(89)90193-2

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 3 (Dec. 2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[31] Serizel, R., & Giuliani, D. (2017). Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. Natural Language Engineering, 23(3), 325-350. https://doi.org/10.1017/S135132491600005X

[32] Shahnawazuddin, S., Adiga, N., Kathania, H.K., Sai, B. T. (2020). Creating speaker independent ASR system through prosody modification based data augmentation. Pattern Recognition Letters, 131: 213-218.
https://doi.org/10.1016/j.patrec.2019.12.019

[33] Shivakumar, P.G., Potamianos, A., Lee, S., Narayanan, S. (2014). Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In WOCCI, 15-19.

[34] Shi, L., Ahmad, I., He, Y., & Chang, K. (2018). Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments. Journal of Communications and Networks, 20(5), 509-518. https://doi.org/10.1109/JCN.2018.000075

[35] Sobti, R., Kadyan, V., & Guleria, K. (2022). Challenges for Designing of Children Speech Corpora: A State-of-the-Art Review. ECS Transactions, 107(1), 9053. https://doi.org/10.1149/10701.9053ecst

[36] Surinderpal Singh Dhanjal,"Speech Analysis And Synthesis Of The Punjabi Language", Ph.D Dissertation. Thapar University, 2014

[37] Vincent Berment, Methods to computerize "little equipped" languages and groups of languages, Theses, Universite Joseph-Fourier - Grenoble I, May 2004, ´ https://tel.archives-ouvertes.fr/tel-00006313.

[38] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, and others. 2015. Human-level control through deep reinforcement learning. Nature 518, 7540 (Feb. 2015), 529–533. https://doi.org/10.1038/nature14236

[39] Yan Liu, Yang Liu, Shenghua Zhong, and Songtao Wu. 2017. Implicit visual learning: Image recognition via dissipative learning model. ACM Trans. Intell. Syst. Technol. 8, 2 (Jan. 2017), 31:1–31:24. https://doi.org/10.1145/2974024

[40] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, and others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144 (Oct. 2016) https://doi.org/10.48550/arXiv.1609.08144

[41] Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E. D., Jin, W., & Schuller, B. (2018). Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Transactions on Intelligent Systems and Technology (TIST), 9(5), 1-28. https://doi.org/10.1145/3178115

[42] Zixing Zhang, Nicholas Cummins, and Björn Schuller. 2017. Advanced data exploitation for speech analysis—An overview. IEEE Sign. Process. Mag. 34 (July 2017). https://doi.org/10.1109/MSP.2017.2699358

[43] Speech Recognition — Feature Extraction MFCC & PLP | by Jonathan Hui | Medium Accessed on 21.01.2024