Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

A NEW APPROACH FOR SPOKEN WORD DETECTION

Miss. Prajakta Kotwal[1], Mr. Vinayak Chavan[2], Mr. Suraj Shete[3]
[1,2]*Dep. Of ENTC,Finolex Academy Of Management and Technology, Ratnagiri*
[3]*Dep. Of ETRX,Finolex Academy Of Management and Technology, Ratnagiri*

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 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 produced and pronounced 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— MFCC, GMM (Gaussian Mixture Model), VAD(Voice Activity Detection), Text-to-Speech

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 it depends on input. So it is essential to develop a system which will check incoming data is correct in pronunciation or not. Here proposed system will do this job. Speech recognition systems have a wide range of applications. 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

2.1 Frame Blocking

The idea of segmentation of the speech wave into frames, or what is known as frame blocking, comes from the fact that the vocal tract moves mechanically slowly, and as a result, speech can be assumed to be a random process with slowly varying properties [3]. Hence, the speech can be divided into frames, over which the speech signal is assumed to be stationary with constant statistical properties.

2.2 MFCC Calculation

The frequently used features for speech processing, also known as the Mel-Frequency Cepstral Coefficients (MFCC), are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies that have been used to capture the phonetically important characteristics of speech.

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency t measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale'. The mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as1000 mels. Typically, the mel-cepstrum is obtained using critical band filters as shown in Fig. 1 [2]. Thus, the parameters of the critical filters, such as center frequencies and bandwidths, may be optimized to reduce the error rate in speech recognition.

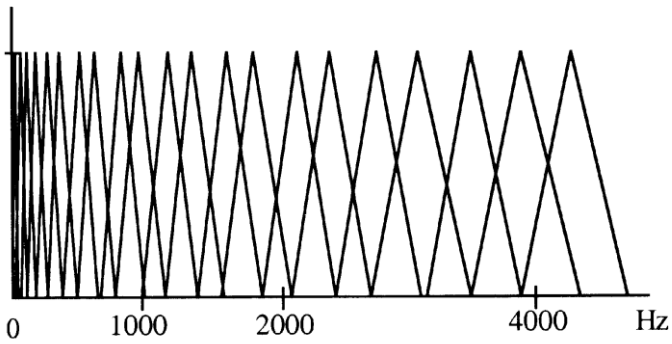

Fig1:critical band filters for computing mel-spectrum

In sound processing, the Mel-frequency cestrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cestrum coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of spectral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

Procedure to find MFCC:

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the Mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum

Studies have shown that human hearing does not follow the linear scale but rather the Mel-spectrum scale which is a linear spacing below 1000 Hz and logarithmic scaling above 1000 Hz. In the final step, the Mel-spectrum plot is converted back to the time domain by using the following equation:

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

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.

2.3 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 vibrate is the frequency of the pitch .e.g. when the vocal fold oscillates at 300 times per seconds, they are said to be producing a pitch of 300Hz.

III. PROPOSED METHODOLOGY

3.1 Objective Of The Project

1. Applying Voice Activity Detection (VAD) and removal of silence segments.
2. Estimation of Mel – Frequency Cepstrum Coefficient (MFCC) .
3. To develop Gaussian Mixture Model.
4. Correct the inappropriate pronounce word and replace it by appropriate pronunciation.

3.2 Methodologies of Implementation:-

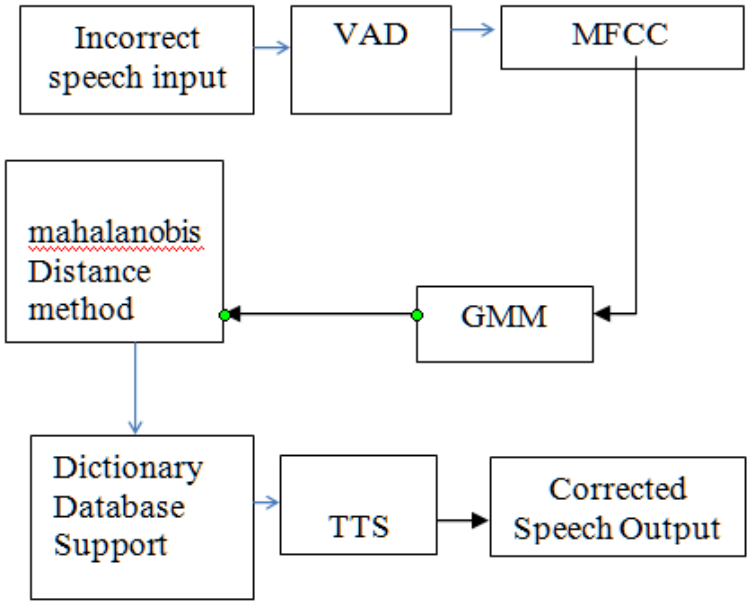

Fig2 Methodology of Implementation

Fig 2 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 perceptual 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 mel-scale. 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.

3.3 Mahalanobis Distance Method

It measures the distance of a point x from a data distribution. The data distribution is characterized by a mean and the covariance matrix, thus is hypothesized as a multivariate gaussian.
The **Mahalanobis distance** is a measure of the distance between a point P and a distribution D. The Mahalanobis distance has the following properties:

1. It accounts for the fact that the variances in each direction are different.
2. It accounts for the covariance between variables.
3. It reduces to the familiar Euclidean distance for uncorrelated variables with unit variance.

For univariate normal data, the univariate z-score standardizes the distribution (so that it has mean 0 and unit variance) and gives a dimensionless quantity that specifies the distance from an observation to the mean in terms of the scale of the data.
The Mahalanobis distance accounts for the variance of each variable and the covariance between variables. Geometrically, it does this by transforming the data into standardized uncorrelated data and computing the ordinary Euclidean distance for the transformed data. In this way, the Mahalanobis distance is like a univariate z-score: it provides a way to measure distances that takes into account the scale of the data.

3.4 Histogram

Colour distribution information can be represented in a number of ways: mean RGB co-variant RGB [8], colour clusters [9±11]), and colour names [12, 13]). However, the most general (and arguably the simplest) representation for colour distributions is the colour histogram. It is most general in the sense that colour statistics such as the mean or colour clusters are usually calculated from the colour histogram. Unfortunately, it is argued that there is a downside to this generality: it is relatively more expensive to match or compare colour histograms. In this paper we show that colour histograms contain highly correlated information and so they can be effectively compressed: they can be represented by a few numbers. As such, colour histogram comparison is no slower than any other colour-based indexing method. Colour histograms are created by partitioning colour space into equal-area regions and counting the number of pixels falling in each region, then allocating that total to the related histogram bin (each histogram has the same number of bins as there are equal-area regions).

IV. BASIC RESULTS

	A	B	C	D	E	F	G	H	I	J	K	L
1	37.38419	38.38419	39.38419	40.38419	41.38419	42.38419	43.38419	44.38419	45.38419	46.38419	47.38419	48.38419
2	-7.15415	-6.15415	-5.15415	-4.15415	-3.15415	-2.15415	-1.15415	-0.15415	0.845846	1.845846	2.845846	3.845846
3	0.860969	1.860969	2.860969	3.860969	4.860969	5.860969	6.860969	7.860969	8.860969	9.860969	10.86097	11.86097
4	-0.48775	0.51225	1.51225	2.51225	3.51225	4.51225	5.51225	6.51225	7.51225	8.51225	9.51225	10.51225
5	0.713679	1.713679	2.713679	3.713679	4.713679	5.713679	6.713679	7.713679	8.713679	9.713679	10.71368	11.71368
6	-1.96126	-0.96126	0.038744	1.038744	2.038744	3.038744	4.038744	5.038744	6.038744	7.038744	8.038744	9.038744
7	0.146534	1.146534	2.146534	3.146534	4.146534	5.146534	6.146534	7.146534	8.146534	9.146534	10.14653	11.14653
8	-4.21715	-3.21715	-2.21715	-1.21715	-0.21715	0.782851	1.782851	2.782851	3.782851	4.782851	5.782851	6.782851
9	-1.82552	-0.82552	0.174483	1.174483	2.174483	3.174483	4.174483	5.174483	6.174483	7.174483	8.174483	9.174483
10	-2.88767	-1.88767	-0.88767	0.112326	1.112326	2.112326	3.112326	4.112326	5.112326	6.112326	7.112326	8.112326
11	-2.09964	-1.09964	-0.09964	0.900358	1.900358	2.900358	3.900358	4.900358	5.900358	6.900358	7.900358	8.900358
12	0.568427	1.568427	2.568427	3.568427	4.568427	5.568427	6.568427	7.568427	8.568427	9.568427	10.56843	11.56843

Fig 3 : MFCC Extraction for 12 filters

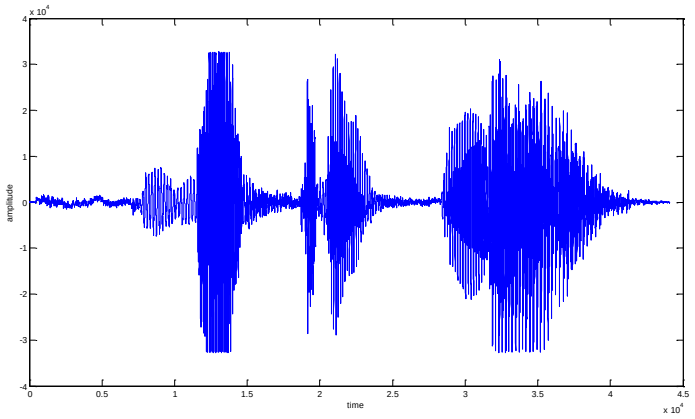

Fig 4 :Signal Plotting for correct pronunciation

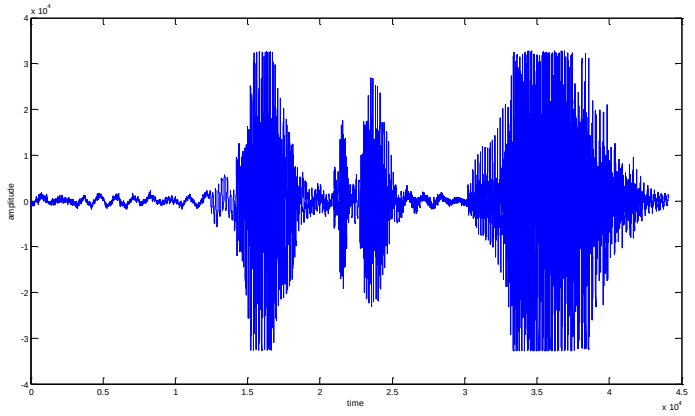

Fig5 : Signal Plotting for incorrect pronunciation

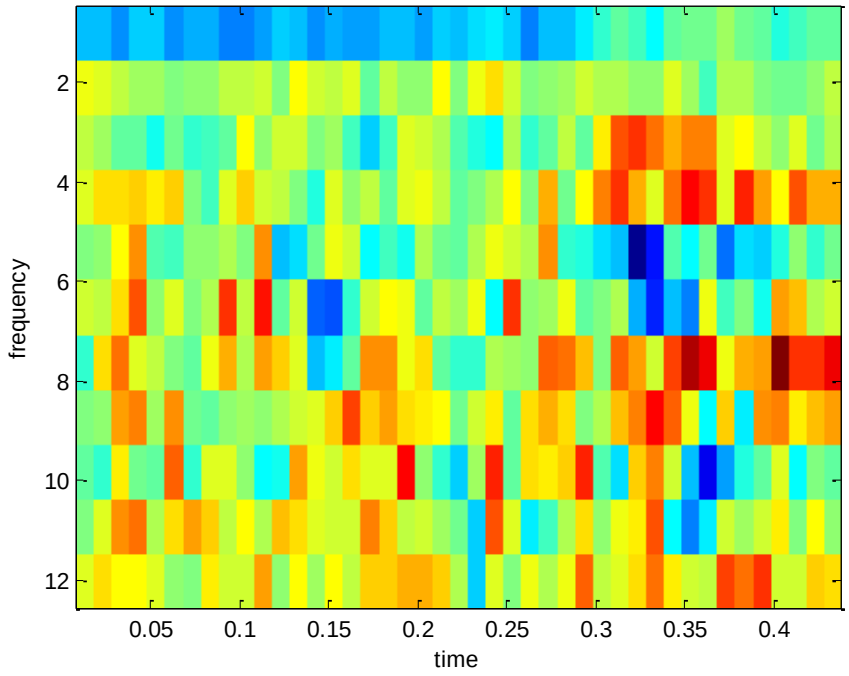

Fig 6:Histogram for word chemistry

Colour distribution information can be represented in a number of ways: mean RGB co-variant RGB [8], colour clusters [9±11]), and colour names [12, 13]). However, the most general (and arguably the simplest) representation for colour distributions is the colour histogram. It is most general in the sense that colour statistics such as the mean or colour clusters are usually calculated from the colour histogram. Unfortunately, it is argued that there is a downside to this generality: it is relatively more expensive to match or compare colour histograms. In this paper we show that colour histograms contain highly correlated information and so they can be effectively compressed: they can be represented by a few numbers. As such, colour histogram comparison is no slower than any other colour-based indexing method. Colour histograms are created by partitioning colour space into equal-area regions and counting the number of pixels falling in each region, then allocating that total to the related histogram bin (each histogram has the same number of bins as there are equal-area regions).

V. 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.

The same application can be run using different methods which can be simple than this one. In future if get the simple methodology then will implement that rather using this method.

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, SubhashMishra , 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] MohamadAdan AL-ALaoui,Lina AL-Kanj,JimmyAzar,and Elias Yaacoub ,"Speech recognition using Artificial Neural Networks and Hidden Markov Model",IEEEmultisciplinary 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 1899

