Dysarthric Speech Characterization and Classification based on Affinity Propagation

Komal Bharti[1], Sandeep agri[2], and Pradip K.Das[3]

[1,2,3]Indian Institute of Technology Guwahati, 781039, Assam, India
{[1]kbharti,[2]sandeep.agri,[3]pkdas}@iitg.ac.in

Abstract. Millions of people are estimated to suffer from speech impairments originating from various causes, including neurological disorders, brain damage, and physical conditions. Dysarthric speech is hard to recognize and understand because of its enormous variability and low intelligibility rate. In today's high-tech environment, speech recognition system has achieved more than 95% accuracy, from which dysarthric speakers are far behind and excluded from the state of art escapade. In this paper, we propose a model to narrow down this gap and help them to get recognized by deep feature analysis and informative feature extraction techniques to get a better characterization of dysarthric speech. Linear Prediction Cepstral Coefficient (LPCC) is used as a feature extraction technique and a pre-clustering-based algorithm, Affinity Propagation (AP), is used to select the best features for a speaker based on the locality of the speech signal. SVM is used to verify the selection of features. UA-Speech digit data set is used for the experiment, and the result is consistent. A significant amount of data is required to build a conventional speech recognition system, but some fields, like medically challenged speech recognition, lack sufficient data, that's why some optimized or alternate techniques are used.

Keywords: Impaired Speech, Dysarthric Speech, Affinity Propagation, SVM, Speech recognition.

1 Introduction

In recent years automatic speech recognition (ASR) systems have been enhanced significantly with the help of Artificial intelligence and Deep Learning Technologies but the state-of-the-art result keeps impaired speech far away. Since ASR systems operate hands-free, they can be used widely by disabled people. However, ASR systems are difficult to use by people who have speech disorders due to various reasons [1]. Cerebral palsy, which is a central nervous system disorder, is one of the causes of speech disorders. This paper is about dysarthric speech in which the speaker leads to slow movements of the muscles in the left hemisphere of the brain while speaking. They can't coordinate with their vocal cords and muscles' actions resulting in unclear and unstable utterances [2]. Dysarthric speakers can often access technology more efficiently and effectively using hands-free or speech-enabled interfaces rather than remote control, equipment switches

and keywords. For those with dysarthria to be able to use ASR effectively, there is a serious need for a reliable system [3]. A particular condition that creates problems in the formation of speech sounds and great difficulties while communicating with others is termed as speech disorder. This may create a problem for others to understand their language. This condition affects a person's ability to speak fluently. There are several kinds of communication disorders broadly categorized into language disorders and speech disorders. People can have both speech and language disorders at a time [4][15].

2 Literature Survey

The classification of vowels is one of the most important components of modeling spoken units since they are steady-state segments. Many studies have been conducted on the phonetic characteristics of vowels in different languages, as well as methods to characterize the spoken units, such as voice activity detection and keyword spotting. Assamese vowels were studied using formant frequencies in RNNs and KNNs by Sharma et al [5]. In another approach, they also classify assamese fricatives using standard deviation, skewness, kurtosis, etc [6]. Matthew and Yossi et al. [7] designed an algorithm that measures vowel duration using a forced aligner based on HMMs and Deep Neural Networks (DNNs).

Affinity Propagation was published by Frey and Dueck in 2007 [8]. In other clustering algorithms like k-means, an initial set of exemplars are taken, which should be close to a good solution in order to get the right clustering. On other hand, "affinity propagation" which takes as input measures of similarity matrix that consist of similarity values between pairs of data points. Jin Wang et al. [9] used affinity propagation in Binaural sound localization to estimate the direction of sound source. They applied the clustering analysis to the similarity matrix based on affinity propagation. Affinity propagation is advantageous over common clustering algorithms such as hierarchical clustering and k-means. In affinity propagation, the number of the clusters does not need to be specified beforehand as done in k-means. It has been used in various fields including speech, such as speaker clustering [10], image processing [11], EEGs [12]. P.Bhagat et al. [13] used PLP coefficient as features for the classification of vowels and used a similar approach of pre-clustering through affinity propagation to select features. A. Smeulders and H. T. Nguyen [14] showed that pre-clustering is an effective way to improve classification by identifying a set of features that contribute to good classification.

3 Methodology

Pre-processing of raw data and feature selection is one of the most important keys for getting an accurate and fast trainable model. First, we take the raw speech signal and do some pre-processing followed by feature extraction using LPCCs, then from extracted features, we select the best features through affinity propagation by taking the best clustering score. After that, we classify the data

and validate the selection of features with SVM. The proposed approach is shown in Fig 1.

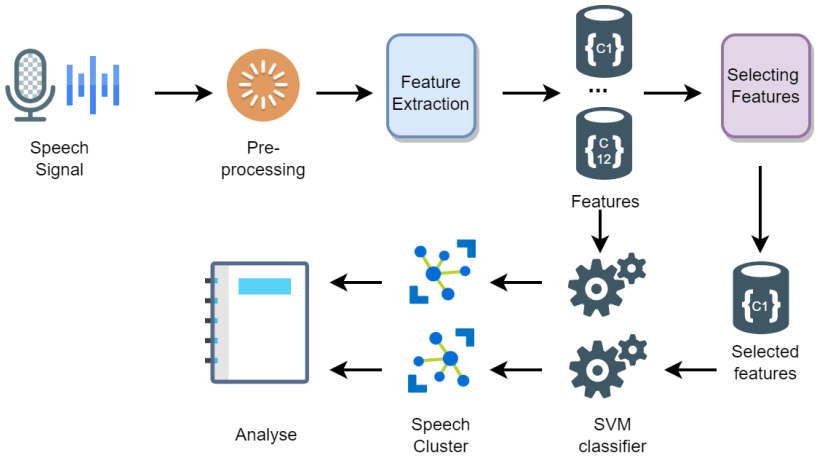

Fig. 1: Proposed Approach

All steps are explained in the next section. The pre-processing, feature extraction and affinity propagation implementation are done in Microsoft Visual Studio 2010 in C language. The pre-clustering analysis and SVM classification are done in Jupyter Notebook in python.

3.1 Pre-processing

We take vowel samples and digit samples from the UASpeech database [15] for all speakers with different severity levels. We correct the DC-shift of speech samples by taking the summation of silence part and averaging it by dividing it by the total samples, which will be the shift. Then we cut the stable frames, i.e., trimming the silence part from the beginning and end of required speech samples. The trimming is done by calculating the Short Term Energy(STE) of all frames of 320 samples each. when the STE increases by 400%, we mark it as the start of a stable frame. We do a similar process for marking the end point of vowel speech utterances. We normalize the speech samples in the range of +5000 to -5000, i.e., multiplying the samples with a factor such that maximum and minimum of sample becomes +5000 and -5000, respectively.

3.2 Feature Extraction

In first stage of the system after pre-processing, the audio signal is converted into some parametric values with distinctiveness which represent speech and

speaker characteristics. The transformation of signal to parameters is called feature extraction and researchers have proposed many different methods for this like MFCC, LPCC, PLP, etc. In this approach, we have used LPCCs.

Framing of the sample is done on the pre-processed signal which is basically cutting of signals into small time slots which can be considered to have a constant property. In this experiment, speech signal is divided into frames of size 320 samples. To avoid the loss of information, frames should be 25% overlapping. Now, we calculated the Ri values through auto-correlation method. It is the degree of similarity/agreement between a given time series(audio signal) and a lagged version of itself over successive intervals of time. It is calculated using equation 1

$$R_i = \sum_{m=0}^{N-i-1} S(m) * S(m+i) \tag{1}$$

In above equation value of i will vary from 0 to 12. So, for each frame of size 320 sample with the above equation we got 13 values naming $R_0, R_1, R_2...R_{12}$. Then we used Levinson and Durbin's algorithm to compute the linear prediction filter coefficients. It uses the auto-correlation method to estimate the linear prediction parameter for a segment of speech.

3.3 Affinity Propagation

Now we got 12 Ci values $C1, C2, C3...C12$ for each frame. We took all C1 from each frame together and treated it as one set of features and similarly, we took other Ci's and got the final 12 sets of different features Fig 2. Now we select the features which get the closest optimal classification so as to reduce the modeling time. For selecting features we find the clustering score of each feature set using Affinity Propagation and select the feature sets which gave the highest score, as the better clustering score implies that the set of features differentiates the classes well.

The Affinity propagation algorithm finds the number of clusters and the mapping of data points to the cluster. It doesn't require the initial centroids or number of clusters. It models each data point as a node in a network.

First, the similarity matrix is generated with the data points by negating the sum of the squares of the differences between data points. The similarity matrix contains information of preferences and similarity. Preferences are the data points that are suitable to be an exemplar. It is stored in the diagonal and the non-diagonal elements represent the similarity values between data points.

As a starting point, we construct an availability matrix with zero elements. Then we use equation 2 to calculate the responsibility matrix.

$$r(i,k) \leftarrow s(i,k) - max_{k`,k` \neq k}\{a(i,k`) + s(i,k`)\} \tag{2}$$

where 'i' is rows and 'k' is column of responsibility matrix. "Responsibilities" r(i,k) is sent from data points to candidate exemplars to indicate how strongly each data point favors the candidate data exemplar.

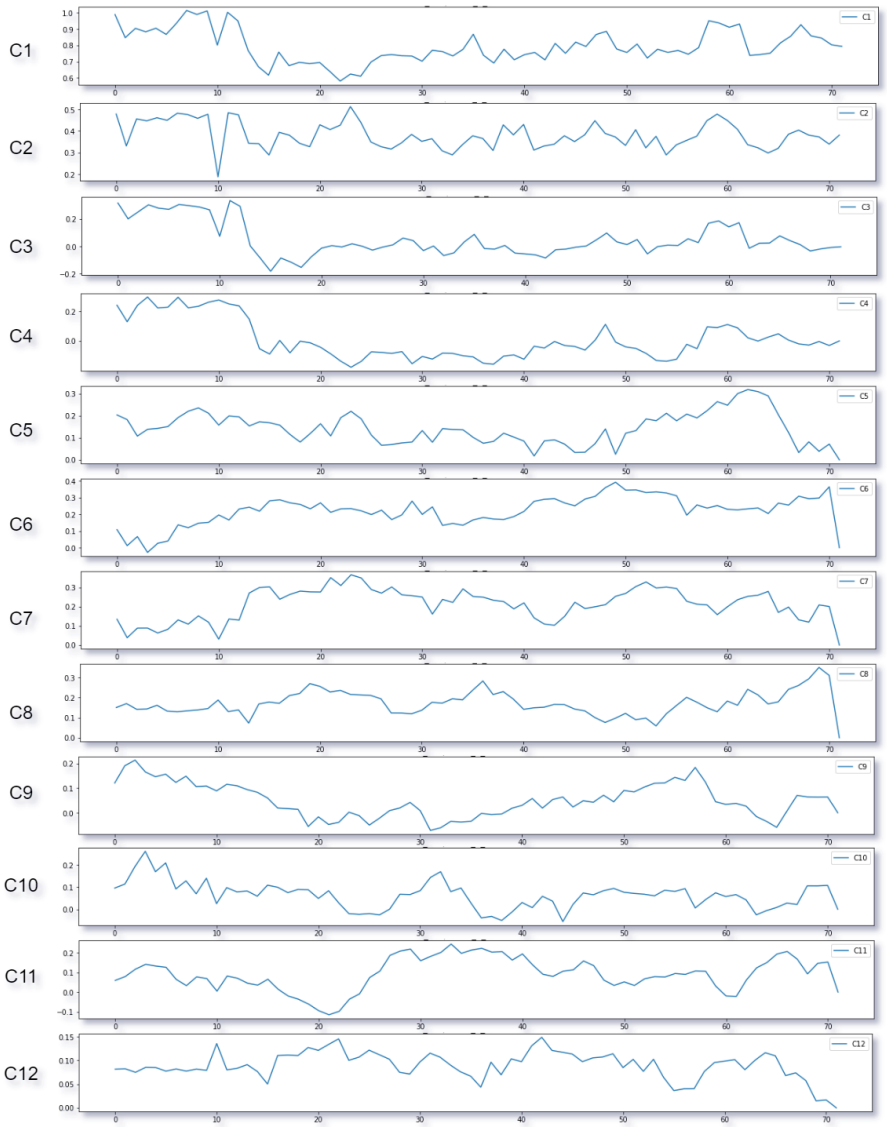

Fig. 2: Features of 1st utterance of vowel /a/

3.4 Classification

Now that we have selected features for labeled data points. We model and classify the vowels and digit data. From the dataset, we took training samples and then tested the remaining testing samples. We also used SVM to validate the selection of features by classifying the data using all feature sets and getting the same accuracy as with selected features.

4 Experiments and Results

We took the vowel and digit data from UASpeech database. English vowels have 5 classes/a/, /e/, /i/, /o/, /u/, and 10 classes of digits 0-9 with 7 utterances of each class. The LPCCs are extracted through self-written C code in Microsoft Visual Studio 2010. Affinity Propagation is done using a C-written code from scratch for understanding the algorithm well and then shifted to python with sklearn library for ease of experimenting with different data and changing parameters quickly. SVM for classification and verification of pre-clustering is also done in Python with sklearn library. Parameters for SVM are shown in Table 1.

No.	Specification	Description
1.	Regularization	1.0
2.	Kernel	RBF
3.	gamma	scale
4.	Cache size	200 MB

Table 1: Parameters of SVM

The set of features for vowel /a/ is shown in the Fig 2. Each feature is the set Ci for each frame from a voice signal. Using pre-clustering we got the clustering score, i.e. random index shown in Table 2. So according to the results shown in table, $c1$ and $c3$ has the highest clustering score and we select them. The selected features were used to classify by SVM and got 75% accuracy.

We also verified the results of selection by taking all the features and applied SVM and got the same accuracy of 75%. In the digit data set we again selected the best features shown in Table 3 on the basis of pre-clustering scores which were c4 and c10. We applied the SVM on the selected features and got 92% accuracy which was then verified using all the features and got the same accuracy of 92%. Fig 3 shows the original cluster v/s the clustering done by Affinity Propagation where blue cluster represents /aa/, green represents /e/, red represents /i/, cyan represents /o/ and magenta represents /u/.

The time taken by the model to train using all the features in vowel data set was 3 ms, which got reduced to 0.9 ms after using the selected features. Similarly, for digit data set the time for modeling was reduced from 18.9 ms to 1.9 ms.

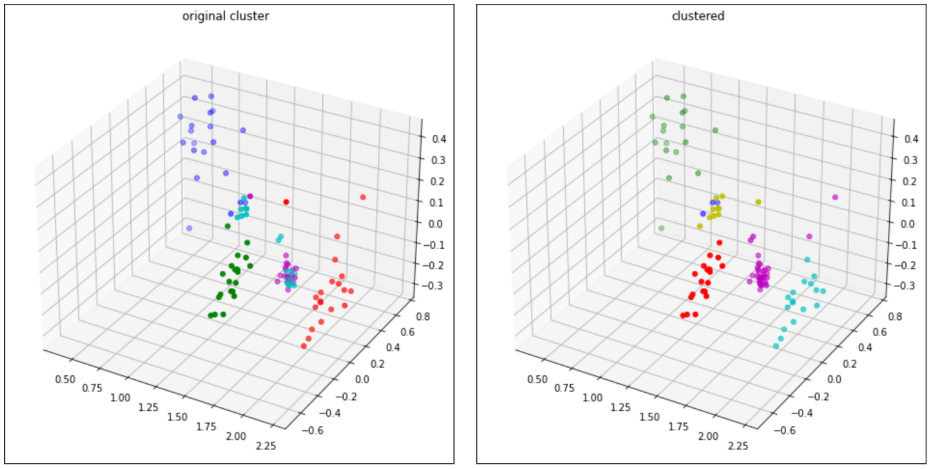

Fig. 3: Original clustering vs Clustering by affinity propagation

Feature	Score
C 1	**0.81**
C 2	0.78
C 3	**0.83**
C 4	0.72
C 5	0.70
C 6	0.73
C 7	0.64
C 8	0.70
C 9	0.75
C 10	0.68
C 11	0.66
C 12	0.69

Table 2: Clustering Score(Random Index) of each feature set for vowel data-set

Feature	Score
C 1	0.87
C 2	0.78
C 3	0.87
C 4	**0.93**
C 5	0.08
C 6	0.83
C 7	0.85
C 8	0.89
C 9	0.84
C 10	**0.90**
C 11	0.86
C 12	0.83

Table 3: Clustering Score(Random Index) of each feature set for dysarthric digit data-set

5 Conclusion and future work

The experiments done in this system focused on using a pre-clustering algorithm technique over English-spoken vowels and digits spoken by a dysarthria patient. We used the LPCCs as feature extraction and did the classification using SVM. We found that pre-clustering through AP works well with LPCCs for feature selection which reduced the modeling time and does not reduce the accuracy. In the future, this method needs to be verified for other data sets. This approach has been experimented for single speaker and can be combined with other feature extraction techniques. The given approach works well for a single speaker and can be used to build a better classification model for dysarthric speech.

References

1. B. Wu et al., "An End-to-End Deep Learning Approach to Simultaneous Speech Dereverberation and Acoustic Modeling for Robust Speech Recognition," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1289-1300, Dec. 2017, doi: 10.1109/JSTSP.2017.2756439.
2. F. Rudzicz, "Articulatory Knowledge in the Recognition of Dysarthric Speech," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 947-960, May 2011, doi: 10.1109/TASL.2010.2072499.
3. Biadsy, Fadi, et al. "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation." arXiv preprint arXiv:1904.04169 (2019).
4. M. Day, R. K. Dey, M. Baucum, E. J. Paek, H. Park and A. Khojandi, "Predicting Severity in People with Aphasia: A Natural Language Processing and Machine Learning Approach," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021, pp. 2299-2302, doi: 10.1109/EMBC46164.2021.9630694.
5. M. Sharma and K. K. Sarma, "Dialectal Assamese vowel speech detection using acoustic phonetic features, KNN and RNN," 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN), 2015, pp. 674-678, doi: 10.1109/SPIN.2015.7095270.
6. C. Patgiri, M. Sarma and K. K. Sarma, "Recurrent neural network based approach to recognize assamese fricatives using experimentally derived acoustic-phonetic features," 2013 1st International Conference on Emerging Trends and Applications in Computer Science, 2013, pp. 33-37, doi: 10.1109/ICETACS.2013.6691390.
7. Y. Adi, J. Keshet and M. Goldrick, "Vowel duration measurement using deep neural networks," 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015, pp. 1-6, doi: 10.1109/MLSP.2015.7324331.
8. Frey, Brendan & Dueck, Delbert. (2007). Clustering by Passing Messages Between Data Points. Science (New York, N.Y.). 315. 972-6. 10.1126/science.1136800.
9. Wang, Jing & Wang, Jin & Qian, Kai & Xie, Xiang & Kuang, Jingming. (2020). Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition. EURASIP Journal on Audio Speech and Music Processing. 2020. 10.1186/s13636-020-0171-y.
10. Xiang Zhang, Jie Gao, Ping Lu and Yonghong Yan, "A novel speaker clustering algorithm via supervised affinity propagation," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 4369-4372, doi: 10.1109/ICASSP.2008.4518623.

11. Zhang, Kang & Gu, Xingsheng. (2014). An Affinity Propagation Clustering Algorithm for Mixed Numeric and Categorical Datasets. Mathematical Problems in Engineering. 2014. 10.1155/2014/486075.
12. Thomas, John & Jing, Jin & Dauwels, Justin & Cash, Sydney & Westover, M Brandon. (2017). Automated epileptiform spike detection via affinity propagation-based template matching. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference. 2017. 3057-3060. 10.1109/EMBC.2017.8037502.
13. P. Bhagath, K. Bharti, A. Kotiya and P. K. Das, "Feature Selection using Pre-clustering via Affinity Propagation for Speech Classification in Low-resource Languages," 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), 2021, pp. 1-6, doi: 10.1109/IICAIET51634.2021.9573696.
14. Hieu T. Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). Association for Computing Machinery, New York, NY, USA, 79. https://doi.org/10.1145/1015330.1015349
15. Kim, Heejin & Hasegawa-Johnson, Mark & Perlman, Adrienne & Gunderson, Jon & Watkin, Kenneth & Frame, Simone. (2008). Dysarthric speech database for universal access research. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 1741-1744.
16. S. Latif, J. Qadir, A. Qayyum, M. Usama and S. Younis, "Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art," in IEEE Reviews in Biomedical Engineering, vol. 14, pp. 342-356, 2021, doi: 10.1109/RBME.2020.3006860.