# Automatic Detection of Speech Disorder in Dysarthria using Extended Speech Feature Extraction and Neural Networks Classification

*T B Ijitona*, J J Soraghan*, A Lowit†, G Di-Caterina*, H Yue**

**Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom*
*†Department of Speech and Language Therapy, University of Strathclyde, Glasgow, United Kingdom*
tolulope.ijitona@strath.ac.uk

**Keywords:** Dysarthria, speech disorder; Centroid Formants, Neural Networks.

## Abstract

This paper presents an automatic detection of Dysarthria, a motor speech disorder, using extended speech features called Centroid Formants. Centroid Formants are the weighted averages of the formants extracted from a speech signal. This involves extraction of the first four formants of a speech signal and averaging their weighted values. The weights are determined by the peak energies of the bands of frequency resonance, formants. The resulting weighted averages are called the Centroid Formants. In our proposed methodology, these centroid formants are used to automatically detect Dysarthric speech using neural network classification technique. The experimental results recorded after testing this algorithm are presented. The experimental data consists of 200 speech samples from 10 Dysarthric speakers and 200 speech samples from 10 age-matched healthy speakers. The experimental results show a high performance using neural networks classification. A possible future research related to this work is the use of these extended features in speaker identification and recognition of disordered speech.

## 1 Introduction

Dysarthria is a neurological motor speech disorder that affects the production of speech due to the weakness of the muscles and nerves involved [1]. These include impairment in the movement of the lips, larynx, vocal cords, tongue and/or nasal air passage [2]. The effects of dysarthria are seen in the speed, variability, consistency or rhythm in speech production [1]. Dysarthria affects the five primary speech subsystems in speech production. These subsystems include respiration, resonance, phonation, articulation and prosody. Dysarthria is generally characterised by slurred speech, slow speech rate, low voice quality, lopsided rhythm, low loudness, facial drooping or/and exertion in moving facial muscles [1]. Common causes of dysarthria include stroke, ALS (Amyotrophic Lateral Sclerosis), Parkinson's disease, multiple sclerosis, degenerative diseases, brain injury, tumours, etc.

Dysarthria in itself is not a life-threatening disorder but affects the standard of living of people with dysarthria (PwD) socially, psychologically and in day-to-day communication. PwD are more likely to depend on others for day to day activities such as communication, socialising and sometimes eating as the severity increases. When dysarthria is early detected, the patients can be put on therapy sessions that will aid their communication and mitigate the effect of the disorder on their standard of living. Early detection of Dysarthria is therefore very curial in Dysarthria management [3].

The first technique involves the use of an invasive tool, such as fiberscope, to physically examine the patient's vocal folds and other speech production organs [4, 5]. This method causes discomfort to patients. Screening sessions using this method can also take a long while to be completed due to the clinical procedures involved. Automatic acoustic analysis for the detection of Dysarthria, on the other hand, is non-invasive and very useful during initial screening.

However, the physical examination is a subjective screening method whose results are based on the medical practitioner's perceptual analysis capabilities and experience [6, 7]. This results in inconsistent and unquantifiable outcomes. At the early onset of dysarthria, it is very difficult to identify certain diagnostic features by perceptual analysis only. This is because there is no quantitative measure of the features that are perceptually analysed. This leads to a high probability of error and thus a possibility of not detecting the speech disorder early. There is, therefore, a need to acoustically detect certain features that characterise the early occurrence of dysarthria. In this research, we are exploring objective ways of automatically detecting dysarthria in speech using a non-invasive acoustic analysis technique.

Objective techniques for voice screening have been proposed in recent studies based on time-domain [8], spectral [9] and cepstral analysis [10], [11]. These techniques include the use of amplitude, pitch, Mel-cepstral frequency, perturbation-shimmer, perturbation-jitter and harmonic to noise ratio. The performance of perturbation measures depends on the accuracy of pitch tracking algorithm which is one of the challenges of disordered speech analysis [12]. Mel-cepstral frequency measures, on the other hand, is a function of the number of Mel-frequency cepstral coefficients (MFCC) used.

In this paper, an extended linear prediction coding (LPC) measurement, called Centroid Formants, is used as an alternative technique as they are independent on pitch estimation accuracy and are less prone to noise, unlike MFCC-based techniques. The next sections of this paper are arranged as follows. The LPC based formants extraction algorithm is described in section 2. The proposed methodology for the automatic detection of dysarthria is discussed in section 3 after which the experimental results and analysis are presented in Section 4. The last section includes the conclusion from this research work and recommendations for future research.

## 2 The LPC Algorithm

Linear Prediction Coding, also known as LPC, is a spectral analysis technique used for encoding a signal in a way that the current value at a particular time t is taken as a linear function of the previous values at a time less than t [13]. The LPC analysis is based on the assumption that the human vocal tract can be modelled as a tube with varying diameter. This results in a mathematical model which is an approximation of the human vocal tract response [14]. In simplifying this model, an optimisation problem is reached which minimises the estimation error over time. This optimisation function is given by:

$$\dot{P}_e = E\{e^2[n]\} = E\left\{\left(x[n] - \sum_{k=1}^{N} a_k x[n-k]\right)^2\right\} \qquad (1)$$

where $a_k$ is the kth LPC coefficient, N is the order of the linear prediction and x[n] is the speech signal.

The speech sample x[n] is represented as a weighted linear sum of Nth previous samples; given that N is the order of the LPC estimation. This results in a prediction system where the next sample is predicted by the sum of N preceding samples. The resulting coefficients of the LPC are used in estimating the formants; the frequency characteristics of a speech signal over time. Formants are also frequencies within the speech spectrum where acoustic energy are concentrated [15].

The linear prediction filter reduces the bit rate of the speech signal thereby reducing the quality of the signal [13]. This results in reduced speech quality. For example, a speech signal sampled at 8kHz and with encoded at 8 bit per sample will have a bit rate of 64kbits/sec. However, performing a linear prediction will reduce the rate to 24kbits/sec. This is one of the limitations of LPC analysis.

Furthermore, research [16], [17] has shown that even though the bit rate is reduced during linear prediction coding, the estimated speech signal remains audible and comprehensible. Due to this attributes, the LPC is useful in speaker identification and also in speech coders with low or medium bit rate [18]. The LPC also offers a robust and reliable way of estimating the main frequency components of speech signals (formants) [18].

In addition, the LPC analysis will be useful in considering the frequency characteristics of the dysarthria speech. The accuracy of the LPC in formants estimation is high compared to other feature extraction techniques [19]. The LPC is robust to noise, unlike the MFCC feature extraction technique. Also, the formants estimation is useful as a tool for measuring the intelligibility and pronunciation features in spoken language [20].

## 3 Automatic Detection Algorithm

The block diagram of the proposed algorithm is illustrated in Figure 1. The first stage is pre-processing followed by feature extraction. After extraction of the speech features, the next stage is classification based on neural network techniques.

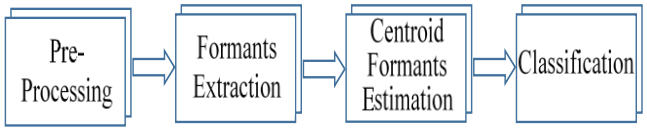


Figure 1: Block Diagram of the Proposed Algorithm

### 3.1 Pre-Processing

Pre-processing is carried out in speech processing to enhance the performance of the speech feature extraction algorithms. This includes resampling, amplitude normalisation, and framing [11]. For this study, all audio samples are resampled at 16 kHz. Due to the variations in speaker volume and microphone distance, the amplitude of the audio samples are normalised such that the dynamic range of the signal lies between -1.0 and +1.0 (without changing the sign of the signal values). Amplitude normalisation is achieved by dividing the signal by its maximum absolute value.

Speech signals are not stationary in nature and thus it is essential to analyse these signals in a short time interval. The process of dividing an audio signal into short interval uniform frames is called framing. The resulting amplitude normalised audio signals sampled at 16 kHz are divided into overlapping frames of 256 samples each with 80% overlap between consecutive frames. Using overlapping frames is targeted towards improvements in the segmentation process.

### 3.2 Formants Extraction

Formants are bands of resonance in the frequency spectrum of a speech signal.[13] These bands of resonance are the significant representation of the signal. The formant extraction algorithm, in this proposed technique, is based on the Linear Prediction Coding (LPC) analysis. The LPC analysis gives a smoothed approximation of the power spectrum of the original signal [13].

The formant extraction is based on the energy distribution of the signal in frequency domain. The formants positions are

chosen in such a way that they match this distribution of energy. These formants are prominent frequencies within the spectrum with bandwidths of less than 400Hz. Therefore, frequency bands with a high concentration of energy and bandwidths less than 400Hz are located as the formants of the speech signal.

Using LPC analysis, the order of the linear prediction is a function of the sampling frequency of the speech signal given by the rule of thumb illustrated in Equation (2).

$$Ncoeff = 2 + \frac{Fs}{1000} \quad (2)$$

where Ncoeff is the order of the LPC and Fs is the sampling frequency.

The estimated LPC coefficients are converted from rectangular form to polar form and the phases of the coefficients with bandwidths less than 400Hz and positive phase are extracted as the bands of resonance of the spectrum. These positive phases are called the formants. Figure 2 and Figure 3 show the formants extracted from ataxic dysarthric speech and health speech respectively for the word defer.

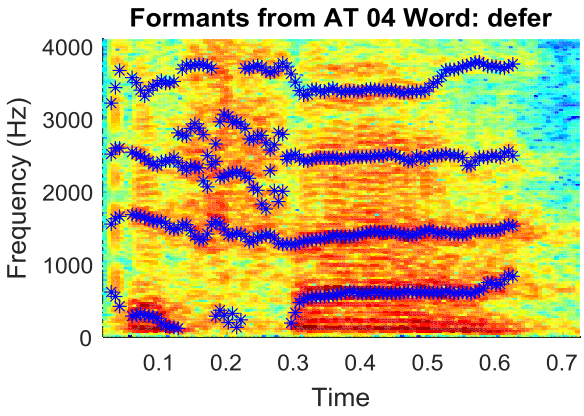


Figure 2: Formants extracted from AT speech

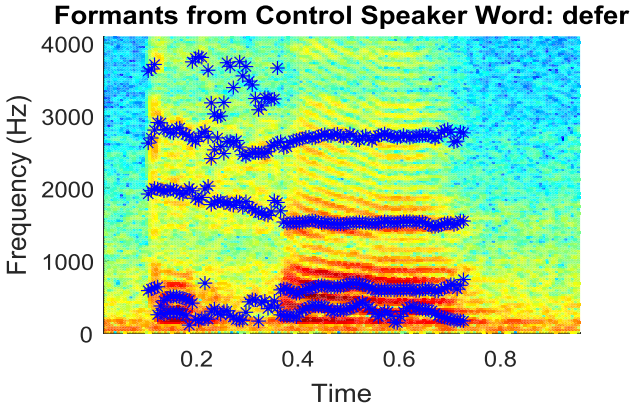


Figure 3: Formants extracted from healthy speech

### 3.3 Centroid Formants Estimation

Centroid formants are the weighted averages of the formants in each frame in the short time frequency spectrum. The formants are weighted by their corresponding formants energy. The centroid formant is a measure of where the power in the frequency spectrum of an audio signal is centralised. For instance, if the majority of the power in the spectrum resides in high-frequency components, then the centroid formant will lie in the high-frequency range. However, if most of the power resides in low-frequency components, the centroid formants will be located at low-frequency range. Figure 4 and Figure 5 illustrates the centroid formants of the audio files showed in Figure 2 and Figure 3 respectively for an ataxic speaker and a healthy speaker.

Given that F1n, F2n, F3n, and F4n (for k=1, 2, 3, 4) are the four formants of the nth frame of an audio signal and the corresponding formants energy are E1n, E2n, E3n and E4n respectively. The centroid formant of the nth frame is given by CFn as illustrated in (3).

$$CF_n = \frac{E_{1,n}F_{1,n} + E_{2,n}F_{2,n} + E_{3,n}F_{3,n} + E_{4,n}F_{4,n}}{4} \quad (3)$$

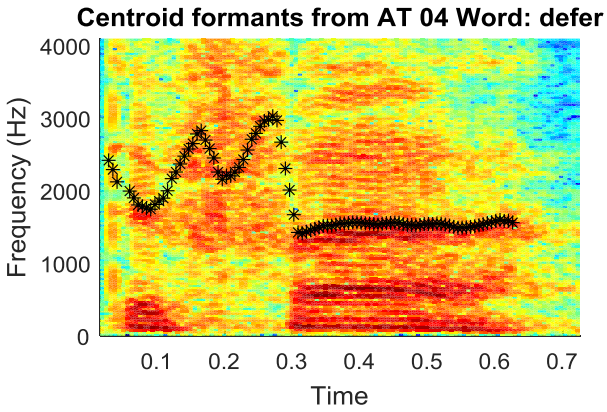


Figure 4: Centroid formants for AT speech

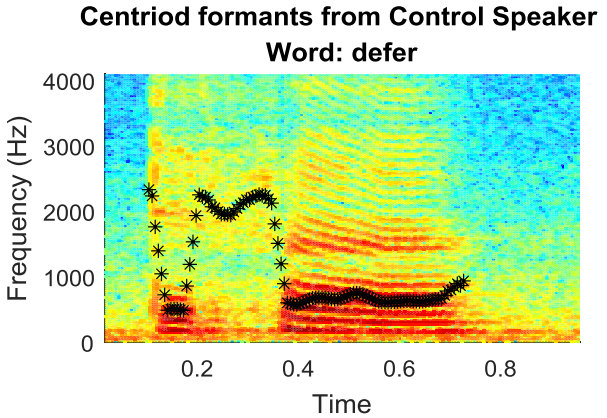


Figure 5: Centroid formants for healthy speech

The centroid formant can be used to measure the rate of change of the formants and the intonation pattern of the audio signal. This is because as the formants changes from low frequency to high frequency, the centroid formants also changes in the same pattern. The weighting of the individual formants also ensures that frequency components with highest power contribution are given the heaviest weight. Therefore, the effects of picking weak peaks as formants will be reduced.

In addition, there is a close relationship between pitch and centroid formants profiles for healthy speech. Considering the fact that formants are harmonics of the fundamental frequency, the pattern of each formant will mimic the shape of the fundamental frequency. If the energy contribution of each formant remains the same within a speech segment, the centroid formant will also give a pattern similar to the fundamental frequency.

However, this similarity with pitch profile is not true for audio signals with rapidly changing intonation patterns, that is, in disordered speech. The centroid formant is very sensitive the rapid changes in pitch and intonation. This means that the high pitch variability in dysarthric speech can effectively and efficiently be tracked using centroid formants. Any sudden change in pitch or intonation is reflected.

### 3.4 Classification

The Artificial Neural Networks is used for classifying the disordered speech. One of the commonly used machine learning methods is the neural network. This classification technique is robust and it combines pattern recognition with acoustic phonetic methods [21]. In this artificial learning technique, knowledge of the acoustic phonetic characteristics of the speech is used to generate rules for classifiers [22] A multilayer neural network with one hidden layer was used for this classification. The excitations (inputs) are the centroid formants and the observations (outputs) indicates whether or not the corresponding audio sample is from the ataxic dysarthric speaker (0) or healthy speaker (1). In our study, single layer neural network with 10 neurons in a hidden layer was used.

## 4 Results and Analysis

### 4.1 Speech Corpus

The dataset used for this study consists of 400 audio samples from 20 speakers, 10 of which are ataxic dysarthric speakers and 10 age healthy control speakers. Each speaker produced 20 single word speech. Each group consist of 5 males and 5 females. The participants in both groups are also age-matched. This corpus was taken from the dataset reported by [23]. The ataxic dysarthric speakers have no cognitive deficiency neither do they have any visual and hearing impairment. The severity of the ataxic dysarthric speakers varied from mild to severe cases as illustrated in Table 1. In addition, all of them were monolingual speakers of Standard Southern British English or Standard Scottish English.

| Participant | Age | Gender | Etiology | Intelligibility Score (%) |
|---|---|---|---|---|
| AT_01 | 46 | M | CA | 74 |
| AT_02 | 60 | F | CA | 67 |
| AT_03 | 28 | M | FA | 6 |
| AT_04 | 52 | F | CA | 25 |
| AT_05 | 28 | F | FA | 9 |
| AT_06 | 65 | F | SCA6 | 58 |
| AT_07 | 72 | M | CA | 19 |
| AT_08 | 51 | M | CA | 44 |
| AT_09 | 56 | M | SCA8 | 82 |
| AT_10 | 57 | F | FA | 80 |

Table 1: Details of participants involved in the study

Moreover, the intelligibility scores for the ataxic dysarthric speakers varied from 6 to 82. These intelligibility scores were estimated from the average scores from five trained listeners during a passage reading task [23]. The etiologies of these participants are either cerebellar ataxia (50%), Friedreich's ataxia (30%) or spinocerebellar ataxia (20%).

### 4.2 Results

The classification was carried out using the Neural Network Toolbox in MATLAB R2016b (Version 9.1) software. Using 10 neurons and a single hidden layer, the audio samples were trained. The audio samples training distribution was as follows; 70% of the audio samples were used for training, 15% for testing and 15% for validation. The audio samples distribution across these 3 groups (training, testing and validation) was done randomly.


Figure 6: Confusion matrix for the neural networks classification

Even though a single hidden layer has been used for this classification, the accuracy overall accuracy recorded was 75.6% using 10 neurons. The confusion matrix for the trained neural network is illustrated in Figure 6. The first two columns of the confusion matrixes indicate the two target classes (0 for normal speech and 1 for dysarthric speech) whereas the third column shows the positive and negative prediction values. Likewise, the first two rows of the confusion matrixes show the two output classes (0 or 1) whereas the third row shows the sensitivity, specificity and accuracy of the network. The training dataset gives an accuracy of 74.3%, the validation dataset gives an accuracy of 80.3%, and whereas the test data set gives an accuracy of 77.0% brings the total to 75.6%.

## 5 Conclusion

In this paper, we have presented an extended speech feature for classification of disordered speech from healthy speech using neural networks. The extended feature, centroid formants, proposed in this paper gave an accuracy of 75.6% with just one hidden layer and 10 neurons. This classification has been carried out across different levels of severity of ataxic dysarthria from mild to highly severe cases. Classification using other artificial intelligence techniques such as Deep Neural Networks (DNN), Support Vector Machine (SVM), LQV and Hidden Markov model has been left for future research work. We intend to investigate how pre and post processing of the extracted feature can be used to increase the performance of the classification algorithm. In addition, the application of centroid formants in speaker identification, speech recognition and emotion detection is yet to be explored. This extended feature can also be combined with other spectral and cepstral features for various classification applications.

## References


[1] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

[2] E. C. Guerra and D. F. Lovey, "A modern approach to dysarthria classification," in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, 2003, vol. 3, pp. 2257-2260 Vol.3.

[3] M. A. Wahed, "Computer aided recognition of pathological voice," in *2014 31st National Radio Science Conference (NRSC)*, 2014, pp. 349-354.

[4] H. F. Robinson, "Assessment of voice problems," *Assessment in Speech and Language Therapy,* pp. 68-84, 1993.

[5] H. Hirose, "Pathophysiology of motor speech disorders (dysarthria)," *Folia Phoniatrica et Logopaedica,* vol. 38, no. 2-4, pp. 61-88, 1986.

[6] P. Wannberg, E. Schalling, and L. Hartelius, "Perceptual assessment of dysarthria: Comparison of a general and a detailed assessment protocol," *Logopedics Phoniatrics Vocology,* vol. 41, no. 4, pp. 159-167, 2016.

[7] B. J. Zyski and B. E. Weisiger, "Identification of dysarthria types based on perceptual analysis," *Journal of Communication Disorders,* vol. 20, no. 5, pp. 367-378, 1987.

[8] M. Novotny, J. Pospisil, R. Cmejla, and J. Rusz, "Automatic detection of voice onset time in dysarthric speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 4340-4344.

[9] J. R. Orozco-Arroyave *et al.*, "Characterization Methods for the Detection of Multiple Voice Disorders: Neurological, Functional, and Laryngeal Diseases," *IEEE Journal of Biomedical and Health Informatics,* vol. 19, no. 6, pp. 1820-1828, 2015.

[10] T. Villa-Cañas *et al.*, "Automatic detection of laryngeal pathologies using cepstral analysis in Mel and Bark scales," in *2012 XVII Symposium of Image, Signal Processing, and Artificial Vision (STSIVA)*, 2012, pp. 116-121.

[11] 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.

[12] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Transactions on Biomedical Engineering,* vol. 59, no. 5, pp. 1264-1271, 2012.

[13] J. L. G. Zapata, D. J. C, x00Ed, M. az, and P. G. Vilda, "Fast formant estimation by complex analysis of LPC coefficients," in *Signal Processing Conference, 2004 12th European*, 2004, pp. 737-740.

[14] 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.

[15] V. S. Selvam, V. Thulasibai, and R. Rohini, "Speech training system based on resonant frequencies of vocal tract," in *Advanced Communication Technology (ICACT), 2011 13th International Conference on*, 2011, pp. 674-679.

[16] N. Kamaruddin, A. W. A. Rahman, and N. S. Abdullah, "Speech emotion identification analysis based on different spectral feature extraction methods," in *Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on*, 2014, pp. 1-5.

[17] J. M. Elvira, F. J. Dickin, and R. A. Carrasco, "A comparison of speech feature extraction employing autonomous neural network topologies," in *Systems and Applications of Man-Machine Interaction Using Speech I/O, IEE Colloquium on*, 1991, pp. 9/1-9/5.

[18] S. B. Magre and R. R. Deshmukh, "A Review on Feature Extraction and Noise Reduction Technique," *International Journal of Advanced Research in Computer*

*Science and Software Engineering,* vol. 4, no. 2, pp. 352 -356, 2014.

[19] V. Narang, D. Misra, and G. Dalal, "Acoustic Space in Motor Disorders of Speech: Two Case Studies," in *Asian Language Processing (IALP), 2011 International Conference on*, 2011, pp. 211-215.

[20] H. Tolba and A. S. El, "Towards the improvement of automatic recognition of dysarthric speech," in *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, 2009, pp. 277-281.

[21] M. Chetouani, A. Hussain, B. Gas, and J. L. Zarader, "Non-Linear Predictors based on the Functionally Expanded Neural Networks for Speech Feature Extraction," in *Engineering of Intelligent Systems, 2006 IEEE International Conference on*, 2006, pp. 1-5.

[22] N. Desai, K. Dhameliya, and V. Desai, "Feature extraction and classification techniques for speech recognition: A review," *International Journal of Emerging Technology and Advanced Engineering,* vol. 13, no. 12, pp. 367-371, 2013.

[23] A. Lowit, A. Kuschmann, J. M. MacLeod, F. Schaeffler, and I. Mennen, "Sentence stress in ataxic dysarthria: a perceptual and acoustic study," *Journal of Medical Speech Language Pathology,* vol. 18, no. 4, pp. 77-82, 2010.