Sains Malaysiana 48(12)(2019): 2737–2747
http://dx.doi.org/10.17576/jsm-2019-4812-15

Automatic Speech Intelligibility Detection for Speakers with Speech Impairments: The Identification of Significant Speech Features

(Pengesanan Kecerdasan Pertuturan Automatik untuk Penutur dengan Ketaksempurnaan Pertuturan: Pengenalpastian Ciri Pertuturan Penting)

FADHILAH ROSDI*, MUMTAZ BEGUM MUSTAFA, SITI SALWAH SALIM & NOR AZAN MAT ZIN

ABSTRACT

Selection of relevant features is important for discriminating speech in detection based ASR system, thus contributing to the improved performance of the detector. In the context of speech impairments, speech errors can be discriminated from regular speech by adopting the appropriate discriminative speech features with high discriminative ability between the impaired and the control group. However, identification of suitable discriminative speech features for error detection in impaired speech was not well investigated in the literature. Characteristics of impaired speech are grossly different from regular speech, thus making the existing speech features to be less effective in recognizing the impaired speech. To overcome this gap, the speech features of impaired speech based on the prosody, pronunciation and voice quality are analyzed for identifying the significant speech features which are related to the intelligibility deficits. In this research, we investigate the relations of speech impairments due to cerebral palsy, and hearing impairment with the prosody, pronunciation, and voice quality. Later, we identify the relationship of the speech features with the speech intelligibility classification and the significant speech features in improving the discriminative ability of an automatic speech intelligibility detection system. The findings showed that prosody, pronunciation and voice quality features are statistically significant speech features for improving the detection ability of impaired speeches. Voice quality is identified as the best speech features with more discriminative power in detecting speech intelligibility of impaired speech.

Keywords: Automatic speech intelligibility detection; speech detection; speech features; speech impairments

ABSTRAK

Pemilihan ciri yang relevan untuk membezakan pertuturan dalam sistem ASR berasaskan pengesanan adalah penting kerana menyumbang kepada peningkatan prestasi pengesan. Dalam konteks ketaksempurnaan pertuturan, kesalahan pertuturan boleh didiskriminasi daripada pertuturan biasa dengan menggunakan ciri pertuturan diskriminatif yang bersesuaian dengan keupayaan diskriminatif yang tinggi antara kumpulan terjejas dan kumpulan kawalan. Walau bagaimanapun, pengenalpastian ciri pertuturan diskriminatif yang sesuai untuk pengesanan ralat dalam pertuturan yang terjejas tidak dikaji dengan baik dalam kajian kepustakawan. Ciri pertuturan yang terjejas adalah sangat berbeza daripada pertuturan biasa, dengan itu, menjadikan ciri pertuturan sedia ada kurang berkesan dalam mengenal pasti pertuturan yang terjejas. Untuk mengatasi jurang ini, ciri pertuturan ketaksempurnaan pertuturan berdasarkan prosodi, sebutan dan kualiti suara dianalisis untuk mengenal pasti ciri pertuturan penting yang berkaitan dengan defisit kecerdasan. Dalam penyelidikan ini, kami mengkaji hubungan antara kecacatan pertuturan akibat lumpuh otak dan kecacatan pendengaran dengan prosodi, sebutan dan kualiti suara. Seterusnya, kami mengenal pasti hubungan ciri pertuturan dengan pengelasan kecerdasan pertuturan dan ciri pertuturan yang penting dalam meningkatkan keupayaan diskriminatif sistem pengesanan kecerdasan pertuturan secara automatik. Hasil menunjukkan bahawa ciri prosodi, sebutan dan suara adalah ciri pertuturan yang signifikan secara statistik untuk meningkatkan keupayaan pengesanan pertuturan yang terjejas. Kualiti suara dikenal pasti sebagai ciri pertuturan terbaik dengan kuasa yang lebih diskriminatif dalam mengesan kecerdasan pertuturan yang terjejas.

Kata kunci: Ciri pertuturan; ketaksempurnaan pertuturan; pengesanan kecerdasan pertuturan automatik; pengesanan pertuturan

INTRODUCTION

According to the American Speech Language Hearing Association (ASHA) guidelines (ASHA 1993), speech impairment refers to oral and verbal communication that is deviant from the norm that it is noticeable or interferes with communication. Characteristics of impaired speech often related with the disturbance and higher variability in speech. Blaney and Wilson (2000) reported that the increase in the variability is highly correlated with severity of impairment that leads to reduction in intelligibility.

Intelligibility refers to the judgement made by a clinician based on how much of an utterance can be

understood by human listeners (Bauman-Waengler 2012). The speech characteristics of speakers with speech impairments are often found to be less intelligible than non-speech impaired speaker. The reduction in speech intelligibility is considered as one of the main characteristics of individuals with speech impairments. Intelligibility varies greatly depending on the extent of the neurological disease or damage (Kent et al. 1989). Severity of speech impairments might differ among individuals, but also differs for a single speaker due to several factors such as fatigue, stress as well as personal and environmental factors (Young 2010).

Automatic speech intelligibility detection is one of the applications that make use of Detection- Based Automatic Speech Recognition (DBASR). The speech intelligibility detector finds the abnormal variation in the speech signal as unintelligible speech. In a standard ASR system, speech feature extraction is common to all classes. However, in detection task, there can be a specific feature extractor for each detector. This is an advantage because it is possible to process and extract relevant speech signals that are optimal for the specific class vs. anti-class problem in each detector (Canterla 2012). The same speech features, however, are possible to be used in all detectors.

In detection-based ASR system, the selection of features is important for discriminating speech. However, not many researches have investigated suitable discriminative speech features for error detection in impaired speech. The speech characteristic of impaired speech is grossly different from regular speech, thus making the existing speech features to be less effective in recognizing impaired speech. As the usual features were not found to be representative of impaired speech, alternative features must then be identified. In this paper, we investigate the relationships of speech impairments caused by cerebral palsy and hearing impairment with related discriminative speech features. In addition, we identify the relation of speech features with speech intelligibility classification and the significant speech features in improving discriminative ability of automatic speech intelligibility detection.

The paper is organized as follows: Next, we provide an overview of speech impairments, automatic speech intelligibility detection and available automatic speech intelligibility detection for speakers with speech impairments. After that, we discuss the discriminative speech features which includes prosodic, pronunciation, voice quality and selection of suitable speech features for detection of impaired speeches. In the following section, we present methods carried out in identifying the speech features. Subsequently, we discuss major findings based on the experiments conducted and lastly, we conclude this study.

BACKGROUND

Speech impairments are categorized into three (3) basic types: articulation impairments, voice impairments, and fluency impairments as shown in Figure 1 (ASHA 1993). Articulation involves the gradual acquisition in moving the articulators in precise and rapid manner (Bauman-Waengler 2012). In other words, articulation is a process of producing speech sounds that involve organs, manners and places of articulation. Thus, articulation impairment is the errors in the production of certain speech sounds characterized by deletions or omissions, substitutions and distortions in the speech that degrade the speech intelligibility (ASHA 1993).

Voice is produced by the vibration of the vocal cords. Air from the lungs sets these muscles into vibration, which is called phonation (Haynes & Pindzola 2012). The voice is varied when it passes through the vocal cords, nose and mouth due to different size and shape spaces, which is called resonance (Haynes & Pindzola 2012). Thus, voice impairments include aspect of phonation and resonance. A voice impairment refers to the abnormal production of speech properties like vocal quality, pitch, loudness, resonance, and/or duration (ASHA 1993).

Fluency is the natural forward flow speech. Fluency impairment refers to the interruption in the flow of speaking, which may be accompanied by excessive tension, struggle behaviour, and secondary mannerisms (ASHA 1993). A person with speech impairments may have problem with articulation, voice or fluency or any combination of these. These impairments lead to the changes of the speech which affect the characteristics of the individual's speech. The speech characteristics of people with speech impairments vary depending on the type of impairment involved.

AUTOMATIC SPEECH INTELLIGIBILITY DETECTION FOR IMPAIRED SPEECH

The speech intelligibility detection is treated as a binary classification problem, classifying words as either intelligible or not. Classification is a task of assigning an object that is characterized by a set of

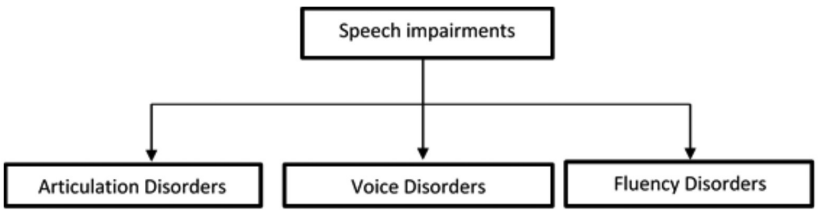


FIGURE 1. Three types of speech impairments (ASHA 1993)

features or parameters to a class or category based on the characteristics similarities of the object (Michie et al. 1994). Figure 2 shows the general framework of the automatic speech intelligibility detection which consists of speech signals, speech feature extraction, classification methods, training and evaluation phase. Speech signals *s(t)* are supplied to the speech feature extraction to extracts the meaningful and significant discriminative speech features *On(t)* in the classification task. Feature extraction is the process of transforming the input speech waveform into a sequence of acoustic feature vector suitable for further speech processing. The objectives of this feature extraction process are (Rosell 2006): The features should extract the important aspects of the speech signal and should be perceptually meaningful; and the features should be robust where the particular task should not be affected by the possible distortions, caused by environmental and/or transmission medium.

The knowledge scores *P(An|On(t))* produced by the classification methods are used later in the training and evaluation phases to detect the speech intelligibility of impaired speakers. In the context of intelligibility discrimination, classification methods are used to classify speech intelligibility according to its classes as intelligible or not intelligible. The most common approaches in classification methods are statistical approach such as k-nearest neighbour (KNN), Linear Discriminant Analysis (LDA) and machine learning algorithms such as Support Vector Machine (SVM), Artificial Neural Network (ANN). However, such as fuzzy logic and formal learning approach Petri Nets and Fuzzy Petri Nets have been considered in recent studies for classification purpose.

For detecting the speech intelligibility of impaired speech, researchers have used features or combination of features that focused on prosodic, pronunciation, phonatory, articulatory, and voice quality. For prosodic, some of the features applied for impaired speech are pitch contour (Kim et al. 2015), Fundamental frequency (F0) and spectral dynamic (Khan et al. 2014). For pronunciation of impaired speech, some of the common features are formant and MFCC (Kim et al. 2015). For voice quality, Kim et al. (2015) propose the use of jitter, shimmer and harmonics to noise ratio. Other form of features used for detecting the speech intelligibility of impaired speech includes LPC, MLFF and PLP features (Fook et al. 2013).

While the use of suitable feature is important for intelligibility detection of impaired speech, adopting an effective classifier(s) is equally important. It is common that more than one type of classifier will be tested by researchers. For example, Kim et al. (2015), have applied Support Vector Machine (SVM), Random Forest (RF), Linear Discriminant Analysis (LDA) and k-nearest neighbour (KNN) classifiers for intelligibility detection of impaired speech, while Fook et al. (2013) studied the classification of the prolongations and repetitions among the speakers with stuttering using the LPC, MLFF and PLP features. The performance of the classifiers varies from one work to another and there is no conclusive evidence as to which classifier is better for detecting the speech intelligibility of impaired speech, though SVM has been associated with better classification accuracy (Fook et al. 2013; Khan et al. 2014).

DISCRIMINATIVE SPEECH FEATURES

Speech signal is converted to parametric representation during the feature extraction process. This parametric representation is a discriminative speech features containing useful information to identify and discriminate speech sounds, which is then used for further analysis and processing. Speech parametric comprises a number of frames derived from extracted speech signal and decomposed to regular interval, for example 10 to 25 milliseconds (ms) per frame. Discriminative speech features must provide a good representation of phonemes and be robust to non-phonetic changes in signal (John 2006).

There are several types of speech features. In pathological speech, we can classify these features according to the aspect of prosodic, pronunciation and voice quality (Kim et al. 2015).

PROSODIC FEATURES

Prosody refers to the structure that organizes sound that comprises tone, loudness, and the rhythm structures of speech (Cutler et al. 1997). Suitable physical representations of prosody include fundamental frequency (pitch), intensity, energy and the normalized duration of syllables. Fundamental frequency (F0) is the lowest frequency that reflects the physiological limits of speech (Colton & Casper 2006), while Intensity refer to the amount of energy transported over a given area of the medium per unit of time (Rosen & Howell 2011). The common energy related features are Signal energy and Zero Crossing Rate (ZCR). Signal energy is a time domain audio feature. The change in energy is computed by dividing speech frames into sub frames of fixed duration.

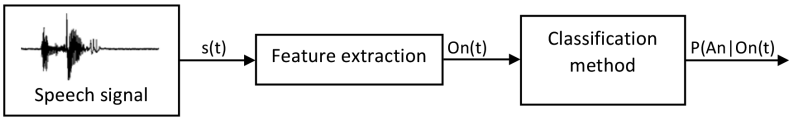


FIGURE 2. General framework of the speech intelligibility detection

PRONUNCIATION FEATURES

Pronunciation feature are spectral based features which usually represents the magnitude properties of speech spectrum (Jurafsky 2009). It is commonly used features in speech processing. In spectral related features, there are many possible feature representations such as LPC, PLP, Rasta and MFCC. By far, the most common in the speech recognition is the MFCC (Jurafsky 2009). Formant frequencies are also common spectral features, which are the concentration of acoustic energy around a particular frequency in the speech wave (Lapteva 2011). The formant with the lowest frequency is labelled as the first formant (F1), the higher is labelled as the second formant (F2), and the highest is the third formant (F3). These formants are closely related to the vowel production where F1 is related to the height of vowel, F2 is related to vowel frontness. F3 is considered to remain relatively constant for speakers (Nolan 2002).

MFCC is a speech feature that is capable to capture the important characteristic of audio signals. MFCC contains time and frequency information of the signal. MFCC has been widely used in the area of speech recognition.

The features related to changes in cepstral features over time are added by adding a delta and double delta features for each 13 features. Overall, 39 MFCC features derived which consists of 12 cepstral and 12 delta ceptral coefficients, 12 double delta ceptral coefficients, 1 energy coefficient, 1 delta energy coefficient and 1 double delta energy coefficient.

VOICE QUALITY BASED FEATURES

The voice quality based features are voicing related features affecting the speech quality, with common features such as jitter and shimmer. Jitter and shimmer are acoustic characteristics of voice signals measured as the cycle-to-cycle variations of fundamental frequency and waveform amplitude, respectively (Farrús et al. 2007). Both features correlate with the hoarseness in speech. There are several types of measurements for jitter and shimmer as described in Table 1.

SELECTION OF DISCRIMINATIVE SPEECH FEATURES FOR IMPAIRED SPEECH

A major concern in the selection of discriminative speech feature is associating suitable speech features to the problem of interest. It is commonly known that a bigger number of features increase the discriminating power of the classifiers. In practice, using more features increase the classification processing time, and classifiers are prone to overfitting. On the other hand, using unrelated speech features degrades the learning performance of the classifiers. One of the objectives of the feature selection research is evaluating the advantages of each feature. According to Kim et al. (2015), speech features can be categorized as prosody, pronunciation, and voice quality. The selection of speech features is reflected to the types of speech impairments, which are articulation impairments, voice impairments, and fluency impairments. Figure 3 shows the mapping of the types of speech impairments to the category or aspect of speech features.

TABLE 1. Types of measurements for jitter and shimmer

Measurement	Description
Jitter (absolute)	The cycle-to-cycle variation of fundamental frequency (Vipperla 2010)
Jitter (relative)	The average absolute difference between consecutive periods, divided by the average period and expressed as a percentage (Vipperla 2010)
Jitter (rap)	The Relative Average Perturbation which is the average absolute difference between a period and the average of it and its two neighbours, divided by the average period (Vipperla 2010)
Jitter (ppq5)	The five-point Period Perturbation Quotient, computed as the average absolute difference between a period and the average of it and its four closest neighbors, divided by the average period (Vipperla 2010)
Shimmer (dB)	The variability of the peak to-peak amplitude in decibels (Vipperla 2010)
Shimmer (relative)	The average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude which is expressed as percentage (Vipperla 2010)
Shimmer (apq3)	the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude (Vipperla 2010)
Shimmer (apq5)	the five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude (Vipperla 2010)

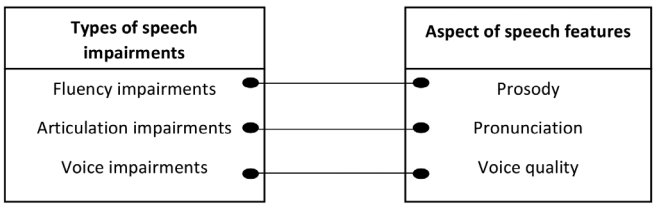


FIGURE 3. Mapping of the types of speech impairments and the aspect of speech features

The three types of impairments are influenced by the speech features of impaired speech. For fluency impairment, the speaking flow is interrupted, which affect the atypical rate, rhythm, and repetition in sounds. The speech features related to prosody such fundamental frequency (F0), intensity, energy, and normalized duration of syllables are related to tone, loudness and rhythm structures. These features are suitable representation of the characteristics of impaired speech with fluency impairments. These features are also used to analyze the variation in harmonic frequencies, which is basically irregular vibration of vocal folds due to un-periodic flow of air through lungs (Butt 2012).

Articulation impairments are correlated with the ability of the articulator to pronounce words. Impaired speeches contain higher pronunciation variations that contribute to intelligibility loss. Therefore, speech features that carry meaningful information related to pronunciation such as MFCC is important to represent the characteristics of the articulation impairments.

Voice impairments, which is related to the abnormal production in voice quality includes aspect of phonation and resonance. Speech features related to voice quality such as jitter and shimmer are important to represent the characteristics of voice impairments. Jitter is the frequency perturbation, while shimmer is an amplitude perturbation. Both features are important in voice quality measurement and serves as index of vocal stability. Excessive jitter and shimmer cause hoarseness, harsh or rough voice quality.

For speech disability caused by cerebral palsy and hearing impairment, the abnormal variation in the impaired speech in comparison with non-impaired speech covers the aspect of prosodic, voice quality and pronunciation of pathological speech as proposed in Kim et al. (2015). As such, the following six speech features are chosen for identifying aspects of speech, which are prosodic (F0, energy and zero crossing rate), voice quality (jitter absolute and shimmer absolute), and for spectral features, MFCC was chosen.

METHOD

This research aims at identifying the best speech features with more discriminative power in detecting speech intelligibility of children with Cerebral Palsy and hearing impairment. The procedures involved in the implementation of the automatic speech intelligibility detection for Malay speaking children with speech impairments due to cerebral palsy and hearing impairments are as follows; Data preparation, speech feature extraction to extract the discriminative features, and speech classification which consists of selected classifiers. These procedures are further discussed as follows.

DATA PREPARATION

The first step in developing the automatic speech intelligibility system is to prepare the speech data that will be used in the training and evaluation phase. In data preparation, there are 3 main stages involved as follows; Data identification: Identifies language and types of speech stimuli to be used, data acquisition: Involves the selection of speakers and speech recording procedures, and speech database: The output of the process of the data preparation These stages are further explained in the following subsections.

DATA IDENTIFICATION

First, we have identified the language of speech data to be acquired. In this research, the speech stimuli were provided in Malay language. It is a phonetic language that belongs to the western subfamily of Malayo Polynesian languages, also known as Austronesian languages (Green 1966; Pawley 1966; Ting et al. 2017). It is used by 500 million people as spoken language, mostly in Malaysia, Indonesia, Brunei, Singapore and southern Thailand (Tan 2012). Malay is also the official language in Malaysia, Brunei, Indonesia and Singapore. Malay language is divided into many dialects. However, this research focuses on the Standard Malay, which refers to the national norm or prestige dialect, which is also designated as the official language in Malaysia (El-Iman & Don 2005).

In terms of types of speech stimuli, they were set to 51 short, simple and meaningful sentences that contain two to five words in each sentence. The sentences were selected after discussions and consultations with the SLPs and teachers. These sentences were designed to suit the speakers' reading abilities and word familiarity. The use of short sentences is also due to the fact that most of the speech impaired children also suffered from physical and cognitive impairments. Thus, they can be easily fatigued, hesitant and tense when they had to utter long or complex sentences. The short, simple, and meaningful sentences were used in this study to provide sufficient features for analysing the speech features.

DATA ACQUISITION

Several tasks were carried out to acquire the speech data that are describe as follows:

Speaker characterization 30 speech impaired children were selected to take part in the recording session from special schools and spastic centre in Petaling Jaya, Kuala Lumpur, Malaysia. There were 16 male and 14 female (aged between 8 and 12 years old; with the mean age of 10 years old. These children were diagnosed with different types of speech impairment. Professional SLPs assessed the children and classified the severity of speech impairment.

Recording environment and apparatus The recording session was carried out in a quiet room with a portable sound booth that has a stand microphone for children to speak. The stimuli were presented on a 17-inch Laptop screen. All speech materials were digitized from the audio playback using a 22 kHz sampling rate at 16-bit sample resolution. The lingWAVES Voice Clinic Suite was used to record the speech. The stand microphone is preferred as the speakers might be uncomfortable with a headset microphone. External hard disk is used as a backup storage.

Recording procedure and design The recording sessions were carried out by placing speakers in a quiet room. The speakers were recorded individually, seated at a desk in front of the microphone. The ling WAVES stand microphone was positioned approximately 4 to 6 inches from the speaker's mouth and the speech stimuli were displayed to the speakers using a laptop screen. The experimenter was seated beside the speakers to assist in the reading to avoid any experimental bias. The session was designed to be fulfilled by the speakers in the corpus with simple meaningful sentences. Each speaker was asked to utter 51 sentences in three repetitions. Three sessions were designed for each speaker and each session was recorded in a different day or big gap of time to reflect intra-speaker variability and avoid the speakers from fatigue and fluctuate emotional state due to long recording sessions. The sessions involving the impaired speakers are of greater frequency and shorter time as compared with the non-impaired speakers. This is because the physical limitations of the impaired speakers. The sentences were pronounced by the experimenter followed by the speaker, and the experimenter will not in any condition correct the children on their speeches. The children were encouraged to speak naturally and clearly.

SPEECH DATABASE

As a result, the total amount of impaired speech samples acquired during the whole process was 4,590 utterances in 3.8 h of recordings including silence. Table 2 summarizes the impaired speech database that has been developed.

Our reference speech corpus consists of speech from 50 unimpaired children (25 males, 25 females) with age ranging from 8 to 12 years old. The recording environment, procedures and speech stimuli were the same with the impaired corpus. The intention was to have a group of speakers that are balanced in terms of age and gender. The selected children were assessed by their teachers to ensure that they are good in literacy. Each speaker uttered the same 51 sentences in one session of recording supervised by the experimenter. The total amount of speech samples acquired during the whole process was 7,650 utterances in 2.5 h of recordings including silence.

SPEECH FEATURE EXTRACTION

The speech features are extracted using the Opensmile toolkit (Eyben et al. 2010). The WAVE signal of speech utterances is segmented into 25 milliseconds with an interval of 10 milliseconds. The speech is parameterized with 13 MFCC (0th to 12th coefficients), normalized log energy, ZCR, F0, jitter local and shimmer local yielding a total of 18-dimensional feature vector. The feature vectors are then used for further process for the classifier. The speech feature extraction files are generated separately in. arff files for both training and evaluation.

TABLE 2. Descriptions of the impaired speech corpus

Speakers' age (years old)	8	6 speakers
	9	6 speakers
	10	6 speakers
	11	6 speakers
	12	6 speakers
Speakers' gender	Male	16 speakers
	Female	14 speakers
Speakers' diagnosis	Cerebral Palsy (CP)	12 speakers
	Hearing impaired	18 speakers
Speakers' severity level	Mild	8 speakers
	Mild-moderate	9 speakers
	Moderate-severe	6 speakers
	Severe	7 speakers

CLASSIFICATION METHODS

In this research, four classifiers are used, which are Support Vector Machine (SVM), Random Forests (RF), Linear Discriminant Analysis (LDA) and K-Nearest Neighbor (KNN). These classifiers are selected as baseline classifier because they have been used in the existing literature for detecting the impaired speech intelligibility as discussed earlier.

LibSVM Matlab toolbox is used for SVM model training and evaluation. The random Forest package is used for training and evaluating the RF which is a MATLAB standalone application. For LDA and KNN, the default parameter provided by MATLAB was used without any modification.

EVALUATION OF CLASSIFICATION METHODS

The evaluation is carried out on the four selected classifiers, SVM, RF, LDA and KNN, which are implemented in MATLAB 2013b. The 10-fold cross validation is used where the speech files were randomly partitioned into 10 equal size subsamples, where nine partitions are set for the training and the remaining one is the test set for evaluating the model. In each run, one of the partitions is used as a test data and the remaining partitions are used as train data. This procedure is repeated 10 times until all 10 subsamples are used as test data. The performance of classifiers is the average classification of the training and testing data.

The evaluation of the baseline system involves speeches from 30 CG and 30 SIG speakers. A total of 2,950 utterances from 1,528 unimpaired utterances and 1,422 impaired utterances were used for the evaluation purposes. These utterances were extracted using Opensmile to extract the significant speech features; energy, F0, ZCR, MFCC 0th coefficient to 12th coefficient, jitter and shimmer.

The results presented for baseline classifiers are measured using the Classification Accuracy, Precision and Recall. The confusion matrix of classification error rate for Type I and Type II are presented as well. There are two types of evaluation being performed. First, the speech data are evaluated in terms of classification rate, classification accuracy, precision and recall for the overall data. Second, the classification accuracy is derived for individual speech feature.

RESULTS

Table 3 presents the confusion matrix of misclassification for SVM, RF, LDA and KNN. RF produces the highest Type 1 (FP) with 10 times, follows by SVM (6), KNN (3) and LDA (2). Meanwhile, for Type II error (FN), SVM and LDA produces the highest which are 18 frequencies, follows by KNN (15), and RF (7).

Table 4 presents the classification accuracy, precision and recall of the selected baseline classifiers in terms of the training and evaluation. The classification accuracy training set for SVM is 96.71%, RF is 99.40% and LDA is 96.87%. KNN produces 100% classification accuracy. In evaluation, KNN produce the highest accuracy with 97.80%, follows by SVM with 96.44%. LDA produces 95.75% and RF with slightly which is 93.22%. Meanwhile, LDA produces the highest precision with 98.62%, follows by KNN with slightly lower, 98.53%, SVM (96.34%) and RF (93.23%). For recall, RF produces highest percentage with 95.67%,

TABLE 3. Confusion matrix of the baseline classification methods

SVM		Prediction		
		Notint	Int	Total
Actual	Notint	TP(147)	FP(6)	153
	Int	FN(18)	TN(124)	142
	Total	165	130	295

RF		Prediction		
		Notint	Int	Total
Actual	Notint	TP(143)	FP(10)	153
	Int	FN(7)	TN(135)	142
	Total	150	145	295

LDA		Prediction		
		Notint	Int	Total
Actual	Notint	TP(151)	FP(2)	153
	Int	FN(18)	TN(124)	142
	Total	169	126	295

KNN		Prediction		
		Notint	Int	Total
Actual	Notint	TP(150)	FP(3)	153
	Int	FN(15)	TN(127)	142
	Total	165	130	295

TABLE 4. The overall classification accuracy, precision and recall of baseline classifiers

Classification method	Accuracy		Precision		Recall	
	Training	Testing	Training	Testing	Training	Testing
SVM	96.71	96.44	98.08	96.34	96.72	89.65
RF	99.40	95.36	99.08	93.23	99.51	95.67
LDA	96.87	95.75	99.15	98.62	96.01	89.32
KNN	100.00	97.80	100.00	98.53	100.00	90.90

TABLE 5. The classification accuracy based on the individual speech features

Speech features	SVM		RF		LDA		KNN	
	Training	Evaluation	Training	Evaluation	Training	Evaluation	Training	Evaluation
Prosody								
F0	56.67	56.73	76.92	70.08	61.59	61.16	73.22	67.23
Energy	51.77	51.71	75.17	74.69	51.59	51.81	56.30	56.35
ZCR	80.58	80.58	86.03	78.89	80.67	80.58	86.00	78.55
Pronunciation	74.52	74.53	82.92	70.85	74.46	74.45	80.04	67.17
Voice quality								
Jitter	73.26	73.50	86.45	75.09	76.33	76.38	86.44	74.21
Shimmer	78.90	78.99	87.75	79.02	82.13	82.04	87.07	77.57

follows by KNN, SVM and LDA with 90.90%, 89.65% and 89.32%, respectively.

Table 5 shows the accuracy for all the baseline classification methods for each individual feature. Mean values are calculated for each aspect of speech features such as prosody, pronunciation and the voice quality aspect.

For prosody, the average classification accuracy of RF is the highest at 74.55%, followed with KNN at 67.38%, LDA and SVM, at 64.52% and 63.01%, respectively. In term of pronunciation, SVM has the highest average classification accuracy at 74.53%, followed by LDA at 74.45%. On the other hand, RF and KNN have average accuracy of 70.85% and 67.17%, respectively. For voice quality, LDA has the highest average classification accuracy at 79.21%, followed by RF, SVM and KNN at 77.06%, 76.25% and 75.89%, respectively.

DISCUSSION

Based on the classification results of baseline classifiers, the discussion of findings is as follows;

THE RELATION OF SPEECH FEATURES WITH SPEECH INTELLIGIBILITY CLASSIFICATION

When comparing Tables 4 and 5, it is clear that the combination of selected speech features has more discriminating power and classification accuracy as compared to the individual speech features. From Table 5, voice quality shows the highest accuracy for all classification methods. This result indicates that jitter and shimmer are significant speech features that contribute to the speech intelligibility deficits among impaired speakers. Best speech features in detecting speech intelligibility

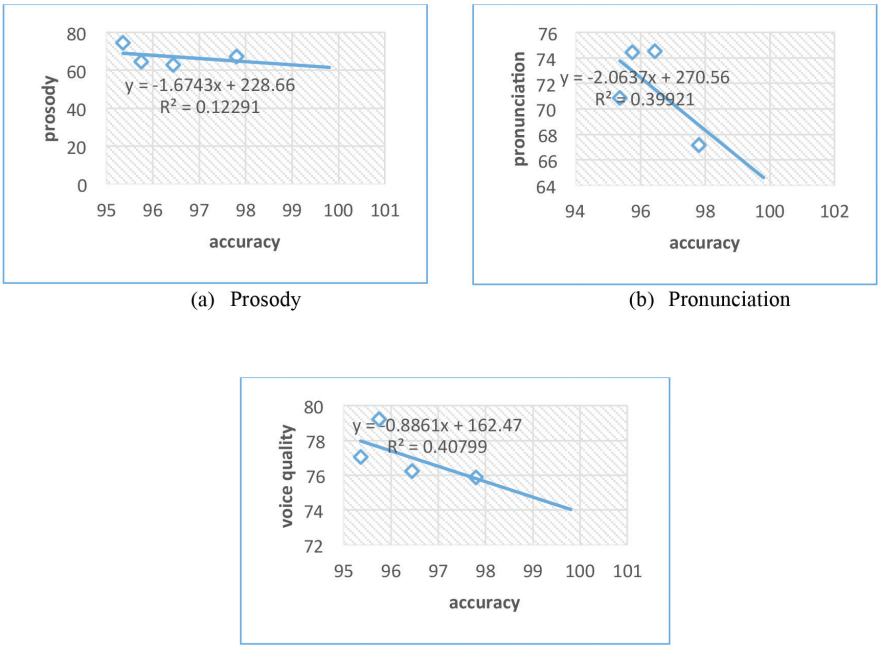
Table 6 shows the average values for prosody, pronunciation and voice quality aspect of all classification methods. Voice quality produces the highest mean score at 77.10%, followed by pronunciation (71.75%), and prosody (67.37%). From Table 6, voice quality has the highest accuracy for all classification methods. This indicates that jitter and shimmer are significant speech features for detecting speech intelligibility of impaired speakers.

Based on Table 6, on average, Random Forest (RF) was found to be the most suitable classifier to build an accurate automatic speech intelligibility detection system for impaired speakers as it has the highest average score at 74.15. FR was found to be the most effective classifier to discriminate both the Prosody and the voice quality (at 74.55 and 77.06, respectively). For discriminating the pronunciation, SVM is the best classifier to be used. However, among the four classifiers, KNN is the poorest performing classifier, and thus may not be suitable to be used for developing an accurate automatic speech intelligibility detection system for impaired speakers.

In addition, a linear regression analysis is performed to determine the effect of prosody, pronunciation and voice quality to intelligibility detection. The purpose is to understand which of the speech aspect statistically significant predictor for intelligibility detection for impaired speech. Figure 4 shows the effect of prosody, pronunciation and voice quality on the classification accuracy.

TABLE 6. The mean values for prosody, pronunciation and voice quality

	SVM	RF	LDA	KNN	Average
Prosody	63.01	74.55	64.52	67.38	67.37
Pronunciation	74.53	70.85	74.45	67.17	71.75
Voice quality	76.25	77.06	79.21	75.89	77.10
Average	71.26	74.15	72.73	70.15	72.07



(a) Prosody (b) Pronunciation

(c) Voice quality

FIGURE 4. Effect of (a) prosody, (b) pronunciation and (c) voice quality on classification accuracy

TABLE 7. Correlation and coefficient of determination of prosody, pronunciation and voice quality

	R	R^2	F	P
Prosody	0.351	0.123	0.280	0.008
Pronunciation	0.632	0.399	1.329	0.012
Voice quality	0.639	0.408	1.378	0.049

Prosody, pronunciation and voice quality are found to be a significant predictors of classification accuracy ($p=0.008$, <0.05), ($p=0.012$, <0.005) and ($p=0.049$, <0.05), respectively. However, we found that prosody is insignificant at $p<0.05$. The variance in classification accuracy can be explained by prosody (12.3%), pronunciation (39.9%) and voice quality (40.8%). Among the three, voice quality is found to explain more on the variation in classification accuracy. Table 7 summarizes the correlation and coefficient of determination of prosody, pronunciation and voice quality.

The finding in this research echoes some of the findings from the existing works in dysarthric speech recognition. In Kim et al. (2015), it was found that the effectiveness of voice quality features for automatic intelligibility assessment is important for TORGO datasets, while prosody-related measures; the composite measure was shown to be a reliable indicator of dysarthric word intelligibility of UA-speech speech database (Falk et al. 2012).

Jitter and shimmer are the two significant aspect of impaired speech due to cerebral palsy and hearing impairment that lead to intelligibility deficits. These two features correlate with the hoarseness in speech, which reduce the quality of speech for impaired speakers (Vipperla 2010). This is because, speech impaired children have speech abnormality that affects the vocal folds, either muscle or neural activity involved with phonation, either lesions that may cause increase in aperiodicity of vocal fold vibration which was reflected in the increased value of jitter (Wertzner et al. 2005). The speech characteristics is also indicated by the reduction of glottic resistance, vocal fold mass lesions and greater noise at production, which are some of the factors that influence shimmer values (Wertzner et al. 2005). Therefore, in this research, we have identified that voice quality that consists of jitter and shimmer have more discriminative power in detecting speech intelligibility of impaired speech compared to prosody and pronunciation aspect.

CONCLUSION

The reduction of intelligibility in impaired speech among children with Cerebral Palsy and hearing impairment can be attributed to several reasons such as the imprecise articulation, severity of impaired speakers and speech variability. It is clear that the speech features play an important role in discriminating speech including impaired speech. This is because these features correlate to the speeches which carry the meaningful information.

Therefore, selecting the relevant speech features is essential for speech intelligibility detection.

We have presented the simulation results using a new set of speech features for the speech intelligibility detection of impaired speeches for children with Cerebral Palsy and hearing impairment. We have established the relationship between the aspect of pathological speech features and the types of speech impairments. From the relationship, we have identified the relevant discriminative speech features for impaired speech to be used in the automatic speech intelligibility detection.

We found that prosody, pronunciation and voice quality features are statistically significant speech features of speech impaired speeches to improve the detection ability. Among the three types of features, voice quality is identified as a best speech features with more discriminative power in detecting speech intelligibility of impaired speech. From this work, we conclude that voice quality is vital speech features for detecting speech intelligibility of children with Cerebral Palsy and hearing impairment.

Many of the existing works did not consider voice quality as essential features for detecting speech intelligibility of individual with Cerebral Palsy and hearing impairment. Though we have proven that voice quality is very important for detecting speech intelligibility of children with Cerebral Palsy and hearing impairment, we cannot assume the same for adult individual. Research on adult individual with Cerebral Palsy and hearing impairment can be an interesting future work. On top of that, the use of effective classifiers such as the Artificial Neural Network (ANN) (Bhushan 2016), Deep Neural Network (DNN) (Ali Bou et al. 2019), and Deep Learning (Zhang et al. 2018) can be considered in future works.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Higher Education, Malaysia under UM High Impact Research Grant UM-MOHE UM.C/HIR/MOHE/FCSIT/05; and Universiti Kebangsaan Malaysia under Young Researchers Incentive Grant GGPM-2017-020. The authors declare that they have no conflict of interest. Informed consent was obtained from all individual participants included in the study.

REFERENCES

Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh & Khaled Shaalan. 2019. Speech recognition using deep neural networks. *A Systematic Review, IEEE Access* 7: 19143-19165.

American Speech and Hearing Association (ASHA). 1993. Dysarthria. http://www.asha.org/public/speech/impairments/dysarthria.htm. Accessed on 4th January 2018.

Bauman-Waengler, J. 2012. *Articulatory and Phonological Impairments: A Clinical Focus.* 5th ed. New Jersey: Allyn & Bacon Communication Sciences and Impairments Series.

Bhushan, C.K. 2016. Speech recognition using artificial neural network - A Review. *Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE)* 3(1) http://dx.doi.org/10.15242/IJCCIE.U0116002.

Blaney, B. & Wilson, J. 2000. Acoustic variability in dysarthria and computer speech recognition. *Clinical Linguistic and Phonetic* 14(4): 307-327.

Butt, A.H. 2012. Speech assessment for the classification of hypokinetic dysarthria in Parkinson's disease (Masters Dissertation). Computer Engineering, Dalarna University (Unpublished).

Colton, R.H. & Casper, J.K. 2006. *Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment.* Baltimore: Lippincott Williams & Wilkins.

Cutler, A., Dahan, D. & Donselaar, W.v. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40: 141-201.

del Hoyo, C. 2012. Design of detectors for automatic speech recognition. PhD Thesis. Department of Electronics and Telecommunications. Norwegian University of Science and Technology (Unpublished).

El-Imam, Y.A. & Don, Z.M. 2005. Rules and algorithms for phonetic transcription of standard Malay. *IEICE Transactions on Information and Systems* (10): 2354-2372.

Eyben, F., Wöllmer, M. & Schuller, B. 2010. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* 2010: 1459-1462.

Falk, T.H., Chan, W.Y. & Shein, F. 2012. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication* 54(5): 622-631.

Farrús, M., Hernando, J. & Ejarque, P. 2007. Jitter and shimmer measurements for speaker recognition. *Proceedings of the International Conference Interspeech.* August 27-31, Antwerp, Belgium. pp. 778-781.

Fook, C.Y. & Muthusamy, H. 2013. Comparison of speech parameterization techniques for the classification of speech disfluencies. *Turkish Journal of Electrical Engineering & Computer Sciences.* doi: 10.3906/elk-1112-84.

Green, R. 1966. Linguistic subgrouping within Polynesia: The implications for prehistoric settlement. *Journal of the Polynesian Society* 75: 6-38.

Haynes, W.O. & Pindzola, R.H. 2012. Motor speech disorders, dysphagia, and the oral exam. In *Diagnosis and Evaluation in Speech Pathology.* 8th ed., edited by Haynes, W.O. & Pindzola, R.H. Upper Saddle River, New Jersey: Pearson Education Inc. pp. 239-266.

Huang, X., Acero, A. & Hon, H.W. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* Upper Saddle River, New Jersey: Prentice Hall.

John, P.H. 2006. 2006. *Hidden Markov Models for Speech Recognition*. Slide presentation, Oregon Health & Science University OGI School of Science & Engineering. Accessed on 23 November 2017. http://cc.cpe.ku.ac.th/~jim/lecnotes/markov_models/articles/Yan2003-HMMSpeech-Recognition%20.pdf.

Jurafsky, D. & Martin, J.H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, New Jersey: Prentice Hall.

Kent, R.D., Weismer, G., Kent, J.F. & Rosenbek, J.C. 1989. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Impairments* 54: 482-499.

Khan, T., Westin, J. & Dougherty, M. 2014. Classification of speech intelligibility in Parkinson's Disease. *Biocybernetics and Biomedical Engineering* 34: 35-45.

Kim, J., Kumar, N., Tsiartas, A., Li, M. & Narayanan, S.S. 2015. Automatic intelligibility classification of sentence-level pathological speech. *Computer Speech & Language* 29(1): 132-144.

Lapteva, O. 2011. *Speaker Perception and Recognition: An Integrative Framework for Computational Speech Processing.* Kassel, Hessen: Kassel University Press.

Michie, D., Spiegelhalter, D.J., Taylor, C.C. & Campell, J. 1994. *Machine Learning, Neural, and Statistical Classification*. New York: Ellis Horwood.

Nolan, F. 2002. The 'telephone effect' on formants: A response. *Forensic Linguistics* 9(1): 74-82.

Pawley, A. 1966. Polynesian languages: A subgrouping based on shared innovations in morphology. *Journal of the Polynesian Society* 75: 39-64.

Rosell, M. 2006. *An Introduction to Front-End Processing and Acoustic Features for Automatic Speech Recognition*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.120.5299

Rosen, S. & Howell, P. 2011. *Signals and Systems for Speech and Hearing*. Leiden, Netherlands: BRILL.

Tan, T.P., Goh, S.S. & Khaw, Y.M. 2012. A Malay dialect translation and synthesis system: Proposal and preliminary system. *International Conference on Asian Language Processing (IALP)*. Hanoi, Vietnam.

Ting, H.N., Bakar, A.R.A., Santhosh, J., Al-Zidi, M.G., Ibrahim, I.A. & Cheok, N.S. 2017. Effects of speech phonological features during passive perception on cortical auditory evoked potential in sensorineural hearing loss. *Sains Malaysiana* 46(12): 2477-2488.

Vipperla, R., Renals, S. & Frankel, J. 2010. Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing* 2010: 525783.

Wertzner, H.F., Schreiber, S. & Amaro, L. 2005. Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological impairments. *Brazilian Journal of Otorhinolaryngology* 71(5): 582-588.

Young, V. & Mihailidis, A. 2010. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology* 22(2): 99-112.

Zhang, Z., Geiger, J., Pohjalainen, J., Amr El-Desoky, M., Jin, W. & Schuller, B. 2018. Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology* 9(5): Article No. 49.

Fadhilah Rosdi* & Nor Azan Mat Zin
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
46300 UKM Bangi, Selangor Darul Ehsan
Malaysia;

Mumtaz Begum Mustafa & Siti Salwah Salim
Faculty of Computer Science and Information Technology
University of Malaya
50603 Kuala Lumpur, Federal Territory
Malaysia

*Corresponding author; email: fadhilah.rosdi@ukm.edu.my

Received: 17 October 2018
Accepted: 2 October 2019