AN ANALYSIS OF NOISE ROBUST TECHNIQUE FOR HMM BASED PUNJABI SPEECH RECOGNITION

Ms. Mandeep kaur
Department of computer science & engineering
Ghuraun ,Punjab
chandigarh university
kambojmandeep30@gmail.com

Navjot Singh
Department of computer science & engineering
Ghuraun ,Punjab
chandigarh university
navjotsingh49900@gmail.com

Abstract— In today's ASR engines, a set of various features extraction and modeling classifier approaches are used. Traditional front end techniques-LPCC, PNCC faces the challenges of performance degradation due to acoustic mismatch conditions. On the other end acoustic classifiers-HMM, GMM, SGMM tackle the issue of training medium vocabulary Punjabi continuous corpora. An extensive study is done to cope with these factors. Various front and back end approaches are analyzed with different baseline and hybrid methodologies. This paper tries to fulfill the gap of actual corpus performance in comparison to synthetic speech corpora. In this paper we analyze variation in acoustic mismatch and modeling information are performed using MFCC, GFCC noise robust approach at front end .

Keywords— ASR, HMM, GFCC, SGMM ,MFCC ,GMM-HMM.)

I. INTRODUCTION

Speech plays a crucial role in making better communication between a machine and a human. To process an input speech signal a human can easily adapt as per signal conditions but in case of a machine its performance degraded. An ASR (Automatic Speech Recognition) based system processes spoken utterance tries to maps them into a corresponding text.

The fundamental process of speech recognition is defined as an ability of a machine to produce output in machine readable form corresponding to a spoken utterance. An ASR structure is reliant on upon two modules, "training" and "testing". An exactness of the ASR scheme is affected by various factors such as environment, distortion, speaker differences, reverberation, etc.

The system framed through these techniques analyze the person's unique voice sample and use it to fine tune that recognizes the person's speech, resulting in more accurate transcription.

II. PUNJABI LANGUAGE

Punjabi language is individual of the 22 official languages. It is spoken by over 100 million people worldwide and around 90 million peoples reside in part of Punjab in India and Pakistan. Punjab is a territory that was divided by the Britishers during the 1947 partition between India and Pakistan. Rest of the Punjabi speaking people are spread in countries including Malaysia, South Africa, the United States, United Kingdom, United Arab Emirates, and the Canada.

In India, Punjabi is the state regional language of Punjab state. One of the problems that are faced in Punjabi language is various geographical regions and the variation of dialect in those regions. This dialectal variation affects language in many ways one of which is change in pronunciation of various words. Punjabi is written in two different scripts one is Gurumukhi and other is Shahmukhi.

III. BASIC STRUCTURE OF ASR

An ASR system mainly consists of two stages: training and testing. A training stage further consists of feature extraction at front end, acoustic feature classification (modeling phase), pronunciation model and language modeling phase at the back end. The second stage employs speech feature vector generation at the front end and knowledge generation through decoding/search phase at the back end as shown in

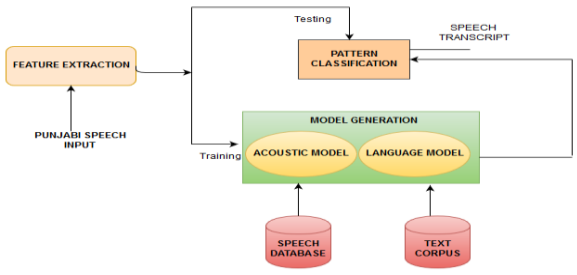

Fig 1 Basic Structure of ASR

Feature Extraction: In feature extraction phase, an input speech signal is processed to generate the non-redundant and unique feature vectors. The extracted feature vectors contain the capability to differentiate each of them from similar feature vectors and also include discrimination properties. Processed vectors provide the knowledge of phonetic class and remove the part of the information that doesn't contain linguistic knowledge (W).

It discards the other characteristics of the signal such as speaker or environment. These feature vectors are further used for classification and knowledge generation purpose in acoustic modeling and decoding phase of the speech system. There are several ways to extract features from a voice signal. Several of these include-

- "Mel Frequency Cepstral Coefficients (MFCC)"
- Gammatone frequency cepstral coefficient(GFCC)

IV. DIFFERENCE BETWEEN MFCC AND GFCC

Broadly speaking, there are two major differences. One of the most used methods for converting signals from the time area to the occurrence area is called GFCC. With the use of a Gammatone filter bank, it uses a cepstral analysis procedure to extract the robustness of the spoken speech. Pre-processing is a phase that is added to the MFCC process that aids in eliminating DC and its first order transformation utilising pre-emphasis. Additionally, it aids in the analysis of the short-term Fourier transformation using the Hamming window or by substituting a different multi-taper spectrum estimation function. Then, a triangular filter is used to conduct the auditory spectral investigation with the Mel filter bank. as shown in figure 2.

"Mel-frequency cepstral coefficients (MFCCs)": A petite-period control range of a wide-ranging is characterized by an MFC, or mel-occurrence cepstrum which are produced from a nonrectilinear "spectrum-of-a-spectrum" cepstral representation of the audio sample.

"Gammatone frequency cepstral coefficients GFCCs": Coefficients based on a collection of Gammatone filter banks are known as gamma-tone frequency cepstral coefficients.

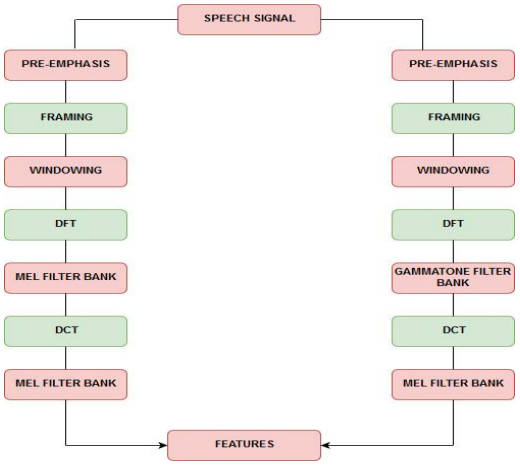

Fig 2 Flow diagram of MFCC and GFCC feature extraction technique

V. LANGUAGE MODEL

Using a Language Model (LM), one may anticipate the word that will be said next by using probability for a word sequence. Silence and filler words described in a filler lexicon are not included in LM. An adjacent group of n words from a particular voice or text sequence makes up an n-gram LM..

VI. ACOUSTIC MODEL

The collected feature vectors are used as an input for categorising auditory data. The categorization technique also aids in the test sample's improved identification of spoken utterances. Modern speech recognition systems rely on the HMM approach for acoustic modelling, which uses the probability distribution of an audio vector. The addition of auditory data is an attempt to better reflect them than HMM states.

To improve performance recognition, the HMM is hybridised with GMM or SGMM. Each state space

of an input voice is modelled using multivariate GMM when GMM is merged on HMM.TRANSCRIPTION FILE

The portrayal of any language in writing or the mapping of audible speech onto written words is known as transcription. Each waveform's information is displayed as text in voice recognition. The transcription file's contents for the Punjabi language are seen in Figure 5.7.

Waveform transcription is created by putting the appropriate words first, followed by the waveform name. The procedure is manual.

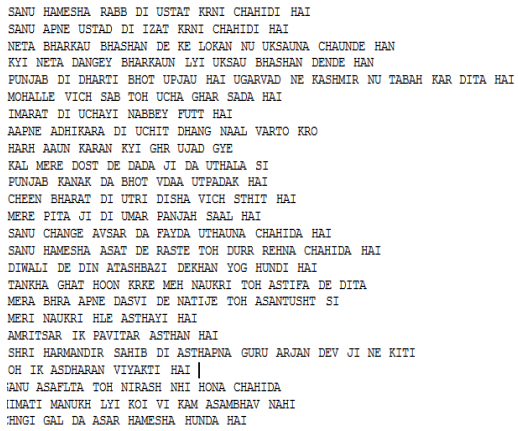

Fig 3 Transcription file

The Lexicon-Dictionary file separates words into the sub-words that are contained in the acoustic model and provides the word pronunciations. Any spoken term that is not recognised or included in a dictionary is referred to as being out of vocabulary (OOV).

It's possible that a single word might have more than one pronunciation. They are distinguished in that situation by a special parenthesis. For instance:

SANU S AE N UW
HAMESHA HH AE M AH SH AH
RABB R AE B
DI D IY
KARNI K AA R N IY

The P-ASR system's experimental findings are reported in this section.. The performance of speech recognition systems is usually specified in terms of accuracy and speed .In this we used native speakers in clean environment .Total numbers of speakers 9 corresponding male and female.The word error is calculated

"Word Error Rate(WER) =(S+D+I)/TW"

Experiment Number	Dataset	MFCC (WER) (%)	GFCC(WER) (%)
Experiment 1	1. Native Speakers 2. Total Speakers = 2(1 M and 1 F) 3. Clean environment	7.11	13.51
Experiment 2	1. Native Speakers 2. Total Speakers = 3 (2 M and 1 F) 3. Clean environment	10.51	23.11
Experiment 3	1. Native Speakers 2. Total Speakers = 4 (2 M and 2 F) 3. Clean environment	13.41	26.68

VII. CONCLUSION

In this work, speaker independent ASR system for Punjabi language has been developed using Kaldi toolkit. MFCC and GFCC are used for Training of large vocabulary system. The proposed System has been implemented with simple Sentences spoken by (5 male and 4 female) 9 Punjabi speakers **.** clean environment, GFCC Provides maximum WER of 26.68% .

REFERENCES

[1.] Bawa P, Kadyan V (2021) Noise robust in-domain children speech enhancement for automatic Punjabi recognition system under mismatched conditions. Appl Acoust 175:107810

[2.] López G, Quesada L, Guerrero LA (2017) Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of

speech-based natural user interfaces. International conference on applied human factors and ergonomics. Springer, Cham, pp 241–250

[3.] Hoy MB (2018) Alexa, Siri, Cortana, and more: an introduction to voice assistants. Med Ref Serv Q 37(1):81–88

[4.] Kumar A, Aggarwal RK (2021) An exploration of semi-supervised and language-adversarial transfer learning using hybrid acoustic model for Hindi speech recognition. J Reliable Intell Environ. https://doi.org/10.1007/s40860-021-00140-7

[5.] Shivakumar PG, Georgiou P (2020) Transfer learning from adult to children for speech recognition: evaluation, analysis and recommendations. Comput Speech Lang 63:101077

[6.] Kumar M, Kim SH, Lord C, Lyon TD, Narayanan S (2020) Leveraging linguistic context in dyadic interactions to improve automatic speech recognition for children. Comput Speech Lang 63:101101

[7.] Leibold LJ, Buss E (2019) Masked speech recognition in school-age children. Front Psychol 10:1981

[8.] Müller T, Speck I, Wesarg T, Wiebe K, Hassepaß F, Jakob T, Arndt S (2019) Speech recognition in noise in single-sided deaf cochlear implant children using digital wireless adaptive microphone technology. Laryngorhinootologie 98(S 02):10859

[9.] Shahnawazuddin S, Bandarupalli TS, Chakravarthy R (2020) Improving automatic speech recognition by classifying adult and child speakers into separate groups using speech rate rhythmicity parameter. In: 2020 International Conference on Signal Processing and Communications (SPCOM). IEEE, pp. 1–5

[10.] Kumar A, Aggarwal RK (2021) Bi-lingual TDNN-LSTM acoustic modeling for limited resource hindi and marathi language ASR. Advances in speech and music technology. Springer, Singapore, pp 409–423

[11.] Shahnawazuddin S, Sinha R (2015) Low-memory fast on-line adaptation for acoustically mismatched children's speech recognition. In: Sixteenth annual conference of the international speech communication association

[12.] Koehler J, Morgan N, Hermansky H, Hirsch HG, Tong G (1994) Integrating RASTA-PLP into speech recognition. In: Proceedings of ICASSP'94. In: IEEE international conference on acoustics, speech and signal processing, vol 1. IEEE, pp. I-421

[13.] Kadyan V, Bawa P, Hasija T (2021) In domain training data augmentation on noise robust Punjabi Children speech recognition. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-021-03468-3

[14.] Zhao X, Wang D (2013) Analyzing noise robustness of MFCC and GFCC features in speaker identification. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, pp. 7204–7208

[15.] Kim C, Stern RM (2016) Power-normalized cepstral coefficients (PNCC) for robust speech recognition. IEEE/ACM Trans Audio Speech Lang Process 24(7):1315–1329