International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Feature Extraction for Speech Classification

Shankari N, Rajashree

Abstract—*Study of phonological Processes, Speech recognition, Speech Synthesis and language learning requires Automatic classification of class of sounds and automatic identification of sound classes. This paper focuses on identifying features efficient in discriminating different classes of sound such as analyzing spectral features such as distinctive frequency components by Linear Productive Coding technique and vocal tract length. Artificial Neural network and Random Forest Classification Technique is used to check effectiveness of identified feature with 10-fold cross validation. The proposed system is also aimed at improving performance of phoneme recognition system.*

Keywords- Artificial Neural Network, Random Forest Classification, Cross Validation

I. INTRODUCTION

A language in general has five parts such as phonemes, syntax, morphemes context and lexemes[1]. The meaning in a language is due to smallest unit called phoneme which has no meaning and is undividable. Phonemes correspond to the sound of the alphabet, although there is not always a one-to-one relationship between a letter and a phoneme. The sounds can be classified into various numbers of parameters by different classification techniques. This paper proposes the study of classification done based on the position and mode involved in articulation. With reference to vocal tract, the narrowest part of producing the sound refers to articulation. Sound is produced always as a result of obstruction formed when two articulators, usually one moving called the active articulator and the other stationary passive articulator come together.

Most of the Indian languages (including the language under study, Hindi) have 5 main categories with reference to the position of articulation, they are:

1. Velar: Uttered by using the part of the tongue (the dorsum) backside against the soft palate, the back side of mouth roof (called as the velum) such as [k], [g] and [ŋ].
2. Palatal: Uttered by using tongue whose body is lifted and hit the hard palate (the central part of the top of the mouth) such as [ch] and [j].
3. Retroflex: Words pronounced using tongue tip which is curled behind the palate such as [t], [d].
4. Dental: Uttered by hitting the tongue against upper jaw of the teeth such as [th], [dh].
5. Labial: Articulated by using both the lips such as [p], [b].

Manuscript published on January 30, 2020.
* Correspondence Author
Mrs. Shankari N*, Assistant Professor, Department of ECE, NMAMIT, Nitte, Udupi district, Karnataka
Ms. Rajashree, Assistant Professor, Department of CSE, NMAMIT, Nitte, Udupi district, Karnataka

There are two types of labial articulations are:

1. Bilabials: Uses both the lips to articulate.
2. Labiodentals: Uttered by using lower lip over upper teeth.[2]

Table 1: Consonant Classification based on position and mode of articulation.

	unaspirated	aspirated	unaspirated	aspirated	
Velar	क ka	ख kha	ग ga	घ gha	ङ ṅa
Palatal	च ca	छ cha	ज ja	झ jha	ञ ña
Retroflex	ट ṭa	ठ ṭha	ड ḍa	ढ ḍha	ण ṇa
Dental	त ta	थ tha	द da	ध dha	न na
Labial	प pa	फ pha	ब ba	भ bha	म ma

II. SYSTEM DESIGN

System design involves the process of defining the architecture or flow diagram.

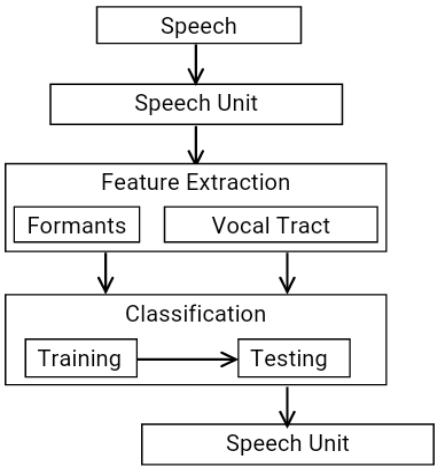

Fig 1: Flow Diagram

The flowchart is shown in Fig.1 which involves speech database namely TIMIT data base. The speech database has recordings of around 630 speakers in eight major dialects of American English from both genders. From each speaker, Ten sentences are recorded which are phonetically rich. Segmentation is performed on this speech data base into five classes of sound such as Velar, Palatal, Retroflex, Dental and Labial [3]. Each sound class are segmented further to obtain constant vowel transition which is required for higher efficiency. Each segmented data are subjected to feature extraction where formants i.e distinctive frequency components of each sound class and vocal tract lengths are estimated using Linear Predictive Coding (LPC) technique.

By estimating the formant frequencies, the equation used to measure the vocal tract length [4] is,

$$l = \frac{c}{4 * f} \qquad (1)$$

Where l is estimated vocal tract length, c is the speed in which air exhaled during speech and f is the first distinctive frequency component. Minimum and maximum lengths of the vocal tract, their median, mode etc. are the statistical variations or features of estimated vocal tract length and are derived. Total feature vector of size 52 is considered for the analysis. Each sound classes with distinct features are classified using Random Forest Classifier and verify the effectiveness of identified feature with 10 fold validation.

III. IMPLEMENTATION

This approach is implemented by using a common phonetic speech corpus TIMIT**.** The training data of speech required for the capturing of acoustic-phonetic knowledge is obtained by TIMIT acoustic-phonetic corpus of speech which also assesses the speech concerned tasks. The training data has soundtracks with 630 orators includes men and women in 8 key vernacular of English language. From individual orator, ten phonetically prosperous lines are documented. 16 bit quantized data are stored which is obtained by sampling the TIMIT database at a rate of 16 KHz**.** From the database, for individual class the consonant sounds namely dental, labial, palatal, velar and retroflex with the vowel trailed are partitioned [5]. The vocal tract length and the formants are obtained after the segmentation process is completed. The path of all the segmented voice clips is provided to the MATLAB code which will then extract all the 52 features of each segmented data. A cell-fea-vector file is generated when we run the code which is a 1X2 matrix which displays 0's and 1's which are the labels. Then run a code for texting and training data where 4 ANN (Artificial Neural network) files are generated wherein we have testing and training data. WEKA tool is used to perform testing and training because it is more efficient than ANN as this produces a single accuracy value whereas ANN produces multiple, and we have to then compute the most accurate out of it. Also, data has to be trained many number of times in ANN which is time consuming. Hence WEKA tool is used which has various classifiers. Random forest classifier is used to get the accuracy[6].

IV. RESULT AND DISCUSSIONS

The project primarily focuses on the important characteristics pertaining to the position of articulation in distinguishing speech component into individual class [7],[8]. Formants, approximation of length of the vocal tract and the numerical deviation are extorted from five sound classes which are categorized depending upon the position of articulation namely dental, labial, retroflex, palatal and velar. The efficiency of features vector of volume 54 is evaluated using artificial neural network (ANNs) and Random Forest (RF) classifier [9]. Initially two sound classes were taken to extract features and are trained and examined to find the classification's accuracy. Higher accuracy is an indication of better features in representing the sound class. The accuracy of random forest algorithm with 10 fold cross validation is given in Table 2.

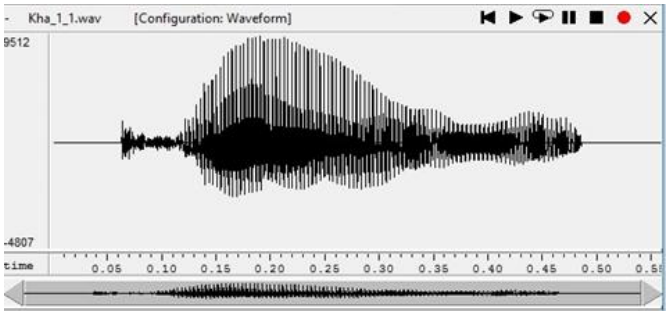

Fig.2: Consonant part ka (velar) extracted from one of the audio files

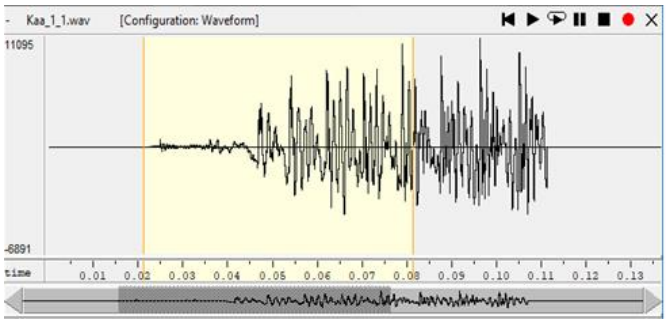

Fig.3: Consonant burst region extracted from the previous segmented ka audio file.

By examining the results obtained, it is clear that the projected characteristics are competent in differentiating the sound classes palatal against dental, labial, retroflex as well as velar. Between the palatal and velar classes the accuracy of the distinction of sounds is 92.9%. This represents that the features based on the position of articulation and type of articulation are efficient in discriminating the palatal and velar sounds. 93.83% accuracy is achieved for palatal & retroflex and accuracy achieved for palatal & labial is 94.07% which signifies that it has a considerable dissimilarity in the position of articulation other class of sounds with palatal.

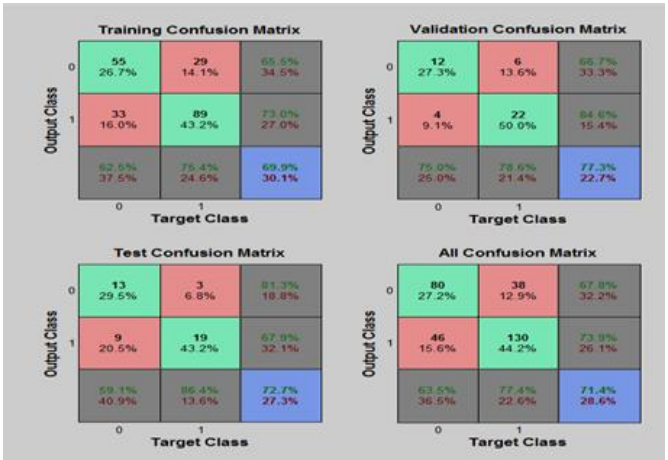

Fig.4: Plot confusion matrix obtained by using the Artificial Neural Network.

Table.2: Accuracy table obtained by Random Forest Method.

Sl. No.	Classes	Accuracy (%)
1	Velar Palatal	92.9
2	Velar Retroflex	76.398
3	Velar Dental	74.089
4	Velar Labial	70.36
5	Palatal Retroflex	93.83
6	Palatal Dental	79.411
7	Palatal Labial	94.07
8	Retroflex Dental	77.43
9	Retroflex Labial	72.18
10	Dental Labial	77.91

The other classes of sounds cannot be discriminated by the proposed method which may achieve an accuracy of 70%. Hence there is a limitation on the expansion of this technique for general use with the suitability of the features proposed in differentiating the all speech classes. The position of tongue hump may be the reason for the reduced accuracy in other classes. Where in some of these classes there will be obstruction of the air and the tongue position may have alike in nature which renders the performance in different classification techniques.

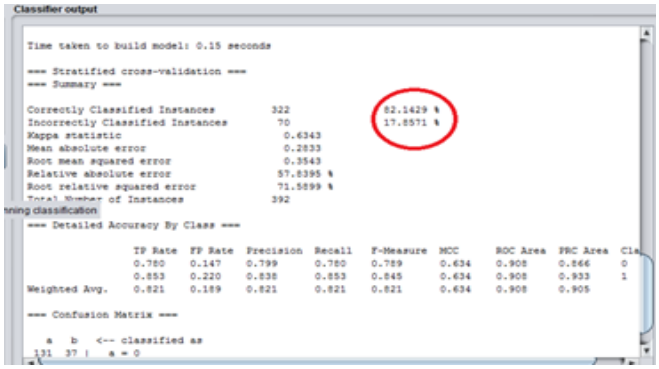

Fig.5: Final accuracy value obtained using the WEKA tool by applying Random Forest Algorithm

V. CONCLUSION

The proposed technique analyses the importance of articulation which is involved in characterizing each unit of speech sound. The Speech is characterized into five different classes namely Palatal, Velar, Dental, Retroflex, and Labial and formants which are the spectral characteristics and length of the vocal tract are extracted. The classification is performed by taking Random Forest (RF) classifier with its non linear nature. The accuracy obtained by classification was supposed to be an indication of how well these features represented the sound classes. The velar & palatal, palatal & retroflex and palatal & labial sounds are fairly discriminated by the proposed system feature which is observed in the result.

REFERENCES

1. Pravin Bhaskar Ramteke, Srishti Hegde, Shashidhar G. Koolagudi. "Chapter 15 Characterization of Consonant Sounds Using Features Related to Place of Articulation", Springer Science and Business Media LLC.(2020).
2. "Smart Computing Paradigms: New Progresses and Challenges", Springer Science and Business Media LLC.(2020)
3. Pravin Bhaskar Ramteke, Shashidhar G. Koolagudi. "Phoneme boundary detection from speech: A rule based approach", Speech Communication.(2019).
4. Pravin Bhaskar Ramteke, Anmol Sadanand, Shashidhar G. Koolagudi, Vidya Pai. "Characterization of aspirated and unaspirated sounds in speech", TENCON 2017 - 2017 IEEE Region 10 Conference.(2017)
5. Manjunath, K.E. & Sreenivasa Rao, K. Int J Speech Technol 19:121doi:10.1007/s10772-015-9329-x. (2016).
6. Keller E., Chollet G., Esposito A., Faundez-Zanuy M., Marinaro M. (eds) "The Analysis of Voice Quality in Speech Processing". In: Nonlinear Speech Modelling and Applications. Lecture Notes in Computer Science, vol 3445. Springer, Berlin, Heidelberg
7. Diego H. Milone, Student Member, IEEE, and Antonio J. Rubio, Senior Member, IEEE, "Prosodic and Accentual Information for Automatic Speech Recognition".
8. Feature Extraction for Speech Recognition by Manish P. Kesarkar, Electronic Systems Group, EE. Dept., IIT Bombay, Submitted November (2003).
9. Dr.Vilas Thakare, Urmila Shrawankar, SGB Amravati University, Amravati, "Techniques for Feature Extraction in Speech Recognition System: A Comparative study".