Turkish Journal of Electrical Engineering and Computer Sciences

Volume 21
Number 7 *Sup. 1*

Article 13

1-1-2013

Comparison of speech parameterization techniques for the classification of speech disfluencies

CHONG YEN FOOK

HARIHARAN MUTHUSAMY

LIM SIN CHEE

SAZALI BIN YAACOB

ABDUL HAMID BIN ADOM

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2013) 21: 1983 – 1994

doi:10.3906/elk-1112-84

Comparison of speech parameterization techniques for the classification of speech disfluencies

Chong Yen FOOK,* Hariharan MUTHUSAMY, Lim Sin CHEE
Sazali Bin YAACOB, Abdul Hamid Bin ADOM
School of Mechatronic Engineering, University Malaysia Perlis, Campus Pauh Putra, Arau, Perlis, Malaysia

Received: 29.12.2011 • **Accepted:** 05.06.2012 • **Published Online:** 24.10.2013 • **Printed:** 18.11.2013

Abstract: Stuttering assessment through the manual classification of speech disfluencies is subjective, inconsistent, time-consuming, and prone to error. The aim of this paper is to compare the effectiveness of the 3 speech feature extraction methods, mel-frequency cepstral coefficients, linear predictive coding (LPC)-based cepstral parameters, and perceptual linear predictive (PLP) analysis, for classifying 2 types of speech disfluencies, repetition and prolongation, from recorded disfluent speech samples. Three different classifiers, the k-nearest neighbor classifier, linear discriminant analysis-based classifier, and support vector machine, are employed for the classification of speech disfluencies. Speech samples are taken from the University College London Archive of Stuttered Speech and stuttered events are identified through manual segmentation. A 10-fold cross-validation method is used for testing the reliability of the classifier results. The effect of the 2 parameters (LPC order and frame length) in the LPC- and PLP-based methods on the classification results is also investigated. The experimental results reveal that the proposed method can be used to help speech language pathologists in classifying speech disfluencies.

Key words: Disfluent speech, mel-frequency cepstral coefficient, linear predictive coding, perceptual linear predictive analysis, support vector machine

1. Introduction

Humans use speech as a verbal means to express their feelings, ideas, and thoughts in communication. In this world, there is 1% of the population having the problem of speech disfluency, and it has been found to affect females and males at a ratio of 1:3 or 1:4 [1,2]. Disfluency and stuttering are a break or interruption of normal speech, such as repetition, prolongation, or interjection of syllables, sounds, words, or phrases, and involuntary silent pauses or blocks in communication [1,3]. Stuttering cannot be completely cured, although it may go into remission for some time [1]. Stutterers can learn to shape their speech into fluent speech with the appropriate speech pathology treatments. Therefore, a stuttering assessment is needed to evaluate the performance of stutterers before and after therapy. Traditionally, a speech language pathologist (SLP) counts and classifies the occurrence of disfluencies, such as repetition and prolongation, in stuttered speech manually. However, these types of stuttering assessment are subjective, inconsistent, time-consuming, and prone to error [1,4–8]. Therefore, it might be good if stuttering assessment can be done through classification of disfluencies using digital signal processing (DSP) and artificial intelligence (AI) concepts. In the last 2 decades, researchers have focused on developing objective methods using DSP and AI concepts to assist the SLP during stuttering

*Correspondence: fook1987@gmail.com

assessment. Table 1 depicts some of the significant research works on the automatic classification of speech disfluencies and/or stuttering recognition.

Table 1. Summary of several research works on stuttering recognition.

First author	Database	Features	Classifiers	Best results (%)
Howell [3]	-	Autocorrelation function and envelope parameters	Artificial neural networks (ANNs)	Approximately 80
Howell [4–5]	12 speakers (UCLASS)	Duration, energy peaks, spectral of word based and part word based	ANNs	78.01
Nöth [6]	37 speakers	Duration and frequency of disfluent portions, speaking rate	Hidden Markov models (HMMs)	-
Geetha [8]	51 speakers	Age, sex, type of disfluency, frequency of disfluency, duration, physical concomitant, rate of speech; historical, attitudinal, and behavioral scores; family history	ANNs	92
Czyzewski [10]	6 normal speech samples + 6 stop-gap speech samples	Frequency, 1st to 3rd formant's frequencies and its amplitude	ANNs and rough set	73.25 and ≥90.0
Prakash [11]	10 normal + 10 stuttering children	Formant patterns, speed of transitions, F2 transition duration, and F2 transition range	-	-
Szczurowska [12]	8 speakers	Spectral measure [fast Fourier transform (FFT) 512]	Multilayer perceptron (MLP), Kohonen	76.67
Wiśniewski [13]	38 samples for prolongation of fricatives + 30 samples for stop blockade + 30 free-of-silence samples	MFCCs	HMMs	70
Wiśniewski [14]	-	MFCCs	HMMs	Approximately 80
Tian-Swee [15]	15 normal speakers + 10 artificial stuttered speech	MFCCs	HMMs	96
Ravikumar [16]	10 speakers	MFCCs	Perceptron	83
Świetlicka [17]	8 stuttering speakers + 4 normal speakers (yields 59 fluent speech samples + 59 nonfluent speech samples)	Spectral measure (FFT 512)	Kohonen, MLP, radial basis function (RBF)	88.1–94.9
Ravikumar [18]	15 speakers	MFCCs	SVM	94.35
Sin Chee [19]	10 speakers	MFCCs	kNN, LDA	90.91
Sin Chee [20]	10 speakers	Linear prediction cepstral coefficients (LPCC)	kNN, LDA	89.77
M. Hariharan [21]	10 speakers	LPC-based cepstral parameters	kNN, LDA	94 and above
Chia Ai [22]	10 speakers	MFCC and LPCC coefficients	kNN and LDA	92.55 and 94.51

From a literature review, it was observed that different feature extraction and classification algorithms have been proposed. In this work, the comparison of 3 speech feature extraction methods is presented for classifying the 2 types of speech disfluencies (repetition and prolongation), using mel frequency cepstral coefficients (MFCCs) and linear predictive coding (LPC)- and perceptual linear predictive (PLP)-based methods. In order to classify the 2 types of disfluencies, the MFCCs and the LPC- and PLP-based cepstral coefficients are extracted to characterize the disfluent speech. Three different classifiers, the k-nearest neighbor (kNN) classifier, linear discriminant analysis (LDA)-based classifier, and support vector machine (SVM), are employed for the classification of the speech disfluencies. A 10-fold cross-validation is applied to assess the reliability of the classifiers results.

The paper is organized as follows. The methodology of the system and database used in the experiment are presented in Section 2. The acoustic feature extraction is presented in Section 3. The fundamentals of the classification algorithms are described in Section 4. The experimental results using 3 acoustic speech features are reported in Section 5. Finally, the conclusions and future work are given in Section 6.

2. Methodology

In this paper, the classification of speech disfluencies consists of 2 important stages, as illustrated in Figure 1: the acoustic feature extraction and classification. MFCCs and LPC- and PLP-based cepstral coefficients are used to characterize the disfluent speech. Three classifiers (kNN, LDA, and SVM) are employed to evaluate the effectiveness of the acoustic features for the classification of 2 types of speech disfluencies (repetition and prolongation). The effectiveness levels of the features are compared to each other. However, it is not easy for us to compare our results with previous works since other researchers used different databases. In this work, stuttered speech samples are taken from the University College London Archive of Stuttered Speech (UCLASS) website [9]. Table 1 tabulates the summary of several research works on stuttering recognition.

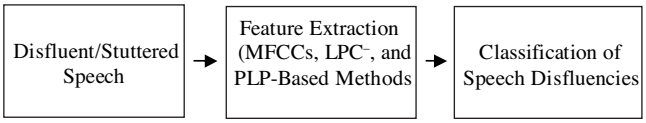

Figure 1. Block diagram of the feature extraction and classification phase.

2.1. Database

The database was obtained from the UCLASS archive [9]. It consists of 3 types of recordings: monologs, readings, and conversation. There are 43 different speakers contributing 107 'reading' recordings. Table 2 shows the distribution of the 'reading' recording database. In this work, only a subset of the available sample, which has 39 samples of speech, is taken from the UCLASS archive [23] for analysis. It includes 2 female speakers and 37 male speakers, whose ages range from 11 years and 2 months to 20 years and 1 month. The samples are chosen to cover a broad range of both age and stuttering rate. Most of the 'reading' samples do not have a text script, and hence only the 'reading' samples with text scripts are chosen for our investigation.

In this study, only 2 types of lexical disfluencies, namely prolongation and repetition, are investigated. Both types of disfluencies can be detected easily in monosyllabic words. Thus, only the speech samples with the content of 'One more week to Easter' and 'Arthur the rat' are selected due to the fact that 90% of its content is monosyllable words [4,5]. Each of the 2 passages consists of more than 300 words. Through listening, the 2

types of disfluencies are identified and segmented manually. A total of 77 speech samples of prolongation and 94 speech samples of repetition are obtained.

Table 2. Age and sex distribution of the reading recordings in the chosen subset of the UCLASS databases.

	Age range	Sex	
		Male	Female
Recording	7 years and 10 months to 20 years and 7 months	92	15
Speaker		38	5

3. Acoustic feature extraction

Acoustic feature extraction plays an important role in speech processing. The LPC-based cepstral parameters. MFCCs, and PLP-based cepstral coefficients are extracted to characterize disfluent speech.

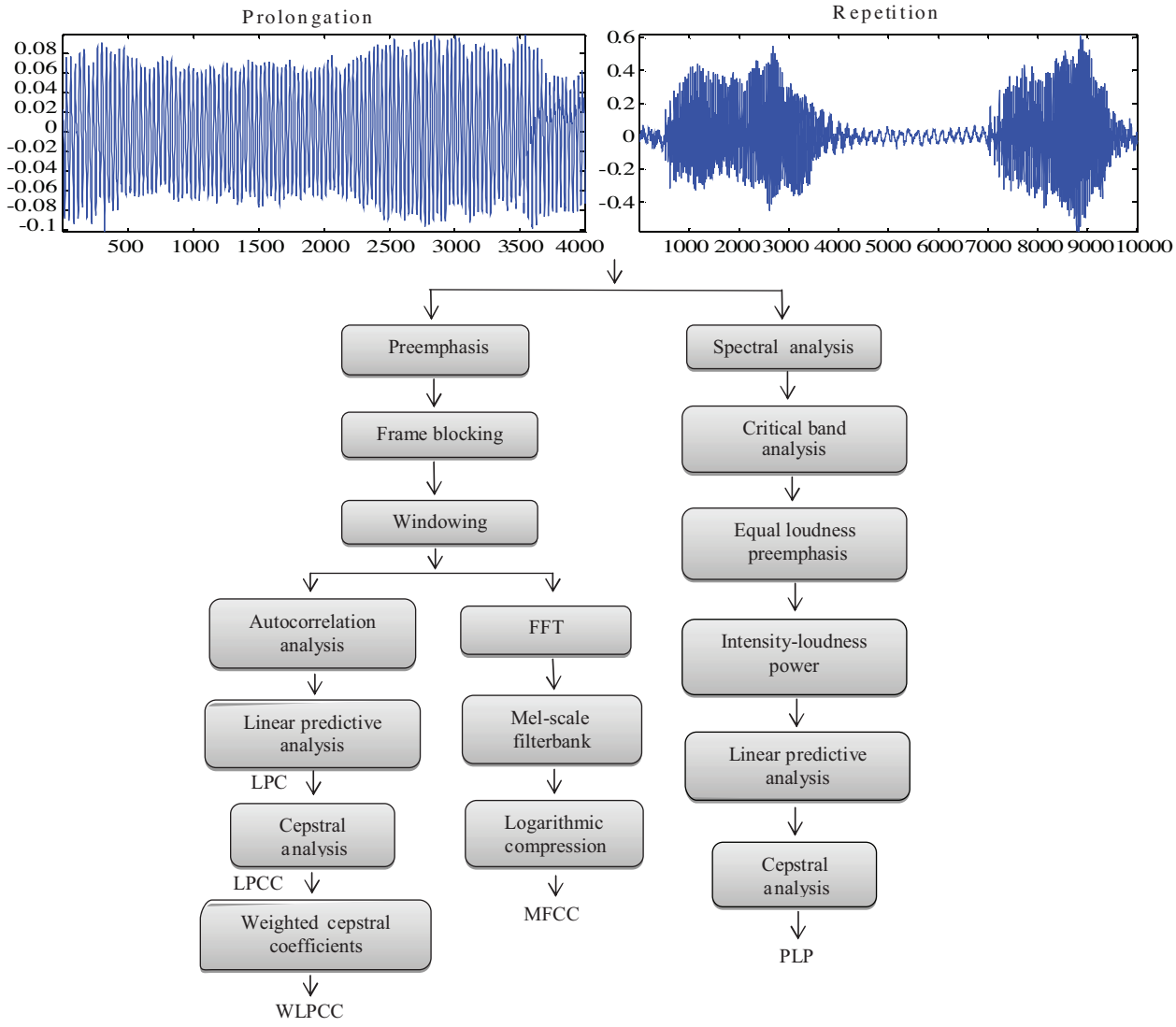

Figure 2. Block diagram of the LPC-based cepstral parameter, MFCCs, and PLP extraction.

Figure 2 illustrates the extraction of the LPC-based parameters (LPC, LPCC, and WLPCC), MFCCs, and PLP from a speech signal (prolongation/repetition). This section briefly explains the background theory of the derivation of the LPC-based cepstral parameters, MFCCs, and PLP-based cepstral features.

3.1. Speech signal preprocessing

The original sampling frequency of the speech samples is 44.1 kHz. For speech processing purposes, each of the speech samples is down-sampled to 16 kHz. This is reasonable for the speech processing task in the present work, because most of the salient speech features are within an 8-kHz bandwidth [24]. Before the stage feature extraction, the speech samples are preemphasized to spectrally flatten the signal, to even the spectral energy envelope by amplifying the importance of the high-frequency components, and for removing the DC component in the signal. First, the disfluent speech samples are passed through a first-order preemphasis filter and its transfer function:

$$H(z) = 1 - \tilde{a} * z^{-1} \quad 0.9 \leq \tilde{a} \leq 1.0, \tag{1}$$

where $\tilde{a}$ is a positive parameter used to control the degree of preemphasis filtering. Normally, the $\tilde{a}$ value is chosen at between 0.9 and 1.0. The commonly used $\tilde{a}$ value is $15/16 = 0.9375$ or 0.95 [25]. In this work, the value of $\tilde{a}$ is set as equal to 0.9375. Preemphasized disfluent speech signals are segmented into frames of N samples with an overlap of $(1/3) \times$ N. Each frame is multiplied with a Hamming window to minimize the signal discontinuities. The frame length is varied from 10–50 ms [21]. The acoustic features are extracted from each frame and used for the classification.

3.2. LPC analysis

In linear predictive analysis, each sample is estimated as a linear combination of the past p samples, where p represents the order of prediction [25]. If $s(n)$ is the present sample, then it is estimated by the past p samples as:

$$\hat{s}(n) = \sum_{m=1}^{p} a_m s(n-m). \tag{2}$$

The prediction error, $e(n)$ is the difference between the actual and the estimated sample value, defined as:

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{m=1}^{p} a_m s(n-m), \tag{3}$$

where a_m are the linear prediction coefficients. Each of the windowed signals is autocorrelated according to the following formula:

$$r(m) = \sum_{n=0}^{N-1-m} x(n)x(n+m), m = 0, 1, \ldots, p, \tag{4}$$

where p is the order of the LPC analysis and p is fixed as 2, 8, 10, and 14 [21,22,26]. To convert the autocorrelation coefficients into LPC coefficients, the LPC analysis is performed and implemented using the Durbin–Levinson recursive algorithm. The final solution for the LPC coefficients is given by Eq. 5.

$$a_j = a_j^p 1 \leq j \leq p \tag{5}$$

These LPC coefficients are then converted to cepstral coefficients using the following recursive method:

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, 1 \leq m \leq p, \tag{6}$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} m > p, \tag{7}$$

where m is the order of the cepstral coefficients. Finally, the LPCC features are obtained. However, the low-order cepstral coefficients are sensitive to the overall spectral slope and the high-order cepstral coefficients are sensitive to noise and other forms of noise like variability [25]. Hence, a standard technique is to weigh the cepstral coefficients by a tapered window so as to minimize these sensitivities. The appropriate weighing is the band pass filter, which is given as follows:

$$w_m = \left[1 + \frac{p}{2} \sin\left(\frac{\pi m}{p}\right)\right], \quad 1 \leq m \leq p. \tag{8}$$

The weighted cepstral coefficients are given by:

$$\hat{c}_m = w_m c_m, \quad 1 \leq m \leq p. \tag{9}$$

The WLPCC features are obtained from every frame of a signal using Eq. (9). In the literature, the WLPCCs have been used as a feature extraction method in other applications [26–27]. In this study, the LPC-based cepstral coefficients are extracted to discriminate between the 2 types of disfluencies (repetition and prolongation). The effects of the different LPC orders and different frame lengths on the classification accuracy are also investigated.

3.3. Mel-frequency cepstral coefficients

MFCCs have been widely used as a feature extraction method for both speech and speaker recognition systems. In recent years, MFCCs also have been proven to be one of the successful feature extraction methods in speech disfluency classification [12,15,16,18,19,22]. According to psychophysical studies, human perception of the frequency contents of sound for speech signals follow a subjectively defined nonlinear scale. The frequency scale-warping to the mel scale has led to the cepstrum domain representation. A block diagram of the MFCC feature extraction method is illustrated in Figure 2.

In this work, the MFCCs are extracted as baseline features for comparison with the LPC-based cepstral coefficients. The MFCCs are computed by applying the fast Fourier transform (FFT) on the windowed signal. The spectrum of each frame is filtered by a triangular bandpass filter, and its center frequencies and bandwidth approximately match the auditory critical band filters, known as the mel scale filter. The mel frequency scale has a linear frequency spacing below 1 kHz and a logarithmic spacing above 1 kHz. The mapping of the linear frequency to the mel frequency is shown in Eq. (10).

$$Mel(f) = 2595 Log_{10}\left(1 + \frac{f}{700}\right) \tag{10}$$

The logarithm is computed on the filter bank output. Finally, a discrete cosine transform is employed to generate the cepstral coefficients. In this study, 25 MFCCs are extracted and used to represent a disfluent speech.

3.4. Perceptual linear predictive analysis

PLP is a combination of spectral analysis and linear prediction analysis. The PLP technique uses concepts from the psychophysics of hearing [28]. In PLP analysis, windowed speech samples are converted into the frequency domain using Fourier transform to compute the short-term power spectrum. This spectrum is subjected to critical band analysis, where a filterbank is designed based on the Bark scale and is preemphasized by a function that approximates the sensitivity of human hearing at different frequencies. The output is compressed through intensity to loudness conversion to approximate the nonlinear relationship between the intensity of a sound and its perceived loudness. The critical band spectrum is converted into a small number of linear predictive coefficients by applying an inverse discrete Fourier transform and finally the linear predictive parameters are usually transformed to cepstral coefficients. The block diagram of the PLP feature extraction is presented in Figure 2.

4. Classification

Three different classifiers, the kNN classifier, LDA-based classifier, and SVM, are employed for the classification of the speech disfluencies. This section briefly describes the fundamentals of the kNN, LDA, and SVM classifiers. A 10-fold cross-validation is used for testing the reliability of the classifiers' results.

4.1. k-Nearest neighbors classifier

In pattern recognition, the kNN algorithm is a method for classifying objects based on the closest training examples in the feature space [29,30]. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer). In the kNN algorithm, the classification of a new test feature vector is determined by the class of its kNNs. Here, the kNN algorithm is implemented using Euclidean distance metrics to locate the nearest neighbors. The number of neighbors (i.e. k) used to classify the new test vector is varied in the range of 1, 2, . . . , 10, and its effects on classification performance are presented in the form of classification accuracy with standard deviation.

4.2. Linear discriminant classifier

Discriminant analysis is a statistical technique to classify objects into mutually exclusive and exhaustive groups based on a set of measurable object's features. Linear discriminants (LDs) [29,30] partition the feature space into the different classes using a set of hyperplanes. The parameters of this classifier model are fitted to the available training data using the method of maximum likelihood. This model assumes that the feature data has a Gaussian distribution for each class. In response to the input features, the LDs provide a probability estimate of each class. The final classification is obtained by choosing the class with the highest probability estimate. Using this method, the processing required for training is achieved by direct calculation and is extremely fast compared to other classifier models. The LDA-based classifier is designed with a 'linear' discriminant function.

4.3. Support vector machine

In this work, SVM is used as a classifier and it is a promising method for solving nonlinear classification problems, function and density estimation, and pattern recognition tasks [7,31]. It was originally proposed to classify samples within 2 classes. It maps the training samples of 2 classes into a higher dimensional space through a kernel function. SVM seeks an optimal separating hyperplane in this new space to maximize its

distance from the closest training point. While testing, a query point is categorized according to the distance between the point and the hyperplane.

Consider a training dataset $\{x_i, y_i\}_{i=1}^{N}$, where $x_i \in R^n$ indicates the input space of the sample and with a corresponding target output of $y_i \in R$ for $i = 1, \ldots, N$. To construct a nonlinear support vector classifier, the inner product (x, y) is replaced by a kernel function $K(x, y)$:

$$f(x) = sign\left[\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b\right]. \tag{11}$$

SVM models are built around a kernel function that transforms the input data into an n-dimensional space, where a hyperplane can be constructed to partition the data. Three kinds of kernel functions, linear kernel, multilayer kernel, and radial basis function (RBF) kernel, are normally used by researchers [7,31]. In this work, the RBF kernel function is used since it gives excellent generalization and a low computational cost. In the RBF kernel, $\sigma 2$ (sig2) is the important parameter and it cause the changes in the shape flexion of the hyperplane.

$$K(x, x_i) = \exp\left(\frac{-||x - x_i||^2}{2\sigma^2}\right) \tag{12}$$

In this work, the LS-SVMLab toolbox [32] is used to perform the classification of the speech disfluencies. There are 2 parameters that are to be chosen optimally, the regularization parameter (γ, gam), and σ^2 (sig^2), which is the squared bandwidth of the RBF kernel to obtain better accuracy. All of the feature extraction and classifications are developed in a MATLAB environment.

5. Results and discussions

The disfluent speech signals are subjected to acoustic feature extraction using LPC- and PLP-based parameterization techniques and MFCCs. The number of PLP and LPC coefficients depends on the order of the LPC. A different number of MFCCs (13, 15, 17, and 24) is used to characterize the disfluent speech. In this work, a 10-fold cross-validation scheme is used to prove the reliability of the classification results. In the 10-fold cross-validation scheme, the proposed feature vectors are divided randomly into 10 sets and the training is repeated 10 times. Three classifiers are employed for classifying the different types of speech disfluencies. In the SVM classifier, the suitable values of the regularization parameter (γ, gam) and $\sigma 2$ (sig2) are chosen optimally as 90 and 0.9, respectively, to obtain better accuracy. In the kNN classifier, different values of 'k' between 1 and 10 are used. In the LDA-based classifier, a 'linear' discriminant function is used. The classification results of the LPC- and PLP-based cepstral coefficients and MFCCs using the kNN, LDA, and SVM classifiers are presented in Tables 3 and 4. The average and standard deviation of the classification accuracies of the different types of speech disfluencies are tabulated. The standard deviation of the classification clearly reveals the consistency of the classifier's results. If the standard deviation is higher, the classification results are inconsistent and it also reveals that the learning parameters of the classifiers affects the performance of the classifiers. From Table 3, it can be observed that the PLP features give better classification accuracy compared to the LPC-based cepstral parameters. The LPC orders are chosen (2, 8, 10, and 14) based on our previous investigation [21]. For low orders, the LPC spectrum may pick up only the prominent resonance peak and yield poor results, which can be observed from Table 3. On the other hand, when the LPC order is higher, it introduces several spurious peaks in the LPC spectrum, which is also gives poor results. Hence, the proper order for the LPC should be chosen

Table 3. Classification of the LPC- and PLP-based cepstral coefficients for different frame lengths.

LPC order	Types of features	10 ms			20 ms			30 ms			40 ms			50 ms		
		kNN	LDA	SVM	kNN	LDA	SVM	kNN	LDA	SVM	kNN	LDA	SVM	kNN	LDA	SVM
2	LPC	84.74 ± 1.42	66.61 ± 1.28	85.96 ± 0.39	84.62 ± 1.35	66.14 ± 1.12	86.20 ± 0.41	84.09 ± 1.80	65.15 ± 1.07	86.02 ± 0.43	84.27 ± 2.18	65.67 ± 0.83	85.79 ± 0.73	83.68 ± 2.13	64.74 ± 1.41	86.55 ± 0.83
	LPCC	88.48 ± 1.35	78.83 ± 0.60	90.53 ± 0.66	88.83 ± 1.52	77.89 ± 0.77	89.82 ± 0.41	89.01 ± 0.99	78.07 ± 0.57	89.94 ± 0.37	88.89 ± 1.43	78.42 ± 0.80	90.00 ± 0.43	87.66 ± 0.75	77.72 ± 1.01	90.00 ± 0.43
	WLPCC	86.14 ± 1.63	65.79 ± 1.41	85.79 ± 0.39	86.32 ± 0.63	65.73 ± 1.04	85.38 ± 0.73	86.73 ± 1.17	65.44 ± 1.01	85.44 ± 0.64	86.84 ± 1.21	65.44 ± 1.42	86.14 ± 0.48	87.49 ± 1.64	65.26 ± 0.69	86.61 ± 0.43
	PLP	94.39 ± 0.92	76.37 ± 0.92	93.39 ± 0.68	94.27 ± 0.46	76.32 ± 0.79	93.63 ± 0.43	93.92 ± 1.18	76.37 ± 0.49	93.98 ± 0.39	94.09 ± 1.12	76.20 ± 0.87	93.92 ± 0.57	94.39 ± 1.00	76.37 ± 0.79	93.51 ± 0.58
8	LPC	90.94 ± 1.89	80.29 ± 0.78	90.76 ± 0.77	89.82 ± 0.79	80.94 ± 0.74	90.82 ± 0.78	91.11 ± 0.86	80.47 ± 0.63	89.71 ± 0.92	90.12 ± 1.64	79.47 ± 0.97	90.29 ± 1.14	90.23 ± 0.73	79.12 ± 0.62	89.42 ± 0.93
	LPCC	88.60 ± 2.05	86.90 ± 0.69	93.68 ± 0.37	88.13 ± 1.61	86.84 ± 0.74	93.92 ± 0.30	87.60 ± 1.63	86.84 ± 0.96	93.57 ± 0.39	87.78 ± 1.62	86.90 ± 1.33	93.68 ± 0.46	88.01 ± 1.36	86.37 ± 0.83	93.22 ± 0.30
	WLPCC	90.12 ± 0.93	88.89 ± 0.39	93.86 ± 0.63	90.06 ± 1.23	88.07 ± 0.88	93.63 ± 0.80	89.88 ± 0.92	88.42 ± 0.95	92.87 ± 0.46	90.41 ± 1.27	88.19 ± 0.86	92.81 ± 0.39	90.06 ± 1.46	87.49 ± 1.14	92.69 ± 0.41
	PLP	92.87 ± 2.15	87.31 ± 0.48	93.63 ± 0.51	93.45 ± 2.02	87.72 ± 0.48	93.80 ± 0.41	93.45 ± 1.37	87.66 ± 0.33	93.39 ± 0.62	93.45 ± 1.45	87.02 ± 0.82	93.22 ± 0.79	93.45 ± 1.60	87.02 ± 0.66	92.98 ± 0.68
10	LPC	89.30 ± 1.79	79.12 ± 0.87	90.47 ± 0.48	87.66 ± 1.05	79.42 ± 1.16	90.76 ± 0.54	88.89 ± 1.32	80.00 ± 1.10	90.29 ± 1.18	88.89 ± 1.41	79.24 ± 1.14	90.53 ± 0.46	87.49 ± 1.11	73.74 ± 1.73	86.49 ± 1.12
	LPCC	87.78 ± 2.36	89.06 ± 1.07	92.16 ± 0.41	87.84 ± 2.68	88.60 ± 1.04	92.57 ± 0.62	88.01 ± 2.21	88.60 ± 0.84	92.87 ± 1.02	87.72 ± 2.36	89.18 ± 1.04	92.63 ± 0.96	86.84 ± 2.23	87.72 ± 0.95	92.28 ± 0.99
	WLPCC	91.29 ± 1.62	89.01 ± 0.54	93.86 ± 0.79	90.70 ± 1.25	89.18 ± 0.69	93.57 ± 0.83	90.35 ± 0.88	89.82 ± 1.00	93.39 ± 0.55	90.99 ± 1.38	89.42 ± 0.64	93.33 ± 0.79	92.40 ± 1.41	89.77 ± 0.63	96.02 ± 0.91
	PLP	91.98 ± 1.29	89.06 ± 0.78	93.98 ± 0.48	92.22 ± 2.12	89.53 ± 0.93	94.50 ± 0.57	92.28 ± 2.17	89.71 ± 1.24	94.39 ± 0.41	92.28 ± 1.63	88.30 ± 0.99	94.15 ± 0.55	92.57 ± 2.37	88.01 ± 0.92	94.33 ± 0.39
14	LPC	86.67 ± 1.43	75.09 ± 1.21	90.18 ± 0.77	83.98 ± 1.91	74.91 ± 1.52	88.48 ± 0.78	83.80 ± 1.70	74.21 ± 1.12	87.78 ± 1.33	83.86 ± 2.57	73.33 ± 1.38	87.31 ± 1.17	82.34 ± 1.95	73.39 ± 1.04	86.61 ± 1.36
	LPCC	89.65 ± 0.48	91.17 ± 1.12	95.03 ± 0.63	89.59 ± 0.86	89.82 ± 1.21	94.50 ± 0.69	89.71 ± 0.92	90.06 ± 1.29	94.62 ± 0.60	89.82 ± 1.21	89.36 ± 0.72	94.04 ± 0.77	90.23 ± 0.87	89.18 ± 0.74	93.27 ± 0.92
	WLPCC	89.59 ± 1.34	89.12 ± 0.63	94.56 ± 0.55	89.47 ± 0.73	89.12 ± 0.74	94.09 ± 0.58	89.59 ± 0.99	89.53 ± 0.51	93.68 ± 0.60	88.89 ± 0.87	89.18 ± 0.50	93.27 ± 0.84	89.12 ± 0.84	89.18 ± 0.50	93.27 ± 1.04
	PLP	91.23 ± 1.03	90.00 ± 0.75	94.50 ± 0.79	91.40 ± 0.83	89.77 ± 0.96	93.98 ± 0.28	91.23 ± 1.10	89.94 ± 0.77	93.68 ± 0.66	90.82 ± 1.17	89.59 ± 0.82	93.33 ± 0.41	91.23 ± 0.96	89.71 ± 0.84	93.04 ± 0.43

to obtain a better classification accuracy [21]. From Table 4, the MFCCs give more than 95% classification accuracy with 13 MFCCs. The SVM classifier gives higher accuracy than the kNN and LDA classifiers, using all of the feature extraction methods (MFCCs, LPC, and PLP). From the results, it is inferred that the SVM is a more suitable and reliable classifier for the classification of speech disfluencies. Different frame lengths also affect the classification performance. The best classification accuracies are obtained for the frame lengths of 10 ms, 20 ms, and 30 ms. The performance of the MFCCs is also similar to the performance of the WLPCC and PLP features. The results of the current work are very promising; however, it is not easy for us to compare our results with previous works since other researchers used different databases and their computations and presentations of the results are not uniform. In order to prove the reliability of the classification results, a 10-fold cross-validation is performed. The results are presented by implementing 3 acoustic feature extraction methods and 3 different classification methods using the speech samples from the UCLASS archive. The results of the present work cannot be compared directly with previous works [16,18], since they used different databases and they implemented their proposed algorithms under different conditions (the classification of normal and stuttered speech). Maximum accuracies of 83% and 94.35% were reported in [16,18]. In [33], the maximum accuracy was reported as 86.19% using 39 MFCCs. However, our proposed methods give better accuracy than earlier works [16,18,33] in classifying 2 types of disfluencies (repetition and prolongation).

Table 4. Classification results of the MFCCs for different frame lengths.

No. of MFCCs	Classifiers	10 ms	20 ms	30 ms	40 ms	50 ms
13	kNN	90.70 ± 1.25	92.51 ± 1.32	91.99 ± 0.83	93.51 ± 1.52	91.40 ± 1.26
	LDA	91.35 ± 0.72	91.75 ± 0.58	90.70 ± 0.80	91.75 ± 0.93	88.30 ± 1.10
	SVM	95.15 ± 0.68	95.73 ± 0.83	93.92 ± 0.30	94.21 ± 0.89	93.51 ± 0.70
15	kNN	88.83 ± 2.87	92.16 ± 1.75	91.35 ± 0.66	92.16 ± 1.36	91.11 ± 1.87
	LDA	91.05 ± 0.55	91.35 ± 1.13	91.05 ± 0.55	91.81 ± 0.48	90.53 ± 0.54
	SVM	95.50 ± 0.62	96.08 ± 0.73	95.79 ± 0.25	96.14 ± 0.49	95.67 ± 0.56
17	kNN	89.12 ± 2.62	92.05 ± 1.21	90.29 ± 0.74	92.28 ± 1.53	90.12 ± 1.88
	LDA	90.47 ± 0.73	91.35 ± 0.91	90.99 ± 0.49	91.87 ± 1.12	90.82 ± 0.39
	SVM	95.73 ± 0.73	96.02 ± 0.82	95.32 ± 0.78	95.73 ± 0.48	95.26 ± 0.85
24	kNN	88.36 ± 2.58	89.94 ± 1.76	90.70 ± 1.31	90.58 ± 1.77	90.35 ± 1.83
	LDA	89.12 ± 1.00	89.30 ± 0.91	89.06 ± 0.68	89.59 ± 1.37	89.53 ± 0.80
	SVM	95.15 ± 1.07	95.38 ± 0.64	94.39 ± 0.96	93.86 ± 0.84	94.09 ± 0.93

6. Conclusion

This paper presented a comparison of 3 speech feature extraction methods for the classification of 2 types of speech disfluencies (repetition and prolongation). Three different classifiers (kNN, LDA, and SVM) were employed. The disfluent speech samples were subjected to feature extraction using MFCCs and LPC- and PLP-based methods. In the LPC- and PLP-based methods, the order of the LPC was varied and its effect on the classification was also investigated. A 10-fold cross-validation was used to test the reliability of the classifier's results. The SVM outperformed kNN and the LDA. SVM gives the best average classification accuracy of above 95% using the WLPCC, PLP, and MFCC features. The classification results indicate that the proposed method could be used as a valuable tool for speech therapists in stuttering assessments. In the future, various feature selection techniques will be used to reduce the number of features and the proposed methods will also be implemented to classify other types of disfluency. Different classification algorithms will also be investigated to improve the classification results of speech disfluencies.

Acknowledgments

This work was supported by grant FRGS-9003-00228 from the Ministry of Higher Education of Malaysia. The authors wish to thank Vice Chancellor Y. Bhg. Brig. Jen. Prof. Dato' Dr. Kamarudin Hussin for his valuable support during the research work.

References

[1] S.S. Awad, "The application of digital speech processing to stuttering therapy", IEEE Transactions on Instrumentation and Measurement, Vol. 2, pp. 1361–1367, 1997.

[2] J. Van Borsel, E. Achten, P. Santens, P. Lahorte, T. Voet, "fMRI of developmental stuttering: a pilot study", Brain and Language, Vol. 85, pp. 369–376, 2003.

[3] P. Howell, S. Sackin, "Automatic recognition of repetitions and prolongations in stuttered speech", Proceedings of First World Congress on Fluency Disorders, pp. 372–374, 1995.

[4] P. Howell, S. Sackin, K. Glenn, "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", Journal of Speech, Language, and Hearing Research, Vol. 40, pp. 1073–1084, 1997.

[5] P. Howell, S. Sackin, K. Glenn, "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", Journal of Speech, Language, and Hearing Research, Vol. 40, pp. 1085–1096, 1997.

[6] E. Nöth, H. Niemann, T. Haderlein, M. Decher, U. Eysholdt, F. Rosanowski, T. Wittenberg, "Automatic stuttering recognition using hidden Markov models", Proceedings of the International Conference on Spoken Language Processing, Vol. 4, pp. 65–68, 2000.

[7] A. Ben-Hur, D. Horn, H.T. Siegelmann, V. Vapnik, "A support vector clustering method", Proceedings of the 15th International Conference on Pattern Recognition, Vol. 2, pp. 2724–2727, 2000.

[8] Y.V. Geetha, K. Pratibha, R. Ashok, S.K. Ravindra, "Classification of childhood disfluencies using neural networks", Journal of Fluency Disorders, Vol. 25, pp. 99–117, 2000.

[9] P. Howell, S. Davis, J. Bartrip, "The UCLASS archive of stuttered speech", Journal of Speech, Language, and Hearing Research, Vol. 52, pp. 556–569, 2009.

[10] A. Czyzewski, A. Kaczmarek, B. Kostek, "Intelligent processing of stuttered speech", Intelligent Information Systems, Vol. 21, pp. 143–171, 2003.

[11] B. Prakash, "Acoustic measures in the speech of children with stuttering and normal non fluency - a key to differential diagnosis", Proceedings of the workshop on Spoken Language Processing, pp. 49–57, 2003.

[12] I. Szczurowska, W. Kuniszyk-Jozkowiak, E. Smolka, "The application of Kohonen and multilayer perceptron networks in the speech nonfluency analysis", Archives Acoustics, Vol. 31, pp. 205–210, 2006.

[13] M. Wisniewski, W. Kuniszyk-Jóźkowiak, E. Smolka, W. Suszynski, "Automatic detection of disorders in a continuous speech with the hidden Markov models approach", Proceedings of Computer Recognition Systems 2, Vol. 45, pp. 445–453, 2008.

[14] M. Wisniewski, W. Kuniszyk-Jóźkowiak, E. Smolka, W. Suszynski, "Automatic detection of prolonged fricative phonemes with the hidden Markov models approach", Journal of Medical Informatics & Technologies, Vol. 11, pp. 293–298, 2007.

[15] T.S. Tan, Helbin-Liboh, A.K. Ariff, C.M. Ting, S.H. Salleh, "Application of Malay speech technology in Malay speech therapy assistance tools", Proceedings of IEEE Conference on Intelligent and Advanced Systems, pp. 330–334, 2007.

[16] K. Ravikumar, B. Reddy, R. Rajagopal, H. Nagaraj, "Automatic detection of syllable repetition in read speech for objective assessment of stuttered disfluencies", Proceedings of World Academy Science, Engineering and Technology, pp. 270–273, 2008.

[17] I. Swietlicka, W. Kuniszyk-Jóźkowiak, E. Smolka, "Artificial neural networks in the disabled speech analysis", Proceedings of Computer Recognition System 3, Vol. 57, pp. 347–354, 2009.

[18] K.M. Ravikumar, R. Rajagopal, H.C. Nagaraj, "An approach for objective assessment of stuttered speech using MFCC features", ICGST International Journal on Digital Signal Processing, Vol. 9, pp. 19–24, 2009.

[19] L. Sin Chee, O. Chia Ai, M. Hariharan, S. Yaacob, "MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA", Proceedings of IEEE Student Conference on Research and Development, pp. 146–149, 2009.

[20] L. Sin Chee, O. Chia Ai, M. Hariharan, S. Yaacob, "Automatic detection of prolongations and repetitions using LPCC", Proceedings of IEEE International Conference on Technical Postgraduates, pp. 1–4, 2009.

[21] M. Hariharan, L. Sin Chee, S. Yaacob, "Classification of speech dysfluencies using LPC based parameterization techniques", Journal of Medical Systems, Vol. 36, pp. 1821–1830, 2012.

[22] O. Chia Ai, M. Hariharan, S. Yaacob, L. Sin Chee, "Classification of speech dysfluencies with MFCC and LPCC features", Expert Systems with Applications, Vol. 39, pp. 2157–2165, 2012.

[23] P. Howell, M. Huckvale, "Facilities to assist people to research into stammered speech", Stammering Research, Vol. 1, pp. 130–242, 2004.

[24] X. Huang, A. Acero, H. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Upper Saddle River, NJ, USA, Prentice Hall, 2001.

[25] L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Upper Saddle River, NJ, USA, Prentice Hall, 1993.

[26] M. Hariharan, L. Sin Chee, S. Yaacob, "Analysis of infant cry through weighted linear prediction cepstral coefficients and probabilistic neural network", Journal of Medical Systems, Vol. 36, pp. 1309–1315, 2012.

[27] H.N. Ting, J. Yunus, S. Salleh, "Speaker-independent Malay syllable recognition using singular and modular neural networks", Jurnal Teknologi, Vol. 35, pp. 65–76, 2001.

[28] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", Journal of the Acoustical Society of America, Vol. 87, pp. 1738–1752, 1990.

[29] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed., San Diego, CA, USA, Academic Press, 1990.

[30] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., New York, NY, USA, Wiley Interscience, 2000.

[31] K. De Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, J.A.K. Suykens, LS-SVM Lab Toolbox User's Guide, Leuven, Belgium, Katholieke Universiteit Leuven, 2010.

[32] J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, Singapore, World Scientific, 2002.

[33] K.M. Ravikumar, S. Ganesan, "Comparison of multidimensional MFCC feature vectors for objective assessment of stuttered disfluencies", International Journal of Advanced Networking and Applications, Vol. 2, pp. 854–860, 2011.