Building an ASR System for Indian (Punjabi) language and its evaluation for Malwa and Majha dialect: Preliminary Results

Vivek Bhardwaj[1], Vinay Kukreja[2], Navjeet Kaur[3] Nandini Modi[4]
[1,2,3,4]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
vivek.bhardwaj@outlook.in

2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579471

Abstract —Automatic Speech recognition (ASR), also referred to as voice recognition or speech-to-text, is one of the most developing field of Natural Language Processing (NLP) that recognizes speech. Speech recognition allowing the human voice to serve as the main interface between the human and the computer. Plenty of progress is to be done in the field of speech recognition for various popular languages and adult users with good recognition rates. But in case of Indian regional languages Punjabi, Telugu, Tamil etc, speech recognition is still at the infant level under disparate acoustic conditions. The objective of this research work is to demonstrate the effectiveness of the pitch acoustic features for improving the recognition performances of the ASR system for different Punjabi dialects. When our developed dialect-based Punjabi ASR system was evaluated for Malwa and Majha dialect speakers, the findings showed WER of 23.25 % and 25.91 %, respectively.

Keywords — ASR, Punjabi Dialect, Acoustic, Kaldi toolkit, MFCC

I. Introduction

Automatic speech recognition is a multidisciplinary field of computer science and computational linguistics that develops methods and technology for computers to recognise and translate spoken languages into text. Due to the complexity and dialect variations of the regional languages such as Punjabi, Tamil, Telugu etc., speech recognition of such languages in India is a very concerned issue. There are very few speech recognition systems for these Indian regional languages with higher recognition accuracy are developed till date. Some of the key areas where Automatic speech recognition systems play an important role are - Speech applications have been implemented in public places such as trains, multiplexes, airport communication, and tourist information, where tourists may get immediate answers to their questions. It would also narrow the gap between people who have difficulty communicating in a different language. Speech Recognition is an eventual inter-face for physically challenged persons because blind individuals find it difficult to read from a screen, and writing on paper every time is inefficient in today's world. Along with this, Children are also an important segment of consumers who will benefit from ASR-based developments in multi-media technologies. Thus, in this work we worked on developing an ASR system for the Punjabi language. Approximately 113 million people speak Punjabi as their first language and mostly spoken in India, Paki-stan, United States, Canada, and the United Kingdom. According to the investigations, the conventional acoustic feature alone cannot accomplish expected performance recognition, but subsequent performance-boosting with the addition of pitch frequency can be achieved. [1] [2] [3]. Due to the lack of a large speech corpus for Punjabi language in terms of speakers of all ages and dialects, a fairly substantial speech corpus for Punjabi language must be developed first. After developing a Punjabi speech corpus, due to the tonal behavior of the Punjabi language, we exploited pitch features in the front end to improve speech recognition rate of the developed system with the help of the Kaldi toolkit

II. Related Work

This section provides a quick overview of previous work in the field of speech recognition.

The authors in [4] presented the analysis of Amazigh speech recognition using the IVR system based on the speaker-independent Amazigh digits ASR system. From the experiments it is cleared that, the recognition accuracy for the speech was rarely affected if SNR in-creases from 15 db and there is a slight decrease in the recognition if the system was tested for the SNR 3 and 9 db in noisy conditions. The best recognition rates are 74.29% and 77.14%. A. Abulimiti and T. Schultz developed an ASR system for the low-resource language, Uyghur, by using different sizes of target language (Uyghur) data in multilingual training. Multilingual training utilising full Uyghur speech corpus (10 hours) and 17 hours of Turkish speech corpus produced the greatest results for the ASR system, with a WER of 19.17 percent [5]. Ishwar and Gayadhar introduced a strategy for increasing children's ASR performance under mismatch acoustic situations by normalising the influence of pitch variabilities [6]. The authors in [7] addressed the various challenges using transfer learning from adult's acoustic models to children's acoustic models in the ASR system developed using Deep Neural Network (DNN) for children. System evaluation was done with the help of multiple children's speech corpuses with a large vocabulary. Evaluations of the ASR system are present-ed on (i) comparisons of DNN Models with the earlier GMM-HMM models. (ii) Standard adaption approaches' effectiveness in transfer learning, (iii) Different variations present in the speech of the children's, in terms of acoustic spectral variations, linguistic constraints and pronunciation variations. Authors analyzed that a large proportion of adaption data is required for the system to operate better at different ages of children. Jyoti

Authorized licensed use limited to: Zhejiang University. Downloaded on March 14,2024 at 10:41:22 UTC from IEEE Xplore. Restrictions apply.

Guglani and A. N. Mishra [8] developed a Punjabi speech recognition system using Kaldi toolkit. The feature vectors from the Punjabi continuous speech signal are extracted using MFCC and PLP methods. he performance of the ASR system employing the tri3 model was better than the tri2 model, while the performance of the tri2 model was better than the tri1 model. Further, it was found that accuracy of the ASR system using MFCC features is higher as compared to the PLP features for the continuous Punjabi speech. For a speech synthesis system (HTS) built with HMM, Kiran Reddy et al. suggested an effective method for collecting pitch information from input voice signals. In the proposed pitch extraction method pitch estimation and voice detection is calculated with the help of the mean signal. CMU Arctic and keele speech databases are used for evaluating the performance of the proposed system. The quality of voice synthesis using the pitch estimation method proposed by the authors in [9] is clearly superior to HMM-based speech systems produced using the Robust Algorithm for Pitch Tracking (RAPT). Hussein [10] built a voice recognition system using SPHINX-IV for Arabic voice recognition. For system testing and training, three Arabic corpuses were created: HQC-1 - Holly Qura'an Corpus (18.5 hours), CAC-1 - command and control corpus (1.5 hours), and ADC - Arabic digits corpus (less than 1 hour). The experimental results indicate that the proposed system accomplished an accuracy of 70.8 %, 98.1 %, and 99.2 % using three different corpora. Hay Mar Soe Naing et al. [11] discussed the issue of robustness in the recognition of children speech using ASR system. Authors proposed the use of nonlinear rectification function and shape of filterbank in the acoustic feature extraction process for children ASR system. The ASR system is built using the Kaldi toolbox, and the speaker adaptative acoustic model is trained using MLLT, LDA, and feature-space Maximum Likelihood Linear Regression (fMLLR) features. After analysing authors concluded that the performance of GFCC feature extraction process is better than the MFCC and BFCC in different noisy conditions and various SNR levels.

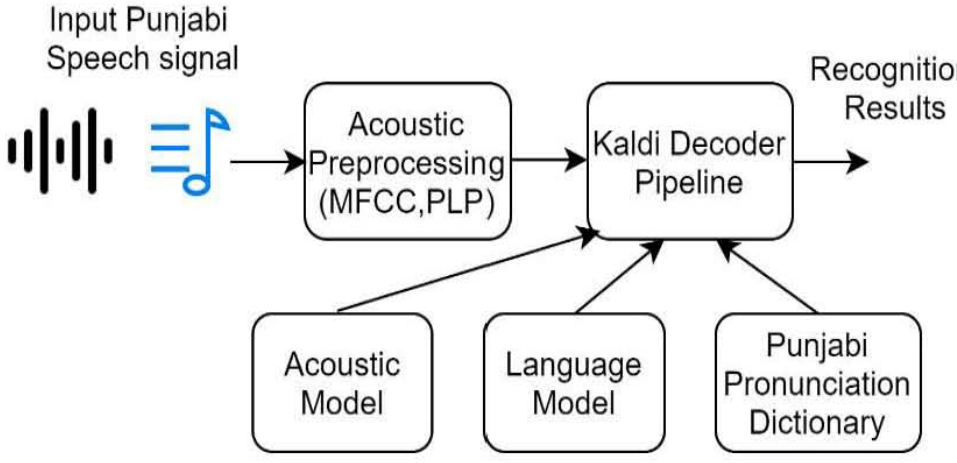

Figure 1: Automatic Speech Recognition (ASR) Framework for Punjabi Language

III. Architecture of Automatic Speech Recognition (ASR) system

Figure 1 depicts the suggested architecture for the speech recognition system.

A. Feature Extraction

Feature extraction is the process of extracting features or to find properties of utterance such as vocal track configuration, power and pitch from the speech signal. It also contains the process of measuring signal characteristic such as energy or frequency response i.e., signal measurement and signal parameterization. So, analysis of speech signals into coefficients is known as feature extraction process. Features from speech signal are to be extracted by using a number of methods. Feature extraction methods are

- PLP - Perceptual linear prediction
- MFCC - Mel Frequency Cepstral Coefficients

Now we have provided a comprehensive description of the various steps involved in MFCC feature extraction.

- Pre-emphasis- is a system process that increases the magnitude of some higher frequencies in relation to the magnitude of other lower frequencies within a frequency band in order to improve the overall signal-to-noise ratio.
- Framing- The signal must be divided into smaller time frames. The standard size of the frame is 25 milliseconds. For a 16kHz signal, this indicates that the frame length is 0.025*16000 = 400 samples [12] [13].
- Windowing- To taper the signal towards the frame boundaries, a hamming window is applied. Hamming window is used in this work. While performing the DFT on the signal, this is done to boost the harmonics, smooth the edges, and lessen the edge effect.
- Each windowed frame is converted into a magnitude spectrum using DFT. The frequency spectrum is determined on each frame using a NN-point FFT, also known as a Short-Time Fourier-Transform (STFT), where NN is commonly 256 or 512
- Applying Mel spectrum Filter Banks- The Mel spectrum is calculated by passing the Fourier converted signal

Authorized licensed use limited to: Zhejiang University. Downloaded on March 14,2024 at 10:41:22 UTC from IEEE Xplore. Restrictions apply.

through a Mel-filter bank of band-pass filters and is a set of 20–40 triangular filters.

- Take Log- The log filterbank energies are then calculated using the log of these spectrogram values.
- Discrete cosine transforms (DCT) - The problem with this spectrogram is that the coefficients of the Filter banks are extremely linked. As a result, these coefficients must be decorrelated. As a result, the DCT is used. The algorithm used to compute PLP features is similar to the MFCC algorithm.

B. Acoustic Modeling

An acoustic model defined as a reference model is a file which contains statistical representation of different sounds that make up a word of a particular language. For every statistical representation a label is assigned called a phoneme. It makes use of audio recordings to produce statistical representations [14]. A large database of speech (called a speech corpus) is required to build an acoustic model and use of different algorithms to create statistical representation for each sound (Phoneme). Acoustic models can be of two kinds (1) Word model and (2) phoneme model. Common models used for acoustic modeling are

• Hidden Markov Model (HMM)

• Deep neural networks (DNNs)

HMM is one of the most common technique for acoustic modeling. Deep neural networks (DNNs) as acoustic models significantly improved ASR system performance. In general, DNN's discriminative power is used for phoneme recognition, while HMM is preferred for decoding task.

C. Language Modeling

Language model is used to assign probability to a sequence of words P (W) by probability distribution. Thus, accurate value probability of a word P (W) is produced by using language model. Large vocabulary ASR systems are developed by using language model. The structural constraints present in the language are used to construct the probability in a language model.

Using equations 1 the speech recognizer recognizes more appropriate words [15].

$$\hat{w} = arg\max_{w} P(w \mid x) \tag{1}$$

The preceding equation is rewritten using Bayes' rule as

$$\hat{w} = \arg\max_{w} \frac{P(X|W)P(w)}{P(x)} \tag{2}$$

Classifier searches for a sequence of words W, X is an acoustic observation, so as to maximize the numerator of the above equation.

IV. Experimental setup

As stated in Table 1, the corpus for ASR system training was created using voice data obtained from individuals who spoke Punjabi language. For testing the ASR system, two different Test_Malwa, and Test_Majha speech corpuses were developed from the speakers of the Malwa and Majha region. All speech files were originally in the microphone channel and sampled at 16kHz. The speech data was also contaminated with noise. Noise was introduced from a variety of contexts, including schools, cars, streets, and restaurants, with signal to noise ratios of 20dB 15dB, 10dB, and 5dB. The MFCC and PLP features are used to train the acoustic model. The Kaldi ASR toolkit [16] [17] was used to create all of the acoustic models. Using MFCC and PLP, we first created GMM-HMM based acoustic models for each dialect. To generate 40 dimensions vectors, a Linear Discriminative Analysis (LDA) transformation is used, followed by a MLLT transformation. We also used the fMLLR approach to apply speaker adaptation [18] [19]. The speech corpus used to train the GMM model for each dialect was also used in DNN- HMM acoustic modelling.

Table 1: Dialect specific testing speech corpus

Dataset	# of speakers		Dialect Spoken	Total Duration (hours)	# of Utterances
	Male	Female			
Train_Punj	25	18	Punjabi (All dialects)	6.2	830
Test_Malwa	8	6	Malwai	1.7	390
Test_Majha	7	5	Majhi	1.4	320

V. Experimental Results

The experimental findings used to evaluate the developed ASR system are presented in this section. The ASR's performance is measured using the Word Error Rate (WER) statistic.

WER= (S (Number of Substitution) + I (Number of Insertion) + D (Number of deletions))/TW (Total Words) (3)

A. Punjabi ASR system performance evaluation of the Malwai Dialect

The results of the ASR system's recognition tests using the Malwai dialect are present-ed in this section. Table 2 demonstrates the system's WER when Pitch features are used at the front end, as well as MFCC and PLP feature extraction algorithms. Ac-cording to the observations, it was found that MFCC features outperform PLP features on the Kaldi toolkit while employing the 3-g Language model (LM). The finest result for the GMM modelling system is 33.75 %, while the WER for the DNN-HMM acoustic modelling system is 23.25%.

Table 2: WER (%) results for the Malwai dialect based - ASR system for MFCC and PLP

Authorized licensed use limited to: Zhejiang University. Downloaded on March 14,2024 at 10:41:22 UTC from IEEE Xplore. Restrictions apply.

Acoustic Model	Testing Dialect	Word error rate (%)			
		Pitch			
		2 g		3 g	
		MFCC	PLP	MFCC	PLP
GMM	Pun_Malwa	35.2	36.09	33.75	34.36
DNN-HMM		26.2	30.2	23.25	25.27

B. *Punjabi ASR system performance evaluation of the Majha Dialect*

The recognition results of the ASR system employing the Majha dialect are shown in this section. Table 3 demonstrates the system's WER leveraging Pitch features and the front end's MFCC and PLP feature extraction algorithms. System Provides 28.33 % WER for Majha Dialect using PLP, 3 g LM and DNN-HMM acoustic modelling. Similarly, we achieved best results of 25.91 % WER for Majha Dialect when the system was evaluated with the Pitch features and MFCC in the front end.

Table 3: WER (%) results for the Majha dialect based - ASR system for MFCC and PLP

Acoustic Model	Testing Dialect	Word error rate (%)			
		Pitch			
		2 g		3 g	
		MFCC	PLP	MFCC	PLP
GMM	Pun_Majha	38.2	39.2	36.21	37.05
DNN-HMM		27.85	28.96	25.91	28.33

Figure 3 shows the bar graph presenting the word error rate using 3-gram Language model for Majha and malwa Punjabi dialect.

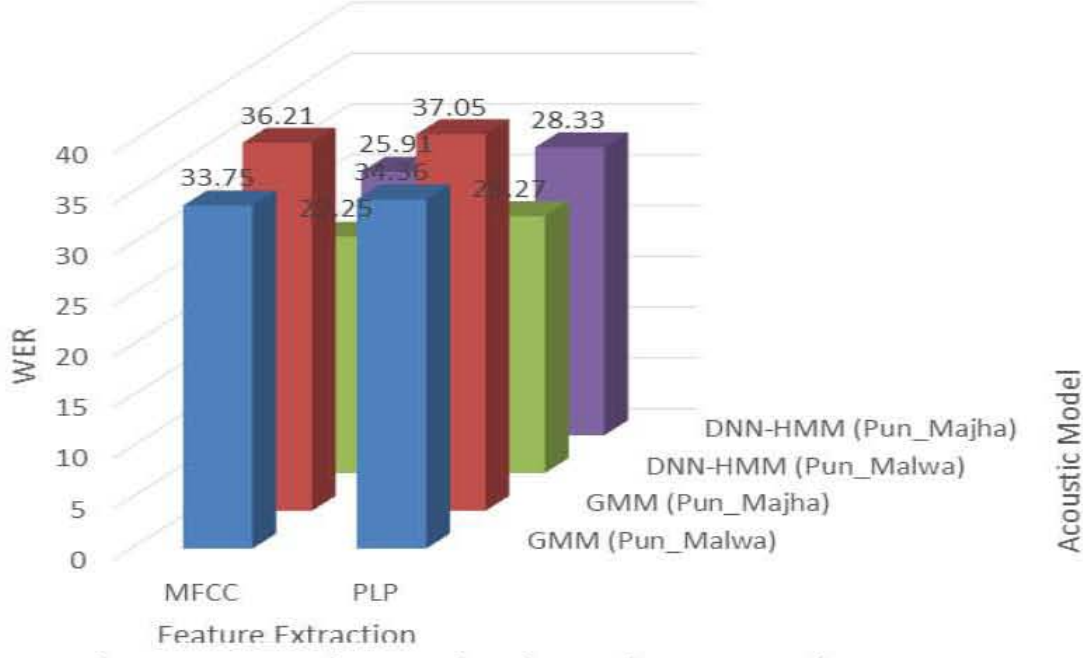

Figure 2: Bar graph presenting the word error rate using 3 gram Language model for Majha and malwa Punjabi Dialect

VI. Conclusion

This research describes the development of dialect-based speech recognition system for Punjabi language. MFCC and PLP feature extraction techniques are used to extract features and Kaldi toolbox is employed to construct the ASR system. The Kaldi Pitch features were also used to improve the recognition of Punjabi dialectal tonal speech. As per the findings, MFCC works out better for Punjabi dialectal speech than the PLP features extraction approach. GMM as well as DNN-HMM are used to create acoustic models. We also used the fMLLR approach to apply the speaker adaptation. Despite the short corpus size used in our approach, we were able to reach a WER of 23.25 percent and 25.91 percent for Malwa and Majha dialects using Pitch features and DNN-HMM acoustic modeling approach.

In the future, the size of the speech corpus will be increased for improving the speech recognition rate of the Punjabi dialectal ASR system.

References

1. C. Li and Y. Qian, "Prosody usage optimization for children speech recognition with zero resource children speech," Proc. Annu. Conf. Int. Speech Commun. Assoc.

Authorized licensed use limited to: Zhejiang University. Downloaded on March 14,2024 at 10:41:22 UTC from IEEE Xplore. Restrictions apply.

INTERSPEECH, vol. 2019-Septe, pp. 3446–3450, 2019, doi: 10.21437/Interspeech.2019-2659.
2. J. Guglani and A. N. Mishra, “Automatic speech recognition system with pitch dependent features for Punjabi language on KALDI toolkit,” Appl. Acoust., vol. 167, p. 107386, 2020, doi: 10.1016/j.apacoust.2020.107386.
3. S. Shahnawazuddin, K. T. Deepak, G. Pradhan, and R. Sinha, “ENHANCING NOISE AND PITCH ROBUSTNESS OF CHILDREN’ S ASR Department of Electronics and Communication Engineering, NIT Patna, India Department of Electronics and Communication Engineering, IIIT Dharwad, India Department of Electronics and Electrical En,” pp. 5225–5229, 2017.
4. M. Hamidi, H. Satori, O. Zealouk, and K. Satori, “Amazigh digits through interactive speech recognition system in noisy environment,” Int. J. Speech Technol., vol. 23, no. 1, pp. 101–109, 2020, doi: 10.1007/s10772-019-09661-2.
5. A. Abulimiti and T. Schultz, “Automatic Speech Recognition for Uyghur through Multilingual Acoustic Modeling,” no. May, pp. 6444–6449, 2020.
6. I. C. Yadav and G. Pradhan, “Significance of Pitch-Based Spectral Normalization for Children’s Speech Recognition,” IEEE Signal Process. Lett., vol. 26, no. 12, pp. 1822–1826, Dec. 2019, doi: 10.1109/LSP.2019.2950763.
7. P. Gurunath Shivakumar and P. Georgiou, “Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations,” Comput. Speech Lang., vol. 63, no., p. 101077, 2020, doi: 10.1016/j.csl.2020.101077.
8. J. Guglani and A. N. Mishra, “Continuous Punjabi speech recognition model based on Kaldi ASR toolkit,” Int. J. Speech Technol., vol. 21, no. 2, pp. 211–216, 2018, doi: 10.1007/s10772-018-9497-6.
9. M. K. Reddy and K. S. Rao, “Robust Pitch Extraction Method for the HMM-Based Speech Synthesis System,” IEEE Signal Process. Lett., vol. 24, no. 8, pp. 1133–1137, 2017, doi: 10.1109/LSP.2017.2712646.
10. H. Hyassat and R. Abu Zitar, “Arabic speech recognition using SPHINX engine,” Int. J. Speech Technol., vol. 9, no. 3–4, pp. 133–150, Oct. 2006, doi: 10.1007/s10772-008-9009-1.
11. H. M. S. Naing, Y. Miyanaga, R. Hidayat, and B. Winduratna, “Filterbank Analysis of MFCC Feature Extraction in Robust Children Speech Recognition,” 2019 Int. Symp. Multimed. Commun. Technol. ISMAC 2019, 2019, doi: 10.1109/ISMAC.2019.8836181.
12. V. Bhardwaj, S. Bala, V. Kadyan, and V. Kukreja, “Development of Robust Automatic Speech Recognition System for Children’s using Kaldi Toolkit,” pp. 10–13, 2020, doi: 10.1109/icirca48905.2020.9182941.
13. Taniya, V. Bhardwaj, and V. Kadyan, “Deep Neural Network Trained Punjabi Children Speech Recognition System Using Kaldi Toolkit,” in 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), 2020, pp. 374–378, doi: 10.1109/ICCCA49541.2020.9250780.
14. V. Kukreja, D. Kumar, A. Kaur, Geetanjali, and Sakshi, “GAN-based synthetic data augmentation for increased CNN performance in Vehicle Number Plate Recognition,” in 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1190–1195, doi: 10.1109/ICECA49313.2020.9297625.
15. G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” IEEE Trans. Audio, Speech Lang. Process., vol. 20, no. 1, pp. 30–42, 2012, doi: 10.1109/TASL.2011.2134090.
16. D. Povey et al., “The Kaldi Speech Recognition Toolkit,” IEEE Signal Process. Soc. 1–4, 2011, Accessed: Jul. 19, 2020. [Online]. Available: http://kaldi.sf.net/.
17. V. Bhardwaj and V. Kukreja, “Effect of pitch enhancement in Punjabi children’ s speech recognition system under disparate acoustic conditions,” Appl. Acoust., vol. 177, p. 107918, 2021, doi: 10.1016/j.apacoust.2021.107918.
18. S. Takaki, S. Kim, and J. Yamagishi, “Speaker Adaptation of Various Components in Deep Neural Network based Speech Synthesis,” 9th ISCA Speech Synth. Work., vol. 2016, pp. 153–159, 2016, doi: 10.21437/ssw.2016-25.
19. V. Bhardwaj, V. Kukreja, and A. Singh, “Usage of Prosody Modification and Acoustic Adaptation for Robust Automatic Speech Recognition (ASR) System,” Rev. d’Intelligence Artif., vol. 35, no. 3, pp. 235–242, 2021.

Authorized licensed use limited to: Zhejiang University. Downloaded on March 14,2024 at 10:41:22 UTC from IEEE Xplore. Restrictions apply.