# *Application of Fisher's ratio technique for speech feature evaluation of vowls/consonants*

**Sarika Hegde**
Department of Computer Science & Engg
NMAM Institute of Technology, Nitte
Udupi
Karnataka, India
sarika.hegde@yahoo.in

**Surendra Shetty**
Department of Masters of Computer Applications
NMAM Institute of Technology, Nitte
Udupi
Karnataka, India
hsshetty4u@yahoo.com

*Abstract* - **Feature extraction technique plays an important role in Automatic speech recognition. There are many techniques for extracting the speech features. Capability of various speech features for categorizing speech sounds differs and can be evaluated after applying classification or clustering model. As an alternative approach, a preliminary study (before classification) can be done about distinguishing capacity of a speech feature. Generally, visualization techniques are used for this purpose. In this paper, we are evaluating the capability of speech feature by computing and analyzing Fisher's ratio for the vowels and consonants. It's a well known technique but not being used widely in the context of Automatic speech recognition. The term Fisher's ratio is computed based on the centroid and variance of feature within a population and across various populations. It gives a picture about the measure of separation of objects of one class with another. The evaluation results are then compared with visualization techniques and classification accuracy.**

**Keywords: Automatic Speech recognition, Fisher's Ratio, Vowels, Consonants, Visualization, Classification**

## I. INTRODUCTION

The Pattern recognition is the most widely used approach for Automatic Speech Recognition (ASR). The objective of pattern recognition is to classify an input pattern into one of several classes based on its similarity to these predefined classes. Let $\boldsymbol{X}$ be a random observation from an information source, consisting of $C$ classes. The classifiers job is to classify each $\boldsymbol{X}$ (unknown observation) into one of the $C$ classes with minimum misclassification error. Automatic Speech Recognition using pattern recognition technique involves the steps like, Speech pre-processing, Feature Extraction, Acoustic Modeling, Pattern Classification and Decision Logic.

Fisher's ratio is mainly used in Fisher's Linear Discriminant for classifying two or more classes. This technique is based on computing a centroid for the populations corresponding to each of the classes. Population centroid information is used to compute a measure called Fisher's ratio also known as F-ratio, and it gives a picture about the measure of separation of one population with others. Fisher's Linear Discriminant attempts to maximize the Fisher's ratio during training and this is done by computing a linear combination of features for each population which maximizes the F-ratio (Johnson & Wichern, 2007). This measure is used by (Patro et al., 2007) to compare the various features in terms of its capability to distinguish one alphabet from others. However, a detailed analysis of using Fisher's ratio for speech feature's evaluation is not found much.

In our paper, every alphabet (i.e. vowel and consonant) is considered as a pattern and capability of various speech features is analyzed to distinguish the patterns successfully. A comparative analysis of various speech features can also be done. After applying exploratory analysis techniques, a speech feature can be discarded if it is found to be poor in representing the pattern. We have mainly used a statistical technique i.e Fisher's ratio for exploratory analysis. There are two advantages of applying exploratory analysis as a preliminary study. It gives an idea about the similarities among various patterns like which vowel sounds are similar or which words are similar in respective of different speech features. Another advantage is that it gives an idea about the weakness and strength of particular speech feature in the context of the required information. The objectives of the content in this paper is as follows,

1. Present a study on analyzing the capability of a speech feature to distinguish different speech pattern using Fisher's ratio.
2. Support the analysis by comparing the results with that of visualization techniques.
3. Support the analysis by comparing the results with classification model.
4. Explore the importance of preliminary study of speech features using statistical techniques.

978-1-5090-5297-4/17/$31.00 ©2017 IEEE

The contents in the paper are organized as follows. In the second section, a detail about the feature extraction, computation of Fisher's ratio is discussed. Third section discusses about the implementation detail including dataset, analysis of Fisher's ratio and the evaluation of speech feature using Fisher's ratio. The conclusions are given in the last section.

## II. METHODOLOGY

The first step in the process is to compute the speech feature for all the audio data. In the second step, Fishers' ratio is computed. In this section, first we describe the various speech feature extraction techniques used in this paper and then followed by the technique for computation of Fisher's ratio.

### *A. Feature Extraction Technique*

Popular feature extraction techniques have been used for speech processing in this paper. In the recent works, many innovations and improvements can be found in these features. But still, these basic features are very strong and any other feature extraction techniques are based on these fundamentals. The speech features computed here are, Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coefficients (LPC, Zero crossing rate (ZCR), Zero Crossing Time Interval (ZCTI), Formant Frequency. The given speech audio is divided into smaller frames of size 30ms with windowing. Then feature is computed for each frame.

#### i. Mel Frequency Cepstral Coefficients (MFCC)

The Mel-Frequency Cepstral Coefficient (MFCC) is a very popular feature used in representing speech data in Automatic Speech Recognition. This feature seeks to emulate human perception of decoding the speech audio to extract the content of spoken words. Before extracting MFCC feature, it is assumed that the speech signal is divided into frames by applying framing, windowing and overlapping techniques. General steps used in extracting MFCC feature for a single frame (Slaney, 1998) are,

i. Compute DFT coefficients
ii. Filtering using triangular filterbank with mel scale frequency
iii. Compute DCT coefficients

An approximate conversion between a frequency value in Hertz (*f*) and $mel$ is given by (Peltonen et al. 2002):

$$mel(f) = 2595log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

The equation for computing DCT coefficients is given by,

$$C_i = \sum_{k=1}^{K}(logS_k).cos\left(\frac{i\pi}{K}\left(k - \frac{1}{2}\right)\right) \; i = 1,2, \dots, K \quad (2)$$

where $C_i$ is the $i^{th}$ MFCC, $S_k$ is the output of $k^{th}$ filter bank channel (i.e. the weighted sum of the power spectrum bins on that channel), $K$ is the number of filterbanks.

#### ii. Linear Predictive Coefficients (LPC)

Linear predictive analysis is one of the most commonly used speech analysis and synthesis tool. Initially, this technique was used as speech coding technique for speech synthesis. But later it was used in speech analysis for speech recognition (Atal & Hanauer, 1971; Gangashetty & Yagnanarayana, 2001). Basic idea behind linear prediction is that the next signal sample is predicted from a weighted sum of $p$ previous samples, given as follows (Peltonen et al., 2002).

$$\hat{s}(n) = \sum_{i=1}^{P} a_i \, s(n - i) \quad (3)$$

#### iii. Zero Crossing Rate (ZCR)

Zero Crossing Rate (ZCR) is given as the number of times the time-domain signal crosses x-axis for a given frame, (McLoughlin 2009). If the signs of two consecutive amplitude values in a signal are same then it means that the signal has not crossed x-axis between these two points. ZCR is computed by counting number of times, signal crosses x-axis and this is done by analyzing signs of pairs of consecutive amplitude values. Computation of ZCR for a signal $x[n]$ of a given frame of size $L$ is given as, (Peltonen et al,. 2002)

$$Z = \frac{1}{2L}\left(\sum_{i=1}^{L-1}\left|sgn\big(x(i)\big) - sgn(x(i-1))\right|\right) \quad (4)$$

#### iv. Zero Crossing Time Interval (ZCTI)

This is one of the speech features which is recently applied for modeling and classification of Malayalam language vowels in Kumar et al., (2005). The importance of zero crossing locations and zero crossing intervals on the intelligibility of the speech is reported in (Niederjohn et al., 1987). The time interval between two successive zero crossing of the signal is considered as the feature. If two successive signal amplitude values differs in their sign (positive amplitude to negative amplitude or vice versa), it indicates that signal has crossed x-axis. But the exact point where signal crosses x-axis can't be estimated accurately. So, mid position between the successive amplitude values is considered as the zero crossing point. For each frame, list of zero crossing time intervals is generated. Finally the mean and standard deviation is computed for the sequence of zero crossing time intervals and used as feature. Kumar et al., (2005) have computed the distribution of such time interval as the feature and have used for modeling the vowels.

### v. Formant Frequency

This feature is related to the shape of the vocal tract of a human when he utters any speech. The shape of the vocal tract changes based on the type of sounds generated. There is a characteristics filter function associated with vocal tract based on its size and shape. The resonant frequency at which the signal passes with maximum energy is computed as formant frequency. In other words, frequency response of these filter function associated with vocal tract is called as formant frequency. Steps used to compute formant frequency are,

i. Compute LPC prediction polynomial
ii. Compute complex roots of prediction polynomial
iii. Convert the imaginary part of roots to Hertz

The linear prediction filter coefficients are computed with same method explained in the previous section. Number of coefficients $p$ is decided with the following formula (rule of thumb for formant)

$$p = 2 + fs/1000 \qquad (5)$$

After computing LPC coefficients, frequency response is calculated using root solving method. The roots are given as list of complex number pair. Formant frequency for each root pair is computed as

$$F = \frac{f_s}{2\pi}\theta\text{Hz} \qquad (6)$$

Number of formant frequencies depends on number of roots. In our work, we have considered up-to first six formant frequencies as feature vector.

## *B. Computation of Fisher's Ratio*

In this paper, F-ratio value is used as a measure of separation of classes (vowels, consonants). For a particular feature, higher F-ratio indicates that the feature contributes more in separating classes. To compute F-ratio, we compute the centroid of given feature and then compute the variance of this centroid within each alphabet (within class variance) and among the alphabets (between class variance).

Let us say, there are $C$ number of samples where each sample belongs to a particular alphabet, which is given as, $(\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_C})$ . For each sample $\boldsymbol{X_j}$ of size $n$, sample mean $\bar{x}_j$ is computed as,

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (7)$$

Sample variance ${s_j}^2$ is computed as ,

$${s_j}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_j)^2 \qquad (8)$$

Then we compute the mean of sample means of all the alphabets denoted as $\mu_{\bar{x}}$ which is given as,

$$\mu_{\bar{x}} = \frac{1}{C}\sum_{j=1}^{C}\bar{x}_j \qquad (9)$$

Using these terms, F-ratio is defined as,

$$F = \frac{\text{between class variance}}{\text{within class variance}}$$

$$F = \frac{\frac{1}{C-1}\sum_{j=1}^{C}(\bar{x}_j - \mu_{\bar{x}})^2}{\frac{1}{C}\sum_{j=1}^{C}{s_j}^2} \qquad (10)$$

For any given feature, high value of $F$ indicates a better separation of alphabets with that feature. This measure can also be used to compare the various features in terms of its capability to distinguish one alphabet from others (Patro et al., 2007). This measure to rank the coefficients of a $p$ dimensional feature based on high F-ratio value. In this case, we have to compute F-ratio for each coefficient separately and then the coefficients in the order of ranking can be considered for further processing.

# III. IMPLEMENTATION AND RESULTS

This section gives a description about the dataset and the detail about various experiments done along with results. The results are discussed in three parts. The first part analyzes the value of Fisher's ratio, the second part evaluates the speech features based on fisher's ratio and the last part discusses the comparison of Fisher's ratio results with visualization and classification.

## *A. Dataset*

Dataset consisting of recorded vowels, consonants sounds and words of Kannada language are used. A database of Kannada language sentences is published by (Prahallad et al., 2012) in the website http://www.iiit.ac.in. This is downloaded and manually segmented into words/vowel/consonant sounds using Audacity, a sound editing software.

List of vowels and consonants sounds considered are given in Table 1. Vowel sounds of ten categories are merged into five groups and consonants (Structured and Unstructured consonants) sounds of 18 categories are used.

Thirty patterns of each category is used for the experiment. All the experiments are done separately for vowel dataset, structured consonant dataset and unstructured consonant dataset.

| **Vowels** | **Structured Consonants** | | | **Unstructured Consonants** | |
|---|---|---|---|---|---|
| /a/ | /k/ | /dd/ | /b/ | /y/ | /h/ |
| /i/ | /g/ | /t/ | /m/ | /r/ | |
| /u/ | /ch/ | /d/ | | /l/ | |
| /e/ | /j/ | /n/ | | /v/ | |
| /o/ | /tt/ | /p/ | | /s/ | |

Table 1. List of vowels and consonant

### *B. Analysis of Fisher's Ratio*

For each of the audio clip in the dataset, first we compute the features in the order MFCC (12 coefficients), LPC (12 coefficients), Zero crossing rate (1 coefficient), Zero crossing time interval mean and standard deviation (2 coefficients), Formant frequency – first six formant frequency (6 coefficients). The total number of coefficients in the feature will be 33. Each coefficient is referred with its index. MFCC coefficients are numbered from 1 to 12. LPC coefficients are numbered from 13 to 24. ZCR index is given as 25 followed by 26 for ZCTI mean and 27 for ZCTI variance. Finally the formant frequency is numbered from 28 to 33.

For each of the coefficients in the feature, fishers' ratio is computed. The above graph in Figure 1, gives the value of Fishers' ratio for 33 coefficients of vowel, structured consonants and unstructured consonants dataset. The x-axis in the graph indicates the different feature and y-axis value gives the Fishers ratio. It can be observed that the ratio value ranges from 0 to 7. But most of the values are in the range of 0 to 2.

The top five coefficients with highest Fisher's ratio and the bottom five coefficients with lowest Fisher's ratio is shown in Table 2. In the table the first column (Highest 5 coeff) in each dataset lists the name of the coefficients with highest Fisher's ratio in that order of high to low. The second column (Lowest 5 coeff) in each column indicates the six coefficients with lowest Fishers ratio in the order of low to high.

Table 2. Feature coefficients with highest and lowest Fisher's ratio

| Vowel Dataset | | Structured Consonant Dataset | | Unstructured Consonant Dataset | |
|---|---|---|---|---|---|
| Highest 5 Coeff | Lowest 5 Coeff | Highest 5 Coeff | Lowest 5 Coeff | Highest 5 Coeff | Lowest 5 Coeff |
| MFCC3 | MFCC7 | ZCR | LPC10 | ZCR | MFCC1 |
| MFCC2 | LPC3 | MFCC1 | LPC8 | MFCC5 | MFCC10 |
| MFCC1 | LPC2 | LPC3 | LPC9 | ZTI(Mean) | MFCC9 |
| MFCC5 | LPC10 | ZTI (Mean) | LPC7 | FF3 | MFCC12 |
| FF1 | LPC8 | FF2 | LPC12 | LPC3 | MFCC2 |

In case of vowel dataset MFCC3 coefficient is having the highest Fishers' ratio given as 6.6638 and the coefficient MFCC7 has the lowest fisher's ratio given as 0.0958. Similarly, the coefficient ZCR is having the highest fisher's ratio (3.0075) in case of structured consonants and LPC10 with lowest fishers' ratio (0.0806). Highest fisher's ratio in case of unstructured consonants is 1.2463 for ZCR and lowest fisher's ratio is for MFCC1 with value 0.06197.

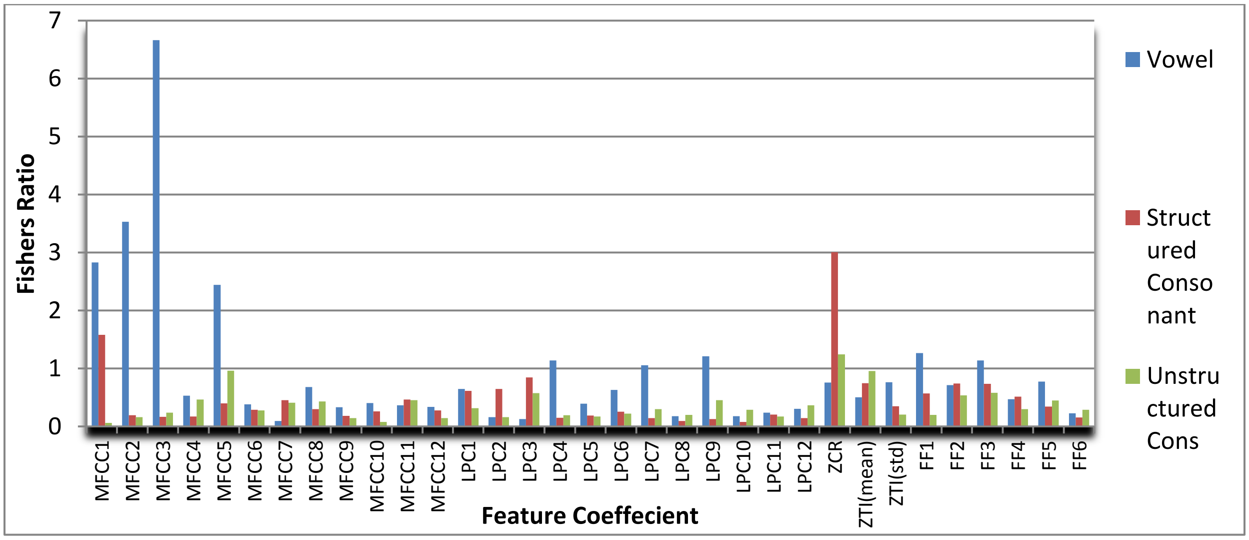

Figure 1. Fishers Ratio for all the coefficients of a feature of Vowel and Consonants dataset

### C. Evaluation of Speech features

By observing the values in the tables, it can be stated that the fishers' ratio varies for different feature coefficient across different dataset. According to the property of Fisher's ratio, the coefficients with higher Fisher's ratio must contribute more in discriminating the alphabet classes compared to the coefficients with lowest Fisher's ratio. For analyzing this property, we have done two types of experiments; one with visualization and another experiment classification.

In visualization, we plot a 3D graph using 3 coefficients with top fishers ratio and similarly another graph with lowest fisher's ratio. The two graphs will be compared visually for analyzing which of the graph is showing alphabets groups or classes more clearly. The graphs in Figure 2, shows the 3 dimensional plots of first three coefficients for vowel dataset. The first graph shows the plot for coefficients with high F-ratio and second graph shows the plot of coefficients with low F-ratio. Similar graphs are plotted for featured coefficients of structured and unstructured datasets as shown in Figure 3 and Figure 4 respectively.

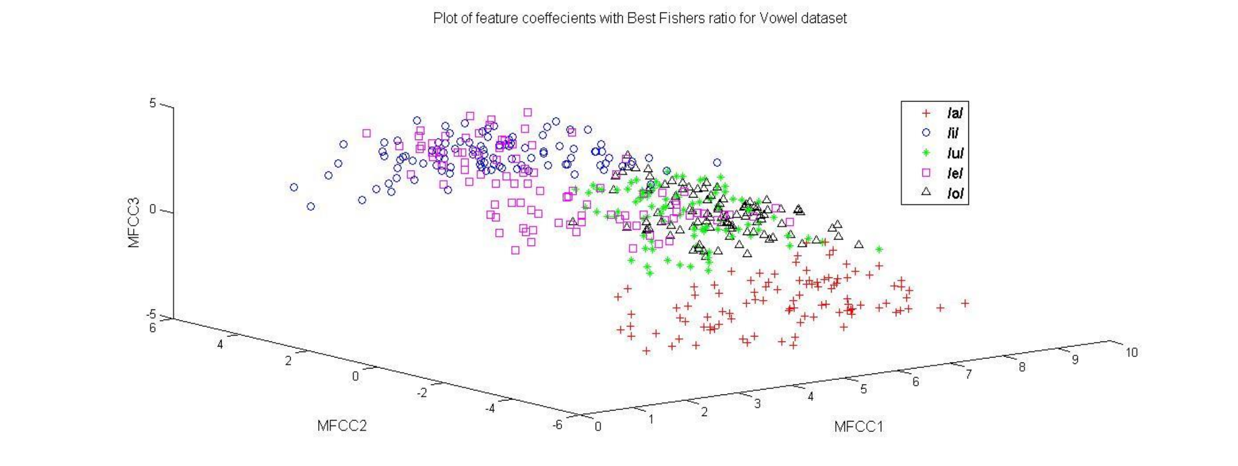

Figure 2(a). A three dimensional plot of coeffecients with highest Fisher's ratio for vowel dataset

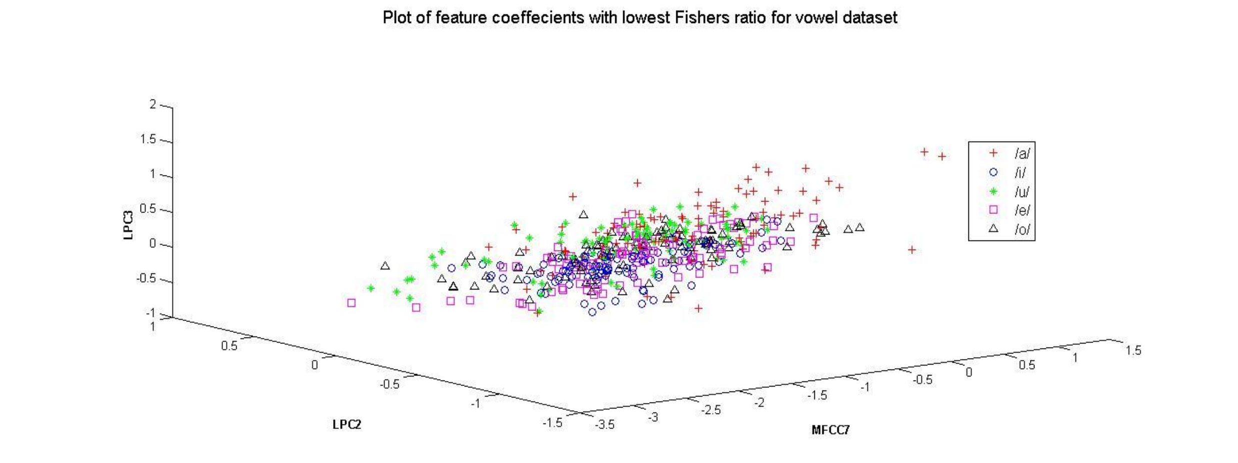

Figure 2(b). A three dimensional plot of coeffecients with lowest Fisher's ratio for vowel dataset

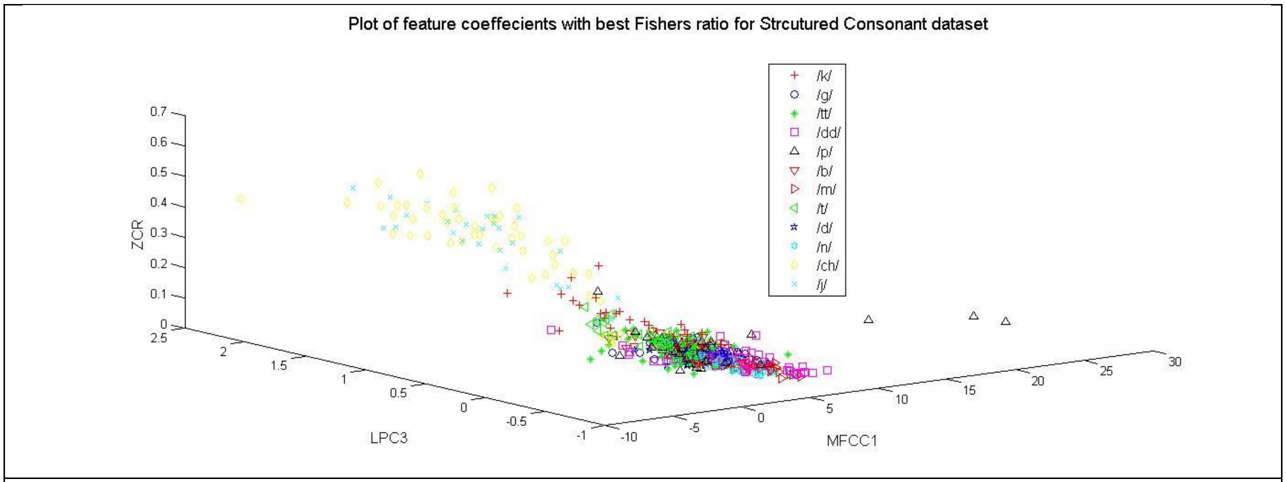

Figure 3(a). A three dimensional plot of coeffecients with maximum Fisher's ratio for structured consonant dataset

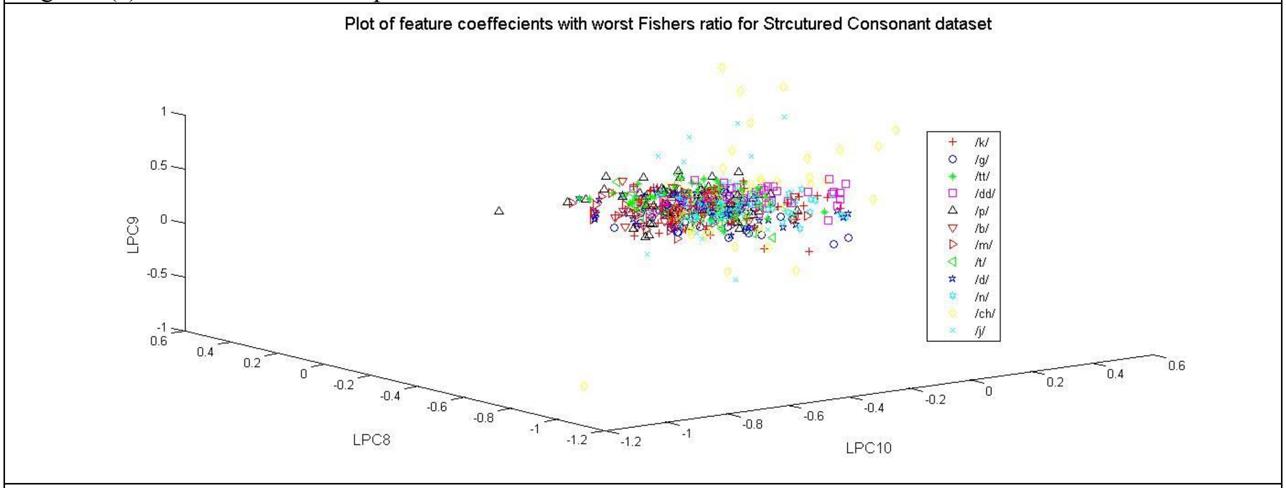

Figure 3(a). A three dimensional plot of coeffecients with minimum Fisher's ratio for structured consonant dataset

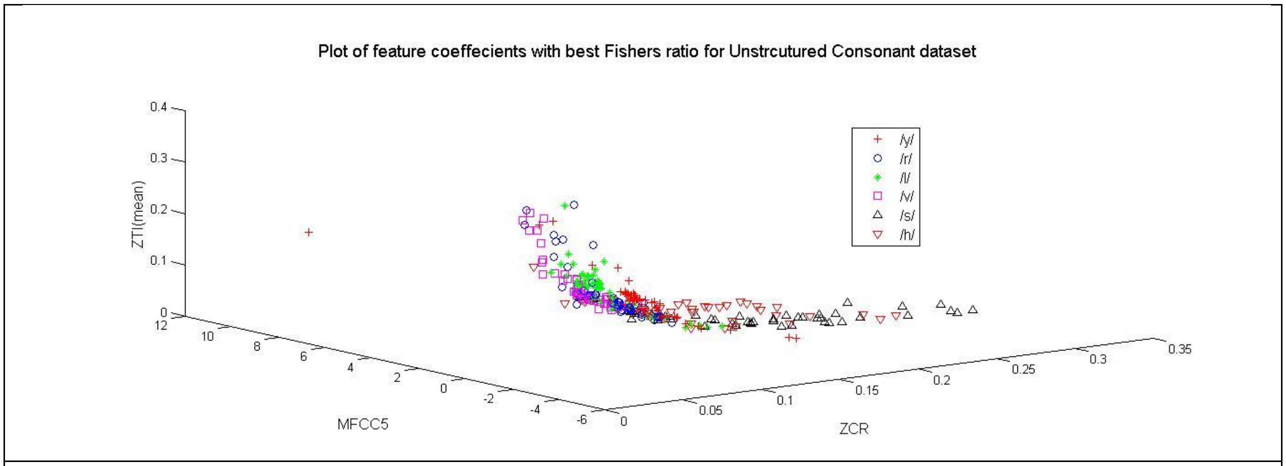

Figure 4(a). A three dimensional plot of coeffecients with highest Fisher's ratio for unstructred consonant dataset

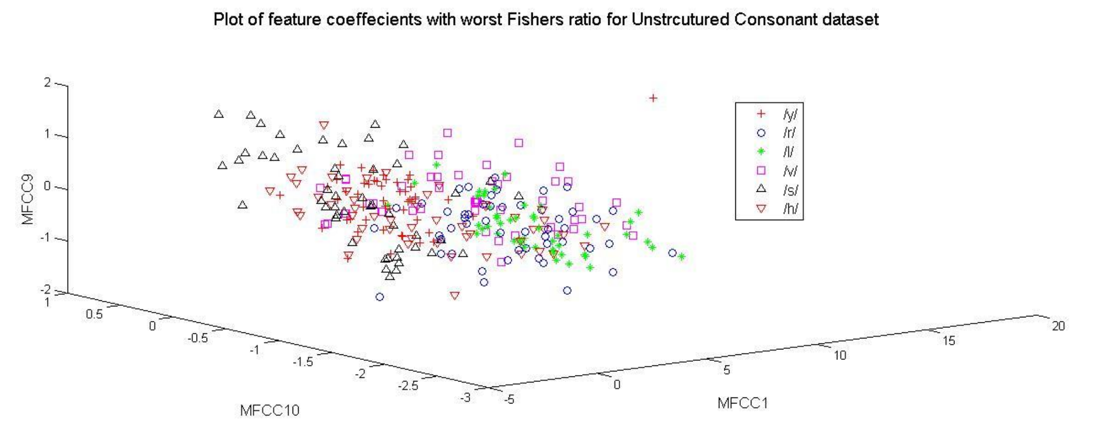

Figure 4(b). A three dimensional plot of coeffecients with lowest Fisher's ratio for unstructred consonant dataset

In all the figures, it can be observed that the first graph is showing the discrimination of classes more clearly than the second graph. Based on this comparison we want to support the claim that higher Fishers ratio for a feature indicates a better capability to discriminate classes and there by Fishers ratio technique can be used at preliminary level to evaluate the speech feature's capability.

The second experiment done was to compare the result of Fishers ratio with that of classification. For this, we selected the top 'n' coefficients with highest Fisher ratio and bottom 'n' coefficients with lowest Fisher ratio from each observation in dataset. The classification model built using SVM classifier is applied on these datasets. Hold out technique is used for generating training and testing sets randomly. The values of 'n' are chosen to be 8, 10, 12, and 14. Twenty iterations of classification (which includes both training and testing) is done and the average classification accuracy is computed.

| | Vowel | | St. Cons | | Unst. Cons | |
|---|---|---|---|---|---|---|
| No of Coeff | Max | Min | Max | Min | Max | Min |
| 8 | 71.43 | 64.29 | 51.79 | 44.64 | 73.81 | 48.81 |
| 10 | 78.57 | 70.14 | 48.81 | 42.02 | 65.48 | 53.37 |
| 12 | 74.13 | 65.71 | 55.95 | 45.83 | 77.38 | 67.86 |
| 14 | 71.43 | 67.14 | 58.00 | 50.33 | 73.81 | 67.86 |

Table 3. Classification accuracy for vowel, structured consonant and unstructured consonant dataset with varying number of coefficients.

The values in table 3, shows the average accuracy of classification (in %) for Vowel dataset, structured consonant and unstructured consonant dataset separately. Each row indicates the number of coefficients considered out of 33 coefficients in a feature. The first column within each dataset indicates the accuracy of classification with coefficients having highest Fishers ratio and second column with lowest Fisher's ratio.

From the above table, it can be observed that for every combination of dataset and number of coefficients, classification accuracy is high for coefficients with maximum F-ratio compared to that of coefficients with minimum F-ratio.

Based on the results of both visualization and classification, we can say that high Fisher's ratio for a speech feature indicates its strong capability of discriminating various classes. However, the threshold value for this ratio is not derived and the measure for the value is decided based on comparison. If there are two coefficients, then we just check which coefficient has high F-ratio compared to the other.

In case of feature with large number of coefficients, this technique can be used at preliminary level for selecting coefficients that contributes maximum for discriminating classes.

## IV. CONCLUSIONS

In classification of any type of data, selection of feature is significant for its success ratio. In this paper, we have analyzed the Fisher's ratio as a technique for assessing the capability of a speech feature to discriminate vowels and consonants. The feature set considered are MFCC, LPC, ZCR and FF. Fishers ratio is computed for all the datasets and evaluated. The higher Fishers ratio indicates a stronger feature and vice versa. The selections of coefficients are done based on maximum and minimum F-ratio. The results of F-ratio analysis is compared with that of Visualization and classification technique. Coefficients with higher F-ratio better classify than coefficients with lower fishers ratio. Based on this experiment, Fishers ratio can be used as a technique for dimensionality reduction of a feature set.

## References

[1]Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B), 637–655.

[2]Gangashetty, S. V., & Yegnanarayana, B. (2001). Neural network models for recognition of consonantvowel (c n v) utterances. *In Neural networks*, 2001. proceedings. ijcnn'01. international joint conference on (Vol. 2, pp. 1542–1547).

[3]Johnson, R. A.,Wichern, D.W., et al. (2007). *Applied multivariate statistical analysis (Vol. 4).* Prentice hall Englewood Cliffs, NJ.

[4]Kumar, R. S., & Lajish, V. (2013). Phoneme recognition using zerocrossing interval distribution of speech patterns and ann. International Journal of Speech Technology, 16(1), 125–131.

[5]McLoughlin, I. (2009). *Applied speech and audio processing: with matlab examples*. Cambridge University Press.

[6]Niederjohn, R. J., Krutz, M.W., & Brown, B. M. (1987). An experimental investigation of the perceptual effects of altering the zero-crossings of a speech signal. Acoustics, Speech and Signal Processing, *IEEE Transactions* on, 35(5), 618–625.

[7]Patro, H., Raja, G. S., & Dandapat, S. (2007). Statistical feature evaluation for classification of stressed speech. *International Journal of Speech Technology,* 10(2-3), 143–152.

[8]Peltonen, V., Tuomi, J., Klapuri, A., Huopaniemi, J., & Sorsa, T. (2002). Computational auditory scene recognition. In Acoustics, speech, and signal processing (icassp), 2002 *IEEE international conference on (Vol. 2, pp. II–1941).*

[9]Prahallad, K., Elluru, N. K., Keri, V., Rajendran, S., & Black, A. W. (2012). The iiit-h indic speech databases. In Interspeech

[10]Slaney, M. (1994). Auditory toolbox: a matlab toolbox for auditory modelling work. *In Tech. rep. 45, apple technical report.* Apple Computer Inc.