Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 13 (2020) Pp-332-337

BBRC
Bioscience Biotechnology
Research Communications

Comparative Analysis of Feature Extraction methods for Kannada Bi-Syllable Words of Dysarthric Speech.

Latha M[1], M Shivakumar[2] and Manjula.R[3]

[1]*Department of Electronics and Communication Engineering, GSSS Institute of Engineering and Technology for Women, Mysuru, Karnataka, INDIA, Affiliated to VTU, Belagavi, Karnataka, India.*

[2]*GSSS Institute of Engineering and Technology for Women, Mysuru, Karnataka, India, Affiliated to VTU, Belagavi, Karnataka, INDIA.*

[3]*All India Institute of Speech and Hearing,Mysuru, India*

ABSTRACT

Recent research works rely on machine learning models in many speech assistance systems. Machine learning based speech assistance models mainly contributes in transforming dysarthric speech to normal speech will be of great help to persons suffering with this aid. For an accurate speech transformation, best set of features need to be extracted from dysarthric speech and machine learning based classifiers need to be trained with those features for translating to normal speech. Present work does a comparative analysis of feature extraction methods for Kannada bi-syllable dysarthric speech. A clustering-based analysis is conducted on feature extraction methods, each separately and in combination is done. Through analysis, best feature set combination suitable for accurate recognition of Kannada dysarthric bi-syllable is identified. While earlier works focused feature analysis only based on classification accuracy, But this present work does cluster analysis to calculate the inter distance between the bi-syllables and identify the region where marginal errors can occur in recognition. MFCC, LPC, PLP, LPCC, PE-SFCC, Prosodic features are the feature extraction methods were analyzed and the combination of the feature extraction methods is compared. The clustering based analysis results that the combination of PE-SFCC + LPC + PLP is found to perform better than other feature extraction methods.

KEY WORDS: MFCC, LPC, PLP, LPCC, PE-SFCC, PROSODIC FEATURES, KANNADA BISYLLABLE WORDS..

INTRODUCTION

Human speech is the common means of communication. Speech production involves the various mechanisms such as respiration, phonation and articulation. When any of these mechanisms is affected results in the distruption of speech or speech disorders.There are numerous reasons behind to have speech disorders in individuals. Speech disorders can affect any individuals.people of all ages.

ARTICLE INFORMATION

**Corresponding Author: latha@gsss.edu.in*
Received 10th Oct 2020 Accepted after revision 25th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31)
A Society of Science and Nature Publication,
Bhopal India 2020. All rights reserved.
Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.13/51

The current paper focuses on one among the various types of speech disorder presented by World's Health Organization i.e Dysarthria. Dysarthria is one of the neurological disorder occurs due to damage of brain which causes muscle weakness in a person's face, lips, tongue, throat or chest. People with Dysarthria experience with following symptoms – slurred speech, mumbling, speaking too slowly or too quickly, soft or quiet speech, difficulty moving the mouth or tongue.

Persons with Dysarthria need better ways to communicate with others. They use other means of communication like hand gestures, writing by hand, computer to translate speech to text, using alphabet boards etc. Various algorithms were developed to translate Dysarthric speech to text. The structured based approaches were used in the traditional speech recognition systems. There is no general frame work designed with respect to speech recognition system that can work common for all dysarthria abnormalities. The normal speakers communicate at a rate of 150 to 200 words per minute while the communication rate with respect to Dysarthria speakers is less than 15 words per minute. Due to this variability in the utterances of dysarthric speech, it has become difficulty to develop a precise model to recognize the desired latent patterns of the speech signal. Thus, developing a speech recognition system involves culmination of efforts from multiple disciplines like speech signal processing, natural language processing and artificial intelligence.

In this paper, comparative analysis of different feature extraction methods for their suitability in developing speech recognition systems for dysarthric speech is presented. The analysis is conducted for Kannada bi-syllable dysarthric speech dataset.

Related Work: The survey is conducted on existing feature extraction and feature analysis methods for dysarthric speech and presented below.

In 2015, N. Souissi and A. Cherif included Mel Frequency Cepstral Coefficients (MFCC) for identification of voice disorders. The study also used first and second derivatives in-addition to different number of MFCC features. The dimensionality reduction is done using Linear Discriminant Analysis. The study concluded that there is no difference between MFCC features and their first and second order derivatives in voice disorder classification.

In 2016, U. N. Wisesty, et.al, analyzed the performance of Linear predictive coding (LPC) and MFCC for Indonesian speech recognition system. The authors concluded that LPC gives better performance than MFCC in differentiating between the voice and unvoiced frames but LPC takes more time than MFCC (Wisesty & Astuti 2015). Log RASTA Perceptive Linear Prediction hybridized with Artificial Neural Network is used for feature extraction in (MeghaRughani & Shivakrishna 2015). 12-Log RASTA PLP method with frame length equal to 12 is selected for dysarthric speech having 25ms of frame size and 10ms of overlap. Frame length is chosen to be equal to maximum length of the utterance. Silence portion is removed from the beginning and end portion based on energy of frame and frame length of each utterance is made equal by appending zeros at the end in order to make number of inputs same for each utterance to neural net. Feature extracted matrix is transferred to array form by appending m+1 column to the end of mth column. So, each utterance is represented by 126 features (13 features per frame x 12 frames). Each feature was assigned to one of the corresponding neuron of the clustering structure of ANN which groups features into 64 different clusters which is sufficient for phoneme classification.

In 2017, T. B. Ijitona and J. J. Soraghan have used speech features called centroid formants for automatic detection of Dysarthria. Formants are the bands of resonance in the frequency spectrum of a speech signal. The concept of centroid formants is helpful in detecting frequency components present in the spectrum of the signal.The location of centroid formants indicates the high frequency range and low frequency renage of the signal. This indirectly presents the variability of pitch and intonation of the speech signal (Ijitona et al. 2017).

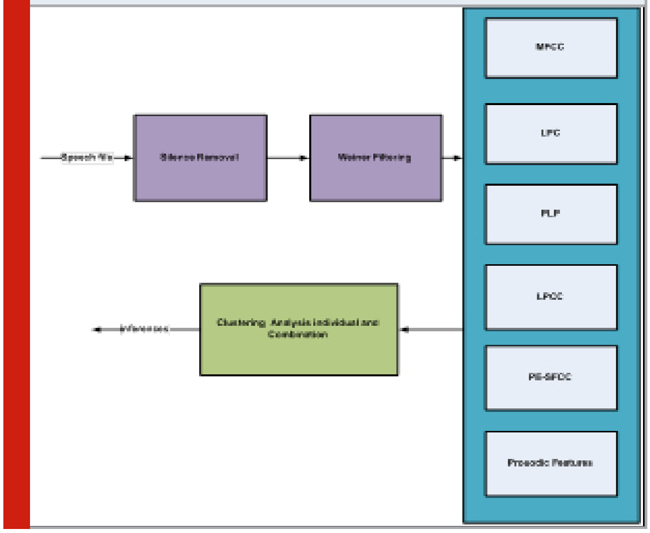
In 2018, N P Narendra and Paavo Alku have used glottal features for dysarthric speech classification. The evaluation of features for classification was done for three categories of non-words, words and sentences. Glottal features when combined with open SMILE features, resulted in higher classification accuracy. A novel sentence-level features are proposed to capture abnormal variation in the prosodic, voice quality and pronunciation aspects in pathological speech.

In 2018, Yılmaz, Emre explored the joint use of articulatory and acoustic features for speech recognition. A fused-feature-map convolutional neural network(fCNN), which performs frequency convolution on acoustic features and time convolution on articulatory features is trained and tested on a Dutch and a Flemish pathological speech corpus and recognition accuracy is higher due to use of joint features (Emre 2018).

In 2019, Krishna Gurugubelli,et.al, have proposed perceptually enhanced single frequency cepstral coefficients (PESFCC) for dysarthric speech detection. PE-SFCC feature set outperformed other state-of-the art-features such as MFCC, PLP, multi-taper MFCC, and CQCC features for dysarthric speech intelligibility assessment. A new feature extraction algorithm called Power Normalized Cepstral Coefficients (PNCC) is proposed in (Kim & Stern 2016). PNCC replaced the traditional log linearity in MFCC with power-law non linearity. Through experiments PNCC is found perform better in recognition accuracy compared to MFCC and PLP in the presence of various types of additive noise and in reverberant environments, with only slightly greater computational cost. Kamil LahceneKadi and Sid Ahmed Selouani used a set of prosodic features selected by LDA on the basis of their discriminative ability, with Wilks' lambda as the significant measure to show the discriminant power. The features used were articulation rate, number of periods, mean pitch, voice breaks, HNR, Jitter, Shimmer, standard pitch, standard period and NHR.

In 2020, VivianaMendoza Ramos and HectorA.Kairuz Hernandez-Diaz have proposed new approach in computing acoustic features for dysarthric speech classification. In this new approach, linear discriminant analysis (LDA) analysis is performed on the speech inputs. from this analysis, it is able to determine the time duration, energy, fundamental frequency through which differences in the utterances of healthy and dysarthria speakers are measured(Mendoza et al. 2020). Yılmaz, Emre & Mitra,et.al, in 2019 have demonstrated gammatone filter bank features for speaker independent ASR systems. They explored the performance of two novel convolutional neural networks using the gammatone filter bank, acoustic and articulate features(Emre et al. 2019).

Figure 1: Feature Analysis Framework



Analysis: The architecture of the proposed feature analysis framework is given in figure 1. From the input speech signal, silence is removed at beginning and end based on the energy of the signal. The speech signal is divided to small segments and energy of each individual segment is calculated as

$$E_s = 10\log(\varepsilon + \frac{1}{N}\sum_{n=1}^{N} S^2(n)) \qquad (1)$$

Where ε is small positive value added to prevent the computing of log 0. E_s for the voiced segment is always higher than that of nonvoiced segment. The function for silence removal is given as

$$f(x) = \begin{cases} E_s \geq E_{th}, & \text{voice segment} \\ E_s < E_{th}, & \text{silence segment} \end{cases}$$

The threshold E_{th} is calculated as

$$E_{th} = \frac{\mu + \omega}{2}$$

Where ω is the minimum energy value of K voiced segments and μ is the mean energy value of K unvoiced segments computed as

$$\mu = \frac{1}{K}\sum_{i=1}^{K} E_{unvoiced}$$

Silence removed speech signal is then enhanced using Weiner filter. From the enhanced signal following features are extracted.

1. MFCC: Mel Frequency Cepstral Coefficient
2. LPC: Linear prediction coefficients
3. PLP: Perceptual linear prediction
4. LPCC: Linear Prediction Cepstral Coefficient
5. PE-SFCC:Perceptually enhanced single frequency cepstral coefficients
6. Prosodic features

Following acoustic features are extracted from the speech signal

1. Number of periods
2. Mean pitch
3. Voice breaks
4. HNR
5. Jitter
6. Shimmer
7. Standard pitch
8. Standard period
9. HNR

The speech signal corresponding to different Kannada bisyllable for both normal and dysarthric speech are passed for feature extraction and all six features considered in this study is extracted.

The clustering analysis is performed for following individual and combination of features

1. MFCC (C1)
2. LPC (C2)
3. PLP (C3)
4. LPCC (C4)
5. PE-SFCC (C5)
6. Prosodic Features (C6)
7. MFCC + LPC (C7)
8. MFCC+ PLP (C8)
9. MFCC + LPCC (C9)
10. MFCC+ Prosodic Features (C10)
11. LPC + PLP (C11)
12. LPC + LPCC (C12)
13. LPC + Prosodic Features (C13)
14. PLP + LPCC (C14)
15. PLP + Prosodic Features (C15)
16. PE-SFCC + LPC (C16)
17. PE-SFCC + PLP (C17)
18. PE-SFCC + LPCC (C18)
19. PE-SFCC + Prosodic Features (C19)
20. PE-SFCC + LPC + PLP + Prosodic Features(C20)
21. PE-SFCC + LPC + LPCC + PLP + Prosodic Features (C21)
22. PE-SFCC + LPC + PLP (C22)

The speech features is then clustered using k- means clustering into N clusters (N corresponding to number of bisyllable). Features corresponding to each bisyllable are clustered into two clusters (normal and dysarthric) speech. The cluster efficiency is validated using following metrics

1. Average Cohesion
2. Average Separation
3. Silhouette coefficient

4. Time taken for Clustering

Cohesion is measure of how close are the objects within the same cluster. A lower within-cluster variation is an indicator of a good compactness (i.e., a good clustering). It is calculated in terms of sum of squares of distances of each point in cluster to the centroid of cluster as given below

$$cohesion = \sum_{i} \sum_{x \epsilon C_i} (x - m_i)^2$$

Separation is an indication of how well-separated a cluster is from other clusters. It is measured as

$$separation = \sum_{i} |C_i| \, (m - m_i)^2$$

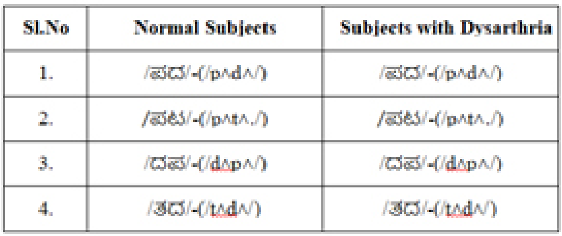
Where $|C_i|$ is the size of the cluster i, and m is the centroid of whole feature set. Higher the separation, is an indicator of good clustering.

The silhouette analysis measures how well an observation is clustered and it estimates the average distance between clusters. Silhouette coefficient is calculated as

$$s = \begin{cases} 1 - \frac{a}{b}, & if\ a < b \\ \frac{b}{a} - 1\ if\ a \geq b \end{cases}$$

For a individual point, a is average distance of i to the points in its cluster and b is minimum of average distance of i to points in another cluster. The value of silhouette coefficient is between 0 and 1 and the value towards 1 is better.

Table 1. Kannada Bi-Syllable words



Sl.No	Normal Subjects	Subjects with Dysarthria
1.	/ಪದ/-(/p^d^/)	/ಪದ/-(/p^d^/)
2.	/ಪಟ/-(/p^t^./)	/ಪಟ/-(/p^t^./)
3.	/ದಪ/-(/dʌp^/)	/ದಪ/-(/dʌp^/)
4.	/ತದ/-(/tʌd^/)	/ತದ/-(/tʌd^/)

The clustering analysis results into 4 clusters through which following metrics are evaluated.The following metrics used in the process are cohesion, separation to other clusters, silhouette coefficient and time taken. The desired values for clustering is calculated. These values are averaged to give the average value of cohesion, average value of separation, average value for silhouette coefficient and average time for clustering. The clustering analysis also results into 2 clusters which represents the Normal and Dysarthric clusters as discussed in results section.

RESULTS AND DISCUSSION

For experimental analysis, the Kannada bi-syllabic words were selected with the combination of dental, bilabial and retroflex components in available speech consonants. The pre-recorded samples of following bi-syllabic words for both Normal subjects and subjects with dysarthria are used for the clustering analysis.

Figure 2: Average cohesion for Kannada bi-syllable words.

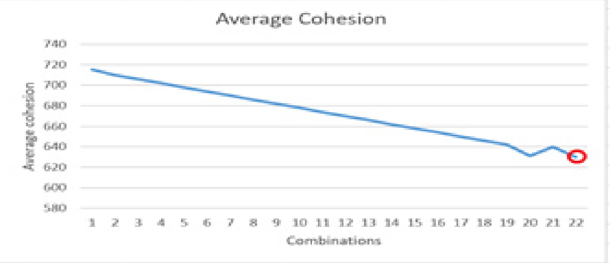
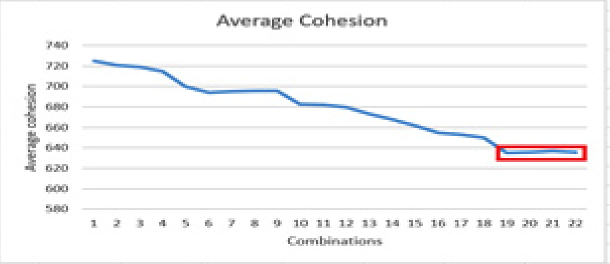
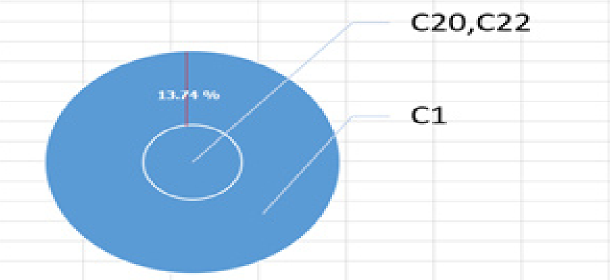


Figure 3: Average cohesion for normal subjects cluster / subjects with dysarthria cluster



Eight training samples are taken in combination of normal subjects and subjects with dysarthria. The results of clustering analysis for Combinations (C1 to C22) for Clustering into N bi-syllable cluster corresponding to each bi-syllable word (N is 4 here) is given in Table 1.

Figure 4: Cohesion radius difference



A. Average Cohesion Cluster:

The lower cohesion values results in best clusters. From the cohesion results for bi-syllable word cluster, combination C22, has the lowest cohesion value. It is 13.49% lower than the highest cohesion value found for bi-syllable word cluster. From the cohesion results for normal/dysarthria cluster, combination C20, C22 almost have close values. They have 13.99% lower cohesion value than the highest cohesion value found for normal/ dysarthria cluster. It be inferred that cluster radius is shrink by average 13.74% in C20, C22 combinations compared to maximum coherence radius.

Figure 5: Average separation for Kannada bi-syllable words.

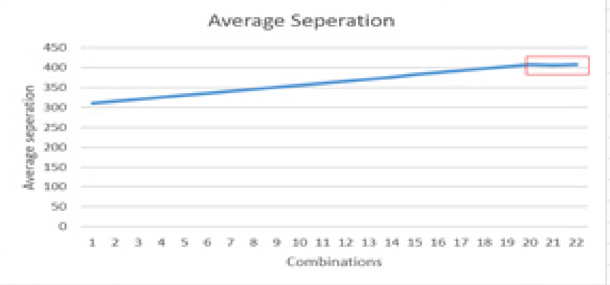


Figure 6: Average separation for normal subjects cluster /subjects with dysarthria cluster

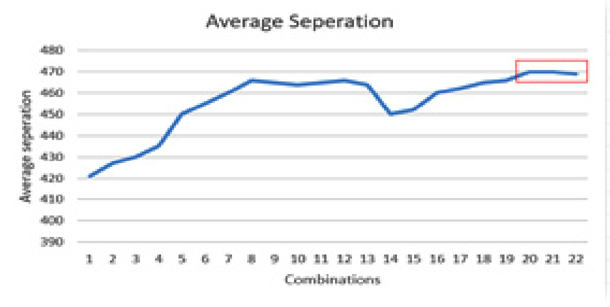
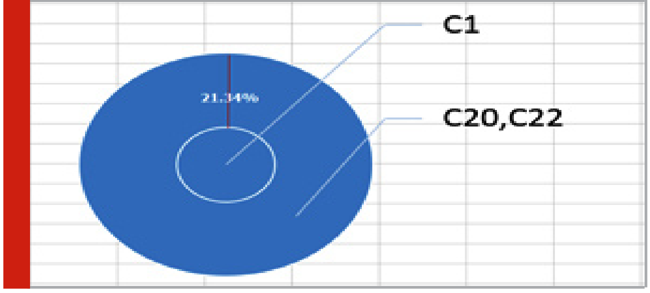


Figure 7: Separation radius difference



Average Separation Cluster: The higher average separation indicates an efficient cluster and lower the misclassification between the words.

From the separation results for bi-syllable word cluster, combination C20, C21, C22 has the highest separation value. It is 31.29% higher than the lowest separation value found for bi-syllable word cluster. From the separation results for normal/dysarthria cluster, combination C20, C21, C22 almost have close values. They have 11.4% higher separation value than the lowest separation value found for normal/dysarthria cluster. It be inferred that cluster radius is increase by average 21.34 % in C20, C22 combinations compared to lowest separation radius

Figure 8: Average Silhouette for Kannada bi-syllable words.

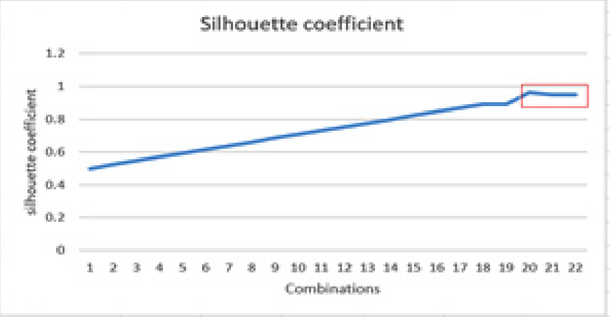


Figure 9: Average Silhouette for normal subjects cluster / subjects with dysarthria cluster

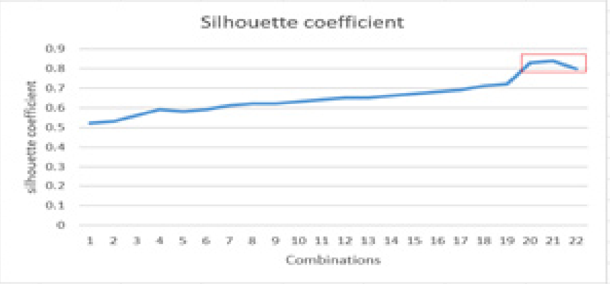


Figure 10: Silhouette radius difference

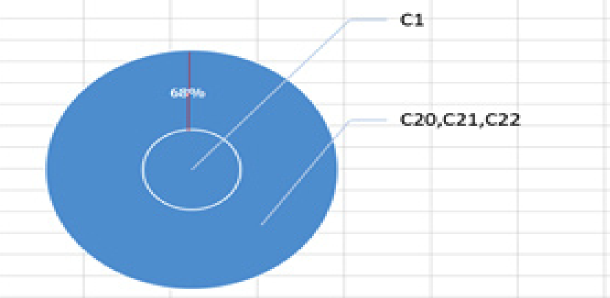
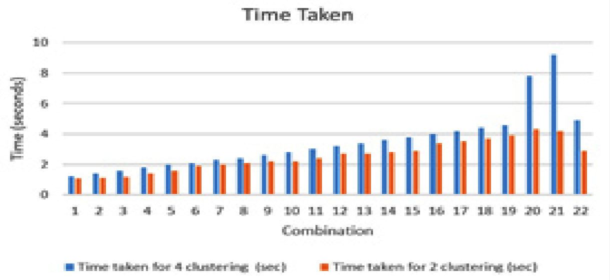


Figure 11: Average timetaken cluster



Average Silhoutte Cluster: From the Silhouette results for bi-syllable word cluster, combination C20, C21, C22 has the highest values. It is 90% higher than the lowest silhouette value found for bi-syllable word cluster. From the Silhouette results for normal/dysarthria cluster, combination C20, C21, C22 almost have close values. They have 68% higher silhouette value than the lowest silhouette value found for normal/dysarthria cluster. It is also inferred that cluster radius is increase by average 79% in C20, C21, C22 combinations compared to lowest silhouette radius. The higher Silhouette coefficient indicates a better cluster.

Average Timetaken Cluster: From the results of time taken for clustering, comparing C20, C21, and C22, the combination C22 takes the lowest time. Even though clustering time is lower in combination C1-C19, they are compressed for having lower silhouette, separation value than other combinations.

From the results, it can be seen that both combination C20 and C22 almost have same performance in terms of cohesion, separation,Silhouette coefficient. But C20 has higher cohesion larger separation and larger silhouette coefficient than other combinations, but time taken is very high compared to C22. Considering time efficiency, and better results for cohesion, separation and silhouette coefficient, C22 (PE-SFCC + LPC + PLP) is more suited for Kannada bi- syllable words. Therefore, C22 is preferred choice for categorization for normal and dysarthric Kannada bisyllable words.

CONCLUSION

Feature extraction is important module in automatic speech recognition for dysarthric speech. This work presented a clustering-based analysis of feature extraction methods for normal / dysarthric bi-syllable Kannada words. Following features of MFCC, LPC, PLP, LPCC, PE-SFCC, Prosody were experimented individually and in combination. Clustering analysis is performed and following metrics - Average Cohesion,Average Separation,Silhouette coefficient, Time taken for clustering (sec) were measured. The combination with a higher value of cohesion, separation and silhouette coefficient and comparatively lower time for clustering is selected as optimal combination feature from which features of Kannada bi-syllable words can be segregated in a better way. From the clustering analysis, combination of PE-SFCC + LPC + PLP features is found to perform better for categorization of bisyallable words in Kannada language with respect to Normal subjects and subjects with dysarthria.

REFERENCES

ArefFarhadipour,HadiVeisi,Mohammad Asgari, (July 2018) Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks",ETRI journal

Chanwoo Kim and Richard M Stern, (2016) Power-normalized cepstral coefficients (PNCC) for robust speech recognition," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 24, no. 7, pp. 1315–1329.

J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, (2015) Automatic intelligibility classification of sentence-level pathological speech," Computer Speech and Language, vol. 29, pp. 132–144.

Jiao, Yishan& Tu, Ming & Berisha, Visar&Liss, Julie. (2018). Simulating Dysarthric Speech for Training Data Augmentation in Clinical Speech Applications. 6009-6013. 10.1109/ICASSP.2018.8462290.

Kamil LahceneKadi, Sid Ahmed Selouani,"Automated Diagnosis and Assessment of Dysarthric Speech Using Relevant Prosodic Features",2014,Transactions on Engineering Technologies

Krishna Gurugubelli, Anil Kumar Vuppala, "Perceptually Enhanced Single Frequency Filtering For Dysarthric Speech Detection And Intelligibility Assessment ",2019 International Conference on Acoustics, Speech, and Signal Processing.

MeghaRughani and D. Shivakrishna, (2015) Hybridized Feature Extraction and Acoustic Modelling Approach for Dysarthric Speech Recognition",arXiv.

N P Narendra, Paavo Alku, (2018) Dysarthric speech classification using glottal features computed from non-words , words and sentences", Interspeech.

N. Souissi and A. Cherif, "Dimensionality reduction for voice disorders identification system based on Mel Frequency Cepstral Coefficients and Support Vector Machine," in 2015 7th International Conference on Modelling, Identification and Control (ICMIC), 2015, pp. 1-6.

T. B. Ijitona, J. J. Soraghan, A. Lowit, G. Di-Caterina and H. Yue, "Automatic detection of speech disorder in dysarthria using extended speech feature extraction and neural networks classification," IET 3rd International Conference on Intelligent Signal Processing (ISP 2017), London, 2017, pp. 1-6, doi: 10.1049/cp.2017.0360.

U. N. Wisesty, Adiwijaya, and W. Astuti, "Feature extraction analysis on Indonesian speech recognition system," in Information and Communication Technology (ICoICT), 2015 3rd International Conference on, 2015, pp. 54-58.

VivianaMendoza Ramos,HectorA.Kairuz Hernandez-Diaz,"Acoustic features to characterize sentence accent production in dysarthric speech",Biomedical Signal Processing and Control,March 2020

Yılmaz, Emre & Mitra, Vikramjit& Bartels, Chris & Franco, Horacio. (2018). Articulatory Features for ASR of Pathological Speech. 10.21437/Interspeech.2018-67.

Yılmaz, Emre & Mitra, Vikramjit&Sivaraman, Ganesh & Franco, Horacio. (2019). Articulatory and Bottleneck Features for Speaker-Independent ASR of Dysarthric Speech. Computer Speech & Language. 58. 10.1016/j.csl.2019.05.002.