International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-4 Number-3 Issue-16 September-2014

Speech Recognition Technology for Hearing Disabled Community

Tanvi Dua[1], Jitesh Punjabi[2], Chandni Sajnani[3], Aditya Advani[4], Shanthi Therese S.[5]

Abstract

As the number of people with hearing disabilities are increasing significantly in the world, it is always required to use technology for filling the gap of communication between Deaf and Hearing communities. To fill this gap and to allow people with hearing disabilities to communicate this paper suggests a framework that contributes to the efficient integration of people with hearing disabilities. This paper presents a robust speech recognition system, which converts the continuous speech into text and image. The results are obtained with an accuracy of 95% with the small size vocabulary of 20 greeting sentences of continuous speech form tested in a speaker independent mode. In this testing phase all these continuous sentences were given as live input to the proposed system.

Keywords

Speech Recognition, Mel-Scale Frequency Cepstral Coefficients (MFCC), Vector Quantization (VQ)

1. Introduction

Speech is the primary and efficient mode of communication between humans. Several applications based on the speech signals exist like speech modification, speech coding, speech enhancement, speech recognition and speaker identification, spoken dialogue processing. speech perception, emotions in speech, Phonetics etc.

Manuscript received September 17, 2014.

Tanvi Dua was graduated as a Bachelor of Engineering in Information Technology from Thadomal Shahani Engg. College She is currently working as a Jr.Software Engineer in BNP Paribhas India Solutions.

Jitesh Punjabi, graduated as a Bachelor of Engineering in Information Technology from Thadomal Shahani Engg. College .He is currently working as a Software Engineer in Capgemini Consulting Pvt Ltd.

Chandni Sajnani Thadomal Shahani Engineering College.

Aditya Advani Thadomal Shahani Engineering College.

Shanthi Therese S. Thadomal Shahani Engineering College.

Based on discrete-time models of speech production, many developments are available on the design of speech analysis and speech synthesis systems. In analysis, speech waveform is converted into acoustic features and in synthesis, based on the parameter estimates, the waveforms are put back to form a speech [1][2].

Overview and Related Work

Dragon is a speech recognition software package developed by Dragon Systems of Newton [3]. Microsoft developed Genie speech recognition software. It is an interactive speech software which enables the user to give commands by speech. AT&T developed navigational software which helps to browse internet through various speech commands [4]. There are different classifications of speech recognition systems exist.

Isolated or Continuous Speech:

Based on the mode of speech, it is classified as Isolated or Continuous speech. Isolated systems recognize the given word. Modeling isolated speech recognizer is comparatively simpler than modeling a continuous speech recognizer. In continuous recognizers, the word sequence (sentences) is built using small units (words). The complexity in building continuous speech recognition system is the representation of all possible inputs.

Speaker dependent versus Speaker Independent system:

In speaker dependent systems, the recognizer is trained to identify a single user whose training samples are used in the construction of speech recognition. Speaker Independent systems must identify wide range of speakers. Speaker dependent system is more accurate than speaker independent since large variability exist in different speakers according to their pronunciation, accent, nativity and other tonal variations.

Small versus large vocabulary systems:

Small vocabulary systems are less than 1000 words. Mid-size systems extend upto few thousands whereas large vocabulary systems can handle many thousands of words.

In this paper, the goal is to assist people with hearing disabilities to communicate. This is possible when their group members are motivating the hearing disabled person to take active part in the interaction by means of human machine interaction. In this way, they can communicate with other people in the outside world of hearing community. This work consists of two parts. The first part transforms speech to text using Automatic Speech Recognition (ASR) which is speaker independent. The second part converts speech into image which will help the hearing disabled community to have a better understanding of the sentences which is spoken into the system. Thus here both audio and visual communication is incorporated. In this project, greeting sentences in continuous word mode of four different speakers are used to develop the speech recognizer.

2. Objectives

Various tasks involved in developing this project are

a) Developing grammar consisting of small to medium size vocabulary using English language. Indian accent (ELIA)
b) Word List consists of 20 greeting sentences which include 32 different words in ELIA.
c) Creating a Picture/ Image database of every sentence.
d) Simple user interface to enable people (both normal Hearing and Hearing Disabled to interact with the system.

3. Proposed Solution

The basic methods that will be used to achieve the project functionalities are explained in this section.

3.1. Speech Recognition and Speech-to-Text (STT) conversion

The ASR engine contains two modules:

a) Speech to Text module that converts the input speech into text.
b) ASR module that recognizes speech and display an appropriate image [5].

ASR is the mapping of input sound wav file into its corresponding text. There are three basic steps in the ASR. They are acoustic parameter estimation, acoustic parameter comparison and final decision making. Different types of feature extraction techniques are Mel Frequency Cepstrum Coefficients (MFCC), Linear Predictive Coding (LPC), Perpetual Linear Predictive Coding (PLP), RelAtive SpecTrA (RASTA)

In this work, to extract the features from the given speech file, most popular MFCC feature extraction technique is used[6]-[8]. The speech signal is assumed to be the output of a Linear Time-Invariant filters (LTI) system. In ASR, fundamental frequency and details of glottal pulse are not important for distinguishing different phones. Instead the most useful information for phone detection is the exact position and shape of the vocal tract. To separate the source and vocal tract parameters efficient mathematical way is cepstrum. The cepstrum is defined as the inverse DFT of the log magnitude of DFT of the speech signal. The mel frequency is a linear frequency below 1000 HZ and a logarithmic spacing above 1000Hz. Given a frequency f in Hz [9][10]

$$\text{mel}(f) = 2595 * \log_{10}(1 + f/700) \quad \ldots\ldots\ldots (1)$$

The acoustic vectors can be used to recognize the speech characteristics. In the speech recognition areas, extracted acoustic features are mapped generally using DTW, HMM and Vector quantization approach. Deshmukh S.D., Bachute M.R. [11], proposed a speech and speaker recognition using MFCC, HMM and VQ. In this work, an approach to the recognition of speech signal using frequency spectral information with Mel frequency for the better representation was tested.Vector Quantization (VQ) is the fundamental and most successful technique used in speech coding, image coding, speech recognition, and speech synthesis, speaker recognition and many pattern recognition applications[12][13]. To make use of vector quantization process, Frames of speech signal are considered as points in a two dimensional space. The classical K-means algorithm based on Euclidean distance is used to construct the code vectors. The Euclidean metric [14] is commonly used because it fits the physical meaning of distance or distortion.
In the training phase, a speech-specific VQ codebook is generated for each known speech by clustering his/her training acoustic vectors. The distance from a vector to the closest codeword of a codebook is called a VQ distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the total distortion is computed. The speech corresponding to the VQ codebook with the smallest total distortion is considered as the recognized output [15].

3.2. Display the appropriate Image

The database consists of image files. For every sentence in the database, there is a corresponding image file. This image file is retrieved along with the text and displayed into the application running on a Hearing Disabled person's device.

3.2.1. Design of ASR System

In speech recognition, the initial step is for the user to speak a word or a phrase as an input to the microphone. The given audio signal through the microphone is digitized by an analog to digital converter and is stored in the memory. To determine the meaning of this voice input, the device attempts to match the input with the digitized voice sample that has been previously saved in wav format. The program contains the input template and attempts to match this template with the actual input using a simple conditional statement. Similarly, since each person's voice is different, the program cannot contain a template for each potential user. So the program must be trained to recognize the words independent of the potential speaker. Fig. 1 Shows the working model of the proposed ASR system.

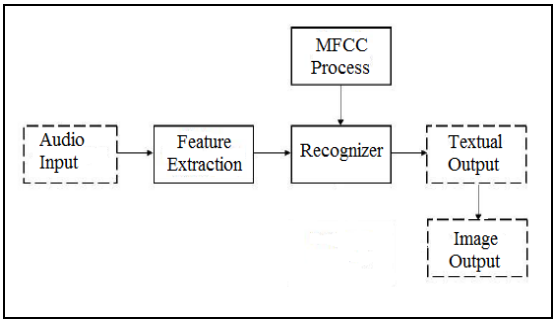

Figure 1: Working Model of proposed ASR

MFCC extraction process includes the steps as shown in Fig 2. The first stage in MFCC feature extraction is to boost the amount of energy in the high frequencies. The goal of feature extraction is to extract spectral features that can help us to build phone or sub phone classifiers. To achieve this hamming window is generally used. Next step is to extract spectral information for windowed signal. The tool for extracting spectral information is DFT. In the next step the frequencies output by the DFT has to be mapped to the mel scale. This spectrum has to be converted into cepstrum by finding inverse DFT of the log magnitude of the DFT of a signal.

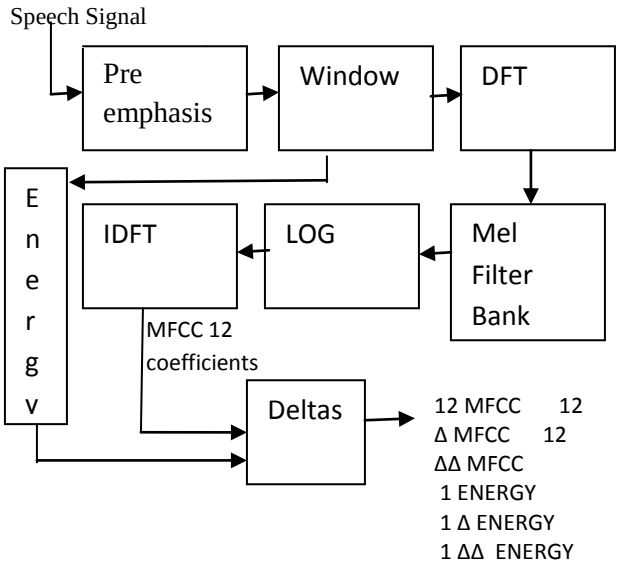

Figure 2: MFCC Extraction Process

The extraction of the cepstrum results in 12 cepstral coefficients for each frame. 13th feature energy in a frame is the sum over time of the power of the samples in the frame is calculated to find 13 MFCC coefficients.

In the training phase, based on the above 39 MFCC features of each speaker for each sentence, code vectors are created by using k-means clustering algorithm. Acoustic modeling using K-Means, cluster the feature vectors.

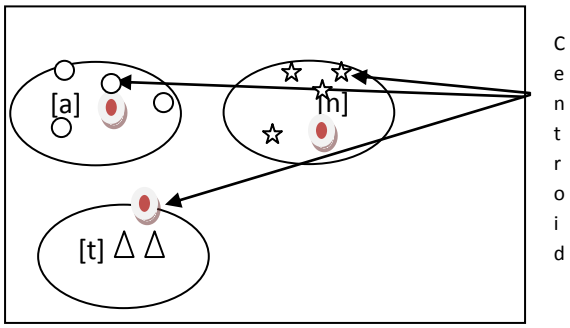

Figure 3: Acoustic Modeling using K-Means

The speech database of 20 sentences of four different users recorded using PRAAT speech analysis tool. [16] All the speech files are recorded at 16 KHz sampling rate using Mono channel.

The system uses MATLAB to efficiently store and retrieve sound files (.wav format).In addition to

displaying the sound id and name, the system also displays corresponding image of the text for effective communication.

3.2.2. User Interface Design

i) Main Menu

Figure 4 shows the plot of speech signal and train folder details of the system.

ii) Addition of Sound File to the database

To add a sound file into the database the user is first asked to enter the sound id, sound name and then the sound duration. Later on the user is supposed to enter

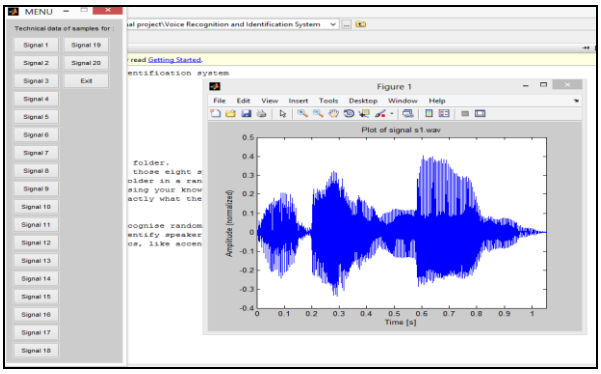

Figure 4: Technical Information of Training files

sample which are normally recommended by the system. The recording starts for the number of seconds mentioned by the user. In that interval of time, the user is supposed to speak into the system through a microphone. Recording stops and the sound gets added to the database.

iii) Final Output

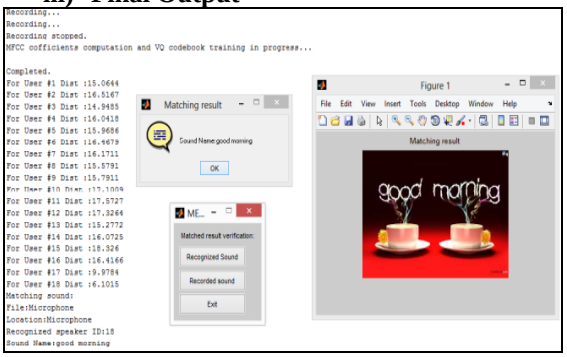

Figure 5: Output of Text and Image

For getting this final output, the user is first asked the duration of time he/she wants to speak into the system. Recording starts and the user inputs his/her voice. Recording stops after the mentioned interval of time and the system calculates the minimum distance/ coefficients out of all the coefficients calculated for every word using MFCC and Vector Quantization. The word with the minimum distortion, based on the input and recorded patterns is selected as the speech recognized. The text corresponding to the feature vector is displayed to the user and the corresponding image file is displayed.

4. Results and Discussions

Each sentence in the database was tested individually for ten different utterances. All these speech utterances were tested during run time. Each greeting sentence is tested by four different speakers. Recognition rate shows the accuracy based on number of times greeting sentences got recognized correctly. Any naïve user is allowed to use the system since it is speaker independent. For untrained users also, the system gives satisfactory performance. Table 1 shows individual sentences and its recognition accuracy. Graph 1 shows the sentence recognition count

1. Speech Recognition Rate

Table 1: Recognition Rate and Accuracy

S.No	Word	Jitesh	Aditya	Tanvi	Chandni	Recognition Rate
1	Good Evening	9	9	10	10	95%
2	Good morning	10	10	10	10	100%
3	Good Bye	10	10	10	10	100%
4	Good day	10	9	9	10	95%
5	Good Luck	10	10	10	10	100%
6	Good Night	9	10	9	10	95%
7	Happy New Year	10	10	10	9	97.5%
8	Hello	10	10	10	10	100%
9	Hey	10	10	10	10	100%
10	Hi	10	10	10	10	100%
11	How Are You	10	10	10	9	97.5%

12	Merry Christmas	10	9	10	8	92.5%
13	Pleased to Meet You	10	9	10	9	95%
14	See You	10	10	10	10	100%
15	See You Later	9	9	9	10	92.5%
16	Sweet Dreams	10	10	10	9	97.5%
17	Thank You	10	9	8	9	90%
18	Welcome	10	9	7	9	87.5%
19	Well Done	10	10	10	10	100%
20	Whatsup	10	10	10	10	100%

2. Graphs

a) Sentence Recognition Count

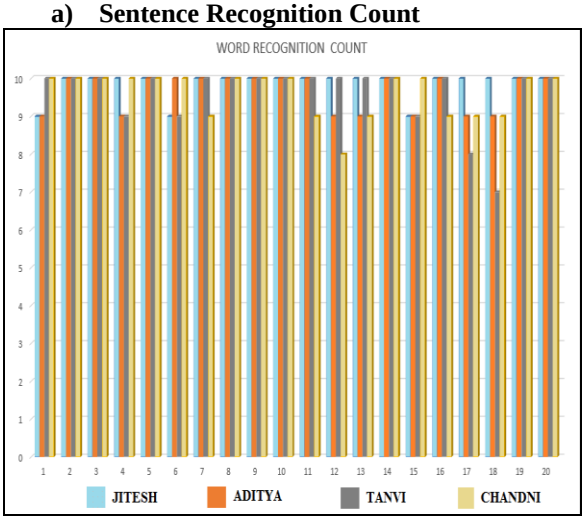

Figure 6: Sentence Recognition Chart

5. Conclusions and Future Work

The main aim of this proposed work was to recognize continuous speech using MFCC and Vector quantization. The speech recognizer recognize the correct sentence and the relevant image was retrieved from the image database. The feature extraction was done using Mel Frequency Cepstral Coefficients (MFCC). The extracted features were stored in a .mat file using MFCC algorithm. A distortion measure based on minimizing the Euclidean distance was used when matching the unknown speech signal with the speech signal database. The experimental results were analyzed with the help of MATLAB.

Some of the various challenges we have faced in this project are as follows.

First, the Automatic Speech Recognition system may not understand the person's pronunciation; the person should speak clearly in a proper English language using Indian accent with a certain appropriate accent. Database challenges: We started with a limited dictionary of most common greeting words. The system showed accuracy of 100% for first 5-7 sentences. As the size of the database was increased up to 12-15 sentences, the accuracy was 95-97%. The accuracy was 90-95% when the database was extended up to 25 sentences. When the database includes more than 50 sentences, the accuracy of the system is a challenging one.

We can add many advanced features in the application for future work such as:

1. Making this application global by appending different languages. This will be a great improvement for this project.
2. Adding well-known vocabulary words and slang to the list of words in the database.Adding a feature of 'text-to-speech conversion'
3. Improving the efficiency and accuracy by increasing the volume of database and the system may also find application in the field of Robotics.

References

[1] Lawrence Rabiner and Biing-Hwang Juang, "Fundamental of Speech Recognition", Prentice-Hall, Englewood Cliffs, N.J., 1993.
[2] Thomas F. quatieri, "Discrete-Time Speech Signal Processing: Principles and Practice", Prentice Hall; 1 edition (November 8, 2001).
[3] http://www.dragonsys.com.
[4] http://www.speech.be.philips.com/index.htm.
[5] Ohood Al-Amoudi, Rouda Al-Kuwari, Sara Qaffaf, Nada Aloraidi, HebaDawoud, Muneera Al-Marri, Tarek El-Fouly, Amr Mohamed, "Assistive Technology for People with Hearing/Speaking Disabilities in Qatar", IEEE Multidiscipilinary Engineering Education Magazine, VOL. 6, NO. 4, Dec 2011.
[6] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk, Speech Recognition using MFCC, International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) pp 135 – 138 , July 28-29, 2012.
[7] Anjali Bala, Abhijeet Kumar, Nidhika Birla Voice Command Recognition System Based on MFCC AND DTW, International Journal of

Engineering Science and Technology, Vol. 2 (12), pp 7335-7342, 2010.
[8] Shivanker Dev Dhingra, Geeta Nijhawan, Poonam Pandit, "Isolated Speech Recognition using MFCC and DTW", IJAREEIE, Vol. 2, Issue 8, August 2013.
[9] Vibha Tiwari, MFCC and its applications in speaker recognition, International Journal on Emerging Technologies vol 1 issue 1: pp 19-22 Feb 2010.
[10] Jr., J. D., Hansen, J., and Proakis, J. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York, 2000.
[11] Deshmukh S.D., Bachute M.R., Automatic Speech and Speaker Recognition by MFCC, HMM and Vector Quantization, IJEIT, Volume 3, Issue 1, July 2013.
[12] F. Soong, E. Rosenberg, B. Juang, and L. Rabiner, "A Vector Quantization Approach to Speaker Recognition", AT&T Technical Journal, vol. 66, pp 14-26March/April 1987.
[13] S. Furui, "Speaker - independent isolated word recognition using dynamic features of speech spectrum", IEEE Transactions on Acoustic, Speech, Signal Processing, Vol. 34, No. 1, pp. 52-59, 1986.
[14] Marcel R. Ackermann, Johannes Blomer, Christian Sohler, "Clustering for Metric and Nonmetric Distance Measures", ACM Transactions on Algorithms, Vol. 6, No. 4, Article 59, pp. 1-26, 2010.
[15] Y. Linde, A. Buzo & R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.
[16] http://www.fon.hum.uva.nl/praat/ used for speech recording.

Tanvi Dua was born on 02-Sept-1992. She graduated as a Bachelor of Engineering in Information Technology From Thadomal Shahani Engg. College She is currently working as a Jr.Software Engineer in BNP Paribhas India Solutions.

Jitesh Punjabi was born on 12-Aug-1992. He graduated as a Bachelor of Engineering in Information Technology From Thadomal Shahani Engg. College .He is currently working as a Software Engineer in Capgemini Consulting Pvt Ltd.

Chandni Sajnani was born on 29-Apr-1992. She graduated as a Bachelor of Engineering in Information Technology From Thadomal Shahani Engg. College .She is currently working as a Software Engineer in Capgemini Consulting Pvt Ltd

Aditya Advani was born on 03-Jan-1993. He graduated as a Bachelor of Engineering in Information Technology From Thadomal Shahani Engg. College .He is currently pursuing Master's degree in computer science at Rochester Institute of Technology, New York,US.

Shanthi Therese S is working as Associate Professor in Information Technology department of Thadomal Shahani Engineering college. Her total teaching experience is of 17 years. Currently pursuing Ph.D from Mumbai University.