Analysis of Speech Signal Data of Mising Vowels using Logistic Regression and K-Means Clustering

Ujjal Saikia[1]
Centre for Computer Science and Applications
Dibrugarh University, Dibrugarh, India

Jiten Hazarika[2]
Department of Statistics
Dibrugarh University, Dibrugarh, India

***Abstract*—In this paper, an attempt has been made to study and analyze speech signal data. Here, the sound or speech data has different attributes like time, pitch, formant frequencies, speaker type, Vowel No etc. The dataset used here is speech signal data which are analog in nature and has been converted to digital format. After converting the data into digital format we want to establish a Logit model to predict the speaker gender on the basis of the pitch signal values which is also considered as fundamental formant frequency. That is our objective is to predict whether a speaker is male or female by looking at the pitch value by using logistic regression. We have applied clustering techniques to visualize and interpret how it works in speech signal data. The logistic model gives us 91% accuracy rate with low and efficient AIC value where as in case of the clustering algorithm we get a 93% accuracy for the whole sample**.

Keywords—Clustering methods; formant frequency; Logit model; pitch; SType

I. INTRODUCTION

The role of statistical techniques in data analysis is vital. Proper use of statistical techniques in data analysis can provide very fruitful result. The key thing is that we have to select proper methodologies and techniques. One of the most frequently used method is the Regression analysis. The regression analysis can give us a very good idea about the whole dataset while applied carefully and selectively. Regression methods like simple linear regression, multiple linear regression are frequently used for data analysis where the relationship between dependent and independent variables are linear in nature. While Logistic regression is preferred when the regressand is a categorical variable. Through the literature Survey it has been observed that various statistical methods like Regression, Classification etc. have been applied in different types of data. Greenstein, J. [1954], shows the effect of television viewing upon elementary school grades using multiple and partial linear regression analysis in educational dataset [1]. Warner and Misra [1996] discussed the use of neural network and compares it with traditional regression analysis. They show that neural network works as a nonparametric regression model and it enables to model complex functional form. They also discussed the advantages and difficulties of using neural network against the use of regression analysis [2].

Gibbs et al. [2006] shows the use of regression analysis to establish the relationship between home environment and reading achievement for institutional cum educational datasets [3]. Pao[2008] made an empirical study on comparison of neural network and multiple regression analysis in modelling capital structure. This study adopted multiple linear regressions and artificial neural networks models with seven explanatory variables of corporation's feature and three external macro-economic control variables to analyse the important determinants of Capital Structure data [4]. Keshavarzi & Sarmadian[2010] compared Artificial Neural Network and Multivariate Regression in Prediction of Soil cation exchange capacity. Investigation of soil properties like Cation Exchange Capacity (CEC) plays important roles in study of environmental research as the spatial and temporal variability of this property have been led to development of indirect methods in soil data analysis [5]. Prica and Sinisa[2010] has done experimental study for recognizing Vowels in continuous speech by using parameters like formant frequency in speech signal data [6]. Ganesan et al. [2010] shows the Application of statistical and machine learning techniques in Diagnosing Cancer Disease using Demographic Data [7]. Raghavendra and Srivatsa[2011] evaluated Logistic Regression and Neural Network Model with Sensitivity Analysis on medical datasets. The goal of this research work was to compare the performance of logistic regression and neural network models on publicly available medical datasets [8]. Al-Shayea [2011] also shows the use of ANN modelling in Medical Diagnosis [9].From the experimental results, it is confirmed that the neural network model with sensitivity analysis gives more efficient result. Different visual plotting techniques has already been used in speech signal data the visually detect the significance of the dataset and its different attributes. Rehman and Hazarika [2014] have done experimental studies for analyzing and recognition of vowels of low resource languages by using speech signal dataset [10]. Another relevant study of using statistical technique is the use of it in the verification and identification of speaker. Zhang [2018], has shown the use of Linear Regression for Speaker Verification [11]. There is also scope in using Statistical methods while studying Acoustic properties of speech signal data. Some previous studies pointed out how to apply basic statistical methods. Saikia et al. [2019] has used in effective data visualization [12]. Jiang et al. [2020] has described statistical methods for Feature Extraction Method for Speaker Recognition very efficiently in their work with proven experimental outcomes and results [13]. Magdiel and Pilar [2021] used standardized domain adaptation techniques for classification in imagined speech recognition [14]. Babak et al. [2021] reviews deep learning approaches in speech emotion recognition by using machine learning techniques.

They have applied these techniques in already available statistical datasets [15].

In our present study we have applied statistical as well as artificial machine learning methods in sound data. We have given our emphasis on building regression model in speech signals with the following objectives.

- To build a model by applying logistic regression to identify male and female speaker on the basis of the pitch values.
- To apply clustering methods to the speech signal data to categorize male and female speakers separately and check the efficiency of the model.

II. Background and Methodologies

A. Speech Signal Data

Speech signal data are presentation of sound data in digital format. It is basically an analogue sound recorded in a closed environment. Closed or sound proof environment is a necessary condition to record sound for analytical purposes. Otherwise it may be affected by other sounds. Noisy environment may hamper to get the actual sound regarding the pronunciation of a particular vowel or any words which is to be prepared for analysis. A sound signal data after converting to digital dataset may have the following attributes. Let us give a brief idea about some of the attributes and parameters of the sound data.

1) Time: It is the duration or time period for which a sound is uttered by human generally represented in seconds during vowel utterance.

2) Speaker type: It is the type of the speaker. For example whether the sound is pronounce by male or female. Or a speaker from native or non-native background etc. This variable is categorical in nature.

3) Pitch (F0): It is known as fundamental frequency. The pitch is defined as the rate of vibration of the vocal chords of the person who is pronouncing the particular sound. This pitch frequency varies across different sounds. It also varies depending on the speaker type.

4) Formant frequencies: One of the distinguishing frequency component of human speeches are the formant frequencies. These values are also obtained from the recorded sound. It is considered as the particular resonance frequency of the vocal tract which is considered to have the maximum frequency during vowel utterance. The formant frequencies are denoted by the symbols F1, F2 etc.

5) Vowel no: It refers to the serial numbers of the Vowels of the language while considering for analysis. Here in Mising language we have a total of 14 Vowels.

B. About Mising Language

Here in our study we have selected sound data which is related to vowel utterance of Mising language. So it is convenient to discuss about the language "Mising". It is a language from North-Eastern India. It is a mixture of Indo-Aryan and Sino-Tibetan family of language. This language is considered as linguistic offshoot of the Tibeto-Burman branch of Sino-Tibetan family. According to Census 2001 only 5 Lakhs people left who speaks this language and day by day it is decreasing consistently. The people who can write this language by using their own script is also very less. It is considered as one of the low resource languages that are spoken by people in Assam. The language is considered as low resource language because of lack of content in internet and online resource. The Mising language has 14 vowels and 15 consonant. In our present study we have included the speech signals of the following 14 vowels of Mising language. They are: /i/ , /i:/, /e/ , /e:/, /a/ , /a:/, /o/ , /o:/ , /u/ , /u:/ , /é/ , /é:/ , /í/, /í:/.

C. Regression Methods

Regression Analysis is one of the most widely used data analysis technique for prediction. It is used to investigate the relationship between dependent and independent variables. The dependent variable is known as target variable while the independent variable are known as predictor variable. The terms "Regressand" and "Regressor" are also used for dependent and independent variables respectively under study. Most common types of regression techniques are outlined below.

1) Simple linear regression: This method is adopted while we have one target variable and one predictor variable. The simple linear regression is the simplest of all the others and can be successfully implemented in different situations in case of two variables when the relationship between them is linear in nature.

2) Multiple linear regression: Multiple linear regression is also known as multiple regression. This method is used when we have to predict the outcome of a dependent variable using several independent or explanatory variables. This method is used in case of more than two variables. The relationship among them is assumed to be linear in nature.

3) Logistic regression: Logistic Regression is a regression model to model a binary dependent variable. The model is known as logistic or logit model and it is used to model the probability of a certain class or event such as Yes/No, Pass/Fail, Male/Female, Affected/Not affected etc. That is, this technique is used when regressand is basically a categorical variable. Here in our current study we have tried to build a logit model for Speaker type when we are given the pitch value. We also determined model accuracy on the basis of true prediction and false prediction for the total number of test data on the basis of the train model. We have checked the model with different parameters like Combination of F0 and VowelNo, F1, Vowel No etc. But all the models based on them is results in higher AIC values. AIC stands for the Akaike information criterion which is an estimator of out-of-sample prediction error, given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.Generally a lower AIC value is expected. Lower the Value of AIC better is the Model. After that we have checked

the same with some clustering algorithm which are based on machine learning.

4) ANN models: ANN stands for Artificial Neural Network. These models are useful in case of building nonlinear and complex relationships among variables under study. The models based on Artificial Neural Network may be used to model complex relationships between inputs and outputs that is for the independent and dependent variables or to find patterns in data and future prediction. A part of the dataset is trained for prediction and on the basis of the trained data outputs are checked for efficient prediction. ANN models are often considered as a computational or mathematical model which inspired by the functioning human brain that is biological neural networks.

5) Clustering techniques: Like regression analysis Cluster Analysis also very much widely used techniques for data analysis. The method is simply known as Clustering. Clustering is defined as a task of grouping of objects or data in such a manner that the objects in the same group are more similar to the each other compared to the objects of the other groups or classes. Clustering is one of the main techniques used in data mining or fact finding in data. It is a common technique in statistical practice which is widely used and applied in various fields like pattern recognition and machine learning. Here in our dataset we have applied K-means clustering method to group our data into different categories. Let us first give a brief idea about K-means Clustering.

6) K-Means clustering: K-means Clustering is one of the most commonly practiced unsupervised machine learning algorithm which is used for partitioning the whole data set into K no. of groups or cluster. In general K is specified prior to the analysis by the data analyst. It classifies the objects or values into two or more groups in such a way that the objects within the same group or clusters are quite similar while the objects from the different clusters are quite dissimilar. Each cluster in K-means clustering is represented by its centroid or centre which corresponds to the mean of the points. The basic idea behind the K-means clustering is discussed below.

In K-means Clustering the total within cluster variation is minimized. There are many K-means algorithms. Here we have used the standard algorithm known as Hartigan-Wong algorithm (1979). This algorithm defines the Euclidean distances between item values and the corresponding centroid.

III. About the Dataset

The dataset used here is secondary in nature while collected in analogue format. It has later been converted to digital format using the application software PRATT for analysis purpose. The dataset looks like as given in Table 1. The table 1 indicates the first 28 samples. The data recorded in a noise free environment inside lab and there after converted to digital format by using PRAAT Software.

In the next page we are showing the sample dataset with basic discussion and about pre-processed data. The attributes like Time, F0, F1, F2 Speaker Types or Speaker Gender (SType) and serial no of Vowels (VowelNo).

1) Discussion about the data: The Mising Data in Table 2 stands for digital dataset which are originally sound signals recorded in the laboratory. That is the data were originally of analogue signals which are converted to digital format by using PRAAT software in computer labs.

TABLE I. Data Dictionary

Serial No.	**Parameters and Detailed Info**
1	The population size 70
2	Sample dataset size 28
3	Total Speakers Male: 28 Total Speakers Female: 42
4	Recording Environment: Laboratory(Closed Noise Free)
5	Recording Equipment Device: PHILIPS Microphone with Noise Cancellation feature
6	Channel: Mono
7	Frequency of the Sampling: 22050 Hz
8	Software: PRAAT
9	Language Spoken: MISING
10	Speakers: Graduates and Post Graduate Students of from MISING community
11	Speaker Type: Native

TABLE II. SAMPLE DATA SET OF FIRST 28 DATA POINTS

The First 28 sample points of Speech signal data of the Low Resource Language MISING					
Time	**F0**	**F1**	**F2**	**SType**	**VowelNo**
0.345669	235.222489	993.258705	1533.717724	1	1
0.282823	237.936687	175.815388	1141.23666	1	2
0.329955	244.257104	648.340679	1213.0288	1	3
0.3928	243.528907	744.776094	1173.473009	1	4
0.34568	197.99159	817.4236	1810.99103	1	5
0.209974	250.077406	836.121161	1754.962313	1	6
0.370915	280.340615	305.037348	1463.17897	1	7
0.298526	274.983809	476.091377	1587.204281	1	8
0.298526	250.441551	396.317328	598.567316	1	9
0.282823	270.121203	532.275868	1729.586239	1	10
0.3928	254.759789	613.597921	1023.284447	1	11
0.23568	240.601283	633.00456	984.396661	1	12
0.361372	271.812947	439.597155	807.948917	1	13
0.267109	254.08066	358.673299	664.206643	1	14
0.408209	127.535411	792.494447	1256.335354	0	1
0.343753	122.173526	743.720154	1209.341717	0	2
0.365238	133.058353	524.762969	1925.038107	0	3
0.408209	143.028187	542.013213	1961.55233	0	4
0.408209	136.462751	514.975377	1411.257475	0	5
0.408209	136.462751	514.976142	1411.259634	0	6
0.451179	143.237093	311.220514	2267.531138	0	7
0.386723	172.180501	325.989119	2336.622676	0	8
0.429694	148.157406	372.217248	1519.504391	0	9
0.451179	166.545805	337.593758	1364.08963	0	10
0.429694	143.768601	452.017997	836.50649	0	11
0.386723	128.887188	428.04571	800.253539	0	12
0.429694	138.032437	319.960304	800.001534	0	13
0.365238	145.755593	372.810288	823.053677	0	14

IV. RESULTS

A. Results of Logistic Regression Method

For our current study we are using logistic regression for model building by considering SType as dependent variable and F0 (pitch) as independent variable. We have used R programming for analysis purpose and following observations were made in our train data set. SType '1' means the speaker is Female and SType '0' means speaker is Male. Results of fitted logistic model are shown in Table 3.

Mylogit is the logistic regression model build for predicting SType from F0 values. We have predicted the value of SType by using the model which gives fruitful result. We have shown the predicted value along with the graphical plot as shown in Table 4.

TABLE III. BASIC RESULTS OF FITTED LOGISTIC MODEL

(R codes for building the model)		
Mylogit <- glm(SType ~ F0 , data = mydata, family = "binomial")		
Coefficients		
(Intercept)	F0	
-19.7270	0.1186	
Degrees of Freedom:		
Total	Residual	
57	56	
Null Deviance:	**Residual Deviance**	**AIC**
77.9	11.92	15.92

TABLE IV. PREDICTED VALUE USING PITCH

Serial No	SType (Observed)	SType (Predicted Value)
1	0	0.01167
2	0	0.03378
3	0	0.06459
4	0	0.08037
5	**0**	**0.50723**
6	1	0.97722
7	1	0.99971
8	1	0.99979
9	1	0.99989
10	1	0.99987
11	1	0.99995
12	1	0.99999

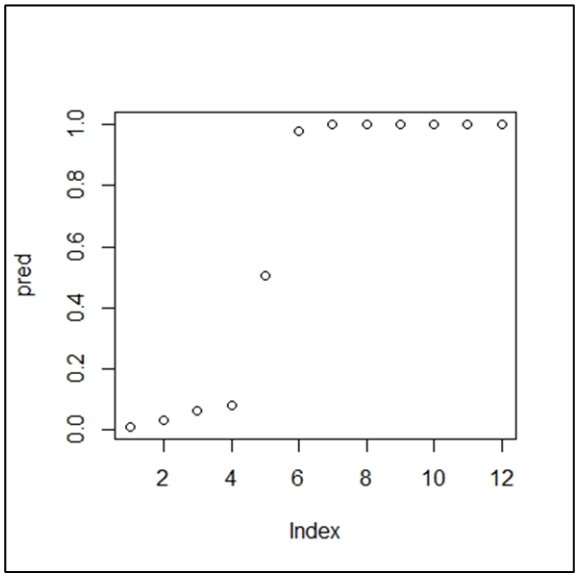

Fig. 1. Graphical Plot of Predicted Value.

B. Results of K-Means Clustering

While applying K-means Clustering, the following observations were made. It divides the whole data set into two groups of sizes 41 and 29. The analytical result gives us the following information. It shows the Cluster mean values for both F0 and SType along with the Clustering vectors and within cluster and sum of squares by cluster as shown below in TABLE 5.

The Cluster Plot was performed by using fviz_cluster () function of R. It give elegant visualization of partitioned data into different groups.

TABLE V. RESULTS BASED ON K-MEANS CLUSTERING

K-means Clustering with 2 Clusters of sizes 41, 29	
Cluster Means	
F0	**SType**
1 234.0461	1.00000000
2 138.4828	0.03448276
Clustering Vector:	
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	**[39]** 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
Within cluster	**Sum of squares by Cluster**
[1] 44696.034	3553.969
(between_SS / total_SS = 76.3 %)	

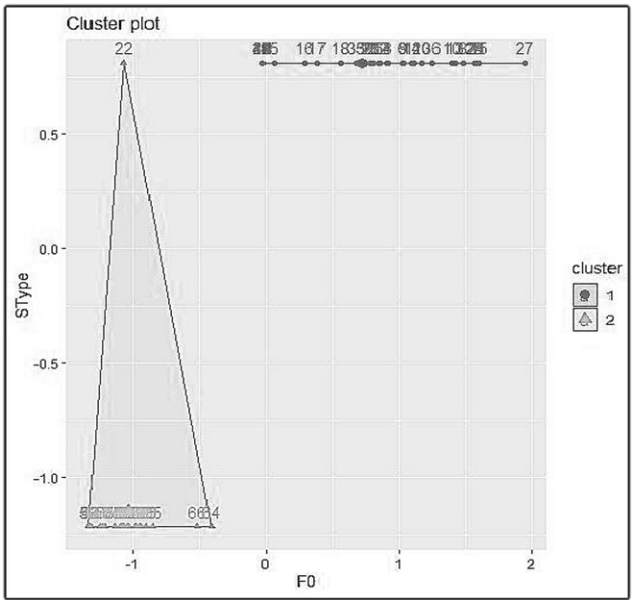

Fig. 2. Cluster Plot of SType and F0.

V. INTERPRETATION

In our experimental study a total of 58 values are used as a train dataset and there after 12 pitch values are used for prediction of gender (SType) from the given pitch (F0). We have seen that almost all values are well predicted except serial no. (5), which shows a critical predicted value 0.50723 against a '0' observed value which may lead to a false prediction if we take the threshold of prediction from 0.5. [TABLE 4].

However if we check from the limited prediction set the logistic model gives us 91% accuracy rate with the Residual Deviance value 11.92 very low and efficient AIC value 15.92. [TABLE 3].

While applying the Hartigan-Wong algorithm on the overall dataset we have observed that out of 70 records two clusters were prepared of sizes 41 and 29. [TABLE 5]. A total of 5 records has been observed which was predicted incorrectly on the basis of the pitch values where 4 incorrect prediction in male data whereas only 1 incorrect prediction in the female voice sample as shown from the Fig. 2 and the earlier tables of reference. Here in case of the clustering algorithm we get a 93% accuracy rate even in the whole sample.

The Clustering Algorithm based on machine learning is slightly ahead in terms of true prediction whereas the Logit model also gives us almost very accurate result in terms of small sample set. The logistic regression helps us in determine the shape of the prediction curve [Fig. 1] and comparing all the other significant factors from the other attributes. In our study we have found out that F0 is the most significant factor which can be used to predict the gender of the speaker.

However for some of the pitch values were incapable of predicting the type of the speaker in both the models due to general tendency of a female speaker having a male like voice and Vice Versa. This phenomenon may occur due to some external factors like noisy environment, recording disturbances etc.

Finally in case of data instances and statistical significance the most vital factor is that as we have mentioned earlier Mising is a low resource language as well as very few speakers are left in our region. It was managed to collect a few samples regarding the vowel pronunciation. We have used K-Means Clustering as it is originally based on signal processing and useful for small samples.

Currently we are collecting more sample voices and planning to use methods like SVM and other ANN models which are supposed to be more reliable.

VI. Conclusion and Future Scope

In the present study we have applied statistical techniques like Logistic regression to analyze the sound data and obtained interesting results as discussed above. We have also applied some other clustering methods based on machine learning. We have compared the results with relative advantages and efficiencies.

It will be helpful in future for comparing result with other data mining techniques as well ANN model and carry forward this work in gender classification in speech signals data. Similarly these works can be carry forwarded for other problems like Speaker identification from native and non-native language, frequency prediction etc. for other low resource languages as well.

Acknowledgment

The research work is based on a primary data set of speech signal data. We sincerely offer our thanks and gratitude to Dr. Rizwan Rehman, Assistant professor of Centre for Computer Science and Applications, Dibrugarh University for providing his data for analytical purposes.

We would also like to offer my sincere thanks to those students who have provided their voice samples for the community research work.

References

[1] Greenstein, J. (1954), Effect of television viewing upon elementary school grades, The Journal of Educational Research, 48, 161-176.

[2] Warner, B. and M. Misra (1996), Understanding Neural Network as Statistical Tool, The American Statistician 50(4),pp. 284-293.

[3] Gibbs Y. Kanyongo, Janine Certo, Brown, I.Launcelot, Using regression analysis to establish the relationship between home environment and reading achievement: A case of Zimbabwe, International Education Journal, 2006, 7(5), 632-641.

[4] Pao,H.T, A Comparison of Neural Network and Multiple Regression Analysis in Modelling Capital Structure, Expert Systems with Applications 35 , 2008, pp.720–727.

[5] Keshavarzi,A. and F. Sarmadian, Comparison of Artificial Neural Network and Multivariate Regression Methods in Prediction of Soil Cation Exchange Capacity, International Journal of Environmental and Earth Sciences , 20111(1),pp. 25-30.

[6] B. Prica and Sinisa, Recognition of Vowels in Continuous Speech by Using Formants, SER.: ELEC. ENERG. vol. 23, no. 3, December 2010, 379-393.

[7] Ganesan,N., Venkatesh,K., Rama, M. A. and A.M.Palani[2010],Application of Neural Network in Diagnosing Cancer Disease using Demographic Data, International Journal of Computer Applications 1(26),pp.81-97.

[8] Raghavendra B.K. & S.K. Srivatsa[2011], Evaluation of Logistic Regression and Neural Network Model With Sensitivity Analysis on Medical Datasets, International Journal of Computer Science and Security 5 (5) ,pp.503-511.

[9] Al-Shayea,Q. K.[2011],Artificial Neural Networks in Medical Diagnosis, International Journal of Computer Science Issues 8(2),pp.150-154.

[10] Rizwan Rehman and Gopal Chandra Hazarika, Analysis and Recognition of Vowels in SHAI`YANG MIRI Language using Formants, International Journal of Computer Applications 89(2):7-10, March, 2014.

[11] Xiao-Lei Zhang, Linear Regression for Speaker Verification, arXiv:1802.04113,v1, Pp.1-10, 2018.

[12] Ujjal Saikia, Rizwan Rehman, Jiten Hazarika, Gopal Ch. Hazarika, Predictive Analysis Using Regression Methods in Low Resource Language "MISING", 2nd International Conference on information systems & management science (ISMS) 2019.

[13] Jiang Lin ,Yi Yumei ,Zhang Maosheng ,Chen Defeng ,Wang Chao ,Wang Tonghan, A Multiscale Chaotic Feature Extraction Method for Speaker Recognition, Complexity, Hindawi, vol. 2020, pages 1-9, December.

[14] Magdiel Jiménez-Guarneros and Pilar Gómez-Gil,, Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition, Pattern Recognition Letters, Volume 141, January 2021, Pages 54-60.

[15] Babak Joze Abbaschian , Daniel Sierra-Sosa , Adel Elmaghraby, Deep Learning Techniques for Speech Emotion Recognition, Sensors 2021, 21(4),1-27.