Speaker's Age Group Classification and Recognition using Spectral Features and Gaussian Mixture Models

Abhinav Sharma[1], Anshu Sharma[2]
[1] Department of Electronics and Communication Engineering
[2]Department of Physics
[1]Graphic Era University, Dehradun, Uttarakhand
[2]Olympus High School, Dehradun Uttarakhand

Abstract-*In this work Age Group Recognition of different Speakers has been performed with the help of the extraction of Mel Frequency Cepstral Coefficients (MFCCs) from various speakers. Gaussian Mixture Models technique has been used for the purpose of Classification. There have been five age groups used for the work namely less than 10 years, 11-20, 21-30,31-40 and more than 41 years. An accurate Age Group recognition of a speaker can play a vital role in the important areas of the society including security identifications, crime investigation etc. The accuracy tests have been performed by varying the number of Mfccs extracted. Extracting more number of features delays the result calculation but may improve the chances of increased accuracy. Results show that accuracy is highest (87.5 %) when 29 number of mfcc were extracted from the speech samples.*

Keywords-Age Group/ Age Recognition, Speaker/Speech Recognition, MFCC, GMM, Speech classification.

I. INTRODUCTION

Automatic speech recognition (ASR) has changed with the advent of digital signal processing hardware and software. But despite of all these advances, machines can not match the performance of their human counterparts in terms of accuracy and speed, specially in case of speaker independent speech recognition.

So today significant portion of speech recognition research is focussed on speaker independent speech recognition problem. The reasons are its wide range of applications, and limitations of available techniques of speech recognition.

Speech recognition system performs two fundamental operations: signal modeling and pattern matching [1].

Signal modeling represents process of converting speech signal into a set of parameters. Pattern matching is the task of finding parameter set from memory which closely matches the parameter set obtained from the input speech signal.

I.1 Motivation for signal modeling [1]

1. To obtain the perceptually meaningful parameters i.e. parameters which are analogous to those used by human auditory system.
2. To obtain the invariant parameters i.e. parameters which are robust to variations in channel, speaker and transducer.
3. To obtain parameters that capture spectral dynamics, or changes of spectrum with time.

The signal modeling involves four basic operations: spectral shaping, feature extraction, parametric transformation, and statistical modeling [1]. Spectral shaping is the process of converting the speech signal from sound pressure wave to a digital signal; and emphasizing important frequency components in the signal. Feature extraction is process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal. Parameter transformation is the process of converting these features into signal parameters through process of differentiation and concatenation. Statistical modeling involves conversion of parameters in signal observation vectors.

I.2 Spectral Shaping

Spectral shaping [1] involves two basic operations: digitisationi.e.conversion of analog speech siganl from sound pressure wave to digital signal; and digital filtering i.e.emphasizing important frequency components in the signal. This process is shown in Fig.1.

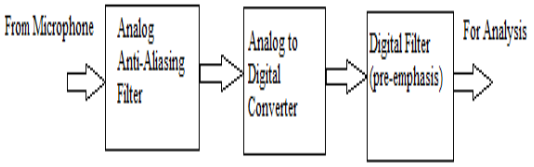

Fig.1- Process os Spectral Shaping

The main purpose of digitisation process is to produce a sampled data representation of speech signal with as high signal-to-noise ratio (SNR) as possible. Once signal

conversion is complete, the last step of digital post filtering is most often executed using a Finite Impulse Response (FIR) filter given as

$$H_{pre}(z) = \sum_{k=0}^{N_{pre}} a_{pre}(k) z^{-k}$$

Normally, a one coefficient digital filter known as pre-emphasis filter, is used

$$H_{pre}(z) = 1 + a_{pre} z^{-1}$$

Advantages of the pre-emphasis filter is-

1. The voiced sections of speech signal naturally have a negative spectral slope (attenuation of approximately 20 dB per decade due to physiology of speech production system [3]. The preemphasis filter serves to offset this natural slope before spectral analysis, thereby improving the efficiency of the analysis [2].

2. The hearing is more sensative above the 1-kHz region of the spectrum. The preemphasis filter amplifies this area of the spectrum. This assists the spectral analysis algorithm in modelling the perceptually important aspects of speech spectrum [1].

II. FEATURE EXTRACTION

II.1-Cepstral Analysis

This analysis technique is very useful as it provides methodology for separating the excitation from the vocal tract shape [2]. In the linear acoustic model of speech production, the composite speech spectrum, consist of excitation signal filtered by a time-varying linear filter representing the vocal tract shape as shown in fig.2

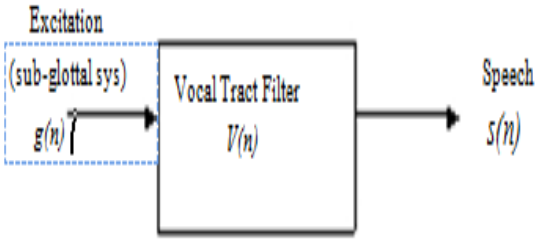

Fig.2-Linear acoustic model of speech production "from [1]"

The speech signal is given as

$s(n) = g(n) * v(n)$
where $v(n)$: vocal tract impulse response
$g(n)$: excitation signal

The frequency domain representation
$S(f) = G(f) \cdot V(f)$

Taking log on both sides,

$$\log(S(f)) = \log(G(f)) + \log(V(f))$$

Hence in log domain the excitation and the vocal tract shape are superimposed, and can be separated. Cepstrum is computed by taking inverse discrete Fourier transform (IDFT) of logarithm of magnitude of discrete Fourier transform finite length input signal as shown in fig.3.

Fig. 3. System for obtaining cepstrum "adapted from[2]".

S(n) is defined as cepstrum. In speech recognition cepstral analysis is used for formant tracking and pitch (0f) detection. The samples of (n) in its first 3ms describe v(n) and can be separated from the excitation. The later is viewed as voiced if (n) exhibits sharp periodic pulses. Then the interval between these pulses is considered as pitch period. If no such structure is visible in (n), the speech is considered unvoiced.

$$S(k) = \sum_{n=0}^{N-1} s(n) \exp(-j2 \pi nk$$

$$\hat{V}(k) = \log(S(K))$$

$$s(n) = (1/N) \sum_{k=0}^{N-1} \hat{V}(k) \exp(j2 \pi/N) nk$$

II.2-Mel Cepstrum Analysis

This analysis technique uses cepstrum with a nonlinear frequency axis following mel scale [3]. For obtaining melcepstrum the speech waveform s(n) is first windowed with analysis window w(n) and then its DFT S(k) is computed. The magnitude of S(k) is then weighted by a series of mel filter frequency responses whose center frequencies and bandwidth roughly match those of auditory critical band filters.

The next step in determining the melcepstrum is to compute the energy in this weighted sequence. This is the energy of filter which normalizes the filter according to their varying bandwidths so as to give equal energy for flat spectrum.

The real cepstrum associated with E mel(n,l) is referred as the mel-cepstrum and is computed for the speech frame at time n as

$$C_{mel}(n,m) = (1/N) \sum_{l=0}^{N-1} \log\{E_{mel}(n,l)\}\cos[2\pi(l+1/2)/N]$$

Such*mel* cepstral coefficients C_{mel}provide alternative representation for speech spectra which exploits auditory principles as well as decorrelating property of cepstrum.

III. FEATURE CLASSIFICATION AND MATCHING

In this work speech samples have been collected from different speakers of different age groups and after spectral feature extraction, speaker's age-group GMMs have been created which were used for the classification. The testing files have been tested against all the age grou's GMMs and recognition has been found out.

In this GMM are used specifically for capturing distribution of data points from the input features. Given a set of inputs, GMM refines the weights of each distribution .Once a model is generated, conditional probabilities can be computed for test patterns (unknown data points).Gausses in the mixture model is known as number of components. They indicate the number of clusters in which data points are to be classified.

The components within each GMM capture finer level details among the feature vectors of each speech command. In this work, GMM's are designed with 64 components and iterated for 100 times to attain convergence of weights.

IV. TESTING, EXPERIMENTAL RESULTS AND DISCUSSIONS

The experimental tests have been performed in the following steps-

1- Collection of speaker speech data of different age groups in different environments.
2-Adding the different prerecorded additive noises to the clean speech data samples.
3-Extraction of distinguished and discriminative features from the collected speech samples using Mel Frequency Cepstral Coefficient (MFCC) technique and producing sets of feature vectors for all 5 age groups.
4-Creating and training the GMM models for all 5 age groups.
5-Testing the performance of GMM models.

IV.1-Speaker's Age-Group Recognition Performance-

For the testing of age groups, speech samples with the trained models for five age group speeches, a folder structure is formed of five folders for five speakers, each containing 10 samples of an age group from different speakers (in all 100 test samples) . A percentage confusion matrix is formed as shown in table-2. A.G.. 1 in the table represents the test folder 1 having 8 test samples of different speakers of age group 1. All its test samples go through the process of feature extraction and the feature vectors are sent to 5 GMM models for five speaker's age group speeches one by one. Now e.g. value 75 in the location (1,1) represents that 75% out of 08 test samples at test folder 1 (Test 1) are recognized as First Age-group.

Table-1 Description of Age Groups used for the work

Age Group Description	Age Group number assigned
11-20 years	1
21-30 years	2
31-40 years	3
Less then 10 years	4
More than 41 years	5

Table-2 Percentage Confusion Matrix (A.G.=Age Group)

A.G	Gaussian Mixture Models				
	1	**2**	**3**	**4**	**5**
1	75	13	0	13	0
2	0	100	0	0	0
3	0	25	75	0	0
4	0	0	0	100	0
5	0	0	0	13	88

According to Table-3 e.g. Now e.g. value 6 in the location (1,1) represents that 6 out of 8 test samples at test folder 1 (Test 1) are recognized as the samples of First age group of speakers.

Table-3 Confusion Matrix (A.G=Age Group)

A.G	Gaussian Mixture Models				
	1	**2**	**3**	**4**	**5**
1	6	1	0	1	0
2	0	8	0	0	0
3	0	2	6	0	0
4	0	0	0	8	0
5	0	0	0	1	7

Table 4 describes the Robustness or the accuracy of the speaker's age group recognition. Here closed test refers that the training speaker speech data and the testing speech data is same. On the other hand open test refers that training speaker speech files and testing speaker speech files are different.

Table-4 Recognition Robustness Performance (%)

Age Group	Closed Test	Open Test
1	88	75
2	100	100

3	88	75
4	100	100
5	100	88
Average	**95.2**	**87.6**

Table 5 gives another description of the testing results. Here numbers of MFCCs extracted from the speech files are varied. Results show that extraction of more number of features may improve the accuracy of the recognition but after a certain set the accuracy may start dropping. In this work the accuracy is achieved best with 29 number of features extracted.

Table-5 Speaker Recognition with varying MFCC's (%)-

Age Group	No. of MFCC's				
	6	**8**	**13**	**21**	**29**
1	50	75	75	75	75
2	88	88	75	88	100
3	88	88	75	88	75
4	100	100	100	100	100
5	88	63	75	75	88
Average	**82.5**	**82.5**	**80**	**85**	**87.6**

V. CONCLUSION AND FUTURE WORK

In this research the age group speech recognition for five age groups was developed. As the future work the recognition of speakers of different nationalities and languages will be achieved. Moreover the boundaries of the age groups can be reduced to analyze the accuracy of recognition technique. e.g. the age group of 11-20 can be replaced by 11-15 and so on.

It is very hard to get the exact matches and high accuracy in many cases, because human voice changes from time to time and most importantly in noisy environments.

Sex of the speaker may affect the recognition results so to cancel the effect of this the database of speech samples should contain speeches of both the sex.

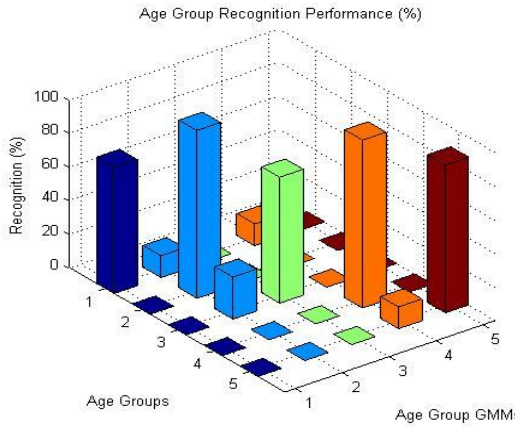

Figure-3 Age Group Recognition Performance (%)

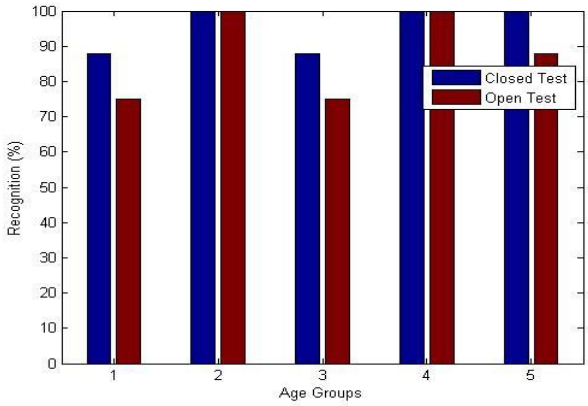

Figure 4-Recognition Robustness Performance (%)

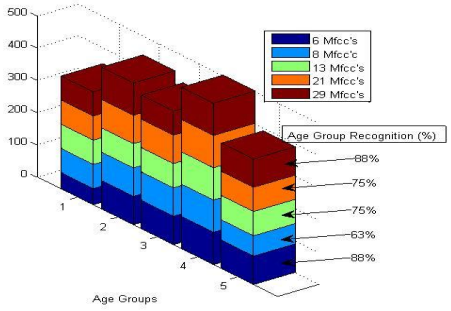

Figure -5 Age Group Recognition with varying no. of MFCC's

REFERENCES

[1] J. W. Picone, "Signal modelling technique in speech recognition," Proc. Of the IEEE, vol. 81, no.9, pp. 1215-1247, Sep. 1993.

[2] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals.Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

[3] D.O. Shaughnessy, Speech Communication: Human and Machine. India:University Press ,2001.

[4] B. Gold and L. R. Rabiner,"Parallel processing techniques for estimating pitch periods of speech in the time domain," J. Acoust. Soc. America, vol.46, pt. 2, no. 2, pp 442-448, Aug. 1969.

[5] Dr.E.Chandra, A.Akila ,"An Overview of Speech Recognition and Speech Synthesis Algorithms", Int.J.Computer Technology &Applications,Vol 3 (4), 1426-1430 ISSN:2229-6093

[6] Ahmad A. M. Abushariah, Reddy Surya Gunawan, " Speech Recognition System using MATLAB ", LAB

LAMBERT Aceademic Publishing GmbH & Co. KG, 2011

[7] Ahmad A. M. Abushariah, Teddy S. Gunawan,Othman O. Khalifa, "English Digits Speech Recognition System Based on Hidden Markov Models", International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia

[8] Susumu Harada, "Harnessing the Capacity of the Human Voice forFluidly Controlling Computer Interface", University of Washington, 2010

[9] James R Evans, Wayne A Tjoland and Lloyd G Allred, "Achieving a Hands-Free Computer Interface using Voice Recognition and Speech Synthesis ",IEEE AES Systems Magazine, 2000

[10] Susumu Harada and James A Landay, "The Vocal Joystick: Evaluation of Voice-based Cursor Control Techniques", Portland, Oregon, USA: ASSETS, 2006

[11] M Abdeen, H Moshammad and M C E Yagoub, "An Architecture for Multi-Lingual Hands Free Desktop Control System for PC Windows", Niagara Falls, Canada : IEEE , 2008

[12] R Maskeliunas, K Ratkevicius and V Rudzionis, "Voice-based Human-Machine Interaction Modeling for Automated Information Services", ISSN 1392-1215 Electronics and Electrical Engineering, 2011

[13] R Norma Conn and Michael McTear, "Speech Technology: A Solution for People with Disabilities", Savoy Place, London WCPR OBL, UK: IEE, 2000

[14] Shashidhar G. Koolagudi, K.Srinivas Rao, "Recognition of Emotions from Speech using Excitation Source Features", International Journal of Speech Technology, June 2012, Volume 15, Issue 2, pp 265-289

[15] Sandeep Kaur , "Mouse Movement using Speech and Non-Speech Characteristics of Human Voice", International Journal of Engineering and Advanced Technology (IJEAT) ,ISSN: 2249 – 8958, Volume-1, Issue-5, June 2012