International Journal of Research in Advent Technology, Vol.7, No.2, February 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Telugu Accent Recognition Using MFCC and GMM

K.V.V.Kunar[1], K Srilakshmi[2], K Srilakshmi[2,] K.Sai Satish Reddy[2], M.L.Teja Harika[2], R.Sridhar[2]
[,1.] *Asst.Professor, Department of ECE, Kallam Haranadhareddy Institute of Technology Chowdavaram, Guntur (DT), A.P, India.*
[2]*B-Tech Students, Department of ECE, Kallam Haranadhareddy Institute of Technology Chowdavaram, Guntur (DT), A.P, India.*

Abstract —In this paper, we introduced a accent recognition method for various different accents in telugu language. MFCC extraction is used for the audio sample feature extraction. Gaussian mixture model is used for the classification. Database is considered of voice samples of 5 speakers with different accents and tested for about 20 words. Accuracy of classification is displayed using a confusion matrix. This accent recognition model is very useful in communication applications.

Keywords—accent, Gaussian mixture model, mfcc

1. INTRODUCTION

Accent is very important issue in order to recognize a speaker. There are various research works in field of biometrics like recognition of fingerprint, voice and face [11] [12] [13]. Here in this paper we have taken accent as base in order to recognize the speaker. Telugu language consists of various dialects. our study lies as a bridge between the speakers of two different dialects[1]. Every speaker have their own speaking style. In order to extract the features for accent detection prosody plays a major role[16]. Prosody is defined as sequence of stress and intonation in language [3][14].

Suppose when consider a situation in customer care office where a person of different dialect called, in such case our accent recognition model will be more helpful to act as a connection between them. In order to extract features Mel frequency ceptral coefficients are used. MFCC are widely used because it will extract features even in present of noise. Once features are extracted, there are several classifiers to classify them into various dialects. There are several researches on classifiers like SVM, HMM, GMM [4][5][6].

Database consists of set of 5speakers with different telugu accents .Both Male and Female speakers are used for training. Each speaker is trained with 20 different words and 5 different words are used for testing and result is displayed using a matrix.

1. FEATURE EXTRACTION USING MFCC

In order extract and select the features of voice signal Mel frequency ceptral coefficients are used [7]. The features are extracted in the form of energy spectrum and are expressed in real logarithmic scale with prior cosine transform and their representation in MFCC will be on mel frequencyscale. Mel frequency ceptral coefficients are more effectual.

Voice signal is applied with pre-emphasis in order to increase the strength of the signal .Then the signal if framed and applied to a hamming window in order to shape the signal according to the sampling rate .Now the signal is applied to fast fourier transform. This is very wide and is not continuous hence they are plotted using mel filter bank processing now obtained mel spectrum is converted to time domain using discrete cosine transform .

Mel-scale frequency cepstral coefficient (MFCC for short) are most commonly used acoustic features for speaker recognition. MFCC gives best speaker recognition as it takes human perception sensitivity with respect to frequencies into consideration,.. Success has been due to representation of the speech amplitude spectrum into compact form. For finding the Mfcc coefficients the main steps are taking the log magnitude spectrum of the windowed waveform which will then be smoothened out by triangular filters and then compute the DCT of the waveform to generate the mfcc coefficients. Since we have performed FFT, DCT transforms the frequency domain into a time-like domain called quefrency domain. The obtained features are similar to cepstrum, thus it is referred to as the mel-scale cepstral coefficients, or MFCC. MFCC alone can be used as the feature for speech recognition. MFCC is a very robust parameter for modeling the speech as it can be modeled as MFCC kind of models the human auditory system and hence makes the reduction of the frame of a speech into the MFCC coefficients a very useful transform as now we have an even more accurate transform to deal with for the recognition of the speakers.

When speech signal is modeled with parameter modeling MFCC gives very high and better accuracy.The pre-emphasis stage basically do the low pass filtering and as speech is not an stationary process ,data will be windowed and in order to acquire accuracy windowing is required. The figure 2 above gives the block diagram for computing the MFCC coefficients.

BASIC STEPS

1) Pre-emphasis

High pass filter is given with a speech signal s(n) :

s_2s2(n)=s(n)-a*s(n-1) (1)

s_2s2(n) represents output of the signal . a bears a value of usually between 0.9 and 1.0.

H(z) gives the ztransform ,which is

H(z)=1-a*z^{-1} (2)

In order to compensate the high frequency part that was suppressed during sound production mechanism of humanbeings. Moreover, it can also amplify the importance of high-frequency formants.

2)Frame blocking

The input speech signal is segmented into frames of 20~30 ms with optional overlap of 1/3~1/2 of the frame size. Usually the frame size (in terms of sample points) is equal power of two in order to facilitate the use of FFT. If this is not the case, we need to do zero padding to the nearest length of powerof two. If the sample rate is 16 kHz and the frame size is 320 sample points, then the frame duration is 320/16000 = 0.02 sec = 20 ms. Additional, if the overlap is 160 points, then the frame rate is 16000/(320-160) = 100 frames per second.

3) Hamming Windowing

Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame (to be detailed in the next step). If the signal in a frame is denoted by s (n), n = 0,…N-1, then the signal after Hamming windowing is s(n)*w(n), where w(n) is the Hamming window defined by the equation below.

w(n,a) = (1- a) - a cos(2pn/(N-1)),0≦n≦N-1 (3)

4) Fast Fourier Transform or FFT

Spectral analysis shows that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore we usually perform FFT to obtain the magnitude frequency response of each frame. When we perform FFT on a frame, we assume that the signal within a frame is periodic, and continuous when wrapping around. If this is not the case, we can still perform FFT but the in continuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem, we have two strategies: Multiply each frame by a Hamming window to increase its continuity at the first and last points.

Take a frame of a variable size such that it always contains a integer multiple number of the fundamental periods of the speech signal. The second strategy encounters difficulty in practice since the identification of the fundamental period is not a trivial problem. Moreover, unvoiced sounds do not have a fundamental period at all. Consequently, we usually adopt the first strategy to multiply the frame by a Hamming window beforeperforming FFT.

5) Triangular Bandpass Filters

We multiple the magnitude frequency response by a set of 20 triangular bandpass filters to get the log energy of each triangular bandpassfilter. The positions of these filters are equally spaced along the Mel frequency, which is related to the common linear frequency f by the following equation:

mel(f)=1125*ln(1+f/700) (4)

Mel-frequency is proportional to the logarithm of the linear frequency, reflecting similar effects in the human's subjective aural perception.

6) Discrete Cosine Transform or DCT

In this step, we apply DCT on the 20 log energy Ek obtained from the triangular bandpass filters to have L mel-scale cepstral coefficients. The formula for DCT is

Cm=Sk=1^{N}cos[m*(k-0.5)*p/N]*Ek, (5)

m=1,2, ..., L

where N is the number of triangular bandpass filters, L is the number of mel-scale cepstral coefficients. Usually we set N=20 and L=13. Since we have performed FFT, DCT transforms the frequency domain into a time-like domain called quefrency.domain.The obtained features are similar to cepstrum, thus it is referred to as the mel-scale cepstral coefficients, or MFCC. MFCC alone can be used as the feature for speech recognition. For better performance, we can add the log energy and perform delta operation.

DELTA MFCC equation is given by :

$\Delta MFCC\ (K) = \Delta MFCC(K) - \Delta MFCC(K-1)$

Here K is the coefficient number i.e K=1 to L.

So delta mfcc is kind of like taking the first derivative of the mel-cepstral coefficients. So it tells us the rate of change of the cepstral coefficients, which will be useful in determining accents and not that useful in speaker verification systems. Still we will be demonstrating the delta mfcc in our proposed framework.

3. GAUSSIAN MIXTURE MODEL

The features extracted from mfcc are trained using gmm. The trained GMM classifier is usually screened with features taken from the check utterances. The complete log likelihood regarding check vectors of just one check utterance regarding this educated GMM matching to help just about every accent-class is usually calculated. The check utterance is regarded as to help belong to which accent-class regarding which the full log-likelihood will become the greatest. Gaussian mixture model (GMM) is a statistical method used to model speaker (in our case accent) specific features. It consistsof a number of individual Gaussians to provide multi-modal density representation for each model. In pattern recognition applications, GMMs are used to generate speaker (dialect) models and also to match different patterns against the trained models. The weighted

sum of M component densities is a Gaussian mixture model which is given by

$$p(\bar{x}|\lambda) = \sum_{i=1}^{M} \mathrm{p_i b_i}(\bar{x}) \quad (6)$$

Here $\bar{x}$is a D-dimensional vector, $b_i(\overline{x})$,
i = 1,…,M, are the component densities and
p_i are mixture weights.Each component density is given by

$$b(\bar{x}) = \frac{1}{(2\Pi)^{D/2}\,|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(-\bar{x}-\bar{\mu_l})^T \sum_1^{-1} \bar{x}-\bar{\mu_l})\} (7)$$

with mean vector modeling μi and covariance matrix Σi . These parameters are represented by,

$$\lambda = \{p_i, \bar{\mu}_i, \Sigma_i\} \quad \text{i=1…..M} \quad (8)$$

Expectation-maximization (EM) algorithm is used to estimate parameters iteratively. The EM algorithm estimates a new model $\bar{\lambda}$ from an initial model λ , so that the likelihood of the new model increases. On each estimation, the following formulae are used,

$$\overline{p_l} = \frac{1}{T}\sum_{t=1}^{T} p(i/\overline{x_t},\lambda) \quad (9)$$

$$\overline{\mu_l} = \frac{\sum_{t=1}^{T} p(i|\overline{x_t},\lambda)\overline{x_t}}{\sum_{t=1}^{T} p(i|\overline{x_t},\lambda)} \quad (10)$$

$$\overline{\sigma_l^2} = \frac{\sum_{t=1}^{T} p(i|\overline{x_t},\lambda)\overline{x_t}^2}{\sum_{t=1}^{T} p(i|\overline{x_t},\lambda)} - \overline{\mu_l^2} \quad (11)$$

Here $\overline{\sigma^2}_i$, $\bar{\mu}_i$ and $\bar{p}_i$ are updated covariance , mean and mixture weights .
Posteriori probability is

$$p(i|\overline{x_t},\lambda) = \frac{p_i b_i(\overline{x_t})}{\sum_{k=1}^{M} p_k b_k(\overline{x_t})} \quad (12)$$

Hence , to identify the accent in a set of s ,as s ={1,2,….s}, is done with gmm modeling by $\lambda_1, \lambda_2, \ldots\ldots\ldots \lambda_s$. By computing the posterori probability,final decision is made for each feature (test sequence) for all accents which are gmm modeled[8][9][10] . Then accent which has more likelihood coefficients are selected as maximum probability and these aremodeled with help of reference data and distance is found using GMM model.

4. RESULTS & DISCUSSIONS

The accent speech data base is considered with speaker belonging to different regions and having different accents such as krishna, telangana, srikakulam, godavari and rayalseema. The data base is generated from the voice samples of both the genders .100 samples have been recorded using text dependent data; we have considered only a word. The data is trained by extracting the voice features MFCC .The data is recorded with sampling rate of 16 kHz..In order to classify the accent and to appropriately identify the speaker the experimentation has been conducted on the database, by considering 75 recordings per training and 25 emotions for testing .we have repeated the experimentation and above 90% over all recognition rate is achieved.

TABLE 1.CONFUSION MATRIX

Region	Telangana	Rayalseema	Krishna	Godavari	Srikakulam
Telangana	10	0	0	0	0
Rayalseema	0	8	0	2	0
Krishna	0	0	10	0	0
Godavari	0	0	0	9	1
Srikakulam	0	0	0	0	10

REFERENCES

[1] Yanli Zheng, Richard Sproat, Liang Gu, Izhak Shafran, Haolang Zhou‡, Yi Su‡, Dan Jurafsky , Rebecca Starr? , Su-Youn Yoon," Accent Detection and Speech Recognition for Shanghai-Accented Mandarin ",University of Illinois, IBM T. J. Watson Research Center† Johns Hopkins University‡, Stanford University.

[2] http://kochanski.org/gpk/prosodies/section1/

[3] Nathan Smith ,"Speech recognition using SVM ",Cambridge university

[4] J. P. Campbell Jr, "Testing with the YOHO CD-ROM voice verification corpus," *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 InternationalConference on*, vol. 1, 1995.

[5] K. Bartkova and D. Jouvet, "Using Multilingual Units for Improved Modeling of Pronunciation Variants," *Acoustics, Speech and Signal Processing, 2006. ICASSP2006 Proceedings. 2006 IEEEInternational Conference on*, vol. 5, pp. 1037- 1040, 2006.

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

[7] L. E. Baum and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554-1563, 1966.

[8] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian

mixture speaker models," *IEEE Trans. Speech and Audio Proc.*, vol. 3, no. 1, pp. 72-83, January 1995.
[9] B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Inform.Theory*, vol. 32, no. 2, pp. 307-309, March 1986.
[10] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *Journal of Acoustical Society of America*, vol. 35, pp. 354-358, 1963.
[11] P. D. Bricker, R. Gnanadesikan, M. V. Mathews, S. Pruzansky, P. A. Tukey, K. W. Wachter, and J. L. Warner, "Statistical techniques for talker identification," *Journal of Acoustical Society of America*, vol. 50, pp. 1427-1454, 1971.
[12] B. S. Atal, "Text-independent speaker recognition," *Journal of Acoustical Society of America*, vol. 52, 1972.
[13] *Speech Accent Archive*, George Mason University, [online] Available http://accent.gmu.edu.
[14] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete – Time Processing of Speech Signals*, NJ: IEEE Press, 2000.
[15] Aditya Bihar Kandali," Emotion Recognition From Assamese Speeches Using Mfcc Features And Gmm Classifier", Indian Institute Of Technology Kharagpur

Authors Bibliography

K. V. V. Kumar born in India 1987. Obtained his B.Tech, from VRS&YRN College of Engineering and Technology Chirala. During 2007-2010. & M.Tech from K.L University in the Specialization of Communication and Radar systems during 2010-2012.He is having More than 7 years teaching experience and having 8 international and national journals / Conference papers Like IEEE and SPRINGER He is Associate member of I.E.T.E and ISSE other bodies Like I.S.T.E. His research interested areas includes Image processing and Signal Processing. and Video Processing.

K.SRI LAKSHMIobtained his B.Tech from Kallam Haranadha Reddy Institute Of Technology Chowdavaram, Guntur Dt, AP in Electronics And Communication Engineering stream.

K.SRI LAKSHMIobtained his B.Tech from Kallam Haranadha Reddy Institute Of Technology Chowdavaram, Guntur Dt, AP in Electronics And Communication Engineering stream.

K.SAI SATISH REDDYobtained his B.Tech from Kallam Haranadha Reddy Institute Of Technology Chowdavaram,Guntur Dt,AP in Electronics And Communication Engineering stream.

M.L.TEJA HARIIKAobtained his B.Tech from Kallam Haranadha Reddy Institute Of Technology Chowdavaram,Guntur Dt, AP in Electronics And Communication Engineering stream

R.SRIDHARobtained his B.Tech from Kallam Haranadha Reddy Institute Of Technology Chowdavaram, Guntur Dt,AP in Electronics And Communication Engineering stream