ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, *January 2015*

Speaker Accent Recognition by MFCC Using K-Nearest Neighbour Algorithm: A Different Approach

Munish Bhatia[1], Navpreet Singh[2], Amitpal Singh[3]

Student, Computer Science, GNDU, Gurdaspur, India [1,2]

Assistant Professor, Computer Science, GNDU, Gurdaspur, India [3]

Abstract: A K-Nearest Neighbour Algorithm involving Mel-Frequency Cepstral Coefficients (MFCCs) is provided to perform Speech signal feature extraction for the task of speaker accent recognition. Mel-Frequency Cepstral Co-efficient is effectively used to perform the feature extraction of the input signal. For each input signal the mean of the MFCC matrix is used for pattern recognition .The K-nearest neighbour algorithm is based on evaluating minimum Euclidean distance measure from input data set to stored data set. Since large number of speakers of different accent are present, they can be grouped together depending upon their accent .Thus each signal coming from different group makes a distinct MFCC vector .In this paper we have compared the MFCC from global group to smaller sub groups**.**

Keywords: MFCC, K-nearest neighbour, Voice Code, Euclidean distance, Global group

I. INTRODUCTION

A speaker's accent distinguishes him/her as a member of a group. These groups have been classified on the basis of geographic areas, by social class or by ethnicity .As human being we are sensitive to accent and can tell whether a speaker belongs to a group or not .For above scenario to be possible ,accent are pattern of speaking that distinguish a group from other group. It is assumed that these stylised patterns affect word rate, the phonologic coding used in lexicon and the prosody of utterances[8][10].Speech technology is still facing problem with pronunciation variation across different accent groups. In automated speech recognition, a variation in accent used in testing and training can lead increase in word error rate. In synthesis of speech, voices are constant in one accent due to large use of corpus based methods operating by single speaker speech [9].

Speech Processing involves capturing of sound and process its features by extracting it .The technique for extracting feature is done by MFCC i.e. Mel FrequencyCepstralCoefficient [1] .MFCC is by far the most successful technique for performing feature extraction .MFCC is depiction of short term power spectrum on the basis of linear cosine transform of log spectrum on non-linear Mel scale of frequency.

II. STEPS FOR COMPUTING MFCC

1.*Pre-emphasis Filtering*

The input speech signal q (n) is sent to high pass filter

$$q_2 (n) = q (n)-a*q (n-1)$$

Where $q_2(n)$ is the output speech signal and a is valued between 0.9 and 1.0.The z-transform is given by

$$T (n) = 1-a/z$$

The goal is to compensate the high frequency portion that was supressed during production of sound.

2. *Framing*

The speech signal then is segmented into frames of ~25 ms with overlap of ~1/3 frame size.

3. *Hamming Window*

Each frame is multiplied with hamming window for the continuity of the first and last point of frame. Hamming windowing is q(n)*w(n) where w(n) is the hamming window.

4. *Fast Fourier Transform*

In this step we perform the FFT to get the magnitude ofthe frequency of each frame assuming that signal inside frame is periodic.

5.*Triangular Band pass Filtering*

Then magnitude of the frequency is multiplied by set consisting of about triangular band pass filter. The positioning of the filteris synchronous with the Mel frequency .Mel frequency is proportional to log of linear frequency.

6.*Discrete Cosine Transformation*

In this step DCT is applied to NlogEnergy resulting from band pass filter. Usually N=20.The result is the features similar to cepstrum that's why these are termed as Mel scale Cepstral Coefficients or MFCC [5].

Along with above, other dynamic features like Del1 (First order derivative) and Del2 (Second Order derivative) are included. As of now 13 MFCC coefficients are computed along with Del1 and DEL2 making total up to 39 coefficients. One important point to be noted that for each signal the MFCC actually form n*p matrix as shown in figure 1 where n is the number of window frames and p is the number of MFCC coefficient i.e. 39.Now the MFCC is passed to pattern recognition techniques[2][3] and for that these matrices are converted to vectors by simply taking the mean of the n column vectors.Fig1 shows an instance of MFCC matrix .Usually as mentioned N=20 .The mean of the column vector is shown in Fig 2.Mean vector is calculated as

Copyright to IJARCCE
DOI 10.17148/IJARCCE.2015.4131

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, *January 2015*

$a_1=(x_1+y_1+\ldots\ldots\ldots z_1)/n$
$a_2=(x_2+y_2+\ldots\ldots\ldots z_2)/n$ and so on.

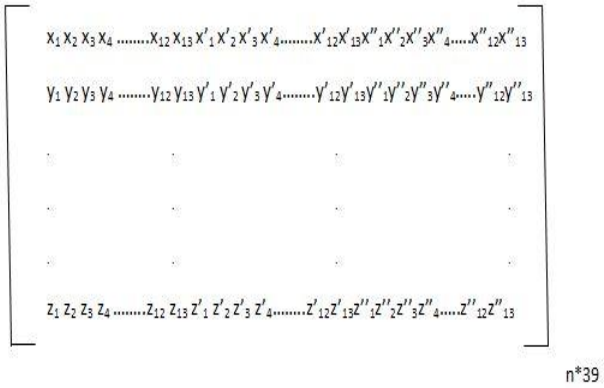

Fig1.MFCC matrix

Mean vector is considered as Voice code (V_C) .The three partition of the vector are termed as VOICE CODE 1(V_C1), VOICE CODE 2(V_C2), and VOICE CODE 3 (V_C3) respectively.

$A= [a_1a_2a_3a_4\ldots\ldots\ldots a_{12}a_{13}b_1b_2b_3b_4\ldots\ldots\ldots b_{12}b_{13}c_1c_2c_3c_4\ldots\ldots\ldots c_{12}c_{13}]$

Fig2MeanVoice Code MFCC Vector

In Fig2. A is the Mean Voice Code Vector, V_C1 is from a_1 to a_{13}, V_C2 is from b_1 to b_{13}, V_C3 is from c_1 to c_{13}.

K –NEAREST NEIGHBOUR ALGORITHM FOR ACCENT RECOGNITION

Basic Assumption

In order to assign a member to a group, initially it is assumed that all the different accent members form a single group which is considered as global group. Within that global group different accent are grouped to form local group with closely similar accents [4]. Now these sub groups are further partitioned into sub subgroups called regions which are different from each other. The number of partitions can be varied. Above scenario can be shown in Fig 3

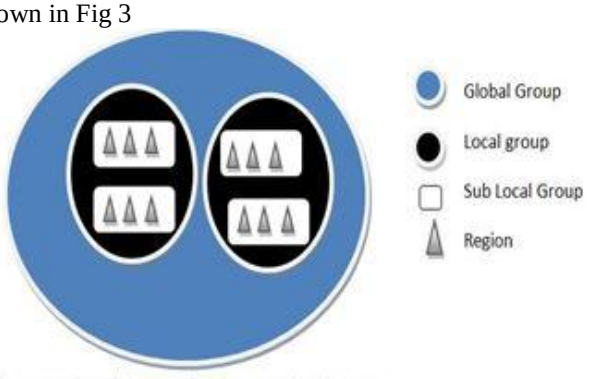

Fig3.Grouping of closely related Members in Sub Groups

Step1

Initially Input signal's V_C1 compared for Euclidian distance as given by Equation 1, with each mean of V_C1 of different local groups of Fig 4 showing global Group C with different sub Groups like C1, C2, C3 etc. These sub groups contain the mean of V_C1 of the stored set.

Equation 1

$$L\_DIST = \min \sum_{i=1}^{N} \sum_{k=1} \left(\left| A_k - Z_{ik} \right| \right)$$

Where L_DIST is the minimum local distance of the global group to local group, A_K is the input signal V_C1 and Z is the stored data set and N is the number of stored data set to compare with. Now as soon as the minimum distance is calculated, it can be estimated that the input signal is closely related the particular local group of minimum global distance. For instance as shown in Fig 4.

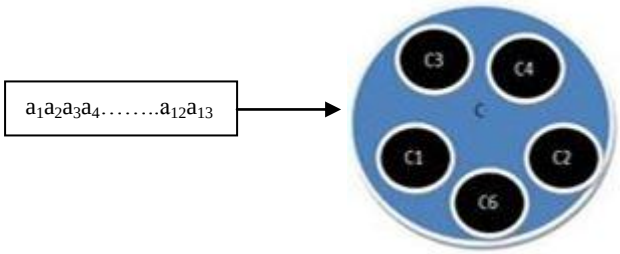

Fig4.Input Signal V_C1 compared with mean V_C1 of local groups in C

Let us assume that the input signal V_C1 is closely related to mean V_C1 of local group C1.

Step 2

After determining the local group, the next step is to determine the sub local group to which the member belongs .For that the comparison of input V_C2 is done with the sub local groups containing their respective V_C2's mean as shown in Fig5. Comparison using Equation 2 gives the minimum distance of local group to sub local group. Variables have their usual meanings.

Equation2

$$SL\_DIST = \min \sum_{i=1}^{N} \sum_{k=14} \left(\left| A_k - Z_{\,i\,k} \right| \right)$$

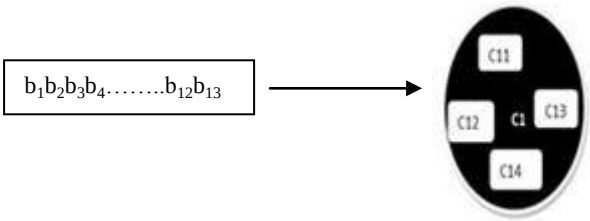

Fig5.Input signal V_C2 compared with mean V_C2 of Sub local Groups in C1

Assume that the input V_C2 is closely related to V_C2 of Sub local group C11.

Step 3

Similarly as above, repeating the step one more time we get the final region to which the member belongs. Finally the V_C3 is compared according to equation 3 which

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 1, *January 2015*

determines the minimum region distance and the final region of the member with the minimum distance given by DIST=L_DIST+SL_DIST+R_DIST

Equation3

$$R\_DIST = \min \sum_{i=1}^{N=39} \sum_{k=27} \left(\left| A_k - Z_{i\,k} \right| \right)$$

Fig 6 shows the final step.

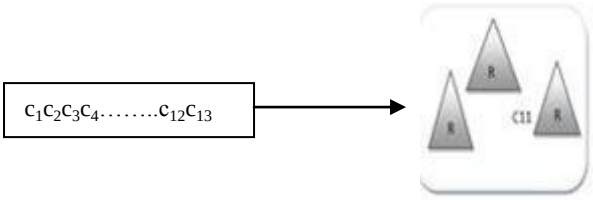

Fig6. Input signal V_C3 compared with mean V_C3 of regions in C11

CONCLUSION

We report the different approach of accent determination by using the K-nearest neighbour algorithm based on MFCC. In this paper we reviewed the classification of accents on the basis of MFCC coefficient obtained from input signal. We also considered the training data must be kept for determination of accent to greater accuracy. Further we are planning to compare different classifiers for different accents.

ACKNOWLEDGMENT

We would like to thank GNDU facility for all the help they have provided.

REFERENCES

[1] Liu Wai Kat Fung, P. "Acoustics, Speech, and Signal Processing, 1999",Proceedings1999 IEEE International Conference.
[2] Zichen Ma, Ernest Fokoué "A Comparison of Classifiers in Performing Speaker Accent Recognition Using MFCCs" published Online June 2014 in SciRes.
[3] SandeshAryal and Ricardo Gutierrez-Osuna "Accent Conversion through Cross Speaker Articulation" in IEEE proceedings. ICASSP 2014
[4] Malay K. Pakhira"A Modified *k*-means Algorithm to Avoid EmptyClusters" in International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009
[5] MFCCofficients Available at http://www.mirlab.org/jang/books /.../speechFeatureMfcc.asp?title=12-2%20MFCC.
[6] Gina-Anne Levow, "Investigating Pitch Accent Recognition in Non-native Speech".
[7] Carlos Teixeira, Isabel Trancoso and Antonio Serralheiro. "Accent Identification".
[8] M Huckvale ,"ACCDIST: a metric for comparing speakers' accents", International Conference on Spoken Language Processing 2004.
[9] Chao Huang, Eric Chang, and Tao Chen "Accent Issues in Large Vocabulary Continuous Speech Recognition (LVCSR)", Microsoft Research.
[10] FadiBiadsy, "Automatic Dialect and Accent Recognition and its Application to Speech Recognition".