Speaker Independent Isolated Word Speech to Text Conversion Using Auto Spectral Subtraction for Punjabi Language

Manjot Kaur, Lakhvir Garcha

North West Institute of Engineering and Technology

Abstract- Speech is one of the important requirements and the well-situated method of communication between people. Real time speech to text is an accurate conversion of uttered words after speaking .STT is very useful tool to interact with people in counselling interviews or conference .Thus the conversion of speech to written language must be accurate and fast so that it can be easily understand by people**.** The fundamental approach of this paper is to develop an algorithm to convert speech to text using Punjabi phonetics. This paper introduces and discusses two popular and different noise reduction techniques (Auto Spectral Subtraction, LBG, MFCC) and presents our simulation result of a noise reduction system. It is shown that the system reduces the noise almost completely by finding the minimum Euclidean distance and keeps the enhanced speech signal very similar to the original speech signal. This paper presents a method to design a speech to text conversion module using JAVA. This method is simple to use and takes less use of memory space.

IndexTerms-AutoSpectralSubtraction,LindeBuzoGrayalgorithm, Mel Frequency Cepstral Coefficient.

I. INTRODUCTION

Speech is the most important form of communication in everyday life in order to make the interaction easier and faster. Speech is the vocalized form of human communication and it is considered to be the primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech .Automatic Speech Recognition provides a path for natural communication between man and machine. A simple alternative to a hardware interface is a software interface i.e. a Speech to Text system. .Speech to Text Conversion or Speech Recognition allows a computer to identify the words that a person speaks into a mike or any other similar hardware and convert it into written words. Basically, the mode of communication between humans takes place in several ways such as facial expressions, gestures, eye contact and speech. Speech to text conversion is very advantageous and used in various applications areas. Human interact with each other in several ways such as facial expression, eye contact, gesture, mainly speech. The speech is primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech. The recognition of speech is one the most challenges in speech processing. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means of Algorithm implemented as a computer program. Speech to text conversion (STT) system is distinguished into two types, such as speaker dependent and speaker independent systems. The main difficulties in implementation of an ASR system are due to different speaking styles of human beings and environmental disturbances. So the main aim of an ASR system is to transform a speech signal into text message independent of the device, speaker or the surroundings in an accurate and efficient manner.

On the basis of way to recognize speech recognition may be Isolated word recognizes utterance to have quiet on both side of sample windows i.e. only one word at a time and word is preceded and followed by silence. This is having "Listen and Non Listen state". Connected word system are same as isolated words but recognizes speech having one or more than one word and these words are divided or separated by small sound to be "run together minimum pause between them. Continuous speech recognizers allows user to talk almost naturally. Thus system recognizes more than one word and words are connected without any silence.

Spontaneous Word speech Recognition: This system recognizes speech that is natural sounding and not be rehearsed. An ASR System with impulsive speech should be able to handle a dissimilar words and mixture of natural speech feature such as words being run together like ums, ahs and others. Motive of Spontaneous Word Speech Recognition is to recognize natural speech.

This paper gives a description of implementation of Speech to Text Conversion System using Auto Spectral technique. The system goes through different steps to accomplish the task of speech to text conversion that are signal preparation, acoustical analysis, Training and Testing. For the purpose of improving accuracy of the system, the system uses a noise reduction technique named Auto Spectral Subtraction. Auto Spectral Subtraction [8] is a simple and efficient noise reduction technique. In this technique, an average signal spectrum and average noise spectrum are estimated in parts of the recording and subtracted from each other, so that average signal-to-noise ratio (SNR) is improved.

The algorithm for the design of optimal VQ is commonly referred to as the Linde-Buzo-Gray (LBG) algorithm, and it is based on minimization of the squared-error measure.

II. Related Work

This section of paper will represent literature review of the works which is similar to the presented work.

An implementation of the Punjabi Speech to text system for connected words has been discussed by Mohit Dua[1]. Hidden Markov model toolkit (HTK) has been used to develop the system. A Java platform based Graphical User Interface (GUI) has been developed to make the system fast and user friendly. The implemented system performs well with Word Recognition Rate (WR) 95.8% and 95.4%, Word Accuracy Rate (WA) 94.1% and 91.6% and Word Error Rate (WER) 5.9% and 8.35% in class room and open environment respectively. The recognition results show that the system performs well with different speakers and in different environments.

Wiqas Ghai analysed Automatic Speech Recognition(ASR) for Indo-Aryan languages and the applicability of techniques applied for other languages so that a concrete work can be initiated for Punjabi language[2].To enhance the performance standards set for other languages, it has been observed that use of techniques Cooperative Heterogeneous ANN Architecture, Maximum Likelihood Linear Regression, Extended MFCC and Learning Vector quantization are helping the researchers to get improved recognition performance of speech recognition systems. So far the work done for Punjabi language is Isolated word speech recognition using Acoustic template matching technique on MATLAB.

Hidden Markov Model (HMM) based Punjabi text to speech synthesis system is presented by Divya Bansal[3]in which speech waveform is generated from Hidden Markov Models themselves, and applies it to Punjabi speech synthesis using the general speech synthesis architecture of HTK (HMM Tool Kit). The developed Text-to-Speech was trained in phase -I on 17 samples with total 61 words all starting with letter ਤ and ਟ and tested for selection of appropriate phoneme sequence on 30 Punjabi words in test 1 and trained for 23 samples containing 81 words containing ਅ and ਆ and tested for 45 selected words in corresponding test-1. Hidden Markov Model Text-to-Speech system approach is very effective for developing Text-to-Speech systems for various languages and can easily implement changes in voice characteristics of synthesized speech with the help of speaker adaptation technique developed for speech recognition. In order to improve efficiency, context-dependent phone models used for synthesis need to be improvised by recording, annotating more Punjabi speech data and applying filters using custom rules/ procedures.

Prachi Khilari gives an overview of major technological perspective and appreciation of the fundamental progress of speech to text conversion and also gives overview technique developed in each stage of classification[4]. A comparative study of different technique is done as per stages and attempts to analyze an approach for designing an efficient system for speech recognition. The development of existing STT system is presented by adding spellchecker module to it for different language. A database has been created from the various domain words and syllables. The desired speech is produced by the Concatenative speech synthesis approach. The system gives the input data from mice in the form of voice, then preprocessed that data & converted into text format displayed on PC. The user types the input string and the system reads it from the database or data store where the words, phones, diaphones, triphone are stored.

This Speech- to-Text conversion system is implemented by Prasanthi using the MFCC for feature extraction and HMM as the recognizers. The system is also used to find the disorder rate of persons affected with Parkinson‟s disease by calculating the efficiency of pronunciation[5]. In audio folder, 105 audio files are recorded and these are analyzed to get feature vectors. These features are initially modeling in the HMM. Then,the test spoken word is addressed by Baum-welch algorithm of HMM. From the simulation results, it can be clearly seen that the average recognition rate of 97.14% is achieved by the number of states (N=3).

Noise Reduction Technique (Auto Spectral Subtraction) has been applied by Karun Verma[6] that improves the accuracy of the system to some extent. This paper clearly gives the comparison of the system at two different aspects. One when no noise reduction technique is applied and other when noise reduction technique is applied. HMM Based Speaker Independent Isolated Word Speech to Text Conversion System is developed. The accuracy obtained in the system is still not 100% accurate. To improve the accuracy, different noise compensation/speech enhancements techniques can be used to make the system more accurate. Accuracy of the system also depends on training process of the system.

Marc[8]analyzed noise removal from noisy speeches. The study includes methods for removing noise from noisy speeches using spectral subtraction. Noisy speech was digitally generated by corrupting the data of the clean speech “Real graph” with the data of the vacuum cleaner noise. The approximation of the noise was obtained by taking the average magnitude of the noise spectrum during non-speech activity. The average magnitude of the noise spectrum during non-speech activity was subtracted from the noisy speech spectrum during speech activity. Our initial noise removal design consisted of no frame averaging with half-overlapped data buffers and 256 points Hamming time windows. Results showed that using any combination of half-overlapped and one-fourth-overlapped data buffers with 128, 256, and 512 points, Hamming windows and three frames or six frames averaging did not improve the performance of the denoising algorithm. However, using one-fourth-overlapped data buffers with 256 points Hamming windows and no frames averaging resulted in the greatest improvement differential *SNR* in the amount of 0.3371 dB, leading to most noise removal from the noisy speech “Real graph”. Thus the goal of denoising noisy speech signals has been successfully achieved.

III. Methodology

1.System Description

The speech to text system is implemented using JAVA for noise reduction. Various techniques has been discussed for natural language processing. Fig 1 shows the GUI interface where there are two phases, one is training (initial step) and other is testing. In training panel, Spectral Subtraction method is used to filter out the acoustic noise. In testing phase speech file is

converted to text usingMel Frequency Cepstral Coefficients and Linde-Buzo-Gray algorithm.

The user first create dataset according to its speech signal and save it in train folder.After that noise from signal is removed and create its spectogram. We find minimum Euclidean distance between training and testing samples to obtain final results as text.

Thus the whole paper is presented by step to steps and gives apparent approach to text punjabi speech to text.

2.Database Preparation

Database is prepared by collecting data having distinct words that is Punjabi counting. 101-200 words have been taken for preparing of dataset.Database preparation involves two steps that is recording of speech files and then labeling of that speech files. In second set of data, Database preparation involves three steps in which first two steps are same as that of first set and third step is noise reduction in speech files.

3.Noise Reduction in Speech files

Noise reduction does not affect quality of speech. It only just lowers the intensity of noise. The system uses converter that allows transformations of many files at a same time. The system also adds speech files to converter to perform noise reduction. After adding the speech files used for database preparation to converter, the system adds command Noise Reduction (Auto Spectral Subtraction). Then converter performs noise reduction technique that is Auto Spectral Subtraction on the speech.

4.Acoustical Analysis

In Acoustical Analysis, the obtained speech files are represented in more efficient way by extracting features of speech files. The system uses the Mel Frequency Cepstral Coefficients (MFCCs) to extract features from speech files.

5.Training

After calculating the MFCCs, training phase compute vector quantization using Linde Buzo algorithm to train the sound files and save the dataset in train folder.

6.Testing

The System is tested for both sets of data used in database preparation. In testing phase, Euclidean distance is calculated and the final results are obtained from which speech is converted to text.

IV. Flowchart of Proposed work

The flowchart shown in figure 1 shows the step by step procedure of the proposed work**.**

1. To study and analyze various techniques for natural language processing.
2. To apply Auto spectral Subtraction for filtering acoustic noise. An average signal spectrum and average noise spectrum are estimated in parts of the recording and subtracted from each other, so that average signal-to-noise ratio (SNR) is improved.
3. To apply Mel Frequency Cepstral coefficient and vector quantization on using Linde-Buzo-Gray algorithm for training the sound files. The algorithm for the design of optimal VQ is commonly referred to as the Linde-Buzo-Gray (LBG) algorithm, and it is based on minimization of the squared-error measure.
4. To perform testing based on Euclidean distance and compute the final results.

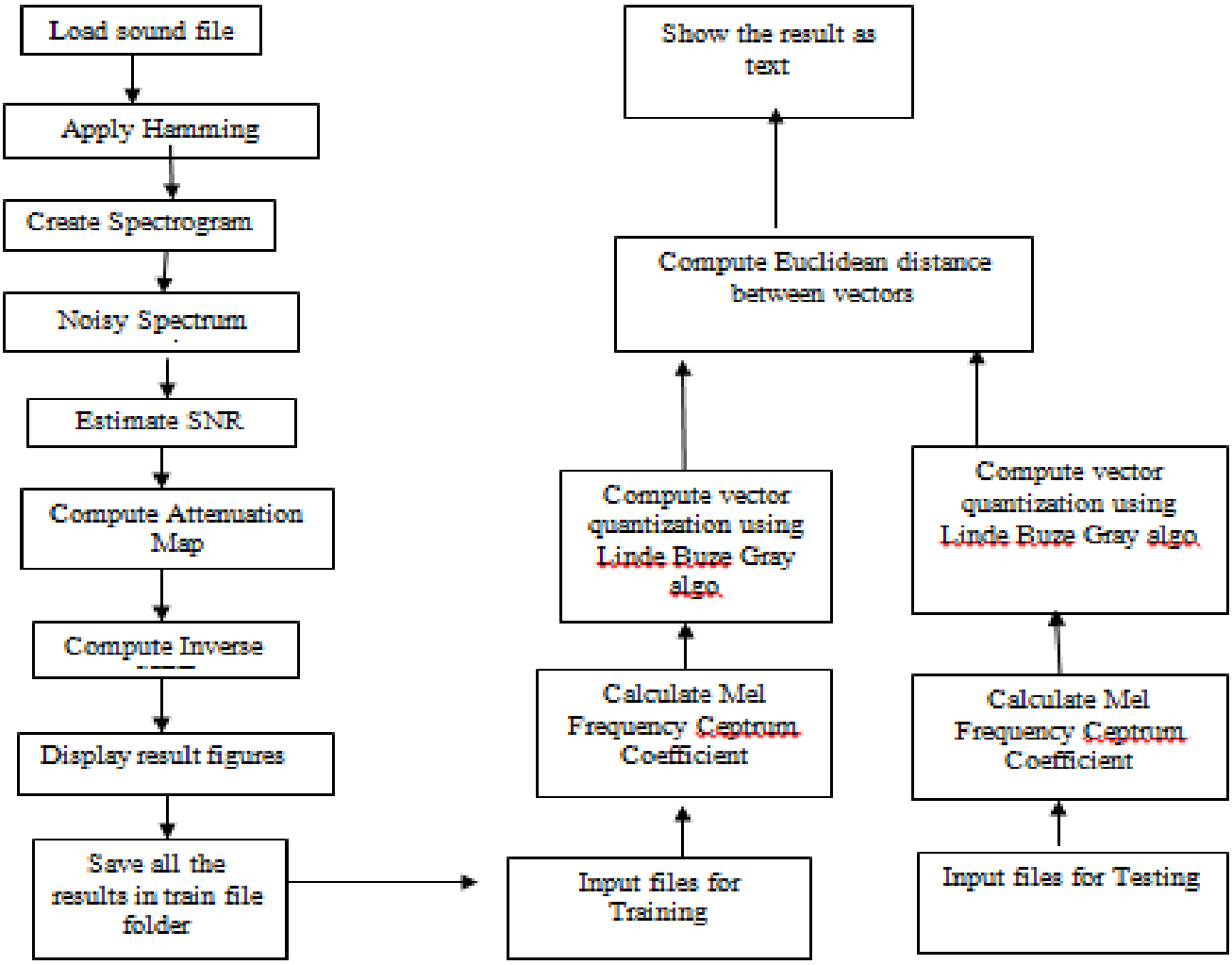

Fig.1 Flow Chart

V. RESULTS

Figure 1 shows the GUI interface of speech to text convertor.

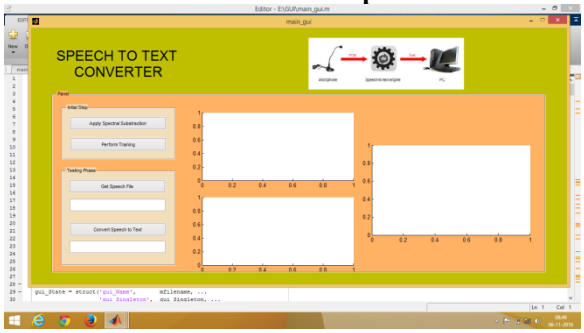

Fig 1. GUI Interface

Figure 2 results in Auto Spectral Subtraction has been applied for training panel.

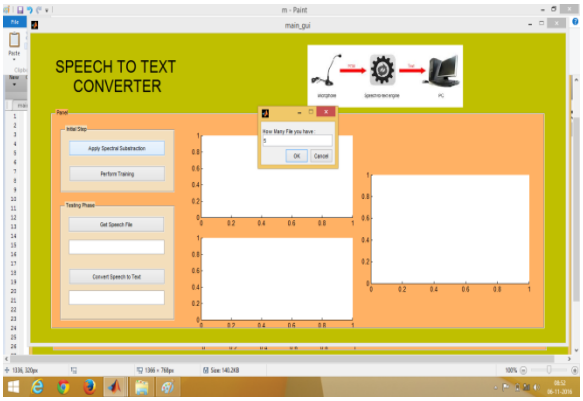

Fig 2. Apply Spectral Subtraction

Figure 3 shows the result of training phase.

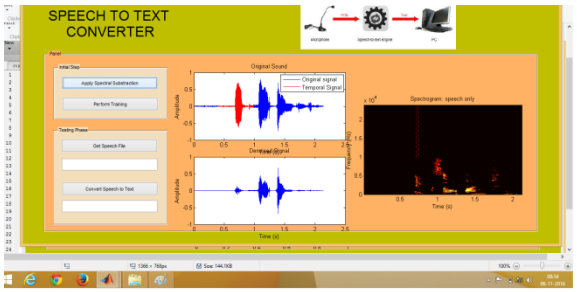

Fig 3. Result

Figure 4 shows the speech file and the result of speech file.

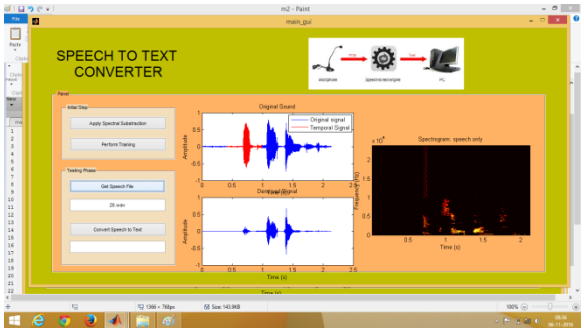

Fig 4. Get speech file and Result

The final result is shown in figure 5 in which speech is converted to text.

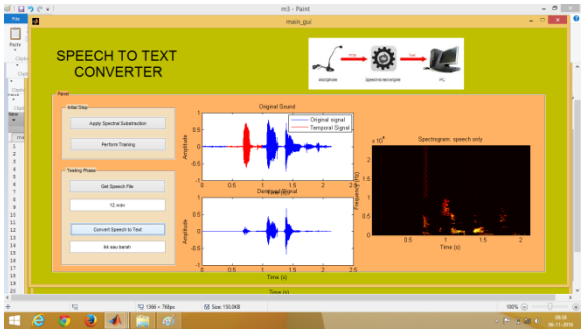

Fig 5. Convert speech to text and result

V. CONCLUSION

In this paper, we developed a algorithm to convert real time punjabi speech signal to text. This system has many applications ,it can be used by dumb and deaf to interact with other persons from society. Speech to text conversion area comes under NLP and human interface. In this paper we use different methods to make it an efficient system. We first use Auto Spectral subtraction to filter out acoustic noise and then we apply Mel Vector quantization using Linde-Buzo-Gray algorithm to perform training and testing of real time punjabi speech files according to user.

The concept not only help physically handicapped challenged people like deaf, dumb and blind but also help speakers in classrooms ,conference halls and parliament to make record of their speech in text format. This paper gives clear approach to text punjabi speech to text in step by step procedure. The user first create dataset according to its speech signal and save it in train folder .After that we remove noise from signal and create its spectogram. We find minimum Euclidean distance between training and testing samples to obtain final results as text.

REFERENCES

[1] MohitDua, R.K.Aggarwal, VirenderKadyan and ShelzaDua: "Punjabi Speech to Text System for Connected Words", pp. 349–355, (2012).

[2] WiqasGhai, Navdeep Singh "Analysis of Automatic Speech RecognitionSystems for Indo-Aryan Languages": Punjabi A Case Study,International Journal of Soft Computing and Engineering (IJSCE) ISSN: pp. 2231-2307, volume-2, issue-1, March 2012

[3] DivyaBansal, AnkitaGoel, Khushneet Jindal, "Punjabi Speech Synthesis System Using HTK",International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.4, (July 2012).

[4] "A Review on speechto text conversion methods"Miss.PrachiKhilari, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) volume 4 Issue 7, July 2012.

[5] G.P.S. Prasanthi, K. Sirisha, G. Ramya,B., "Speech to Text Conversion Using HMM",International Journal of Advanced Research in Electronics andCommunication Engineering(IJARECE),volume5, issue 3, March 2016.

[6] "Speaker Independent Isolated Word Speech to Text Conversion using HTK for Punjabi Language" ShwetaMittal ,KarunVermaComputer Science and Engineering Department, Thapar University, Patiala.

[7] "Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM)"
Su Myat Mon, HlaMyoTun,International Journal Of Scientific & Technology Research ISSN 2277-8616,vol4,issue 06, JUNE 2015.

[8] Karam, M., Khazaal, H.F., Aglan, H. and Cole, C. (2014) Noise Removal in SpeechProcessing UsingSpectral Subtraction. Journal of Signal and Information Processing, **5**, 32-41,Feb 2014.

[9] Verteletskaya, E. and Simak, B. "Speech "Distortion Minimized Noise Reduction Algorithm" Proceedings of the World Congress on EngineeringandComputer Science, Vol. I, San Francisco, 20-22 October 2010.

[10] Sim B, Tong Y, chang J., Tan C. "A parametric formulation of the generalized spectral subtraction method "IEEE. Trans. Speech Audio Process,6(4), 328-337.

[11] Jurafsky, D., Martin, J. H.: "Speech and Language Processing. Dorling Kindersley Pvt. Ltd., India. Name of chapter: HMMs and Speech Recognition", 261-309,2000.

[12] Young, S. J., Woodland, P.C., Byrne, W. J.: "Spontaneous speech recognition for the credit card corpus using the htk toolkit. IEEE transaction on Speech and Audio Processing". Volume 2, 615-621, Oct 1994.

[13] Hain, T., Woodland, P. C., Evermann, G., Povery, D.: "New features in the cu-htk system for transcription of conversational telephone speech" ,57-60,2001.

[14] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Transactions on Acoustics, Speech, and Signal processing, vol. 32, no. 6, pp. 1109-1121, December 1984.

[15] Hu Y., Bhatnager M. Loizou P. “A crosscorellation technique for enhancing speech currupted with correlated noise”,Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing1.pp 673-676, 2001.

AUTHORS

First Author – ManjotKaur, North West Institute of Engineering and Technology

Second Author – LakhvirGarcha, North West Institute of Engineering and Technology