IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Review On Developing Speech Recognition System For Dogri Language

[1]Shiv Kumar

[1]Assistant Professor

[1]Department of Computer Application

[1]Government Degree College Boys Kathua, Jammu and Kashmir, India

Abstract: The voice interface is a new buzz in the market and applications such as Google Assistant, Cortona, Alexa, and Siri making it popular. While talking about product design, we talk about inclusive design and that inclusive design is not possible if we do not consider the low resource languages or regional languages such as Dogri, Kumauni, etc. This paper examines the present status of the Dogri language in the context of creating a Speech Recognition System (SRS). It also discusses the work done in various Indian regional languages. The overview of the speech recognition system and challenges to building the Dogri language speech recognition is discussed. This paper discusses the various techniques for developing SRS systems such as the conventional approach (Acoustic/Phonetic), Pattern Recognition (Artificial Intelligence/ machine learning (AI/ML), HMM/GMM) approach, and hybrid approach.

Index Terms **-** Acoustic-Phonetic, Artificial Intelligence, HMM/GMM, Speech Recognition System, Voice Interface.

I. INTRODUCTION

"The speech recognition system is a domain of Natural Language Processing" [1]. In the last few years, speech recognition has made tremendous progress, with services like Google Voice Search covering over 100 languages. Academics and businesses alike are eager to see it expand its scope to include other languages from around the globe. The digital signal processor and digital programmable logic devices have greatly improved, making speech processing systems a viable option for voice recognition, control, and communication systems. Speech processing systems must have a low computational complexity in order to be commercially successful. Improved digital processors and reduced memory chip costs have made speech processing systems increasingly popular for voice recognition as well as communication.

Humans communicate with each other through speech, which is the most common form of human language. There are several different components to speech, including time, frequency, and amplitude, all of which fluctuate. Depending on the frequency band, transitions might occur at different periods. A Speech Recognition System (SRS) is a piece of software designed to extract, recognize, and translate speech features utilizing a computerized smart device (SRS). It is the primary goal of SRS to build technology that enables humans to interact with machines in our natural language in a real-time environment independent of vocabulary quantity, noise, speech features, or even accents. It is possible to use SRS for word recognition in isolation, word recognition in conjunction with other words, or continuous speech recognition. It has applications ranging from language development in young children, telecommunication areas, for people with impaired hearing [2].

Dogri is a Pahari language, albeit the Dogra people do not generally use this nomenclature. To them, Pahari means "mountain" in its original sense. Locals call Dogras further up in the mountains "Pahari people" who speak "Pahari language." In linguistic terms, Pahari refers to a group of Indo-Aryan languages spoken in the lower Himalayas from Nepal to Kashmir. The Dogri language has recently developed fast as a literary language. It was added to the Indian Constitution's eighth schedule in December 2003[3].

Very little work has been done for Indian languages compared to non-Indian languages [4]. The amount of work in Indian regional languages has not yet reached a critical level to be used as a real communication tool, as already done in other languages in developed countries. Dogri is a Northern Indo-Aryan language spoken in Jammu and Kashmir. Isolated word speech recognition in native languages spoken in different parts of India is an ongoing topic of research. Speech Recognition Systems are used in various application areas such as Education, Agriculture, Smart homes, etc. To develop such applications and open opportunities to the J&K technology sector, there is a need for having a speech recognition system for Dogri Language. This paper has five sections. Section 1 is an introduction to the Dogri language and speech recognition system. Related work in recent years is discussed in Section 2 Related work. The overview of the speech recognition system which is necessary to understand before implementing any speech recognition system is described in Section 3. Section 4 discusses the challenges while building the speech recognition system. The paper is concluded with Section 5 conclusion.

II. RELATED WORK

Virender Kadyan et al. [5] did a comparative study of the Punjabi Automatic Speech Recognition (ASR) system. They have used MFCC and GFCC feature extraction techniques and HMM-GMM and HMM-DNN for the classification. The experiment was performed on a continuous and connected Punjabi speech corpus. The experimental setup demonstrates a significant improvement of 4–5 percent and 1–3 percent for connected and continuous speech corpus with respect to HMM-DNN over HMM-GMM.

Ravindra P. Bachate and Ashok Sharma compared various approaches such as KNN, SVM, ANN, DNN, and DBN for developing Marathi SRS [6]. The MFCC algorithm was used for the feature extraction process. They have compared various classification approaches using a confusion matrix. The results show that the Deep Belief Network approach works better compared to the KNN, SVM, ANN, and DNN. They have listed one limitation of DBN is that DBN computation cost is more compared to other listed classification approaches.

There are issues with HMM such as the number of states found by Samira Hazmoune et al. during the implementation of the Speech Recognition System [7]. For each type of data, they propose employing a different set of HMMs, each with a different number of states and then using the KNN rule to determine the K nearest models and the most represented class by comparing Viterbi likelihood. The UCI Spoken Arabic Digit dataset is used to test the suggested architecture. Whether compared to the HMM and KNN baselines or prior efforts on the same dataset, the acquired findings indicate the usefulness of there technique.

Jyoti Guglani and A. N. Mishra implemented Punjabi ASR System using Kaldi framework [8]. They have used Mel frequency cepstral coefficients (MFCC) features and perceptual linear prediction (PLP) algorithms for the feature extraction process. N-gram language model-based automated speech recognition (ASR) systems for both monophonic and triphonic models are described. The word error rate (WER) of an ASR system was used to measure its performance (WER). When comparing the ASR model with the tri phone model, a significant drop in WER was found. Using the tri3 model, ASR performance improves over that of the tri2 model, and using the tri2 model improves over that of the tri1 model. For continuous Punjabi speech, the MFCC feature outperforms PLP features in terms of speech recognition accuracy.

M. Kalamani et al. [9] proposed a Tamil language continuous ASR system. They have proposed FCM with the EM-GMM approach for the Tamil ASR system. The performance of the proposed approach is measured using WER. When compared to existing algorithms in various noisy situations, the suggested approach enhances recognition accuracy by reducing WER from
1.6 to 5.47 percent, increasing it from 1.2 to 4.4 percent.

Prashant Upadhyay et al. [10] proposed a Kaldi-based Hindi ASR system. Using the Kaldi automatic speech recognition toolkit, they have suggested Context-Dependent Deep Neural- network HMMs (CD-DNN-HMMs) for Hindi speech with a huge vocabulary. Compared to a standard triphone model, experiments on the AMUAV database show that CD-DNN-HMMs outperform the more traditional CD-GMM-HMMs model in terms of word mistake rate by 3.1 percent.

The transfer learning for children from an adult has been proposed by Prashanth Gurunath Shivakumar and Panayiotis Georgiou [11]. They have implemented and compared GMM-HMM and DNN approaches for developing the SRS system. The experiment performed for transfer learning proved that DNN gave good results over the GMM-HMM approach.

The Marathi SRS system has been implemented using the HTK toolkit by Supriya and Dr. Handore [12]. They have used the MFCC algorithm for feature extraction and the HMM-GMM approach for pattern recognition. But the issue with the research work is, they have used only 3 male and 3 female speakers with 910 sentences which is not sufficient the perform the experiments.

III. Overview of Speech Recognition System

To develop any speech recognition system, it is important to understand the various aspects of the Speech Recognition System such as states of speech, classification of speech recognition systems, techniques used to develop speech recognition systems and so on which is shown in Figure 1. This section discusses above-mentioned point to get the overall view of the development of the speech recognition system.

3.1 States of Speech

People's primary means of exchanging ideas and information is speech and language. Speech is the ability to express ideas and thoughts through the use of articulate vocal sounds, or the ability to do so.

- Silence: Silence which is defined as the absence of any speech.
- Unvoiced: A type of speech in which the vocal cords do not vibrate, resulting in a speech waveform that is either periodic or random in character.
- Voiced: In this the vocal cords are strained and, as a result, vibrate periodically as air is expelled from the lungs, resulting in a quasi-periodic waveform in the ensuing speech waveform.

The primary objective of any speech recognition system is to identify the state of speech which helps to recognize the speech.

3.2 Classification of Speech Recognition Systems

The classification of the speech recognition systems has been done based on the context of the use of the SRS system. “The SRS is mainly classified into two types – Speaker dependent and speaker-independent [4]”. The approaches used in the design of various SRS systems varied, as do the applications for which they are used. The majority of SRS systems are intended for personal use, whereas speaker-independent systems are developed for broad usage.

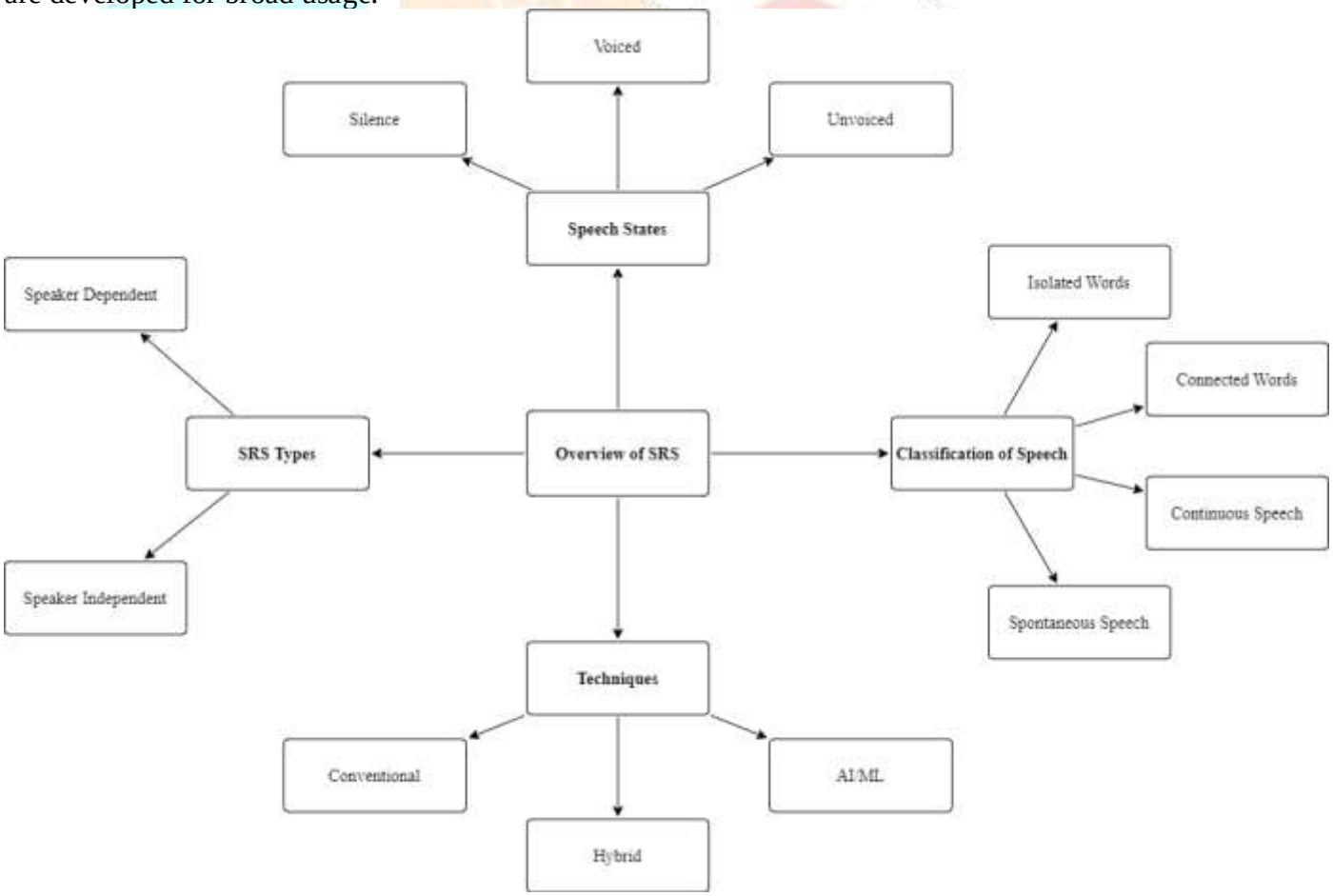

Figure 1: Overview of Speech Recognition System

3.3 Techniques used to develop Speech Recognition Systems

The techniques of speech recognition are classified into main three types - a) Conventional Approach (Acoustic/Phonetic) b) Pattern Recognition (AI/ML, Template-based, HMM/GMM), and Hybrid Approach.

A. Acoustic-Phonetic Approach:

Acoustic phonetics argues that there are a limited number of distinct phonetic units in spoken language, according to the theory. The phonetic units are distinguished by a variety of acoustic features that can be seen throughout the duration of or within the spoken signal. It is the first stage of the acoustic phonetics technique that the spectrum analysis of speech and features identification, which transforms spectral data into a collection of characteristically broad-based acoustic characteristics of the various phones, performs. In the subsequent stage, the speech signal is divided into stable acoustic areas, and one or more phonetic labels are assigned to each segmented area, resulting in a characterization of the phoneme lattice of the Speech. Following the phonetic label segmentation and labeling, the final step in this methodology is to identify a legitimate word (or word string) from the sequences generated by this methodology.

B. Pattern Recognition Approach:

In the pattern recognition approach, before pattern recognition, the speech data pre- processing and feature extraction process has been done. The pre-processing deals with digitizing the speech samples and reducing the noise present. Later the features are extracted from speech samples using various algorithms such as MFCC, GFCC, Spectral features, etc. [13]. After the feature extraction process, Principle Component Algorithm (PCA) was applied to select the useful features from the available.

Then various pattern recognition techniques such as Artificial Neural Network (ANN). Deep Neural Network (DNN), HMM-GMM, Template-based, etc. [14]. Nowadays, researchers are preferring to use the neural network approach as a pattern recognition approach as it is proved as the best approach for implementing speech recognition systems and it is becoming popular.

C. Hybrid Approach:

There are many research articles where the hybrid approach for developing a speech recognition system is used. This approach combines the acoustic-phonetic and pattern recognition approach by adopting the best things from both these approaches [15][12].

3.4 Challenges in Developing Dogri Speech Recognition System

Many challenges remain unsolved in the speech recognition system, which has been under development since 1920. Here are some things that can affect your ability to speak clearly.

a) Language Resource Availability:

For building any SRS system, it is important to have sufficient corpus for training and testing the developed system. That speech corpus must be collected by considering different aspects of speech such as different environmental conditions, geographical locations, different genders, a variety of age groups, etc. For low resource language such as the Dogri, we can implement speech synthesis to create the speech corpus artificially, but genuine speech corpus can give us more accuracy compared to the corpus generated using speech synthesis. There are various techniques such as Wavenet is used for implementing raw speech using the speech synthesis technique [16]

b) Speech Type:

A variety of characteristics, including voice tone, accents, speaking tempo, voice pitch, and the production of phonemes, are all involved in the development of a particular speech style. Voice timbre, ranging from normal to screamed. For the same language, the attention shifts from person to person and location to location. Phonemes are generated in accordance with the language in which they are spoken. Depending on the type of speech, words might be separate, linked, continuous, or spontaneous. ASR systems are often unable to decipher the words because of voice pitch changes.

c) Environment:

“One of the most difficult obstacles to speech recognition is the environment. It is possible that the environment is represented by background noise, room acoustics, and channel conditions” [17]. These settings introduce noise and signal interference to the audio.

d) Speaker characteristics:

Variability in speech is dependent on the speaker's characteristics. They include things like the speaker's age, sex, and articulation variation. A speaker's mental state, tension, emotions, etc. are all factors in articulation diversity.

To develop an ASR system for a certain language, it is necessary to use different approaches because each language has a unique structure. Each language has its unique grammatical and phonetic statements.

IV. CONCLUSION

This paper has discussed the importance of the Dogri language Speech Recognition System for inclusive design. Here, the different research work on the other language SRS is discussed to understand the different techniques and pros and cons of the techniques. The overview of the SRS gave an idea about the things which must consider while developing the Dogri SRS. The challenges, including low resource for developing the SRS system has been discussed.

REFERENCES

[1] R. P. Bachate and A. Sharma, "Acquaintance with Natural Language Processing for Building Smart Society," E3S Web Conf., vol. 170, p. 02006, 2020.
[2] N. D. Londhe, "Recognition for Chhattisgarhi," 2018 5th Int. Conf. Signal Process. Integr. Networks, pp. 667–671, 2018.
[3] S. Dutta and B. Arora, "Parts of Speech (POS) Tagging for Dogri Language," Lect. Notes Networks Syst., vol. 203 LNNS, pp. 529–540, 2021.
[4] R. P. Bachate and A. Sharma, "Automatic speech recognition systems for regional languages in India," Int. J. Recent Technol. Eng., vol. 8, no. 2 Special Issue 3, pp. 585–592, 2019.
[5] V. Kadyan, A. Mantri, R. K. Aggarwal, and A. Singh, "A comparative study of deep neural network based Punjabi-ASR system," Int. J. Speech Technol., vol. 22, no. 1, pp. 111–119, 2019.
[6] R. P. Bachate and A. Sharma, "Comparing different pattern recognition approaches of building marathi asr system," Int. J. Adv. Sci. Technol., vol. 29, no. 5, pp. 4615–4623, 2020.
[7] S. Hazmoune, F. Bougamouza, S. Mazouzi, and M. Benmohammed, "A new hybrid framework based on Hidden Markov models and K-nearest neighbors for speech recognition," Int. J. Speech Technol., vol. 21, no. 3, pp. 689–704, 2018.
[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.
[9] M. Kalamani, M. Krishnamoorthi, and R. S. Valarmathi, "Continuous Tamil Speech Recognition technique under non stationary noisy environments," Int. J. Speech Technol., vol. 22, no. 1, pp. 47–58, 2019.
[10] P. Upadhyaya, S. K. Mittal, O. Farooq, Y. V. Varshney, and M. R. Abidi, "Continuous Hindi Speech Recognition Using Kaldi ASR Based on Deep Neural Network," Adv. Intell. Syst. Comput., vol. 748, pp. 303–311, 2019.
[11] P. Gurunath Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," Comput. Speech Lang., vol. 63, 2020.
[12] S. Supriya and S. M. Handore, "Speech recognition using HTK toolkit for Marathi language," in IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, ICPCSI 2017, 2018, pp. 1591–1597.
[13] F. S. Cabral, H. Fukai, and S. Tamura, "Feature extraction methods proposed for speech recognition are effective on road condition monitoring using smartphone inertial sensors," Sensors (Switzerland), vol. 19, no. 16, 2019.
[14] S. K. Saksamudre and R. R. Deshmukh, "A Review on Different Approaches for Speech Recognition System," no. September, 2015.
[15] S. Huang and S. Renals, "Hierarchical Bayesian language models for conversational speech recognition," IEEE Trans. Audio, Speech Lang. Process., vol. 18, no. 8, pp. 1941–1954, 2010.
[16] A. van den Oord et al., "WaveNet: A Generative Model for Raw Audio," pp. 1–15, 2016.
[17] R. P. Bachate and A. Sharma, "Automatic speech recognition systems for regional languages in India," Int. J. Recent Technol. Eng., vol. 8, no. 2 Special Issue 3, pp. 585–592, Jul. 2019.