Comparative Analysis of Various Feature Extraction Techniques for Classification of Speech Disfluencies

Nitin Mohan Sharma[1,2], Vikas Kumar[1,2], Prasant Kumar Mahapatra[1,2], Vaibhav Gandhi[3]

[1]CSIR-Central Scientific Instrument Organisation (CSIR-CSIO), Chandigarh-160030, India;
[2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad-201002, India;
[3]Middlesex University London,UK.
{nitin.csio17a, vikas.csio18j}@acsir.res.in, prasant22@csio.res.in, V.Gandhi@mdx.ac.uk

Abstract— **Speech plays a vital role in communication, from expressing oneself, to utilizing speech-based platforms, speech is a necessity. Any disruption in speech is referred to as disfluency, and can impact one's quality of life. This paper presents an experimental study on various techniques for the detection and classification of speech disfluencies. Six different types of disfluencies are examined in this paper, namely Interjection, Sound Repetition, Word Repetition, Phrase Repetition, Revision and Prolongation (6 classes). However, this paper also goes a step further by including the clean speech signals as an added class alongside the six disfluencies, thereby making this work more robust with 7 classes. Various machine learning approaches have been investigated on the University College London Archive of Stuttered Speech (UCLASS) dataset; a standard disfluency dataset generated by University College London (UCL). Five different feature extraction techniques viz. Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), Gammatone Frequency Cepstral Coefficients (GFCC), Mel-filterbank energy features, and Spectrograms have been used. Comparative analysis of various classifiers shows that MFCC, GFCC, and Spectrograms achieved greater than 90% accuracy on both 6 and 7 classes with the kNN classifier. As a future scope to this study, the authors aim to focus on tackling the challenges of detecting multiple disfluencies present simultaneously in a speech sample.**

Index Terms— **Disfluency, Speech Recognition, Feature Extraction, Speech Signals.**

1. Introduction

Speech is an effective way to express ideas, feelings, and thoughts by humans. Speech is vital to effective communication. However, speech is not always without disruptions. Any disruption in speech is referred to as disfluency. Disfluencies in speech make it challenging for individuals to express themselves. Most people are disfluent to some extent in speaking. Disfluency, seen in about 5% children and in about 1% adults of the world population, is also referred to as stuttering (Bloodstein, 1969; Esmaili et al., 2016a). The male population is 4 times more likely to have disfluency in comparison to females (Awad, 1997; Chee et al., 2009b).

Determining the cause of stuttering is still a major challenge, however several aspects such as genetic, psychological, and neurological factors have been explored (Conture, 2001; Johnson and Supplement, 1961; Watkins et al., 2008; Weber-Fox et al., 2013). Same type of genes shared by family members and usually the family environment, have a huge impact on the stuttering of individuals (Drayna and Kang, 2011). However, there is no consistent proof of sharing stuttering with genes (Smith and Weber, 2016). Generally, Psychological factors are seen in the young children which can often be recognized as anxiety, lack of confidence, fear, hesitation etc. (Smith and Weber, 2016). These factors also play a significant role in speech disfluencies.

Most commonly used classification of disfluencies was proposed by Johnson (Johnson, 1961) and it has been used ever since by researchers and clinicians. The purpose of this study by Johnson was to explore the implications from the obtained data and to have clarification of the fundamental nature of stuttering. According to Johnson's classification, the types of disfluencies are: 1. Incomplete Phrase (To change the words), e.g. HE SHOULD…Where is he going? 2. Revision (To change the words), e.g. Where are THEY we going? 3. Interjection (Addition of words that are meaningless or irrelevant to the sentence), e.g. Where are uh-ummm we

going? 4. Phrase Repetition (Repeating a Phrase), e.g. WHERE ARE Where are we going? 5. Word Repetition (Repeating a whole word), e.g. WHAT-WHAT-what are you doing? 6. Part-word Repetition (Repeating a sound or syllable), e.g. W-W-W-what are you doing? 7. Prolongation (Holding a sound for a longer duration), e.g. What are wwwwwwwe doing? 8. Broken Word (Inserting pause inside a word) What are we DO-(pause)-ING?

The researchers (Yairi and Ambrose, 1999) have attempted to provide a grouping scheme for the listed symptoms based on which symptoms are Most Typical and Less Typical . Conture's (Conture, 2001) scheme considers symptoms that happen within words (classification types 5–8) as a sign of stuttering falling under Most Typical. Yairi and Ambrose's (Yairi and Ambrose, 1999) scheme places these same symptoms into the Most Typical. Wingate's (Wingate, 2001, 2002) scheme further divides the Most Typical symptoms (types 6–8) from hesitation-type Less Typical symptoms (types 1–5). Thus, all three schemes place symptoms 1–4 in the Less Typical, and symptoms 6–8 in the Most Typical.

Classification types are quite helpful in diagnostic assessment of speech disfluencies. Generally, the Speech Language Pathologist (SLP) manually performs a disfluency assessment by counting the number of disfluent words as a proportion to the total words in a passage (Yaruss, 1997). This also considers the time duration of the disfluencies and the total time taken for reading a passage. However, correct assessment has a huge dependency on the expert speech language pathologists, which may lead to low agreements between different SLPs. Even the work of experts on disfluency assessment utilizes a manual approach which includes recording of the speech, transcribing the speech and counting the different kinds of speech disfluencies (Stein-Rubin and Fabus, 2011). Hence, these stuttering assessments are usually time consuming, inconsistent, subjective, and prone to error (Curlee, 1981; Young, 1975). An automated assessment of speech disfluencies could be an effective solution. The automated model can evaluate the performance of the disfluent people before and after therapy, thereby acting as an important addition to speech therapy. It can also reduce the tedious work of SLPs by correctly identifying the disfluencies in the speech. This will help SLPs in focusing more on the therapy session rather than counting the disfluencies in the speech.

Automatic Speech Recognition (ASR) systems do not focus on handling disfluencies in speech (Mullin, 2016). Users suffer because the disfluent speech is not recognized accurately by the ASR systems. One of the reasons for the inaccuracy of ASR systems is that they assume that every sound generated by a user is an intended sound. However, this is not true for people with speech impediments. Hence, it leads to reduced effectiveness in proper utilization of the ASR systems. ASR systems will also be benefitted by automatic assessment of the disfluent speech as it will help the disfluent users to effectively utilize the ASR system. The task of automatic identification and classification of disfluencies in human speech is generally complex due to the lack of properly annotated datasets and occurrence of multiple disfluencies simultaneously.

In this paper, the main focus was to include a greater number of speech disfluencies which has not been the case with any study with machine learning approaches. The authors explored different acoustic features such as MFCC, LPCC, GFCC, Mel-filterbank energies and Spectrogram with different machine learning algorithms such as kNN, Decision Trees, LDA, SVM, Ensemble

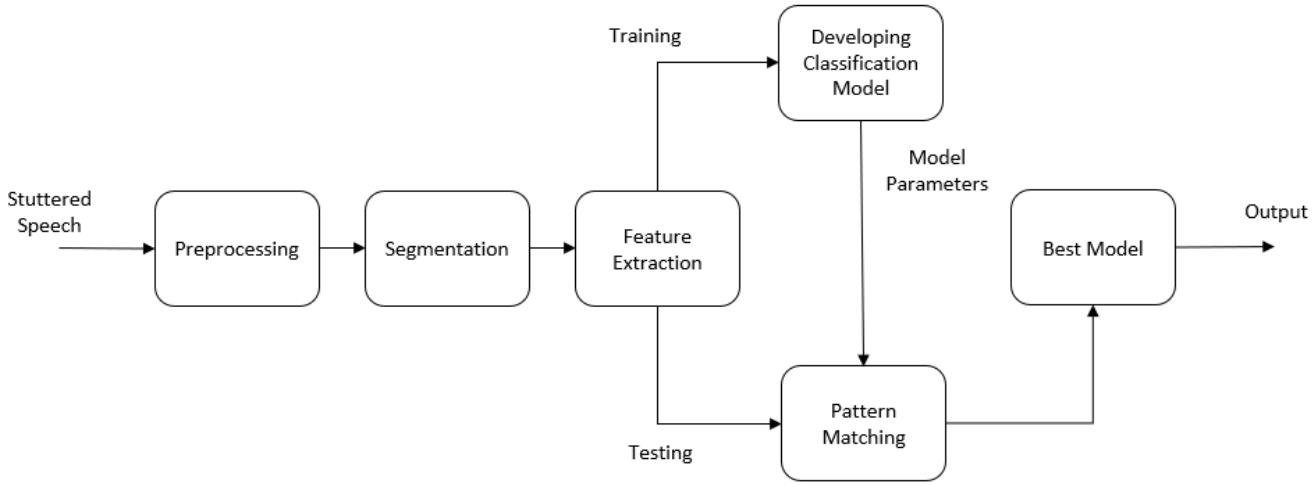

Fig. 1. General Framework for Disfluency Classification

algorithms. The MFCC, GFCC and Spectrogram features performed well with more than 90% accuracy. However, the MFCC features were better than the other two features in terms of a smaller number of coefficients used and more precision, recall and F1 score. The novelty of this paper lies in classifying a greater number of disfluency classes as well as including clean speech to be trained alongside disfluencies based on the assumption that the inherent patterns of disfluent classes match the clean speech class too. Hence, training our model on clean speech would make the model more robust. This paper is organized as follows. Section 2 presents an overview of previous studies on automatic recognition and classification of speech disfluencies. Section 3 discusses various techniques used for pre-processing, feature extraction and classification of speech disfluencies. Section 4 presents the results of the techniques used in this research. Section 5 concludes the paper with future scope.

2. Related Work

Various studies have been carried out on the detection and classification of disfluencies in speech. The general framework (Fig. 1) for speech disfluency classification is to first train a set of machine learning and/or deep learning classifiers on a set of features obtained from a given dataset. Numerous speech corpora are available for research purposes having the recorded speech of volunteers. The most widely used corpus for disfluency recognition is University College London Archive of Stuttered Speech (UCLASS) (Howell et al., 2009) database. Most of the studies presented here have incorporated speech samples from UCLASS. UCLASS is a stuttered speech database developed by speech group in Psychology department of University College London. Two releases of UCLASS have been developed in English language. Release-I was developed in 2004 which included a total of 138 participants (120 Male & 18 Female) with age ranging between 5 years 4 months and 47 years 0 months. This release contained only monolog audio samples recordings. Release-II of UCLASS was introduced in 2008. It contained 3 types of audio samples: Monolog, Reading, Conversation. A total of 82 samples of monolog recordings are presented in release-II, out of which 76 samples are of male and 6 samples are of female participants. Similarly, 108 samples (93 male & 15 female) of reading and 128 samples (110 male & 18 female) are available in the corpus. Age of the participants in release-II is between 5 years 4 months and 20 years 7 months. Audio recordings are presented in WAV, MP3 and SFS (Speech Filing System) format.

Limited number of transcriptions are provided for the data in both release-I and release-II, which is a major drawback of the dataset as it is difficult to validate the research findings without required annotations. Another drawback of UCLASS is the mismatch between the male and female participant ratio. Sound quality of the recordings is also variable as the recordings were not achieved at the same location. This may be a challenge for ASR systems to cope up with the sound quality variability.

The general approach used by every study is to segment the speech files according to the annotations. Features are then extracted from segmented speech files. The extracted features with the corresponding labels are then fed into the specified learning models.

The first attempt in the recognition of the stuttered events was made by Howell & Sackin (Howell and Sackin, 1995). They trained Artificial Neural Networks (ANN) for the two stuttering events Repetitions and Prolongations. Autocorrelation function and envelope parameters were used as input features.

Autocorrelation functions performed slightly better than envelope parameters with a hit rate of 82% for prolongation and 77% for repetition, where envelope parameters recorded a hit rate of 79% and 71% for prolongation and repetition, respectively. Other studies were performed by different authors for recognition of disfluencies in speech using ANN (Geetha et al., 2000; Howell et al., 1997a, b; Savin et al., 2016) with highest reported accuracy of 87.39% by incorporating data from 4 speakers (Savin et al., 2016).

Until recently most researchers used ANN as a classifier for recognition of stuttering events in speech. However, recent studies have focused on other algorithms such as Hidden Markov Models (HMM) (Nöth et al., 2000; Tan et al., 2007; Wiśniewski et al., 2007a; Wiśniewski et al., 2007b), Support Vector Machines (SVM) (Arbajian et al., 2017; Ravikumar et al., 2009), k-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA) (Ai et al., 2012; Chee et al., 2009a; Fook et al., 2013; Hariharan et al., 2012; Lim et al., 2009), Multilayer Perceptron (MLP) (Ravikumar et al., 2008; Świetlicka et al., 2009; Świetlicka et al., 2013; Szczurowska et al., 2014) etc.. The highest accuracy of 90% was reported using HMM on a dataset consisting of 10 artificially generated stuttered speeches (Tan et al., 2007). An accuracy of 95% was reported using SVM and 94% using k-NN and LDA on UCLASS (Howell et al., 2009) dataset. An accuracy of 91% was reported using MLP on 8 stuttering speakers and 4 normal speakers (Świetlicka et al., 2009). However, majority of these studies have included only two types of disfluencies for

recognition i.e., Repetition and Prolongation and the training of the classification algorithms were performed on a limited number of speech samples.

More recent studies have included deep learning architectures for the recognition of the disfluent speech. Kourkounakis *et. al.* (Kourkounakis et al., 2020a) investigated the Bi-Directional LSTM with spectrograms as features to produce overall 91.15% accuracy in classifying 6 classes. Authors again extended their work by developing an architecture named FluentNet, which includes ResNET with Bi-Directional LSTM and attention layer (Kourkounakis et al., 2020b). Their work achieved 91.75% accuracy on UCLASS dataset and 86.70% accuracy on manually developed dataset LibriStutter.

Sheikh *et. al.* (Sheikh et al., 2021) explored time delay neural networks with MFCC as features on UCLASS dataset, achieving only 50.79% overall accuracy. Lea *et. al.* (Lea et al., 2021) introduced a disfluency dataset named SEP-28k generated from public audio podcasts with 5 different types of disfluencies. They explored ConvLSTM on the dataset by achieving 83.6% weighted accuracy. The authors incorporated Mel-filterbank energies, pitch, pitch delta, articulatory features and phoneme probabilities as features for the study. The trend of incorporating different types of features have also changed over the years. Initial reported studies have used features such as autocorrelation function and envelope parameters (Howell and Sackin, 1995), duration, energy peaks, spectral of word based and part word based, frequency of dysfluent portions, speaking rate etc. (Czyzewski et al., 2003; Geetha et al., 2000; Howell et al., 1997a, b; Nöth et al., 2000; Szczurowska et al., 2014). Recent studies have incorporated features as Mel Frequency Cepstral Coefficients (MFCC) (Chee et al., 2009a; Fook et al., 2013; Ravikumar et al., 2009; Ravikumar et al., 2008; Tan et al., 2007; Wiśniewski et al., 2007a; Wiśniewski et al., 2007b), Linear Predictive Cepstral Coefficients (LPCC) (Ai et al., 2012; Hariharan et al., 2012), Linear Predictive Coding (LPC) (Fook et al., 2013; Hariharan et al., 2012), Perceptual Linear Predictive (PLP) (Esmaili et al., 2016b; Fook et al., 2013), Spectrograms (Kourkounakis et al., 2020a; Kourkounakis et al., 2020b) etc. These features provide information about the spectra of the speech signal which is quite helpful in estimating the phonetic information of the signal. Most of the studies with machine learning techniques have used this spectral information to classify the disfluencies. The recent deep learning studies(Kourkounakis et al., 2020b) have exploited the spectral as well as temporal information of the speech signals in order to better understand the complex features.

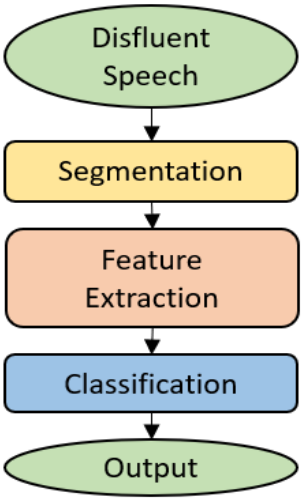

Fig. 2. General Block Diagram

Almost every deep learning architecture (Kourkounakis et al., 2020a; Kourkounakis et al., 2020b; Lea et al., 2021) has introduced LSTM to better understand the temporal structure of the signal.

Studies have also included manually generated datasets, which have been developed and utilized in different languages viz. Polish (Czyzewski et al., 2003; Świetlicka et al., 2013; Szczurowska et al., 2014), Malay (Tan et al., 2007), Mandarin (Jiang et al., 2012), Kannada (Geetha et al., 2000) etc. However, these datasets have been utilized for small studies and are not available as a public dataset to be used for research purposes. As far as the authors are aware, this study is the first with a separate class which includes a clean speech, unlike any other study.

3. Methodology

The approach followed in this study is presented in **Fig. 2** and focused on five types of speech disfluencies: Interjection, Sound Repetition, Word Repetition, Phrase Repetition, Revision and Prolongation. Speech data for the same was obtained from UCLASS release-I. A total of 25 speech samples were selected from the reading data for the experimental study according to the availability of the orthographic transcriptions. These samples contained monologues from the participants. Out of the 25 speech samples, 2 were of female speakers and 23 were of male speakers. Age of the speakers was in the range of 8 years to 18 years.

3.1. Segmentation

Speech disfluencies were recognized and then segmented based on the annotation provided by

Table 1 MFCCs 7 Class

Classifier	Number of MFCC Features	Validation Accuracy	Test Accuracy	Precision	Recall	F1 Score
Fine kNN	13 MFCCs	89.6%	91.04%	84.03	85.84	84.9
Fine kNN	21 MFCCs	93.0%	94.6%	91.03	91.44	91.24
Fine kNN	**31 MFCCs**	**94.7%**	**96.4%**	**94.12**	**94.88**	**94.5**
Fine kNN	40 MFCCs	93.3%	95.2%	91.79	91.93	91.8

Table 2 MFCCs 6 Class

Classifier	Number of MFCC Features	Validation Accuracy	Test Accuracy	Precision	Recall	F1 Score
Fine kNN	13 MFCCs	88.3%	90.23%	97.02	97.06	97.04
Fine kNN	21 MFCCs	92.1%	94.35%	98.29	98.4	98.35
Fine kNN	**31 MFCCs**	**93.9%**	**95.9%**	**98.89**	**98.70**	**98.79**
Fine kNN	40 MFCCs	92.0%	94.42%	98.39	98.20	98.29

Kourkounakis *et. al.* (Kourkounakis et al., 2020a). A total of 1272 speech samples were extracted from the 25 speech files. The extracted samples were then labelled into 7 classes (clean speech, interjections, sound repetitions, word repetitions, phrase repetitions, revisions and prolongations) according to the annotations.

3.2. Feature Extraction

The speech files had a sampling frequency of 44.1 kHz, which were down-sampled to 16 kHz. This is because the most salient features required for the processing of speech are present within 8 kHz frequency range and to fulfill Nyquist criteria the sampling frequency must be at least 16 kHz.

MFCC features were the first features extracted from the dataset. Over the past few decades, the MFCC has been widely used in the speech recognition and speaker identification (Chee et al., 2009a; Fook et al., 2013; Ravikumar et al., 2009; Ravikumar et al., 2008; Tan et al., 2007; Wiśniewski et al., 2007a; Wiśniewski et al., 2007b). The sound generated by human beings depends upon the shape of the vocal tract which includes tongue, teeth etc. If there is a way to determine the shape of vocal tract, then any produced sound can be represented correctly. Vocal tract is represented by the envelope of the time power spectrum of the speech signal and MFCC accurately represents this envelope. Mel-scale present in the MFCC extraction is a scale that relates the perceived frequency of the signal to the actual measured frequency as it can capture the minor changes in the frequency. It scales the frequency closer to a human ear can hear.

MFCC feature extraction approach generally includes pre-emphasizing the signal as a first step, which includes applying a filter to the speech signal to provide compensation for the suppressed high frequency component during speech production mechanism of human beings. The signal is then divided into frames and windowing is applied to the signals before applying FFT to it. Mel-scale filters are then applied to extract

Table 3 LPCCs 7 Class

Classifier	Number of LPCC Features	Validation Accuracy	Test Accuracy	Precision	Recall	F1 Score
Weighted kNN	13 LPCCs	80.2%	80.53%	78.24	37.22	50.44
Weighted kNN	**21 LPCCs**	**82.3%**	**83.0%**	**85.65**	**46.22**	**60.04**
Weighted kNN	31 LPCCs	82.1%	82.77%	89.38	45.69	60.47
Weighted kNN	40 LPCCs	81.4%	82.06%	88.55	42.76	57.67

Table 4 LPCCs 6 Class

Classifier	Number of LPCC Features	Validation Accuracy	Test Accuracy	Precision	Recall	F1 Score
Weighted kNN	13 LPCCs	67.7%	69.08%	86.15	90.05	88.06
Weighted kNN	**21 LPCCs**	**72.1%**	**73.43%**	**88.42**	**91.55**	**89.96**
Weighted kNN	31 LPCCs	71.3%	72.56%	87.07	91.09	89.03
Weighted kNN	40 LPCCs	69.9%	71.58%	87.12	90.22	88.64

Table 5 GFCCs 7 Class

Classifier	Number of GFCC Features	Validation Accuracy	Test Accuracy	Precision	Recall	F1 Score
Fine kNN	13 GFCCs	85.0%	85.99%	84.73	60.73	70.75
Fine kNN	21 GFCCs	88.6%	90.25%	81.58	83.27	82.41
Fine kNN	**31 GFCCs**	**90.5%**	**92.4%**	**86.15**	**86.78**	**86.4**

Table 6 GFCCs 6 Class

Classifier	Number of GFCC Features	Validation Accuracy	Test Accuracy	Precision	Recall	F1 Score
Fine kNN	13 GFCCs	81.9%	84.10%	94.94	94.89	94.92
Fine kNN	21 GFCCs	87.1%	89.25%	96.36	96.78	96.57
Fine kNN	**31 GFCCs**	**89.4%**	**91.57%**	**96.99**	**97.66**	**97.32**

frequency bands. Discrete Cosine Transform (DCT) is performed as the last step to generate MFCC features.

Linear Predictive Cepstral Coefficients (LPCC) are computationally less expensive as they do not contain steps like computing the Fourier transform of the signal in the initial stages to convert the signal from time to frequency domain. The Cepstral coefficients can be directly derived from Linear Predictive coding (LPC). The first 3 steps of extracting LPCCs namely pre-emphasis, framing and windowing are the same as MFCCs. Autocorrelation analysis performed in LPC is based on the idea that any given speech segment in the time domain can be represented as a linear combination of the previously present segments values.

Third speech extractor used is Mel-filterbank energies which are closely related to MFCC features. Steps required to generate Mel-filterbank energies are similar to MFCCs except the last step. In general, the coefficients generated after applying the Mel-scale are highly correlated and hence some machine learning algorithms tend to suffer performance loss. Hence, DCT is applied to Mel-filterbanks to generate uncorrelated MFCCs with reduction of dimensions. However, Mel-filterbank energies do give us the human perception of speech signals. Even though MFCC features provide better performance with less coefficients than Mel-Filterbank energy features, however, some studies (Miranda et al., 2019) have reported the better performance of Mel-filterbank energies due to more end to end machine and deep learning approaches. Hence, the Mel-filterbank energy features were included in the study.

Gammatone Frequency Cepstral Coefficients (GFCC) are more recent features in processing of the human speech. These are based on the filtering mechanism of human cochlear (Patterson et al., 1987).The main idea behind GFCC was to develop a model that resembled the psychophysical observations of auditory periphery (Jeevan et al., 2017). The scaling used in GFCC is equivalent rectangular bandwidth (ERB) while MFCC uses Mel scale. Motivation behind including the GFCC features is its ability to provide more noise robustness than MFCC features.(Zhao and Wang, 2013)

Table 7 Mel Filterbank 7 Class

Classifier	Number of Mel FB Features	Validation Accuracy	Test Accuracy	Precision	Recall	F1 Score
Weighted kNN	13 MFBs	79.1%	79.69%	69.30	37.06	48.29
Weighted kNN	21 MFBs	79.9%	80.7%	71.54	41.71	52.69
Weighted kNN	31 MFBs	80.1%	80.6%	76.85	37.96	50.81
Weighted kNN	**40 MFBs**	**80.1%**	**80.8%**	**76.96**	**37.43**	**50.36**

Table 8 Mel Filterbank 6 Class

Classifier	Number of Mel FB Features	Validation Accuracy	Test Accuracy	Precision	Recall	F1 Score
Weighted kNN	**13 MFBs**	**62.9%**	**64.22%**	**86.77**	**84.71**	**85.73**
Weighted kNN	21 MFBs	62.3%	63.90%	86.66	83.57	85.09
Weighted kNN	31 MFBs	62.5%	64.14%	85.37	84.23	84.80
Weighted kNN	40 MFBs	61.9%	63.60%	84.62	84.55	84.58

Table 9 Spectrogram 7 Class

Classifier	Number of Spectrogram Features	Validation Accuracy	Test Accuracy	Precision	Recall	F1 Score
Fine kNN	13 Spectrograms	75.8%	76.11%	57.72	8.31	14.53
Fine kNN	21 Spectrograms	77.9%	78.41%	75.71	22.66	34.88
Fine kNN	31 Spectrograms	79.9%	80.78%	80.88	37.23	50.98
Fine kNN	40 Spectrograms	83.4%	85.85%	73.69	76.36	75.00
Fine kNN	**257 Spectrograms**	**93.8%**	**96.35%**	**94.18**	**94.65**	**94.41**

Table 10 Spectrogram 6 Class

Classifier	Number of Spectrogram Features	Validation Accuracy	Test Accuracy	Precision	Recall	F1 Score
Fine kNN	13 Spectrograms	54.0%	54.83%	78.69	81.17	79.91
Fine kNN	21 Spectrograms	62.8%	65.14%	86.69	85.67	86.17
Fine kNN	31 Spectrograms	72.1%	74.67%	91.14	90.43	90.78
Fine kNN	40 Spectrograms	77.5%	80.63%	93.65	93.31	93.48
Fine kNN	**257 Spectrograms**	**89.6%**	**92.09%**	**97.38**	**97.45**	**97.41**

The process of GFCC features extraction includes converting the signal to frequency domain by applying Fast Fourier Transform (FFT) and multiplying it with the gammatone filterbank. The signal is then again converted back to the time domain using inverse Fourier transform. Signal is decimated to a lower frequency, to reduce the effect of noise. Then non-linear rectification process is applied to the absolute values of the decimated signal. The rectification process uses cubic root operation as opposed to a log operation used in MFCC. Then finally, discrete cosine transform is applied to obtain the GFCC features.

A Spectrogram is a visual representation of the strength of the signal. In other words, it is a representation of the amplitude of various frequencies present in a particular signal over time (Khan et al., 2021). The visual representation of spectrograms displays a brighter region where there is a heavy concentration of the sound and a dark or less bright region where there is very less or no sound.

To compute spectrograms, the audio is first split into overlapped frames or windows. Then a short time Fourier transformation is applied to each window and its absolute value is obtained. The resulting windows include the information about magnitude vs frequency. Logarithm is applied to the windows to convert to decibel scale. This provides a better image of sound structure.

Various studies (Ai et al., 2012; Fook et al., 2013) have included a variation in the number of coefficients of the feature extractors. However, the work presented in our paper goes a step further with the approach and has included 4 different number of coefficients. Authors extracted 13, 21, 31, and 40 coefficients of each feature extractor for our study to investigate the impact of number of coefficients on overall performance of the classification algorithms. The authors decided to experiment with different combinations of number of coefficients. Authors started by including 13 as the lowest number for all the coefficients and eventually increased the number of coefficients (21, 31, 40). The decision to stop the experiments for each feature was based on when the increase in coefficients did not improve the results.

3.3. Classification

The extracted features were given as input to the various machine learning classification algorithms. Algorithms used for the classification were Decision Trees, Linear and quadratic discriminant analysis, Support Vector Machines, k-Nearest Neighbor, and Ensemble Algorithms (Bagged Trees). Different variants of these algorithms were used along with a heuristic set of parameter values. In total, 15 classification algorithms were evaluated in this experimental study.

The kNN variants used were: 1) Fine kNN for which the value of k was 1 and distance was calculated using Euclidean distance. 2) Medium kNN for which the value of k was 10 and the distance was calculated using Euclidean distance. 3) Coarse kNN for which the value of k was 100 and the distance was calculated using Euclidean distance. 4) Cosine kNN the value of k was 10 and the distance was calculated using a cosine distance metric. 5) Weighted kNN for which the value of k was 10 and the distance was calculated using a distance weight.

Different variants used for decision trees were fine trees, medium trees and coarse trees for which the only difference was the maximum number of splits of the leaves, which were 100 splits for fine trees, 20 splits for

Table 11 Best results

Classifier	Number of Classes	Number of Features	Validation Accuracy	Test Accuracy
Fine kNN	**7 Class**	**31 MFCCs**	**94.7%**	**96.4%**
Fine kNN	7 Class	31 GFCCs	90.5%	92.4%
Fine kNN	7 Class	257 Spectrograms	93.8%	96.35%
Fine kNN	**6 Class**	**31 MFCCs**	**93.9%**	**95.9%**
Fine kNN	6 Class	31 GFCCs	89.4%	91.57%
Fine kNN	6 Class	257 Spectrograms	89.6%	92.09%

medium trees and 4 splits for coarse trees and the split criteria used for all three variants was Gini's diversity index.

LDA creates linear boundaries between classes whereas the quadratic discriminant creates non-linear boundaries. Hence, linear discriminant and quadratic discriminant was used for LDA and quadratic discriminant respectively. The ensemble algorithms combine different machine learning algorithms to achieve high accuracy. Algorithm used was: Bagged trees which combines random forest with decision tree learners. Method used was bootstrap aggregation and the number of learning cycles was set to 30, whereas the number of splits for decision tree used was 60000. Linear SVM, Fine gaussian SVM, Medium gaussian SVM, Coarse gaussian SVM were used as variants of the support vector machines. Kernel function used for linear SVM was linear while for the other variants gaussian kernel function was used. The Regularization parameter (C) used in all the variants was 1 whereas the gamma parameter used was 1, 0.87, 3.5 and 14 for Linear SVM, Fine gaussian SVM, Medium gaussian SVM, Coarse gaussian SVM respectively.

All the classifiers were investigated on five different feature extractors with variations in the number of features. This led to a large number of results, but in this paper only the best results for each classifier (and parameter combination) are presented.

3.4. Evaluation Metrics

Four parameters are used for the evaluation of disfluency detection; Accuracy, Precision, Recall and F1 Score. Accuracy gives the measure of correctly classified data samples over total data samples. Precision gives the proportion of positive predictions that are actually correct. Recall measures the proportion of actual positives that are predicted correctly. F1 score gives the harmonic mean between precision and recall. Better the F1 score, more balanced is the classification model.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

$$\text{F1} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision+Recall}}$$

4. Results and Discussions

Most studies included in the literature have explored only two classes, repetition and prolongation for classification. The work presented in this manuscript has included a total of 6 types of disfluencies alongside the clean speech, unlike any other previous studies. The chance accuracy with two classes would be 50% while the chance accuracy with our study involving 6 classes is 16.66%. This means our model can discover/classify more patterns from within a disfluent speech than other studies. The main reason for the inclusion of clean speech alongside other disfluencies was to increase the robustness of the classifiers. The authors feel that the patterns of disfluencies such as repetitions (character, word, phrase) are somewhat similar to the clean speech, and so it would be sensible to include clean speech alongside the 6 disfluencies. Hence, each classifier was examined on 7 classes (clean speech and 6 disfluencies) as well as 6 class (6 disfluencies only) data. The 6-class data was included only for the comparison with previous studies which also have not included clean speech.

The extracted features data was divided into a training set and a testing set in the ratio of 70% and 30% respectively. Classification algorithms were trained on the training dataset with 5-fold cross-validation. Test data was not used in the training stage.

Table 1 gives the results on 13, 21, 31 and 40 MFCC features for 7 class data. Fine kNN performed well here. A validation accuracy of 94.7% and test accuracy of 96.4% was achieved on 31 MFCC features, which was the best accuracy in this study. Even the precision, recall and F1 score is 94.12, 94.88 and 94.5 respectively, providing a very low misclassification rate. Fine kNN achieved greater than 89.6% validation accuracy and greater than 91.04% test accuracy for all number of features. Similar results were obtained with 6 class data on kNN classifier as shown in Table 2. kNN achieved 93.9% validation accuracy and 95.9% test accuracy on 31 MFCC features.

Table 12 Confusion Matrix 6 Class

Disfluency	Interjection	Sound Repetition	Word Repetition	Phrase Repetition	Revision	Prolongation
Interjection	**9773**	197	88	48	28	25
Sound Repetition	169	**14990**	170	66	49	90
Word Repetition	65	109	**3975**	27	13	11
Phrase Repetition	55	59	21	**3101**	9	5
Revision	23	55	24	7	**1678**	8
Prolongation	14	67	16	8	5	**2684**

However, the precision, recall and F1 score produced here was higher than MFCC with 7 class making it more accurate than the previous.

LPCC features performed quite poorly on the given dataset. Table 3 and Table 4 show the results of 7 class and 6 class data. Weighted kNN achieved 83% test accuracy with 7 class data on 21 features but with a less recall score. Which means that the number of False Negative (FN) values are more in the predicted values. However, with 6 class, 73.43% test accuracy was achieved but the precision, recall and F1 score was 88.42, 91.55, 89.96 respectively, which shows that misclassification was much less than 7 class data with LPCC features.

Gammatone Frequency Cepstral Coefficients also performed well with both 6 and 7 class data with test accuracy of more than 91% with fine kNN. Table 5 and Table 6 show the results. Validation accuracy of 90.5% and test accuracy of 92.4% for 7 class data and 89.4% validation accuracy and 91.57% test accuracy was achieved for 6 class data on 31 GFCC features with good precision, recall and F1 score. GFCC features were extracted with only 13, 21 and 31 features. The reason for not including 40 features was that the maximum number of permitted features were 32 because of the low sample rate of 16 kHz. This limits the number of features produced by GFCC.

Mel-filterbank energy features also performed poorly on both 7 and 6 class data. Test accuracy of 80.8% (on 40 features) and 64.22% (on 13 features) was achieved with 7 and 6 class data respectively as shown in Table 7 and Table 8 respectively. 7 class data performed very poorly on recall and F1 score while 6 class data performed well comparatively on all three precision, recall and F1 score.

Table 9 and Table 10 show the results of spectrogram features for 7 and 6 class data respectively. Spectrogram features gave very low results from the start but as the

Table 13 Confusion Matrix 7 Class

Disfluency	Clean Speech	Interjection	Sound Repetition	Word Repetition	Phrase Repetition	Revision	Prolongation
Clean Speech	**115353**	510	1124	305	206	127	124
Interjection	590	**9450**	73	35	27	5	7
Sound Repetition	1171	84	**14114**	58	18	26	35
Word Repetition	292	26	36	**3984**	10	10	7
Phrase Repetition	164	18	19	10	**3040**	5	2
Revision	112	9	26	11	6	**1694**	4
Prolongation	138	13	38	2	2	4	**2605**

Table 14 Comparison with existing studies

Author(s)	Dataset	Features Used	Number of classes	Classifier	Best Accuracy
Fook *et. al.*	UCLASS	LPC, PLP, MFCC	2 (Repetition and Prolongation)	kNN, LDA, SVM	95.7%
Izabella *et. al.*	Manual	--	2 (Fluent vs Disfluent)	MLP, RBF	92%
Chee *et. al.*	UCLASS	MFCC	2 (Repetition and Prolongation)	kNN, LDA	90%
Chee *et. al.*	UCLASS	LPCC	2 (Repetition and Prolongation)	kNN, LDA	89.77%
Chia *et. al.*	UCLASS	MFCC, LPCC	2 (Repetition and Prolongation)	kNN, LDA	94.51%
Kourkounakis *et. al.*	UCLASS	Spectrogram	6	ResNet	91.15%
Kourkounakis *et. al.*	UCLASS	Spectrogram	6	FluentNet	91.75%
Proposed approach	**UCLASS**	**MFCC, LPCC, GFCC, Mel filterbank Energy, Spectrogram**	**6 and 7**	**kNN, LDA, SVM, Decision Trees, Bagged Trees**	**96.4% for 7 Class with Fine kNN and MFCC, 95.9% for 6 class with Fine kNN and MFCC**

number of features started increasing, the results started improving. Spectrogram performed lowest on 13 features with 76.11% and 54.83% test accuracy for both 7 and 6 class with very low precision, recall and F1 score for 6 class. But with each increment in number of features the results improved. So, the number of features were increased to a higher value (257) and as expected 96.35% test accuracy was achieved for 7 class data and 92.09% test accuracy was achieved for 6 class data. The achieved accuracy maybe more for 7 class data but 6 class data produced more in terms of precision (97.38), recall (97.45) and F1 (97.41) score.

The kNN classifier have produced the best results for each feature extractor. However, other classifiers were also investigated for the given problem. But every classifier other than kNN gave poor results. For example, for MFCC 7 class data 13 features, Decision Trees, LDA, Bagged Trees and SVM produced 76.27%, 75.76%,82.83% and 75.80% test accuracy. But the important aspect of it was a very high misclassification rate. The values of precision, recall and F1 score were very less and mostly algorithms tend to have skewed results by classifying disfluent features as clean speech features. Similarly, for every other feature extractor with 7 and 6 class data, the results produced were very poor. For the sake of simplicity, the results produced by other classifiers have not been included here.

Table 11 presents the best results from all the feature extractors with 7 and 6 class data. MFCC features proved to be the best features for disfluent speech detection. With both 7 class and 6 class data, MFCC achieved the highest accuracy to detect disfluencies. MFCCs features accurately represent the envelope of the time power spectrum of a speech signal, correctly predicting the shape of vocal tract that produces the sound.

The number features required for the best accuracy were 31. Experimenting with 40 features resulted in comparatively less accuracy, precision, recall and F1 score.

For LPCC features, the best accuracy for both 7 and 6 class data was achieved at 21 features. However, for 7 class the misclassification was comparatively higher. Even though the statistical measures (precision, recall and F1 score) improved with 6 class data, the overall accuracy was not comparable to MFCCs.

GFCC features showed positive results towards the given problem with accuracy and statistical measures comparable to MFCC features. However, in the end MFFC features still had an upper hand. Like MFCC, the best results of GTCC also came when number of features were 31 for both 7 and 6 class data. Mel-filterbank energy features produced the poorest results of all 5 feature extractors. For 7 class data, Mel-filterbank features produced highest accuracy at 40 features and for 6 class data, these features produced highest accuracy at 13 features.

Spectrogram features started with very poor results with 13 features. However, with the increase in the number of features the results started gradually improving. The best results were achieved at 257 features for both 7 and 6 class data. The results produced by spectrogram features were comparable to MFCC features.

Out of 5 feature extractors only 3 performed well with an accuracy of more than 90%. However, both MFCC and GFCC utilized 31 features for the best results, still the overall accuracy of MFCC was better. Spectrogram

features produced an accuracy comparable to MFCC features for 7 class data. But utilized a greater number of coefficients to produce the same accuracy. Hence, spectrogram features were computationally expensive. As a result, the experimentation was stopped at 257 features. However, based on the previous results of spectrogram the results may still improve if one chooses to move beyond that.

Precision determines how many positive samples are being predicted than the negative ones whereas the Recall determines how many positive samples are being correctly predicted out of total positive samples. The high precision values in the results clearly shows that the number of false positives (FP) are being predicted less by the model. Similarly, the higher Recall value of the results gives an insight about the model being able to predict less false negatives (FN). F1 score is a harmonic mean of Precision and Recall. Higher the F1 score, higher is the correct predictive performance of the model.

Table 12 and Table 13 displays the confusion matrix of best-case results in both 6 class and 7 class scenarios respectively. The matrix also shows the correlation between the different disfluencies. For example, in confusion matrix 6 class, for the interjection, the highest number of similar samples are of sound repetition, showing a certain level of similarity between these two classes. Similar conclusions can be drawn for other disfluencies too. However, in confusion matrix 7 class table, for all 6 disfluencies the highest number of similar samples are with clean speech (class 7), which justifies the original claim of including clean speech to make the classifier more.

Table 14 gives a comparison of the results produced by other studies vs the study presented in this paper. As discussed, most studies included 2 types of disfluencies whereas the study presented here included 6 types of disfluencies and also clean speech. This makes our study more robust than the previous studies. Also, the two studies which included 6 classes have significantly lower accuracy than the models presented in this work.

The main reason behind the MFCC features performing better may be that MFCC can correctly identify the sound produced by the human vocal tract and can provide with the correct shape of the power spectrum of sound (spectral envelope) which resembles the sound produced by a human (Deshwal et al., 2019).

5. Conclusion and future scope

This study presented different algorithms and feature extractors for the detection and classification of disfluencies present in speech signals. Out of all the algorithms, kNN outperformed every other algorithm single handedly. The similarity-based prediction capability of the kNN gave it an edge over other algorithms, as this algorithm can find similarities for new data in any range between 1 to 100 neighbors.

MFCC features too produced the best results, mainly because MFCCs describe the overall shape of spectral envelope which correctly resembles the sound produced by human vocal tract. MFCC features not only outperformed LPCC, GFCC, Mel-Filterbank energies, but also were computationally less expensive than spectrogram features as MFCC features took approximately 3.6 hours to train on 7 class data to produce best results whereas spectrogram features took approximately 33 hours for the same.

As the real time speech signals may include multiple disfluencies simultaneously in a speech sample, the future scope in this field is to develop an automated segmentation technique for the disfluent speech in real time. Data from different speech dialects should be incorporated in this study so that a robust learning algorithm can be investigated for the classification of disfluencies. An approach also needs to be developed for automatic correction of the disfluent speech signals. This will help to improve the flow of speech signal as well as have a meaningful impact on the speech understandability.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge director CSIR-CSIO, India for providing necessary infrastructure and motivation for the work. The authors are thankful to Dr. Tarandeep Singh, PhD scholar at AcSIR, for providing insights while working on machine learning techniques. The authors are also thankful to Dr. Peter Howell from Division of Psychology and Language Sciences for his support for this work in providing necessary material for speech annotations. The authors are also thankful to Tedd Kourkounakis and Dr. Ali Etemad from Electrical and Computer Engineering Department of Queens University Canada for providing the annotation to selected files of UCLASS dataset.

References

Ai, O.C., Hariharan, M., Yaacob, S., Chee, L.S.J.E.S.w.A., 2012. Classification of speech dysfluencies with MFCC and LPCC features. 39, 2157-2165.

Arbajian, P., Hajja, A., Raś, Z.W., Wieczorkowska, A.A., 2017. Segment-Removal based stuttered speech remediation,

International workshop on new frontiers in mining complex patterns. Springer, pp. 16-34.
Awad, S.S., 1997. The application of digital speech processing to stuttering therapy, IEEE Instrumentation and Measurement Technology Conference Sensing, Processing, Networking. IMTC Proceedings. IEEE, pp. 1361-1367.
Bloodstein, O., 1969. A handbook on stuttering.
Chee, L.S., Ai, O.C., Hariharan, M., Yaacob, S., 2009a. MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA, 2009 IEEE Student Conference on Research and Development (SCOReD). IEEE, pp. 146-149.
Chee, L.S., Ai, O.C., Yaacob, S., 2009b. Overview of automatic stuttering recognition system, Proc. International Conference on Man-Machine Systems, no. October, Batu Ferringhi, Penang Malaysia, pp. 1-6.
Conture, E.G., 2001. Stuttering: Its nature, diagnosis, and treatment. Allyn & Bacon.
Curlee, R.F., 1981. Observer agreement on disfluency and stuttering. Journal of Speech, Language, and Hearing Research 24, 595-600.
Czyzewski, A., Kaczmarek, A., Kostek, B.J.J.o.I.I.S., 2003. Intelligent processing of stuttered speech. 21, 143-171.
Deshwal, D., Sangwan, P., Kumar, D.J.W.P.C., 2019. Feature extraction methods in language identification: a survey. 107, 2071-2103.
Drayna, D., Kang, C.J.J.o.n.d., 2011. Genetic approaches to understanding the causes of stuttering. 3, 374-380.
Esmaili, I., Dabanloo, N.J., Vali, M., 2016a. Automatic classification of speech dysfluencies in continuous speech based on similarity measures and morphological image processing tools. Biomedical Signal Processing and Control 23, 104-114.
Esmaili, I., Dabanloo, N.J., Vali, M.J.B.S.P., Control, 2016b. Automatic classification of speech dysfluencies in continuous speech based on similarity measures and morphological image processing tools. 23, 104-114.
Fook, C.Y., Muthusamy, H., Chee, L.S., Yaacob, S.B., Adom, A.H.B.J.T.j.o.e.e., sciences, c., 2013. Comparison of speech parameterization techniques for the classification of speech disfluencies. 21, 1983-1994.
Geetha, Y., Pratibha, K., Ashok, R., Ravindra, S.K.J.J.o.f.d., 2000. Classification of childhood disfluencies using neural networks. 25, 99-117.
Hariharan, M., Chee, L.S., Ai, O.C., Yaacob, S.J.J.o.m.s., 2012. Classification of speech dysfluencies using LPC based parameterization techniques. 36, 1821-1830.
Howell, P., Davis, S., Bartrip, J.J.J.o.S., Language,, Research, H., 2009. The university college london archive of stuttered speech (uclass).
Howell, P., Sackin, S., 1995. Automatic recognition of repetitions and prolongations in stuttered speech, Proceedings of the first World Congress on fluency disorders. University Press Nijmegen Nijmegen, The Netherlands, pp. 372-374.
Howell, P., Sackin, S., Glenn, K.J.J.o.S., Language,, Research, H., 1997a. Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for lexical dysfluency classifiers. 40, 1073-1084.
Howell, P., Sackin, S., Glenn, K.J.J.o.S., Language,, Research, H., 1997b. Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: II. ANN recognition of repetitions and prolongations with supplied word segment markers. 40, 1085-1096.
Jeevan, M., Dhingra, A., Hanmandlu, M., Panigrahi, B., 2017. Robust speaker verification using GFCC based i-vectors, Proceedings of the International Conference on Signal, Networks, Computing, and Systems. Springer, pp. 85-91.
Jiang, J., Lu, C., Peng, D., Zhu, C., Howell, P.J.P.o., 2012. Classification of types of stuttering symptoms based on brain activity. 7, e39747.
Johnson, W., 1961. Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers. Journal of Speech & Hearing Disorders. Monograph Supplement.
Johnson, W.J.J.o.S., Supplement, H.D.M., 1961. Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers.
Khan, A.S., Ahmad, Z., Abdullah, J., Ahmad, F.J.I.A., 2021. A spectrogram image-based network anomaly detection system using deep convolutional neural network. 9, 87079-87093.
Kourkounakis, T., Hajavi, A., Etemad, A., 2020a. Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory, ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6089-6093.
Kourkounakis, T., Hajavi, A., Etemad, A.J.a.p.a., 2020b. FluentNet: end-to-end detection of speech disfluency with deep learning.
Lea, C., Mitra, V., Joshi, A., Kajarekar, S., Bigham, J.P., 2021. Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter, ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6798-6802.
Lim, S.C., Ooi, C.A., Hariharan, M., Sazali, Y., 2009. MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA.
Miranda, I.D., Diacon, A.H., Niesler, T.R., 2019. A comparative study of features for acoustic cough detection using deep architectures, 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 2601-2605.
Mullin, E., 2016. Why Siri Won't Listen to Millions of People with Disabilities. Scientific American. Retrieved January 8, 2018.
Nöth, E., Niemann, H., Haderlein, T., Decher, M., Eysholdt, U., Rosanowski, F., Wittenberg, T., 2000. Automatic stuttering recognition using hidden Markov models, Sixth International Conference on Spoken Language Processing.
Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1987. An efficient auditory filterbank based on the gammatone function, a meeting of the IOC Speech Group on Auditory Modelling at RSRE.
Ravikumar, K., Rajagopal, R., Nagaraj, H., 2009. An approach for objective assessment of stuttered speech using MFCC, The international congress for global science and technology, p. 19.
Ravikumar, K., Reddy, B., Rajagopal, R., Nagaraj, H.J.P.o.w.a.s., engineering, technology, 2008. Automatic

detection of syllable repetition in read speech for objective assessment of stuttered disfluencies. 36, 270-273.
Savin, P., Ramteke, P.B., Koolagudi, S.G., 2016. Recognition of repetition and prolongation in stuttered speech using ANN, Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics. Springer, pp. 65-71.
Sheikh, S.A., Sahidullah, M., Hirsch, F., Ouni, S., 2021. StutterNet: Stuttering Detection Using Time Delay Neural Network, 2021 29th European Signal Processing Conference (EUSIPCO). IEEE, pp. 426-430.
Smith, A., Weber, C., 2016. Childhood Stuttering–Where are we and Where are we going?, Seminars in Speech and Language. NIH Public Access, p. 291.
Stein-Rubin, C., Fabus, R., 2011. A guide to clinical assessment and professional report writing in speech-language pathology. Nelson Education.
Świetlicka, I., Kuniszyk-Jóźkowiak, W., Smołka, E., 2009. Artificial neural networks in the disabled speech analysis, Computer Recognition Systems 3. Springer, pp. 347-354.
Świetlicka, I., Kuniszyk-Jóźkowiak, W., Smołka, E.J.C.S., Language, 2013. Hierarchical ANN system for stuttering identification. 27, 228-242.
Szczurowska, I., Kuniszyk-Jóźkowiak, W., Smołka, E.J.A.o.A., 2014. The application of Kohonen and Multilayer Perceptron Networks in the speech nonfluency analysis. 31, 205-210.
Tan, T.-S., Ariff, A., Ting, C.-M., Salleh, S.-H., 2007. Application of Malay speech technology in Malay speech therapy assistance tools, 2007 International Conference on Intelligent and Advanced Systems. IEEE, pp. 330-334.
Watkins, K.E., Smith, S.M., Davis, S., Howell, P.J.B., 2008. Structural and functional abnormalities of the motor system in developmental stuttering. 131, 50-59.
Weber-Fox, C., Wray, A.H., Arnold, H.J.J.o.f.d., 2013. Early childhood stuttering and electrophysiological indices of language processing. 38, 206-221.
Wingate, M.E., 2001. SLD is not stuttering. Journal of Speech, Language, and Hearing Research.
Wingate, M.E., 2002. Foundations of stuttering. Acoustical Society of America.
Wiśniewski, M., Kuniszyk-Jóźkowiak, W., Smołka, E., Suszyński, W., 2007a. Automatic detection of disorders in a continuous speech with the hidden Markov models approach, Computer Recognition Systems 2. Springer, pp. 445-453.
Wiśniewski, M., Kuniszyk-Jóźkowiak, W., Smołka, E., Suszyński, W.J.J.o.M.I., Technologies, 2007b. Automatic detection of prolonged fricative phonemes with the hidden Markov models approach. 11.
Yairi, E., Ambrose, N.G., 1999. Early childhood stuttering I: Persistency and recovery rates. Journal of Speech, Language, and Hearing Research 42, 1097-1112.
Yaruss, J.S., 1997. Clinical measurement of stuttering behaviors. Contemporary Issues in Communication Science and Disorders 24, 27-38.
Young, M.A., 1975. Observer agreement for marking moments of stuttering. Journal of Speech and Hearing Research 18, 530-540.
Zhao, X., Wang, D., 2013. Analyzing noise robustness of MFCC and GFCC features in speaker identification, 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, pp. 7204-7208.