AUGMENTATIVE AND ALTERNATIVE COMMUNICATION METHOD BASED ON TONGUE CLICKING FOR MUTE DISABILITIES

MUHAMMAD AMIRUL AMIN AZMI, NIK NUR WAHIDAH NIK HASHIM* AND HAZLINA MD. YUSOF

Department of Mechatronics Engineering, Faculty of Engineering,
International Islamic University Malaysia, PO Box 10,
Kuala Lumpur 50728, Malaysia.

**Corresponding author: niknurwahidah@iium.edu.my*

(Received:22nd Nov 2018; Accepted:22nd Jan 2019; Published on-line: 1st June 2019)

https://doi.org/10.31436/iiumej.v20i1.1021

ABSTRACT: This paper presents a pilot study for a novel application of converting tongue clicking sound to words for people with the inability to speak. 15 features of speech that are related to speech timing patterns, amplitude modulation, zero crossing and peak detection were extracted. The experiments were conducted with three different patterns using binary Support Vector Machine (SVM) classification with 10 recordings as training data and 10 recordings as development data. Peak size outperformed all features with 85% classification rate for pattern P1-P3 whereas multiple features produced 100% classification rate for P1-P2 and P2-P3. A GUI based system was developed to validate the trained classifier. Multiclass SVM were constructed based on the best features obtained from binary SVM classification outcome, namely peak size and skewness amplitude modulation, and then tested on 15 recordings. The GUI based multiclass SVM obtained a satisfying performance of 67% correct classification of the test data set.

***ABSTRAK*:** Kertas ini membentangkan panduan kajian kepada aplikasi terkini dalam menukar bunyi klik pada lidah kepada perkataan untuk orang yang mempunyai kehilangan upaya dalam bertutur. 15 ciri khas berkaitan pertuturan adalah pola masa, modulasi nilai tertinggi, tiada titik persilangan dan nilai terpilih yang dikesan. Eksperimen telah dijalankan dengan tiga corak berlainan menggunakan perduaan Mesin Vektor Sokongan (SVM) klasifikasi dengan 10 rakaman sebagai data terlatih dan 10 rakaman sebagai data yang dibina. Saiz tertinggi yang melebihi semua ciri-ciri pada 85% kadar klasifikasi dilihat pada corak P1-P3, sedangkan ciri-ciri pelbagai telah terhasil pada 100% kadar klasifikasi P1-P2 dan P2-P3. Sistem berdasarkan GUI telah dibina bagi menilai ciri terlatih. Kelas pelbagai SVM telah dibina berdasarkan ciri-ciri terbaik dan dihasilkan daripada klasifikasi perduaan SVM, iaitu saiz tertinggi dan modulasi saiz tertinggi tidak linear, dan telah diuji dengan 15 rakaman. Kelas pelbagai SVM yang didapati melalui GUI ini adalah memberangsangkan iaitu 67% klasifikasi adalah tepat pada set data yang diuji.

KEYWORDS: *mute; speech; alternative communication; classification; features*

1. INTRODUCTION

Normally, people who suffer from hearing disability will also experience the inability to talk (mute). According to the world federation of the deaf, it is estimated that there are 70 million people with hearing and mute disabilities in the world [1]. In Malaysia, according to the Ministry of Women, Family and Community Development, the number of people with hearing and mute disabilities is around 3,771 people [2].

In the context of face-to-face conversation, there are several conventional methods used by people with hearing and mute disabilities to communicate with the public, such as American Sign Language (ASL), mobile applications, and writing or typing. ASL is a complex communication method used by the deaf and mute people through hand and body gestures. The disadvantage of ASL is that the resulting signals are often different from one speaker and another due to different backgrounds of hearing loss. The use of sign language also limits the transfer of information due to the components required to use sign language being limited to hand gestures, arm gestures, and facial expressions. The use of sign language is also limited in dark conditions as the receiver cannot see any movement of the signal using the finger or hand clearly, in turn, further affecting the smoothness of the communication being carried out. Although mobile applications are considered to be a popular alternative communication method, the majority of the population that suffers the disabilities cannot afford to own a decent smart phone. Transmission of information using a mobile application will not be as efficient as a normal conversation when communicating face to face. Similarly, communication through typing and writing are not effective due to limited transmission of information at one time and the inability to communicate at farther distances or remotely.

The aim of this project is to convert tongue clicking sounds to speech. Each and every physiological characteristic of human beings is unique and will never be identical to one another, this fact leads to the foundational hypothesis of this research, which is that each tongue clicking sounds are unique in their own way. This study aimed to achieve the following objective. First, to identify the features in tongue clicking sounds that could distinguish different patterns. Second, to classify the different tongue clicking sounds and associate them with a particular speech. Third, to develop a system that converts tongue clicking sounds to speech. Based on these objectives, the following question arises: what is the best feature to use for a speaker recognition application? This question will be explored further throughout this research.

This research will not only benefit the deaf and mute but also benefits the people who will be communicating with them. Companies or organizations will be more confident to rely on the deaf and mute in certain areas and it also encourages them to actively engage with the surrounding communities. This proves that communication aids to people with disabilities are indispensable and will greatly benefit everyone if realized.

2. PREVIOUS WORK

People with conditions related to disorders of motor speech control and language disorders due to brain injury require some kind of Augmentative and Alternative Communication (AAC). However, depending on the type of disorders and the level of physical and cognitive capabilities, the use of AAC varies from one person to another. The development of AAC is currently at the stage of producing, at most, 25 words per minute with inefficient response time. This section briefly reviews the existing method of AAC.

2.1 ASL to Speech Conversion

In work by [9], the idea to translate finger spelling (sign) to speech using recognition and synthesis technique was proposed. Two modules will be used in this project to make sure that a finger spelling (gesture) recognition module and a text-to-speech synthesis module can perform properly.

This research identified that the sensor-based system is more accurate than the vision-based system. The sensor-based system only uses tactile sensors, flex sensors, and an accelerometer that significantly lower the cost as compared to the vision-based system that requires a camera with image processing as a portable device. The sensor-based system also provides a fast response in recognizing the gestures and thus reduces the computational time in real-time applications.

2.2 ASL Identification using Microsoft Kinect

The research by [7] incorporates Microsoft Kinect as a vision-based approach for utilizing a picture processing technique as well as an additionally implanted framework-based approach where a sensor is placed on a hand glove for acknowledgment of the hand motions. They describe the process as independent of lighting conditions, which means that the image can be captured in full light or low light due to the processing data coming from the depth sensor, which is used for skeletal data.

2.3 Breathing to Speech Conversion

A breath control device was first initiated by [3] at the Google Science Fair where a prototype using an arduino microcontroller was developped. The device translates breathing into electrical signals using a MEMS microphone where it detects varying intensity level or timing patterns of the user's exhale. These variations were represented as dots and dashes that were then associated with the Morse code to spell out words that were synthesized into voice.

Another study in [4] designed a device specifically for patients on ventilator support with the aim to support those with severe speech impairment and loss of motor functionality. The device captured breathing patterns and extracted features such as frequency variations, intensity, and phase. These feature patterns were then used to train the system and once recognized, each pattern was associated with synthesized words and phrases. The device was tested on seven healthy people with no speech impairment with each person producing 10 repetitions of three unique breathing patterns. The device was reported to have a mean reliability of 90% with the range of values being between 73% and 100%.

Researchers in [10] introduced communication by breathing device driven by a belt-mounted breath-mouse that was improvised based on the Dasher mouse gesture language. Dasher is an alphabetical library based on the language model that predicts the probability of each letter in a given context. It allows a user to write any sentence using a mouse, touchscreen, or gaze tracker as the steering command. Novice Dasher users were reported to write 6.0 ± 1.3 words per minute with an average of 2.0% error in spelling.

2.4 Eye Blink-Based Alternative Communication

A device that converts eye blinks to English alphabet Morse code was developed by [5]. The device used an IR led-sensor module attached to an eyeglass to recognize eye blinks. The sensor detected two levels of signals based on the amount of reflected light when the eyes were closed and open. Measurements were fed into an Arduino ATMEGA 328P. Short blinks were translated into dots and longer blinks were translated into dashes. The device took an average of three seconds to convert the blinking patterns to an English alphabet Morse code.

3. METHODOLOGY

3.1 Data Collection and Pre-Processing

As an initial investigation, we would like to see if it is possible to distinguish between different tongue clicking patterns. Thus, for database A, five different tongue clicking patterns from a single person were recorded using a TASCAM DR-05 with a 44.1 kHz sampling rate. The patterns differed based on the number of clicks but with consistent pauses. For example, pattern one is one clicking, pattern two is a double click, and so on. We collected 30 recordings for training data and 30 recordings for development data to validate the classifiers trained using the training data. To further validate the classifier, we collected 15 recordings from 5 different people where each recorded three patterns of tongue clicking. Table 1 shows the number of recorded tongue clicking patterns. The recordings were made in a closed room with each sound being recorded within 10 to 40 seconds depending on the tongue clicking patterns.

Table 1: Database for collected tongue clicking patterns

Clicking pattern	**Number of clicks**	**Number of training data**	**Number of development data**	**Number of test data**
Pattern1 (P1)	4	10	10	5
Pattern2 (P2)	1	10	10	5
Pattern3 (P3)	5	10	10	5

Figure 1 shows the recorded tongue clicking sound for five different patterns. Speech signals were edited manually using Audacity and saved as a separate .wav files labeled P1_1 to P1_10 and repeated for all P2 and P3 recordings. In the pre-processing stage, each speech signal can be decomposed in the form of voiced, unvoiced, and silence segments. Each signal was then divided into 40 ms frames to satisfy quasi-periodic analysis where each frame was determined to be either a voiced, unvoiced, or silence segment using the method in [8].

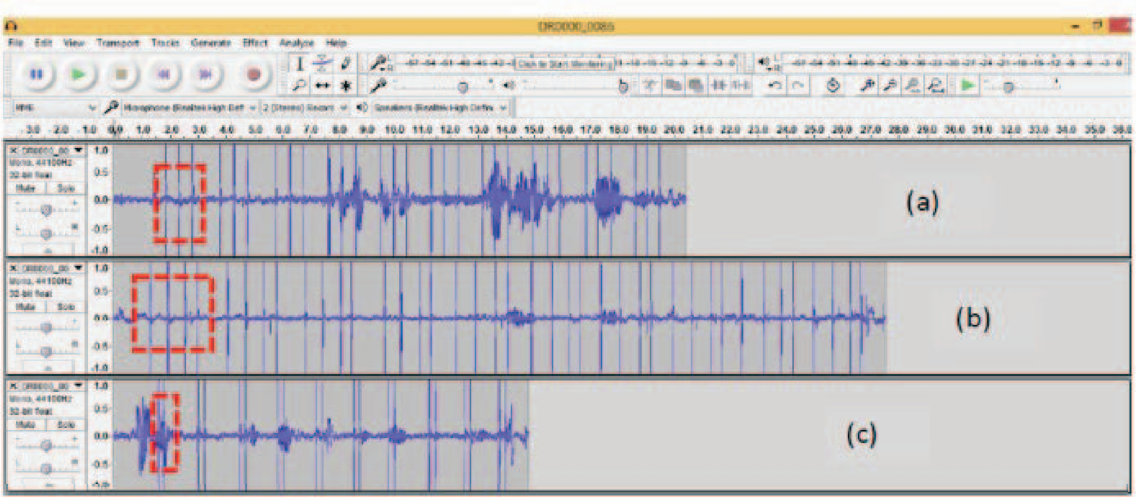


Fig. 1: Tongue clicking signals for each pattern. (a) P1, (b) P2, and (c) P3.

3.2 Feature Extraction

A total of 15 features of speech that are related to speech timing patterns, amplitude modulation, zero crossing and peak detection were extracted.

a. Speech Timing Pattern (Transition Matrix)

Audio consisting of voiced, unvoiced, and silence segments were marked as 1, 2, and 3, respectively. The variation happens during the timing pattern of the speech. Here, the variations were captured in in the form of transition from one state to another state. Interestingly, these states were interchangeable depending on the set of probabilities concerning the states. The probability is simply estimated using a discrete-time Markov process as further discussed in [6]. The transition matrix elements were sequenced into one row vector {t11, t12, t13, t21, t22, t23, t31, t32, t33}. For example, t12 represents the probability of transitioning from the voiced state to unvoiced state within one signal. Only pause and voiced transition parameters, which are {t11, t13, t31 and t33}, were used hereafter due to the fact that they are most related to the tongue clicking sound.

b. Root Mean Square Amplitude Modulation (RMS AM)

The envelope of the waveform was determined based on the analysis of the root mean square amplitude modulation, also known as 'the square-law envelope detector', as shown in Fig. 2. The input signal was squared and sent through an averaging represented by a low-pass filter. An averaging is a crude low-pass filter with a gain of 1. By squaring the signal, the input signal was demodulated using itself as the carrier wave. The square root was then taken in order to reverse the scaling distortion from squaring the signal and to characterize a more accurate statistical measure. Seven RMS AM statistical measurements were extracted for each signal which included minimum, maximum, range, variation, average, skewness, kurtosis, and coefficient of variation.

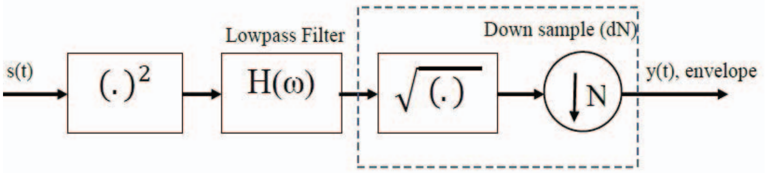


Fig. 2: Block diagram representing the square-law envelope detector.

c. Zero Crossing Rate

Zero crossing rate calculates the value of times in a given time interval/frame that the amplitude of the speech signal crosses through a value of zero.

d. Absolute Peak Detection

The MATLAB built-in function for peak detection produced peak detection for both positive and negative peaks as demonstrated in Fig. 3. For each signal, two features were gathered. The peak size represents the total absolute peak detection and peak average represents the average absolute peak.

In summary, the 15 features were:

- 4 features from transition parameters {t11, t13,, t31, t33}
- 7 features from root mean squared amplitude modulation {minimum, maximum, range, variation, average, skewness, kurtosis, and coefficient of variation}
- Zero-crossing rate
- Peak average and peak size

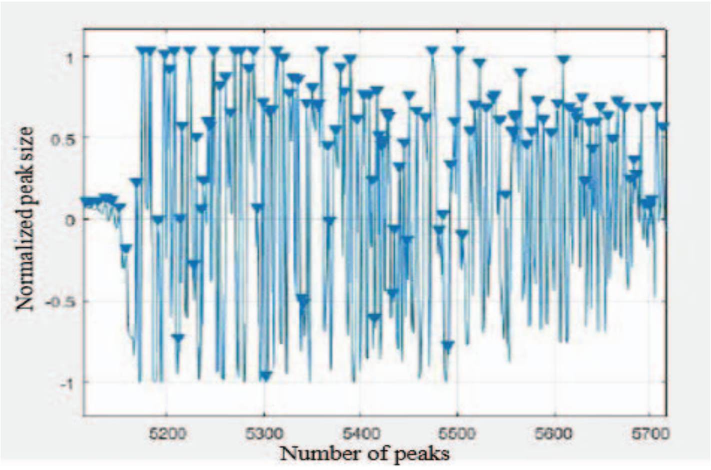


Fig. 3: Example of peak detection for one tongue clicking pattern.

3.3 Construction of Multiclass SVM using Binary SVM

The analysis of classification for Database A were conducted using the built-in binary SVM function in MATLAB (svmtrain) that accepts two parameters, training data and class label, to identify the optimum feature combination that would produce the highest pairwise classification rate. The binary SVM classification was then used to construct a multiclass SVM classifier to classify between three tongue clicking patterns of P1, P2 and P3. Figure 4 shows the block diagram for the constructed multiclass SVM using a binary SVM classifier.

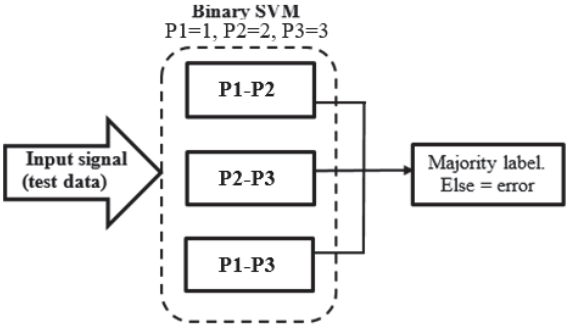


Fig. 4: Construction of multiclass-SVM using binary-SVM.

Each binary classifier used features that were shown to have the highest percentage when trained and tested using database A. The features selected were peak size and skewness. The result produced either output P1 (labeled=1), P2 (labeled=2), or P3 (labeled=3). Then the majority of the output of the three binary classifiers was calculated. Each pattern was then associated with a certain sentence and speech sound such as ‘hello’ for P1, ‘good morning’ for P2, and ‘nice to meet you’ for P3.

4. RESULTS AND DISCUSSION

4.1 Analysis Using Training and Development Data

The binary SVM classifier was trained using the training data and the performance using the development data was analyzed. The best combination of features was limited to only one and two features in order to prevent over-modelling with the small number of samples. Table 2 shows the result of the best feature for one combination binary-SVM classification for P1-P2, P2-P3 and P1-P3 pairwise analysis.

Table 2: Binary-SVM percentage of correctness for one feature classification

Class	Classification %	Features
P1	100	Peak size, t31, Max AM, Range AM, Skewness AM, Kurtosis AM
P2	100	
(%ALL)	100	
P2	70	Peak size
P3	100	
(%ALL)	85	
P1	100	Peak size, Peak average, Max AM, Range AM, Kurtosis AM
P3	100	
(%ALL)	100	

For the two-feature combinations, due to a high classification percentage that produced a greater than 90% correct classification, the results are represented in the form of highest frequency of each feature as a contributor to the high percentage of correctness. Figures 5, 6 and 7 show the results for each pairwise classification of P1-P2, P2-P3 and P1-P3.

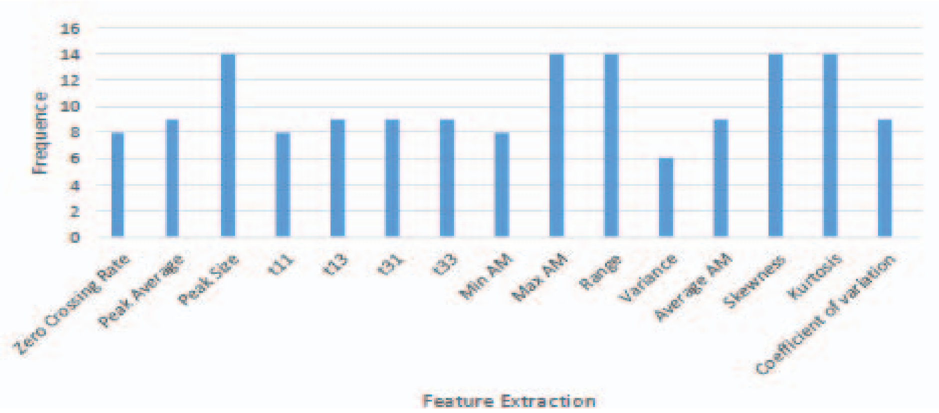


Fig. 5: Result for each feature contribution to the highest two-feature combination for P1-P2.

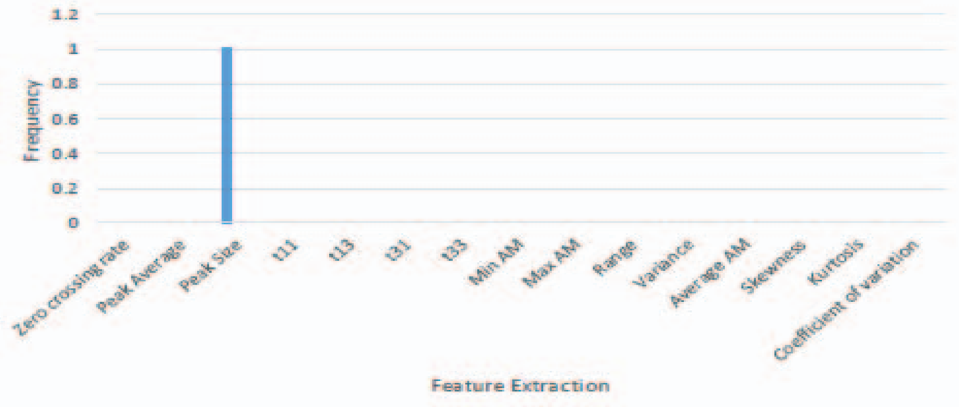


Fig. 6: Result for each feature contribution to the highest two-feature combination for P2-P3.

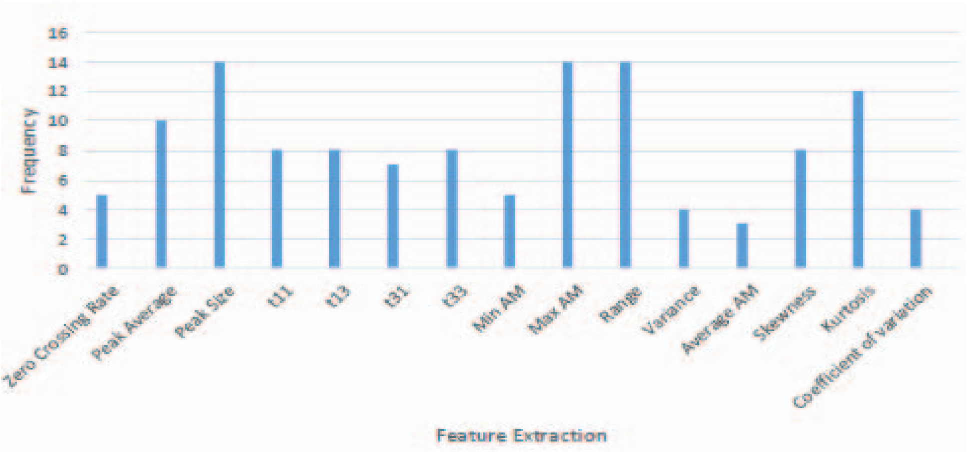
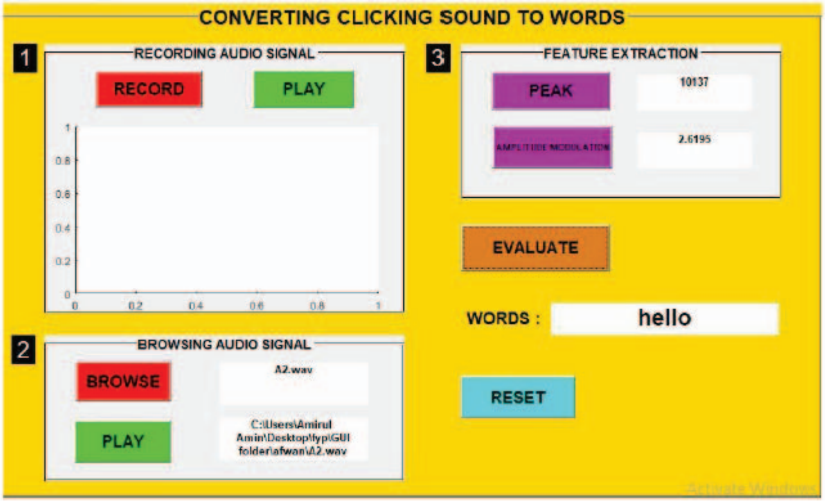


Fig. 7: Result for each feature contribution to the highest two-feature combination for P1-P3.

The X-axis represents the 15 features extracted and Y-axis represents the frequency of occurrence for the feature to contribute the highest percentage of binary-SVM classification (90% and above) when combined with the other features.

4.2 Analysis Using Testing Data

After performing the analysis for one- and two-feature combinations using the binary SVM classification, the best two features that were chosen for use in the multiclass SVM classification were peak size and skewness. For differentiating P2-P3 and P1-P3, peak size was used. Peak size was observed to produce good results when analyzed using the binary SVM classification, whether by itself or combined with another for two-feature combination, in order to differentiate the pattern of clicking sound. For differentiating P1-P2, the combination of peak size and skewness was used. These features were randomly chosen among the other features that produce good classification for creating the GUI. In this part, tongue clicking patterns produced by five subjects were used as a testing input. Results for the test classification are shown in Fig. 9(a) and 9(b). A Graphical User Interface (GUI) was developed for the conversion of tongue clicking sound to words as shown in Fig. 8.



10137

2.6195

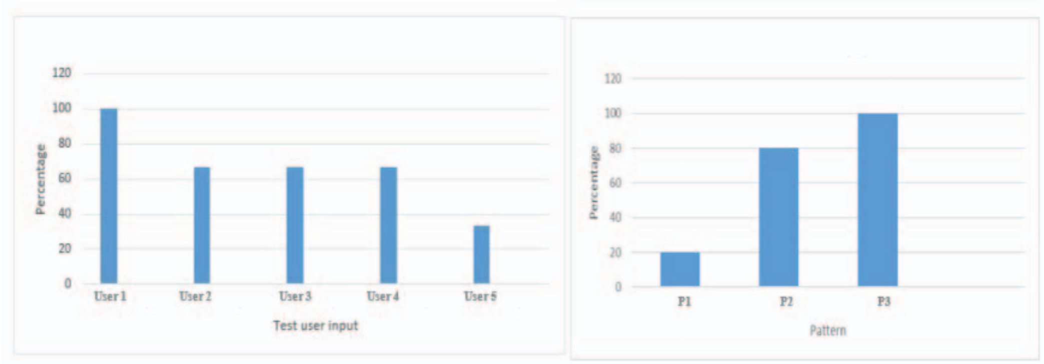
WORDS : hello

A2.wav

C:\Users\Amirul Amin\Desktop\fyp\GUI folder\afwan\A2.wav

Fig. 8: The GUI for the conversion of tongue clicking sound to words.

There are three main sections that have been designed in the GUI to operate this system. The first section is an input by real-time recording from the user using the microphone of the laptop. The second section is an input by uploading the tongue clicking sound of the user that is already stored in the file. The third section is the two features chosen from previous analysis. When the user clicks on *peak* and *amplitude modulation (skewness)*, the values are the feature extracted from the input provided by the user through section one or two.



(a) (b)

Fig. 9: Successful rate on multiclass SVM classification for (a) five user inputs and (b) each pattern.

In Fig. 9, the trained classifier was able to recognize 100% of input from user 1, 67% correct identification from user 2 to 4 and 33% correct identification for user 5. A moving average was then applied to the peak size threshold to improve the classification result and the success rate for all five user inputs produced 67% correct identification as illustrated in Fig. 10(a) and 10(b).

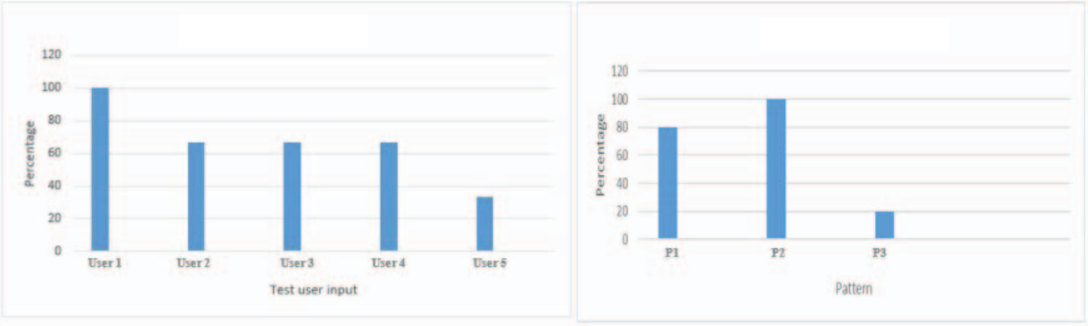


Fig. 10: Successful rate on multiclass SVM classification with moving average filter for (a) five user inputs and (b) each pattern.

Even though the overall percentage result for all five user input is the same, the success rate on the pattern is different for P1 and P3. P3 gives a higher success rate without moving average filter while P1 gives higher percentage successful rate with moving average filter.

5. CONCLUSION

In conclusion, the method of converting tongue clicking sounds to speech is novel in the field of AAC. However, a higher number of samples and a pre-trained tongue clicking session before collecting samples might increase the rate of detection. Besides using only two features (peak size and skewness) other features have yet to be analyzed and used for training the classifier.

ACKNOWLEDGEMENT

This work is supported by the Fundamental Research Grant Scheme (FRGS16-030-0529) from the Ministry of Education (MOE), Malaysia.

REFERENCES

[1] World Federation of Deaf, Retrieved December 10, 2018, from https://wfdeaf.org/

[2] Statistik Pendaftaran OKU. Retrieved December 10, 2018, from http://www.data.gov.my/data/ms_MY/dataset/statistik-pendaftaran-oku

[3] Dilbagi A. (2014). Talk breathes new life into the alternative communication device market. Available at: http://newatlas.com/talk-morse-code-speech/33042/.

[4] Kerr D, Bouazza-Marouf K, Gaur A, Sutton A, Green R. (2016). A breath controlled AAC system. Journal of Communication Matters, 30(3): 11-13.

[5] Mukherjee K, Chatterjee D. (2015). Augmentative and Alternative Communication Device Based on Eye-Blink Detection and Conversion to Morse-Code to Aid Paralyzed Individuals, Int. Conf. On Communication, Information and Computing Technology (ICCICT), pp. 1-5. doi: 10.1109/ICCICT.2015.7045754

[6] Nik Nur Wahidah NH, Wilkes M, Salomon R, Meggs J, and France DJ. (2016). Evaluation of Voice Acoustics as Predictors of Clinical Depression Scores, Journal of Voice, 31(2): 256.e1-256.e6.

[7] Nagori NP, Malode V. (2016). Communication Interface for Deaf-Mute People using Microsoft Kinect. Int. Conf. Autom. Control Dyn. Optim. Tech., pp. 640-644.

[8] Ozdas A, Shiavi RG, Silverman SE, Silverman MK, Wilkes DM. (2004). Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. IEEE Trans. Biomed. Eng., 51(9): 1530-1540.

[9] Vijayalakshmi P, Aarthi M. (2016). Sign language to speech conversion. Fifth Int. Conf. Recent Trends Inf. Technol., pp. 1-6.

[10] Shorrock TH, MacKay JC, Ball CJ. (2004). Efficient Communication by Breathing, Proceedings of the First International Conference on Deterministic and Statistical Methods in Machine Learning, pp. 88-97. doi: 10.1007/11559887_5