Learning Assistant in Educational Field Using Automatic Speech Recognition

Prajakta Kotwal, Prof. M. R. Dixit

Abstract- **-Automatic Speech recognition is the translation of spoken words into text. It takes speech data as input and divides it into small time domain frames. Speech signal processing considering speech signals stationary for a small time interval. From point of view speech signals are divided into small units Morphims or Phonims. Any speech data can be sorted as word uttered followed by voice and silence intervals. Voice activity detection can be are employed to detect voiced and unvoiced part of speech. Speech processing consists of speech recognition, speech synthesis, speaker recognition, understanding of speech with reference to context, speech coding, speech enhancement, speech transmission, speech to text conversion & text to speech conversion etc. In general speech to text conversion system will convert input speech data to output text data. If the input speech data is inappropriate with some errors then there is a possibility to get incorrect output data.**

The proposed system contains options for correction of inappropriate input data so that the output text and speech data produce and pronounce is correct. The proposed system will be employed as learning assistance in educational field for students to learn correct pronunciation of words. The proposed system will also help tourists for conversation in local language.

Keywords: - HMM (Hidden Markov Model), GMM(Gaussian Mixture Model), ANN(Artificial Neural Network), VAD(Voice Activity Detection)

I. INTRODUCTION

The task of speech recognition is to recognize input speech. Currently used Speech Recognition software available in market takes speech as input though it is correct in pronunciation or not and proceeds on it to produce output which may not give us 100% accurate result .Obviously its depends on input. So it is essential to develop a system which will check incoming data is correction pronunciation or not. Here proposed system will do this job. Speech recognition systems have a wide range of applications from the relatively simple isolated-word recognition systems. Like any pattern recognition problem, the fundamental problem in speech recognition is the speech pattern variability. In general the sources of speech variability are as follows: Duration variability, Accent, Speaker variability, Noise etc. So proposed system based on speaker independent system and recording of input is taken in noise-free environment only.

II. THEORETICAL ANALYSIS

A. MFCC Calculation:

Generally speaking, a conventional automatic speech recognition (ASR) system can be organized in two blocks: the feature extraction and the modeling stage.The feature extraction is usually a non-invertible (lossy) transformation. Making an analogy with filter banks, such transformation does not lead to perfect reconstruction, i.e., given only the features it is not possible to reconstruct the original speech used to generate those features. Computational complexity and robustness are two primary reasons to allow loosing information. Increasing the accuracy of the parametric representation by increasing the number of parameters leads to an increase of complexity and eventually does not lead to a better result due to robustness issues. The greater the number of parameters in a model, the greater should be the training sequence.

Speech is usually segmented in frames of 20 to 30 ms, and the window analysis is shifted by 10 ms. Each frame is converted to 12 MFCCs plus a normalized energy parameter. The first and second derivatives (D's and DD's) of MFCCs and energy are estimated, resulting in 39 numbers representing each frame. Assuming a sample rate of 8 kHz, for each 10 ms the feature extraction module delivers 39 numbers to the modeling stage..

B. Pitch Calculation:

The human perception of the frequency contents of sound for speech signals does not follow linear scales. Thus for each tone with an actual frequency measured in Hz. A subjective pitch is measured on a scale .A person pitch originates in the vocal cord and the rate at which the vocal folds vibrates is the frequency of the pitch .e.g. when the vocal folds oscillates at 300 times per seconds, they are said to be producing a pitch of 300Hz.

Manuscript received November 15, 2013

PrajaktaKotwal, Research Candidate, Kolhapur Institute of Technology, Kolhapur, Maharashtra,India.
(e-mail:- kotwalprajakta77@gmail.com)
Prof.M.R.Dixit,Dept.Of Electronics, Kolhapur Institute of Technology,Kolhapur,Maharashtra, India.
(e-mail:- mrdixit@rediffmail.com)

III. PROPOSED METHODOLOGY

A. OBJECTIVE OF THE PROJECT :-

a) Applying Voice Activity Detection (VAD) and removal of silence segments.
b) Estimation of Mel – Frequency Cepstrum Coefficient (MFCC) .
c) To develop Hidden Markov Model.
d) To develop Gaussian Mixture Model.
e) To develop artificial neural network.
f) Correct the inappropriate pronounce word and replace it by appropriate pronunciation.

B. METHODOLOGIES OF IMPLEMENTATION :-

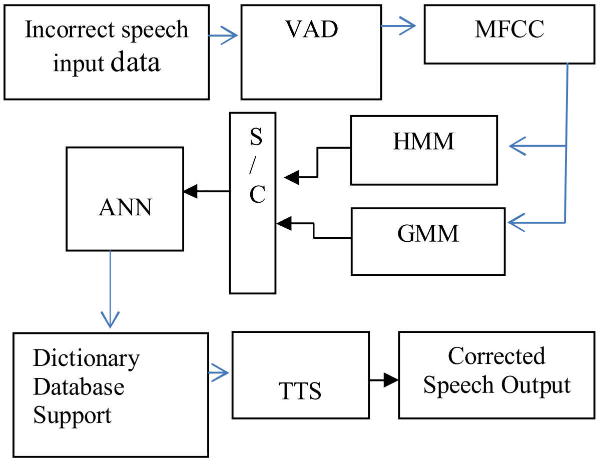


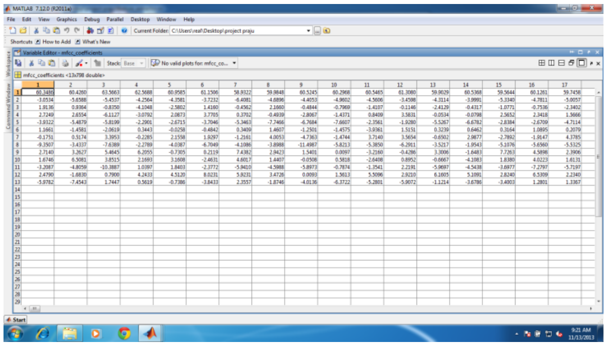
Fig1 : Methodology of Implementation

Fig 1.- Shows design of a system which convert incorrect speech input data into correct speech output data. The proposed system includes workflow as – recording of speech input data using sound recorder of PC and recording tools. Speech input data can be recorded over appropriate frequency range which can be applied to voice activity detection (VAD) also known as speech activity detection. This technique can be used in speech processing in which presence or absence of human speech is detected. Output speech signal from VAD is use for calculating Mel Frequency.

Mel Frequency ceptrum coefficient is based on human peripheral auditory system. The human perception of the frequency contents for sound of speech signal does not follow a linear scale. Thus for each tone with an actual frequency measured in Hz, a subjective pitch is measured on a scale called melscale. Isolated speech recognition can be performed by hidden markov model and Gaussian mixture model ,from them both will compare and the best output that model will be used to give input data for artificial neural network, which will classify speech segments as voiced, unvoiced, nasal / frative / plosive etc.The available development tools like Colea, Dragon, Natural Reader will be applied for speech to text conversion. The output speech will be represent the appropriate or corrected utterance of the input speech data.

Output signal based on MFCC coefficient.

BASIC RESULS:


MATLAB 7.12.0 (R2011a)
File Edit View Graphics Debug Parallel Desktop Window Help
Shortcuts How to Add What's New
Stack: Base
No valid plots for: mfcc_co...
Workspace
Command Window
Start

Current Folder: C:\Users\real\Desktop\project praju

Variable Editor - mfcc_coefficients

mfcc_coefficients <13x798 double>

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	60.3486	60.4260	63.5663	62.5688	60.9585	61.1506	58.9322	59.9848	60.5245	60.2968	60.5465	61.3080	59.9029	60.5368	59.5644	60.1261	59.7458
2	-3.0534	-5.6588	-5.4537	-4.2564	-4.3581	-3.7232	-6.4081	-4.6896	-4.4053	-4.9602	-4.5606	-3.4598	-4.3114	-3.9991	-5.3340	-4.7811	-5.0057
3	1.9136	0.9364	-0.8350	-4.1048	-2.5802	1.4160	-0.4562	2.1660	-0.4844	-0.7969	-1.4107	-0.1146	-2.4129	-0.4317	-1.0771	-0.7536	-2.3402
4	2.7249	2.6554	-6.1127	-3.0792	2.0873	3.7705	0.3702	-0.4939	-2.8067	-1.4371	0.8409	3.5831	-0.0534	-0.0798	2.5652	2.3418	1.5666
5	-3.9322	-5.4879	-5.8199	-2.2901	-2.6715	-3.7046	-5.3463	-7.7466	-6.7684	-7.6607	-2.3561	-1.9280	-5.5267	-6.6782	-2.8384	-2.6709	-4.7114
6	1.1661	-1.4581	-2.0619	0.3443	-0.0258	-0.4842	0.3409	1.4607	-1.2501	-1.4575	-3.9361	1.5151	0.3239	0.6462	0.3164	1.0895	0.2079
7	-0.1751	0.5174	3.3953	-0.2285	2.1558	1.9297	-1.2161	4.0053	-4.7363	-1.4744	3.7140	3.5654	-0.6502	2.9877	-2.7892	-1.9147	4.3785
8	-9.3507	-3.4337	-7.6389	-2.2789	-4.0387	-6.7049	-4.1086	-3.8988	-11.4987	-5.8213	-5.3850	-6.2911	-3.5217	-1.9543	-5.1076	-5.6560	-5.5325
9	2.7140	3.2627	5.4645	6.2055	-0.7305	0.2119	7.4382	2.9423	1.5401	0.0097	-3.2160	-0.4286	3.3006	-1.6483	7.7263	4.5898	2.3906
10	1.6746	6.5081	3.8515	2.1693	3.1608	-2.4631	4.6017	1.4407	-0.0508	0.5818	-2.6408	0.8952	-0.6667	-4.1083	1.8380	4.0223	1.6131
11	-3.2087	-4.8059	-10.3887	1.0397	1.8403	-2.3772	-5.9410	-4.5988	-5.8973	-0.7874	-1.3541	2.2191	-5.9697	-4.5438	-3.6977	-7.2797	-5.7197
12	2.4790	-1.6830	0.7900	4.2433	4.5120	8.0231	5.9231	3.4726	0.0093	1.5613	5.5096	2.9210	6.1605	5.1091	2.8240	6.5309	2.2340
13	-5.9782	-7.4543	1.7447	0.5619	-0.7386	-3.8433	2.3557	-1.8746	-4.0136	-6.3722	-5.2801	-5.9072	-1.1214	-3.6786	-3.4003	1.2801	1.3367
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	
24																	
25																	
26																	
27																	
28																	
29																	

Fig 2 :MFCC Extraction

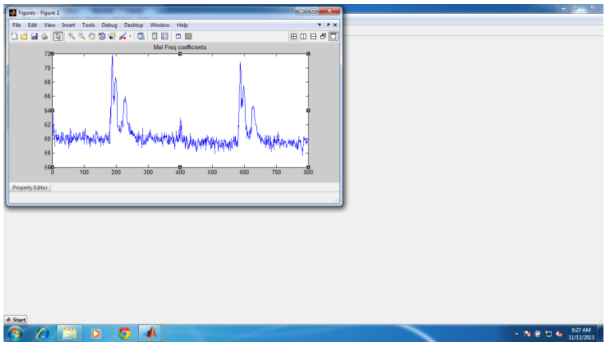


Fig 3: Signal Plotting

Pitch calculation=253 for first correct input voice.

IV. CONCLUSION / FUTURE SCOPE

Feature extraction is done by MFCC (Mel Frequency Cepstral Coefficient) which represents audio based on perception of human ear.This result is useful in speech recognition application. The same techniques can be used in different applications. There are lots of techniques likes GMM, HMM, DTW, ANN etc .which can be used as per requirement of application. Dictionary can be modified when it needs to change.

REFERENCES

[1] Nelson Morgan,"Deep and Wide: Multiple Layers in Automatic Speech Recognition",IEEE Transactions on audio,speech and language processing,vol.20, no.1, January 2012.

[2]ArchanaShende,Subhash Mishra, Shiv Kumar, "Comparison of different parameters used in GMM based automatic speaker recognition", International Journal of Soft Computing and Engineering (IJSCE)

Volume-1, Issue-3, July 2011

[3]Mohamad Adan AL - ALaoui, Lina AL-Kanj, JimmyAzar,and Elias Yaacoub ,"Speech recognition using Artificial Neural Networks and Hidden Markov Model", IEEE multisciplinary engineering education magazine.vol.3, no.3.september 2008.

[4]Chulhee Lee, Donghoon Hyun, Euisun Choi, Jinwook Go, and ChungyongLee, "Optimizing Feature Extraction for Speech Recognition",IEEE transactions on speech and audio processing, vol. 11, no. 1, january 2003.

[5]Harry Printz and Isabel Trancoso ,"Editorail", IEEE transactions on speech and audio processing, vol. 10, no. 8, november 2002.

[6] Alexandros Potamianos, Member, IEEE, and Petros Maragos, "Time-Frequency Distributions for Automatic Speech Recognition", IEEE transactions on speech and audio processing, vol. 9, no. 3, march 2001.

[7] VibhaTiwari,"MFCC and its applications in speaker recognition", International Journal on Emerging Technologies 1(1): 19-22(2010).

[8]D.B.Paul,"Speech Recognition using Hidden Markov Model",The Lincoln Laboratory Journal vol.3, no.1,1990.

[9]Lawrence R.Rabiner, "A-Tutorial on Hidden Markov Models and selected applications in speech recognition",vol.77,no.2, Feb 1999.

[10]Yang Liu,"Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies"IEEE Transactions on audio, speech and language processing, vol. 14, Sept 2006