THE EFFECT OF CHANGES IN SPEECH FEATURES ON THE RECOGNITION ACCURACY OF ASR SYSTEM: A STUDY ON THE MALAY SPEECH IMPAIRED CHILDREN

[1]F. Rosdi, [2]M. B. Mustafa, [3]S. S. Salim, and [4]B.A. Hamid

[1,2,3]Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
[4]Department of Audiology and Speech Science, Faculty of Health Sciences, Universiti Kebangsaan Malaysia, 50300 Kuala Lumpur, Malaysia

Email: fadhilah.rosdi@um.edu.my[1], mumtaz@um.edu.my[2], salwa@um.edu.my[3], badrulhamid@ukm.edu.my[4]

ABSTRACT

Speech impairments refers to disability that causes the human speech production to deviate from the norm. Although there have been several researches undertaken to identify the differences between non-impaired and impaired speech, little is known about their effects on the speech intelligibility and the performance of ASR systems in recognizing impaired speech of children. This study investigates the speech features of impaired speech in relation to intelligibility deficits and degradation in ASR performance; which includes, formant frequencies, intensity, fundamental frequency (F0) and perturbation features such as jitter and shimmer. As there is no existing speech database for performing the evaluation, we have developed a speech database of speech impaired children and have analysed the impaired speech features. We have identified significant differences in the selected features. We also have identified the relationship between the ASR system's Word Error Rate (WER) of impaired speeches with the speech features. The results show that there are significant differences in F0, jitter and shimmer across the Control Group (CG) and the Speech Impaired Group (SIG). This paper explains the differences between impaired speeches and non-impaired speeches that can be used in developing automated speech recognition system. We have observed that F0 affects the ASR performance and was found to be a significant predictor that influences the accuracy of vowel phonemes /e/ and /u/.

Keywords: Speech impairments, Speech impaired children, Speech intelligibility, Speech features, Automatic Speech Recognition

1.0 INTRODUCTION

Over the past decades, Automatic Speech Recognition (ASR) system offers invaluable contributions to the field of speech and language therapy in improving speech, language and communication skills among impaired speakers [1] [2] [3] [4] [5]. However, developing an ASR system becomes more challenging for children with speech impairments due to inability of speakers to speak fluently, affecting the production of speech sound.

Several studies have reported that speech-impaired speakers produce a higher number of speech errors such as substitution, omission, distortion and addition [6] [7] [8] [9]. These errors result in low intelligibility of speech compared with regular speakers. Speech intelligibility is a common measure of the severity level of impairment, where it is calculated as the ratio of words understood by the listener to the total number of words articulated [10].

Common examples of speech impairments are Dysarthria and Apraxia [11]. Such disabilities affect the placement, timing, direction, pressure, speech and integration in the movement of the lips, tongue, velum or pharynx [12]. Thus, speech produced in impaired speaker is deviates from the normal speech.

Human speech is represented by various features, and measurement of relevant speech features of production in speech impairment were reported in existing literatures [13] [14]. Wertzner et al. [14] state that the most important vocal speech features for clinical use are the measurement of vocal extension profile such as frequencies and intensity, noise, acoustic spectrograph; fundamental and formant frequencies and perturbation index; jitter and shimmer. In this paper, we have considered formant frequencies (F1, F2), fundamental frequency (F0), intensity, jitter and shimmer to evaluate the effect of changes in these selected features on the intelligibility deficits and ASR performance.

2.0 RESEARCH BACKGROUND

Several studies [19] [20] [21] [22] have investigated the characteristics of speech features in speakers with speech impairments. Speech features of speech impairments vary with the changes within the vocal organs [20]. The selected features are defined as follows:

- Formant frequencies is the concentration of acoustic energy around a particular frequency in the speech wave [15].
- Fundamental frequency (F0) is the basic vibratory rate of the vocal folds, influenced by the length, mass, and tension of the vocal folds [16].
- Intensity is the amount of energy that is transported past a given area of the medium per unit of time [17].
- Jitter is a measure of the cycle-to-cycle variation of the pitch period. It correlates with the hoarseness in speech [18].
- Shimmer is a measure of variability of the peak-to-peak amplitude of the signal which also correlates with hoarseness in speech [18].

Selection of relevant features is important to understand the characteristics of impaired speech which could potentially influence the performance of ASR systems. The next section presents the speech features studied in available literature in relation to the speech quality and ASR performance.

Table 1. Relationship between the changes of speech features and speech quality

Speech features	Changes in speech	Effect on speech quality
Formant	High / increase	Brighter sound [27]
	Low / decrease	Darker sound [27]
F0	High / increase	Louder sound [15]
	Low / decrease	Softer sound [15]
Jitter	High / increase	Creaky, hoarseness in speech Breathy sound Rough sound [28]
	Low / decrease	Smooth sound [28]
Shimmer	High / increase	Decreasing voice loudness [28] Hoarseness in speech Softer voice
	Low / decrease	Increase voice loudness [28] Louder voice
Intensity	High	Intense, loud sound [15]
	Low / Decrease	Weak, soft sound [15]

Formant frequencies are determined by the shape of the vocal tract. The vocal tract above the larynx is constantly changing shape as once speak, which in turn, changes the quality of the vowel [23]. Impaired speakers face the problems of controlling the tongue movement in obtaining the desired shape of the vocal tract, thus affecting the formant values. Fundamental frequency (F0) and its harmonic components produced by the vocal cords as it vibrates during speech productions [24]. However, the ability of the vocal cords to vibrate diminish due to speech impairment such as dysarthria [11]. When the vocal cord is unable to vibrate properly, it causes problem in speech generation, as well as breathing and swallowing problems [25]. Instability or lack of control in vocal cord vibration increases the jitter [26]. Meanwhile, shimmer is affected due to reduction of glottis resistance and mass lesions in the vocal folds [14]. This will produce creaky, hoarse and breathy sound, as well as limited pitch and loudness variations. Table 1 summarises the findings from literature related to the relationship between the changes of speech features and speech quality.

Several studies were conducted to investigate the effect of speech impairments on the ASR system's performance. Changes of speech features and speech quality degradation in impaired speeches lead to poor ASR performance. Table 2 shows the existing research carried out to determine the effect of the speech impairments on the ASR system's performance. Ferrier et.al [29] have determined the relationship between the speech intelligibility and ASR accuracy whereby high intelligibility leads to high recognition accuracy. There has also been growing interest among the researchers to explore the speech characteristics of impaired speech towards the development of ASR system which can recognize impaired speech. Kain et al. [30], Kain et al. [8] and Rudzidc, [31] modified the speech features of dysarthria to more closely match the non-dysarthric speaker. The study reported that the intelligibility of dysarthric speech can be improved up to 20%.

Table 2. Effect of speech features of impaired speakers on ASR performance

Reference	Database	Features studied	Effect on the ASR performance
Kain, et.al., [30]	English	F0, Formant, Intensity	Dysarthric speech can be modified to improve intelligibility from 68% to 87%.
Kain, et al, [8]	English dysarthric speakers	F0, Formant, Intensity	Improving the intelligibility of dysarthric vowels of one speaker from 48% to 54%
Rudzidc, [31]	TORGO - English dysarthric speakers	Formant, F0	The correction of phoneme errors results in the greatest increase in intelligibility of dysarthric speech

From the literatures, it was found that the changes of speech features in impaired speech influence the speech production quality and the ASR system's performance. However, speech impairment's features influence towards the ASR system's performance leaves an open question as to which features are significant and how they influence the recognition accuracy of impaired speech. The aim of this study is twofold: first, to analyse the speech features of children's impaired speech that cause intelligibility deficits; and second, to understand the effect of speech features that causes intelligibility deficit (as identified in experiment one) on the performance of ASR systems. This study identified the significant differences of speech features of speech impaired children, such as F1, F2, F0, Intensity, Jitter, Shimmer and ASR system performance in comparisons with unimpaired children's speech features.

This paper is organised as follows: Section 2 describes the methods and materials used to carry out this study; Section 3 presents the results of the study, Section 4 the major findings; and finally, Section 5 concludes the study.

3.0 METHODS

In this research, we investigate how the significant speech features of impaired children's speeches influence the recognition accuracy of ASR systems. We have performed statistical analysis to determine the speech features that are significantly different between impaired and unimpaired speeches. We also examine the correlation between these speech features and the recognition accuracy of ASR systems. This section describes the participants, speech database, procedures, statistical analysis and evaluation of the speech features on the recognition accuracy of ASR systems.

3.1 Participants

We have selected 30 children with speech impairment to take part in the recording session from a special school and spastic centre in Petaling Jaya, Kuala Lumpur, Malaysia. The participants comprise 15 males and 15 females with age ranging from eight to 12 years old, with the mean age of ten years. We have screened the participants based on selection criteria, which are: (1) aged between 8 years and 12 years; (2) native Malay speaker; (3) balanced for gender; and (4) able to understand simple instructions. The children were diagnosed with different types of speech impairments. A professional speech language pathologist (SLP) assessed the children and classified the severity of speech impairment. The severity level was measured using the Percentage of Consonant Correct (PCC) from the narrow phonetic transcription [32] [33]. Details of the impaired speakers are shown in Table 3.

Table 3. Details of Impaired Speakers

Speaker	Gender	Age	Diagnosis	PCC (%)	Severity
SIG01	Male	8	Hearing impaired	98	Mild
SIG02	Male	8	Cerebral palsy	47	Severe
SIG03	Male	8	Cerebral palsy	30	Severe
SIG04	Male	9	Hearing impaired	99	Mild
SIG05	Male	9	Hearing impaired	71	Mild-moderate
SIG06	Male	9	Cerebral palsy	26	Severe
SIG07	Male	10	Hearing impaired	67	Mild-moderate
SIG08	Male	10	Hearing impaired	76	Mild-Moderate
SIG09	Male	10	Cerebral palsy	38	Severe
SIG10	Male	11	Hearing impaired	96	Mild
SIG11	Male	11	Cerebral palsy	52	Moderate-Severe
SIG12	Male	11	Hearing impaired	99	Mild
SIG13	Male	12	Hearing impaired	97	Mild
SIG14	Male	12	Cerebral palsy	86	Mild
SIG15	Male	12	Cerebral palsy	51	Moderate-Severe
SIG16	Female	8	Cerebral palsy	63	Moderate-Severe
SIG17	Female	8	Cerebral palsy	47	Severe
SIG18	Female	8	Hearing impaired	71	Mild-Moderate
SIG19	Female	9	Hearing impaired	70	Mild-Moderate
SIG20	Female	9	Hearing impaired	56	Moderate-Severe
SIG21	Female	9	Hearing impaired	92	Mild
SIG22	Female	10	Cerebral palsy	77	Mild-Moderate
SIG23	Female	10	Cerebral palsy	67	Mild-Moderate
SIG24	Female	10	Hearing impaired	70	Mild-Moderate
SIG25	Female	11	Cerebral palsy	62	Moderate-severe
SIG26	Female	11	Cerebral palsy	58	Moderate-Severe
SIG27	Female	11	Hearing impaired	72	Mild-Moderate
SIG28	Female	12	Hearing impaired	92	Mild
SIG29	Female	12	Cerebral palsy	29	Severe
SIG30	Female	12	Cerebral palsy	48	Severe

Apart of 30 impaired children, we have recruited 30 unimpaired children (15 males, 15 females) with a similar age range as controlled group. The selected children were assessed by their teachers to ensure they are good in literacy. The demographic of the unimpaired children are shown in table 4.

Table 4. Number of Male and Female Unimpaired Speakers by Age

Age	Male	Female	Age	Male	Female	Age	Male	Female
8	3	3	10	3	3	12	3	3
9	3	3	11	3	3			

3.2 Speech Corpus

We have prepared 51 short, simple, and meaningful Malay sentences for speech recording to provide sufficient features for analysing the intelligibility deficits. The sentences were constructed after discussions and consultations with the SLPs and teachers to suit the speakers' reading abilities and word familiarity. The use of short sentences is justified due to the children's physical and cognitive impairments, making them easily fatigued, hesitant and tense when they had to utter long or complex sentences.

All children were seated in turn with a desk in front of the recording equipment, individually, and the experimenter sat beside the children to assist them. The lingWAVES Voice Clinic Suite was used to record the speech. The stand microphone of lingWAVES was positioned approximately 4-6 inches from a speaker's mouth. A laptop was used the display the sentences to be read by the children. For children that cannot read, the researcher read the text to the children and asked them to repeat the sentences. The children were asked to utter the selected sentences with three repetitions. The children were encouraged to speak with clear pronunciation as he/she would normally speak. The entire procedure of testing and recording was repeated for each participant. The speech of each candidate was sampled at the rate of 16 kHz in accordance to HTK manual book [34]. Later, the speech samples were normalized to 70 dB using Praat. This is to make sure that all the different sound file were scaled to the approximately the equal loudness. The total amount of impaired speech samples acquired during the whole process was 4,590 utterances in 3.8 hours of speech including silence.

We have also built a speech corpus of unimpaired children as a reference corpus for comparison purpose. For unimpaired children, the recording environment, procedures and speech stimuli were the same with the impaired children. The total amount of speech samples acquired during the whole process was 9,180 utterances in 2.5 hours of recorded speech including silence.

3.3 Instrument: ASR System

For performing the WER analysis on the recognition of the CG and SIG speeches, we have built a standard ASR system based on the HTK toolkit [35]. The baseline acoustic model was trained from 60 unimpaired speakers (CG). We have used the triphone Hidden Markov Models (HMMs) acoustic models to study the effect on the word and phoneme level recognition. In triphone acoustic model, we use crossword triphone HMMs on the unimpaired speech containing 464 tied states with 12 Gaussian mixtures per state. The recognition features used are 39-dimensional mel-frequency Cepstral coefficient (MFCC) vectors. The lexicon size was 133 words and bigram language model were employed. For the test data, we used 765 utterances from CG children and 765 utterances from SIG children as well.

3.4 Data Analysis

This section describes the procedures for performing analysis on the acoustic and ASR system performance.

- *Acoustic Differences between CG and SIG*

We performed the analysis of selected speech features of SIG (speech Impaired Group) and CG (Control Group) speakers focusing only on Malay vowels as depicted in Fig 1.

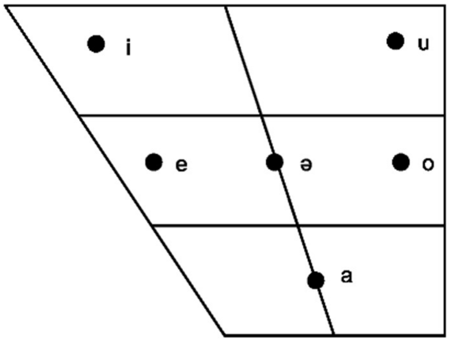

Fig 1. Vowel Phonemes of Standard Malay

The analysis of F1, F2 (measured in Hz), F0 (Hz), intensity (measured in dB), jitter (as a %) and shimmer (as a %) were performed with the 6 Malay vowels /a/, /e/, /i/, /o/, /u/ and /ə/ extracted from the selected short sentences that were uttered by 30 SIG and 30 CG speakers. Six simple semantically meaningful Malay sentences were chosen from the recorded speech sample randomly from the recorded speech as shown in Table 5. Each vowel sound was segmented into 150 milliseconds. The Windows-based version of Praat software was used to perform the acoustic analysis. In total, 540 vowels from 30 SIG and 540 vowels from 30 CG were used for analysis.

Table 5. Sample Selected Sentences and Words for Vowels Extraction

Vowel	Phone	Sentences	Words selected	IPA	Phoneme
/a/	aa	Dia main ***bola*** di padang *He plays with a ball in the field*	Bola (*ball*)	Bola	b-ow-l-**aa**
/e/	ey	***Leher*** zirafah panjang *The giraffe's neck is long*	Leher (*neck*)	Leher	l-ey-h-**ey**-r
/i/	ih	Boboi sedang gosok ***gigi*** *Boboi is brushing his teeth*	Gigi (*teeth*)	Gigi	g-ih-g-**ih**
/o/	ow	Ibu siram ***pokok*** bunga *Mother is watering the flowers*	Pokok (*tree*)	Pokok	p-ow-k-**ow**-k
/u/	uw	***Itu*** gajah *That is an elephant*	Itu (*that*)	Itu	ih-t-**uw**
/ɘ/	er	***Kuda*** lari laju *The horse ran fast*	Kuda *(horse)*	Kudɘ	k-uw-d-**er**

- *ASR System Performance*

The recognition accuracy of ASR system was evaluated using the Word Error Rate (WER) which is the standard measure of error in ASR research. WER is computed as 100(I + D + S) / N, where I, D and S are the total number of phoneme insertions, phoneme deletions, and phoneme substitutions and N is the total number of reference words.

3.4 Statistical Analysis

We have performed statistical analysis to determine the significant group mean differences of F1, F2, F0, intensity, jitter and shimmer as well as WER of ASR system between the CG and SIG using the ANOVA. All analysis was performed using the Windows-based IBM SPSS Statistics version 21. We have classified the subject group independent variables as CG and SIG. The dependent variables are the speech features and WER. We aim to study the effect of one or more group means of speech features on the number of groups; CG and SIG. Specifically, it tests the null hypothesis (H_0):

$$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \cdots \mu_\kappa \tag{1}$$

where μ = group mean and k = number of groups. We begin with the assumption that the H_0 is TRUE, which is there were significant differences in speech features between CG and SIG. Otherwise, we accept the alternative hypothesis (H_a) which is there are at least 2 group means that are significantly different from each other.

To understand the effect of the changes of speech features in SIG speech on the ASR performance, we used the Linear Regression test to determine the correlation between the significant features and WER in SIG speech. We built the regression model for each speech features to predict whether that feature is a significant predictor of WER. Here, the speech features are a set of predictors or independent variables, while the WER is a dependent variable. The linear regression model is to express the dependent variable, y as a linear function of p, predictor variables x_i (i = 1,..., p) and an error term ε [36] :

$$y = c_0 + c_1x_1 + \cdots + c_px_p + \varepsilon \tag{2}$$

The linearity is actually between the dependent variable y and the coefficients c_i. For a set of n data observations x, the linear regression model can be expressed in matrix form [36]:

$$y = cX + e \tag{3}$$

4.0 RESULTS

The acoustic analysis and the word error rate (WER) for impaired speeches (SIG) and unimpaired speeches (CG) reveal several similarities and differences.

- *Formant Frequencies (Hz)*

Table 6. The Mean and SD of F1 and F2 for CG and SIG

Group	Features	Vowels						Overall
		/a/	**/e/**	**/i/**	**/o/**	**/u/**	**/ə/**	mean
CG	F1 (Hz)	825.80 ± 92.51	595.50 ± 75.20	462.60 ± 99.52	647.22 ± 72.29	487.22 ± 64.33	576.56 ± 79.72	599.15 ± 80.59
	F2 (Hz)	1667.43 ± 235.73	2324.59 ± 365.61	2229.95 ± 498.76	1161.32 ± 184.96	1187.86 ± 170.96	1812.15 ± 192.16	1730.55 ± 274.70
SIG	F1 (Hz)	888.94 ± 135.87	658.36 ± 115.74	622.03 ± 138.03	671.93 ± 127.08	618.39 ± 112.20	720.47 ± 140.76	696.69 ± 128.28
	F2 (Hz)	1594.53 ± 231.67	1921.50 ± 372.73	2085.20 ± 378.06	1286.20 ± 223.80	1381.29 ± 205.10	1596.41 ± 205.27	1644.19 ± 269.44

The mean values of F1 and F2 for speeches from CG and SIG are shown in Table 6. The overall means and standard deviations (SD) of F1 for CG, and SIG are 599.15 ± 80.59, and 696.69 ± 128.28, respectively; and the values for F2 are 1,730.55 ± 274.70, and 1,644.19 ± 269.44, respectively. Differences between F1 and F2 were found to be insignificant at p < 0.05, (F = 2.088, p = 0.179; F = 0.131, p = 0.725).

- *Fundamental frequency (pitch) and Intensity (dB)*

Table 7 shows the mean and the SD of F0 and intensity for each group. The overall mean and SD of F0 and intensity for the CG and SIG are 256.04± 41.59, 223.10± 66.03 and intensity for the CG and SIG are 60.58± 6.34, 57.44± 8.04, respectively.

Table 7. Mean and SD of F0 and Intensity

Group	Features	Vowels						Overall
		/a/	**/e/**	**/i/**	**/o/**	**/u/**	**/ə/**	mean
CG	F0	257.56± 37.84	257.09± 46.56	237.85± 51.01	247.20± 44.95	275.60± 28.52	260.96± 40.63	256.04± 41.59
	Intensity	66.26± 5.49	61.54± 6.28	50.78± 5.31	6257± 5.88	60.02± 9.17	62.29± 5.93	60.58± 6.34
SIG	F0	208.51± 66.32	245.82± 56.76	215.94± 72.81	211.92± 59.09	229.79± 64.99	226.60± 76.19	223.10± 66.03
	Intensity	59.60± 7.37	60.21± 7.70	51.79± 8.54	58.42± 7.13	56.37± 7.20	58.23± 10.32	57.44± 8.04

There were significant differences between CG and SDG in F0 at p< 0.05 (F = 18.279, p = 0.002), while intensity was found insignificant difference between CG and SIG at p< 0.05 (F = 1.613, p = 0.233).

- *Jitter and Shimmer*

Table 8 summarises the means of jitter and shimmer for all groups. For the CG, the mean and SD-values of jitter and shimmer are 0.63±0.34, 3.78±1.64and 14.47±5.22, respectively. For the SIG, the mean and SD-values of jitter and shimmer are 1.78±1.43, 8.78±4.53 and 11.90±4.98, respectively.

Table 8. The Mean and SD-Values of Jitter and Shimmer for the CG and SIG

Group	Features	Vowel						Overall
		/a/	/e/	/i/	/o/	/u/	/ə/	mean
CG	Jitter (%)	0.48±0.32	0.51±0.40	0.54±0.27	0.65±0.31	0.87±0.40	0.72±0.4	0.63±0.34
	Shimmer (%)	3.04±1.34	3.67±1.76	4.25±1.90	4.02±1.87	3.88±1.50	3.79±1.5	3.78±1.64
SIG	Jitter (%)	1.55±1.41	1.37±0.91	2.18±1.51	2.00±1.71	1.72±1.41	1.8±1.6	1.78±1.43
	Shimmer (%)	8.48±4.56	7.25±3.58	9.34±4.33	9.73±4.30	8.11±4.71	9.8±5.7	8.78±4.53

Jitter and shimmer values for the CG are much lower compared with the SIG. It shows that fewer perturbation values are found in the speech of normal children compared with the speech of speech impaired children. There are significant differences between the ratings of the CG and SIG in jitter (F = 71.894 p = 0.000) and shimmer, (F= 125.830, p = 0.000) at p < 0.050.

- *Differences between CG and SIG*

Table 9 concludes the differences between CG and SIG for each speech feature.

Table 9. Differences between CG and SIG for each Feature

Features	CG	SIG	Mean difference	p-value
F1	599.15	696.69	97.54	0.179
F2	1730.55	1644.19	86.36	0.725
F0	256.04	223.10	32.94	0.002**
Intensity	60.58	57.44	3.14	0.233
Jitter	0.63	1.78	1.15	0.000**
Shimmer	3.78	8.78	5.00	0.000**

*p<0.05, **p < 0.00

There is a high increment of Jitter and Shimmer in SG, which are 1.15% (182%) and 5.00% (132%), respectively. F1 for SIG increases 97.54Hz (16%). Meanwhile, F0, F2 and intensity reduce in SIG speech, 86.36Hz (5% reduction), 32.94Hz (13% reduction) and 3.14 (5% reduction) respectively. Overall, we can conclude that there are some differences in speech feature values between the SIG and CG.

3.2 ASR system Performance

Fig. 2 shows the difference of WER where SIG was 50.03% higher than CG. There is a strong negative correlation, which the WER increase with the decrease of PCC value or the degree of severity impairment. The correlation between severity of impairments with WER is significant at p<0.05, (r =-0.95, p= 0.00). It was found that the WER for SIG speakers increase with the decrease of intelligibility in speech.

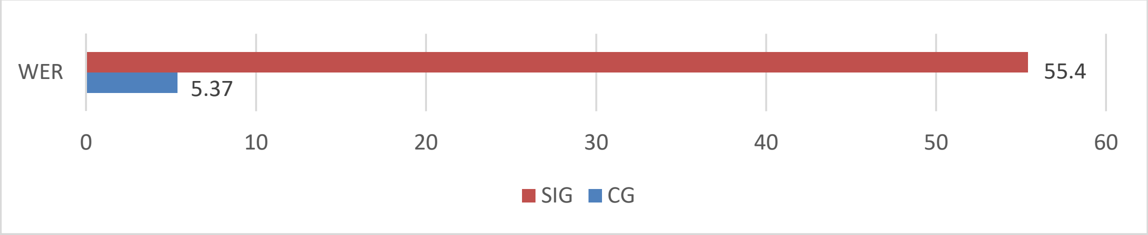

Fig. 2: Comparison of WER between SIG and CG

From the acoustic analysis, F0, jitter and shimmer were shown to be the significant features that contribute to the intelligibility deficits in impaired speech. Fig. 3 shows the effect of F0, jitter and shimmer on WER. All features were found to be insignificant predictors of WER. Decreased F0, increased in jitter and shimmer were associated with the increase in WER. As shown in Fig. 3(a), increase in F0 reduces the WER, indicating that high pitch actually reduces the WER. On the other hand, increase in Jitter (Fig. 3b) and shimmer (Fig. 3c) resulted in higher WER when recognizing impaired speech.

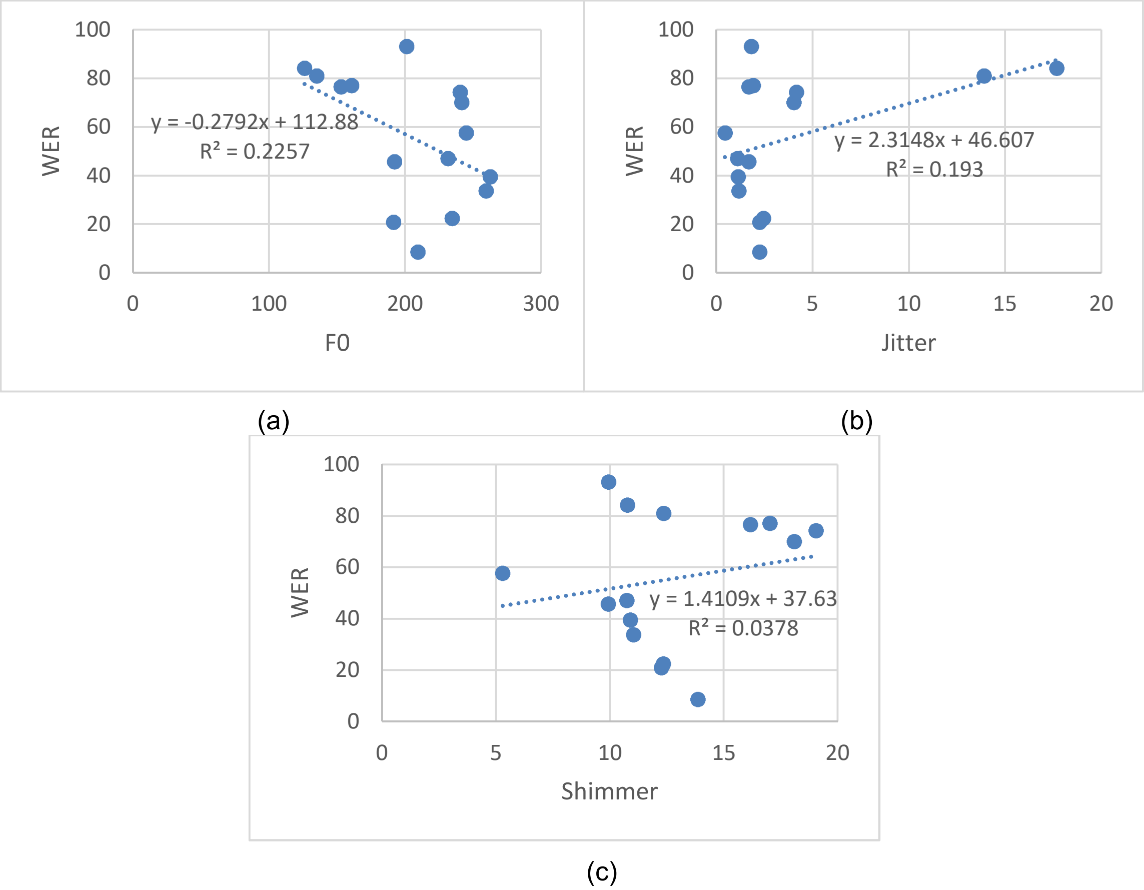

Fig. 3: Effect of Speech Features (a) F0 (b) Jitter (c) Shimmer on WER

The variance in WER can be explained by F0 (22.6%), Jitter (19.3%) and Shimmer (3.8%). Among the three, F0 was found to explain more on the variation in WER. From the analysis F0, Jitter and shimmer were found to be significant at p <0.05. Table 10 summarises the results.

Table 10. Correlation and Coefficient of Determination of the Three Features

Feature	R	R^2	F	p
F0	0.475	0.226	4.972	0.031
Jitter	0.439	0.193	4.811	0.034
Shimmer	0.195	0.038	4.214	0.048

We have also measured the recognition accuracy of each individual vowel phoneme. Fig. 4 shows that the recognition accuracy for the SIG are consistently lower for all phonemes as compared to CG. We observed that recognition of /e/ and /u/ in SIG have significant decrement and most affected with 50% drop in recognition rates.

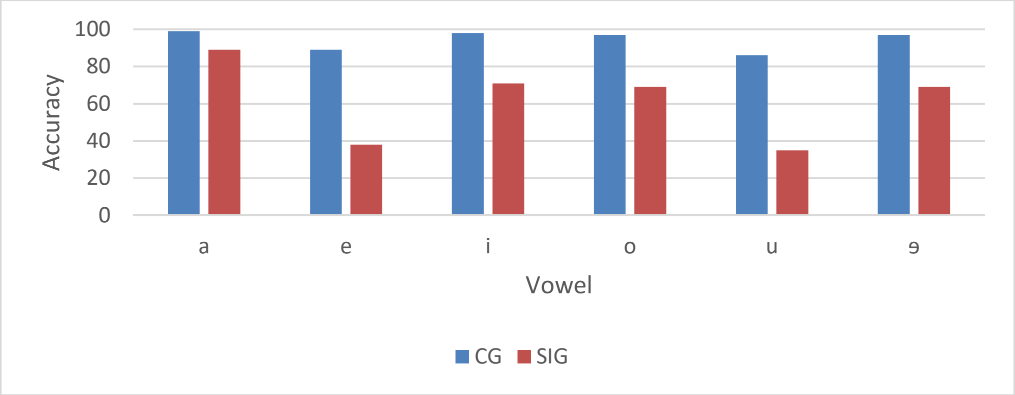

Fig. 4: Phoneme Recognition

We also analysed the effect of F0, jitter and shimmer on each vowel phoneme. We have observed that all the features of F0, jitter and shimmer are significant predictors of vowel /e/ accuracy at p<0.05. However for vowel /u/, only F0 was found to be the significant predictor for error rate.

Table 11. Linear regression of each phoneme

Vowel	**Features**	**Results**
/a/	F0	R = 0.369, R square = 0.136, F = 2.051, p = 0.176
	Jitter	R = 0.414, R square = 0.171, F = 1.238, p = 0.324
	Shimmer	R = 0.422, R square =0.178 , F = 0.795, p = 0.522
/e/	F0	R = 0.691, R square = 0.477, F = 11.878, p = 0.004*
	Jitter	R = 0.696, R square = 0.484, F = 5.635, p = 0.019
	Shimmer	R = 0.782, R square = 0.612, F = 5.774, p = 0.013
/i/	F0	R = 0.270, R square = 0.073, F = 1.025, p =0.330
	Jitter	R = 0.274, R square = 0.075, F = 0.486, p = 0.627
	Shimmer	R = 0383, R square = 0.146, F = 0.629, p = 0.611
/o/	F0	R = 0.297, R square = 0.88, F = 1.255, p = 0.283
	Jitter	R = 0.328, R square = 0.108, F = 0.726, p = 0.504
	Shimmer	R = 0.347, R square = 0.121, F = 0.503, p = 0.688
/u/	F0	R = 0.552, R square = 0.304, F = 5.689, p = 0.033*
	Jitter	R = 0.555, R square = 0.308, F = 2.669, p = 0.110
	Shimmer	R = 0.607, R square =0.369 , F = 2.140, p = 0.153
/ɘ/	F0	R = 0.279, R square = 0.078, F = 1.096, p = 0.314
	Jitter	R = 0.333, R square = 0.111, F = 0.748, p = 0.494
	Shimmer	R = 0.379, R square = ,0.143 F = 0.614, p = 0.620

* Significant predictors

5.0 DISCUSSION

5.1 The significance features of impaired speech that contribute to low speech intelligibility

We have identified the speech features that contribute to the intelligibility deficits in impaired speech among children. F0, jitter and shimmer were found to show significant differences in impaired speech.

- F0

In this study, the statistical analysis shows that there is significant differences in F0 between the CG and SIG. The results of the F0 reduction in the SIG are similar with [37], where F0 tend to be lower for impaired speech. However, some studies have reported that there is no significant difference in F0 decrement in impaired speech [21]. This is because individuals with speech impairment can still control some prosodic features in their speech, even though they lose intelligibility in the vowel production [21] [10].

- Jitter and Shimmer

SIG speakers have higher jitter and shimmer compared to CG. The statistical analysis shows that there is a significant difference in jitter and shimmer between the CG and SIG, which is similar to the work in [38]. However, our findings contradict those of other studies [14] which claimed that there are no differences in jitter and shimmer between impaired and non-impaired children. This is because the speech impaired children did not present any 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 reflect the increased value of jitter [14]. As to shimmer, the existing study indicates that the characteristics such as reduction of glottic resistance, vocal fold mass lesions and greater noise at production, are some of the factors that influence shimmer values [14], were found to be irrelevant in this research.

Table 12. Comparison of Findings in Acoustic Analysis of Impaired Speech

Authors	Features studied					
	Formant	F0	Intensity	Jitter	Shimmer	Duration
In this research	Not significant	Significant	Not significant	Significant	Significant	-
White, [22]	-	Not significant	-	Not significant	Not significant	-
Saz et. al., [21]	Significant	Not significant	Not significant	-	-	Not significant
Jeng et al., [37]	-	Significant	-	-	-	-
Wertzner et. al, [14]	-	Significant for vowel /e/	-	Not significant	Not significant	-
Hartl et al., [38]	-	-	-	Significant	Significant	-

5.2 The Effect of Impaired Speech on ASR System's Performance

We have identified the effect of the significant speech features on ASR system's recognition accuracy. From Table 10, it was found that the F0 has the highest positive correlation with the WER, followed by Jitter, while shimmer did not show any significant correlation with the WER. We also found that the variation in F0 explain 22.6% variation in WER. The three features were found to be very significant for vowel /e/, but not for other vowels although F0 has a significant influence on the accuracy of vowel /u/. Another reason why vowel /e/ has high positive correlation is because the vowel is pronounced with high pitch. F0 is a significant indicator as excessive variations in F0 may resemble excess patterns of word stress, and may accentuate general problems with speech timing. This could be due to the inability of impaired speakers to control the tongue movements where it plays an important role in F0 deviation. Vowel /e/ involves the movement of middle tongue and /u/ involves the high tongue which mainly depend on the jaw and tongue movement in producing speech. Based on these results, impaired speakers face problem in producing precise articulation and phoneme that related to jaw and tongue movement, and this greatly influence the ASR accuracy.

6.0 CONCLUSION

This study has analysed the speech features of children's impaired speech. We have performed acoustic and WER analysis to understand the differences in ASR performance on impaired and unimpaired speech among children. We have also investigated the effect of significant speech features on the ASR performance. The acoustic analysis revealed that F0, jitter and shimmer are significant features in contributing to the low intelligibility of impaired speech. However, these features do not affect the ASR performance significantly. Phoneme accuracy results show that /e/ and /u/ are more affected than other phoneme as a result of speech impairments. The findings from this research provide a better understanding of impairment related changes in speech, which can contribute to the further improvements of ASR based speech assistive tools in clinical studies for individuals with speech impairments.

As the focus of the current research is on speeches of speech impaired children with dysarthria, there are several research that can be undertaken in the future. First of all, similar research could be carried out on other types of speech impairment such as apraxia. Research can also be carried out to determine speech features of speech impaired adults that influence the recognition accuracy of ASR systems.

Acknowledgment

This research is supported by UM High Impact Research Grant UM-MOHE UM.C/HIR/MOHE/FCSIT/05 from the Ministry of Higher Education Malaysia and University of Malaya Research Grant (UMRG) Grant no.: RG284-14AFR.

REFERENCES

[1] M, Shahin et al., "Tabby Talks: An automated tool for the assessment of childhood apraxia of speech", Speech Communication, vol. 70, 2015, pp. 49–64.

[2] D. Sztahó et al., "A Computer-Assisted Prosody Pronunciation Teaching System", *In Proceedings of the 4th Workshop on Child Computer Interaction (WOCCI)*, Singapore, 2014.

[3] O. Saz et al., "Tools and Technologies for Computer-Aided Speech and Language Therapy" Speech Communication, vol. 51, no. 10, 2009b, pp. 948-967.

[4] O. Bater et al., "Wizard-of-Oz Test of ARTUR – A Computer-Based Speech Training System with Articulation Correction", *In Proceedings of the Conference on Computers and Accessibility,* Baltimore, 2005, pp. 36–43.

[5] Bacha Rehmam, Zahid Halim, Ghulam Abbas & Tufail Muhammad (2015), Artificial Neural Network-Based Speech Recognition Using DWT Analysis Applied On Isolated Words From Oriental Languages, Malaysian Journal of Computer Science, 28 (3): 242 – 262.

[6] C. Coleman and L. Meyers, "Computer recognition of the speech of adults with cerebral palsy and dysarthria", Augmentative Alternative Communication, vol. 7, no. 1, 1991, pp. 34–42.

[7] K. Hux et al., "Accuracy of three speech recognition systems: Case study of dysarthric speech", Augmentative Alternative Communication, vol. 16, no. 3, 2000, pp. 186–196.

[8] A. Kain et al., "Improving the Intelligibility of Dysarthric Speech", Speech Communication, vol. 49, no. 9, 2007, pp. 743-759.

[9] T.M. Leszcz, "The effect of multi talker background noise on speech intelligibility in Parkinson's disease and controls", Master's thesis. The School of Graduate and Postdoctoral Studies. The University of Western Ontario, 2012.

[10] R. Patel, “Phonatory control in adults with cerebral palsy and severe dysarthria”, Augmentative Alternative Communication, vol. 18, no. 1, 2002, pp. 2-10.

[11] F.L. Darley et al., “Differential diagnostic pattern of dysarthria”, Journal of Speech Hearing Research, vol. 12, 1969, pp. 246-269.

[12] L. Nicolosi et al, Terminology of Communication Disorders: Speech-Language-Hearing, Ed. 5, 2004, Philadelphia.

[13] S.B. Davis, Acoustic characteristics of normal and pathological voices. 1978, Report Haskin Laboratories.

[14] H.F. Wertzner et al, “Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders”, Brazilian Journal of Otorhinolaryngology, vol. 71, no. 5, 2005, pp. 582-588.

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

[16] L.M. Justice, Communication Sciences and Disorders: An Introduction. 2006, Upper Saddle River, New Jersey, Pearson.

[17] S. Rosen et al., Signals and Systems for Speech and Hearing, Ed. 2, 2011, BRILL

[18] R. Vipperla et al, “Ageing voices: The effect of changes in voice parameters on ASR performance”, EURASIP Journal on Audio, Speech, and Music Processing, vol. 2010, no. 5, 2010.

[19] Y. Ikui et al., “Acoustic characteristics of ataxic speech in Japanese patients with Spinocerebellar Degeneration (SCD)”, International Journal of Language Communication Disorders, vol. 47, no. 1, 2011, pp. 84-94.

[20] G. Niedzielska, “Acoustic analysis in the diagnosis of voice disorders in children”, International Journal of Pediatric Otorhinolaryngology.; vol. 57, no. 3, 2001, pp. 189–193.

[21] O. Saz et al., “Analysis of acoustic features in speakers with cognitive disorders and speech impairments,” EURASIP Journal on Advances in Signal Processing, 2009a.

[22] K. White, “Acoustic characteristics of Ataxic Dysarthria”, Honors Thesis, The Department of Speech, Language and Hearing Sciences, University of Florida, 2012.

[23] N. Ashley, “Fundamental Frequency, Harmonics, and Formant Frequencies. The Ohio State University”, Retrieved from http://underlingsosu.wordpress.com/tag/formant-frequency/

[24] S. Lemmetty, “Review of Speech Synthesis Technology”, Master’s thesis, Helsinki University of Technology, 1999.

[25] American Speech Language and Hearing Association (ASHA), Retrieved from http://www.asha.org/public/speech/disorders/vfparalysis/

[26] K.A. Wilcox et al., “Age and changes in vocal jitter”, Journal of Gerontology, vol. 35, no. 2, 1980, pp. 194–198.

[27] R. Parncutt, The Science and Psychology of Music Performance: Creative Strategies for Teaching and Learning: Creative Strategies for Teaching and Learning, 2002, Oxford University Press.

[28] Voice Academy, University of Iowa. Voice Academy Glossary (VAG), Retrieved from http://www.uiowa.edu/~shcvoice/glossary.html

[29] F. Ferrier et al., "Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition", Augmentative Alternative Communication, vol. 11, no. 3, 1995, pp. 165–175.

[30] A. Kain et al., "Intelligibility of modifications to dysarthric speech", *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[31] F. Rudzicz, "Acoustic transformations to improve the intelligibility of dysarthric speech", *In Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, 2011, pp 11–21.

[32] L.D. Shriberg et al., "Phonological Disorders II: A conceptual framework for management", Journal of speech and Hearing Disorders, vol. 47, 1982a, pp. 242-256.

[33] L.D Shriberg et al., "Phonological Disorders III: A Procedure for Assessing Severity of Involvement", Journal of speech and Hearing Disorders, vol. 47, 1982b, p. 256-270.

[34] S. Young et al., The HTK Book (for HTK Version 3.4), Retrieved from http://htk.eng.cam.ac.uk/docs/docs.shtml

[35] Cambridge University Engineering Department (CUED), "The Hidden Markov Model Toolkit (HTK) version 3.4.1", Retrieved from http://htk.eng.cam.ac.uk/

[36] G. Hu et al., "Multivariate Regression Modelling for Home Value Estimates with Evaluation Using Maximum Information Coefficient", Studies in Computational Intelligence, vol. 443, no. 2013, 2013, pp. 69-81.

[37] J.Y. Jeng et al., "Production and perception of mandarin tone in adults with cerebral palsy", Clinical Linguistics & Phonetics, vol. 20, no. 1, 2006, pp. 67–87.

[38] D.A. Hartl et al., "Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia", Eur Arch Otorhinolaryngology, vol. 260, no. 4, 2003, pp. 175-82.