AN ANALYSIS OF THE LING SIX SOUNDS USING THE MFCC AND DTW FOR A SPEECH-REHABILITATION SYSTEM IN PATIENTS WITH COCHLEAR IMPLANTS

Kanut Puengsiricharoen[1], Jarinratn Sirirattawan[2], and Direk Sueaseenak[1]
[1]Department of Biomedical Engineering, Faculty of Engineering, Srinakharinwirot University, Thailand
[2]Department of Otolaryngologist, Faculty of Medicine, Srinakharinwirot University, Thailand

ABSTRACT

The Ling Six Sounds is a basic tool to measure the hearing performance of patients with cochlear implants or hearing aids, and they are also basic sounds for pronunciation. These sounds include “ah,” “ee,” “oo,” “mm,” “sh,” and “ss”. The frequency spectrum of a sound signal ranges from 250 Hz – 4000 Hz. This paper presents an analysis of the Ling Six Sounds of the Thai people who have a cochlear implant using the Mel-frequency Cepstral Coefficient (MFCC) as a features extraction in our speech-rehabilitation system. Dynamic Time Warping (DTW) was used to investigate the similarity index of each MFCC value of the Ling Six Sounds. The results indicated that the Ling Six Sounds with MFCC features can be effectively used in our speech-rehabilitation system because the significant-differences index of the DTW distance involving similar words was below 3000, and was over 4000 when involving different words. In the next step of our research, the speech- rehabilitation system will be developed for Thai people who have a cochlear implant to recover their communication ability.

Keywords: Ling Six Sounds, cochlear implants, hearing aids, Mel-frequency Cepstral Coefficient, Dynamic Time Warping.

1. INTRODUCTION

In 2012, Susan, Scollie et al. [1] conducted a study on the Ling Six Sounds, which are the sounds for audiometry developed by Dr. Daniel Ling. The sounds are “ah,” “ee,” “oo,” “mm,” “sh,” and “ss”. They cover sound frequencies from 250-4000 Hz and the high, mid, and low frequencies of speech. The frequency characteristics of the Ling Six Sounds are shown in Figure 1.

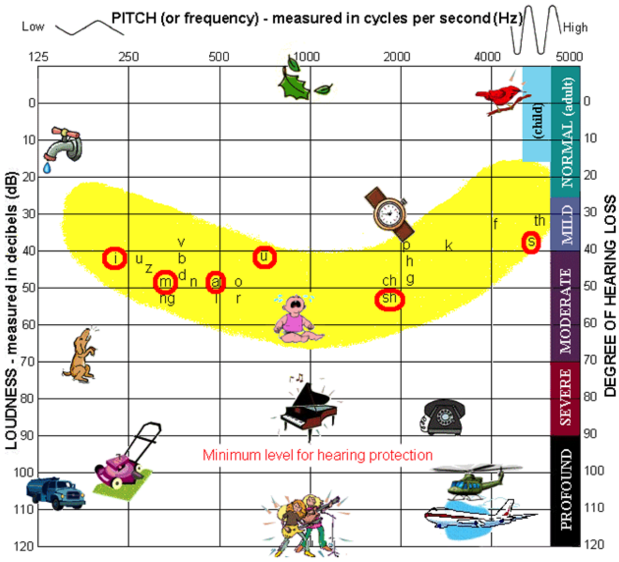

Figure 1. The frequency characteristics of the Ling Six Sounds [1]

The Ling Six Sounds [2, 3] is a basic tool. It measures the hearing performance of patients with cochlear implant surgery or hearing aids. The “oo” sound is used to measure how much very low-frequency sounds can be heard, including every low-pitched vowel. The “mm” sound is used to measure how much a low-frequency sound can be heard, such as vowels in every word we hear. The “ah” sound is used to measure how much a mid-frequency sound can be heard. Those who cannot hear this sound may therefore hear words in conversation that are unclear. The “sh” sound is a high-frequency sound. People with severe hearing loss may not be able to hear this sound if they are not wearing a hearing aid. The “ss” sound is a very high-frequency sound. People with severe hearing loss may not be able to hear it without a hearing aid. The “ee” sound can be both high and low in the speech-frequency range. If individuals pronounce it correctly, they can hear both high and low frequencies well. When imitating a sound, if they say “ss,” they have low-frequency hearing problems, and if they say “oo,” i.e. they have a high-frequency hearing problem [4, 5].

However, the advantage of the Ling Six Sounds are not just for auditory testing but are also basic sounds for pronunciation. For example, the “mm” sound is used to

Manuscript received on January 20, 2021; revised on March 19, 2021, accepted on April 9, 2021

*Corresponding author E mail: direks@g.swu.ac.th, Department of Biomedical Engineering, Faculty of Engineering, Srinakharinwirot University, Thailand.

pronounce the word " me," the "oo" sound is used to pronounce the word "boo," the "ah" sound is used to pronounce the word "car," the "ee" sound is used to pronounce the word "see," the "sh" sound is used to pronounce the word "wish," and the "ss" sound is used to pronounce the word "us" [6]. People who have received cochlear implant surgery have hearing and speech problems after the procedure. They should practice hearing and speech skills with a speech-language pathologist at the hospital.

This paper proposed an analysis of the Ling Six Sounds of Thai people with a cochlear implant. The Mel-frequency Cepstral Coefficient (MFCC) was used as a feature extraction. Dynamic Time Warping (DTW) was used to compare the similarity of each MFCC value of the Ling Six Sounds to determine whether it can be used to build a speech- rehabilitation system for people with a cochlear implant.

2. LITERATURE REVIEW

In 2010, Phokharatkul et al. [7] presented Thai speech recognition by integrating two approaches: the neural network and double-filter banks. The dataset recorded direction words and the numbers 0 to 10 in the Thai language. They used the Mel-frequency Cepstral Coefficient (MFCC) for feature extraction, and Euclidean distance was used to compare the similarities between the input and the reference vectors. The results were that the average accuracy rate is 96.3%.

In 2014, M. Bhadragiri Jagan and N. Ramesh Babu [8] presented a comparison of the features of the words. The MFCC is used as feature extraction and the DTW is used for pattern matching. The database recorded five separate words in the English language. The results showed that, for similar words, the DTW distance was below 100, while, for different words, the DTW distance was more than 300. Therefore, the threshold for separating words can be set due to the obvious DTW distance differences.

In 2015, Srijiranon et al. [9] compared the Neuro-Fuzzy system with Gaussian Mixture Models (GMM), the Support Vector Machine (SVM), the Decision-tree, and Byes to inspect the performance of recognizing the human voice using Nero-Fuzzy. They used the Perceptual Linear Predictive (PLP) for feature extraction. The database recorded nine words in the Thai language in several environments. The results showed that the accuracy of Neuro-Fuzzy was higher than with other popular algorithms, but GMM and SVM also gave a relatively high accuracy.

In 2016, Gupta K and Gupta D [10] compared features of speech using the Mel-frequency Cepstral Coefficient (MFCC), Relative Spectral Filtering (RASTA), and Linear Predictive Coding (LPC) in the Automatic Speech Recognition System. The results were the advantages and limitations of each feature and a comparison of the results from other research [11-13]. The advantages of LPC are good accuracy and robustness to noise and the limitations of LPC are it can't separate words with similar pronunciation. The advantages of RASTA are high accuracy and useful for capturing low modulated frequencies. The limitations of RASTA for better performance should be combined with PLP. The advantages of MFCC are very high accuracy with low complexity, the limitations of MFCC are the accuracy of MFCC can be reduced by noise. Imtiaz MA and Raja G [14] created an automatic speech-recognition system. The MFCC is used for the extraction of a feature of the speech. The DTW is used for matching the speech features. K-Nearest Neighbor (KNN) is used to classify the words. The database recorded ten isolated words in the English language. The results revealed that the accuracy was 98.4% and the error rate was 1.6%.

In 2018, Swedia ER et al. [15] presented a way of recognizing the speech digits of the Indonesian language by using the Long Short-Term Memory (LSTM) algorithm and then comparing the accuracy using different features, such as the LPC and MFCC. The dataset recorded the numbers 0 to 9 in the Indonesian language. The results showed that the accuracy of the MFCC was 96.58% and that of the LPC was 93.79%. Tantisatirapong et al. [16] proposed comparing several feature extractions - i.e., the MFCC, the Spectrogram (SPT), Energy Spectral Density (ESD), and Power Spectral Density (PSD) for the Thai speech-recognition system from the northeastern, southern, and central regions. The dataset recorded the numbers zero to nine in the Thai language from 30 female and 30 male speakers. The results showed that the accuracy of the MFCC feature was higher than with the ESD, SPT, and PSD.

3. METHODOLOGY

3.1. The Ling Six Sounds Recording Process

The database was obtained by recording the Ling Six Sounds from ten cochlear implant subjects (five men and five women) ages 5 to 60 from the Speech- Rehabilitation Unit at HRH Princess Maha Chakri Sirindhorn Medical Center (MSMC), with ethics approval obtained from the Srinakharinwirot University committee (SWUEC-177/2562E). Recorded speech files were in the *.wav file format using a sampling rate of 44.1 kHz and a resolution of 16 bits. The time duration of the recording was two seconds. A sample of the data collection is shown in Figure 2.

Figure 2. The ling six sounds recording process involving patients

3.2. The Mel-Frequency Cepstral Coefficients (MFCC)

The MFCC converts the conventional frequency to the Mel Scale to optimize the appropriate frequencies of sensitivity for human perception [17-19]. Therefore, it is quite suitable for speech-recognition tasks.

A cepstral is a discrete cosine transform (DCT) of a logarithm from a short signal spectrum. The cepstral coefficient on the Mel Scale is an improved technique of cepstral by adjusting the spectrum scale on a scale suitable for human hearing, based on the characteristics of the sound signals. Low-frequency audio signals are more important than high-frequency ranges, so a spectrum scale was designed to capture more detail from the low-frequency signal. This design is called the Mel Scale. The steps for calculating the cepstral coefficient on the Mel Scale are as follows:

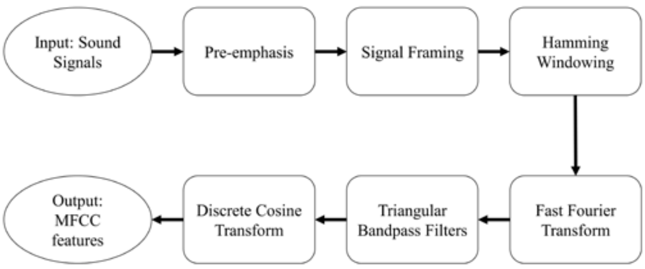

Figure 3. The Mel Frequency Cepstral Coefficients characteristic vectors extraction flow

- Pre-emphasis: The magnitude of the signal at high frequencies is emphasized by passing the signal through a filter.
- Signal framing: The signal will be framing into short frames. The duration of one frame is 20 - 40 milliseconds (ms).
- Windowing: The Hamming window is the most common type of window. The purpose of this stage is to avoid spectral leakage.
- Fast Fourier Transform (FFT): The Fast Fourier Transform (FFT) is used to convert the signal from the time domain into the frequency domain.
- Mel-scaled Filterbank: The signal will pass through a filter. The most common type of filter is Triangular Bandpass Filter. The signal with a frequency higher than 1000 Hz will be passed through a logarithmic frequency and the signal with a frequency lower than 1000 Hz will be passed through a linear frequency.
- Discrete Cosine Transform (DCT): The Discrete Cosine Transform (DCT) is used to convert the signal from the frequency domain into the time domain.

The Mel spectrum [17-19] is calculated by transmitting the Fourier transformed signals through a Mel-filter bank. The Mel scale is a measurement based on the frequencies which humans can perceive. It does not correspond to the physical frequency of the linear tones because the hearing system of humans does not have a linear pitch. The measuring of the Mel from physical frequency can be shown as (1).

$$f_{Mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (1)$$

when f_{Mel} refers to the perceived frequency and f refers to the physical frequency in Hz.

From an MFCC calculation, filter banks are commonly used in the frequency domain. Usually, the middle frequencies of the filters are regularly spaced on the frequency axis. The warped axis is used according to the expression of a nonlinear function in (1) to mimic the human ear's perception. The triangular filter shape is the most commonly used for a filter shape. Figure 4 shows the triangular filter banks with a Mel frequency warping.

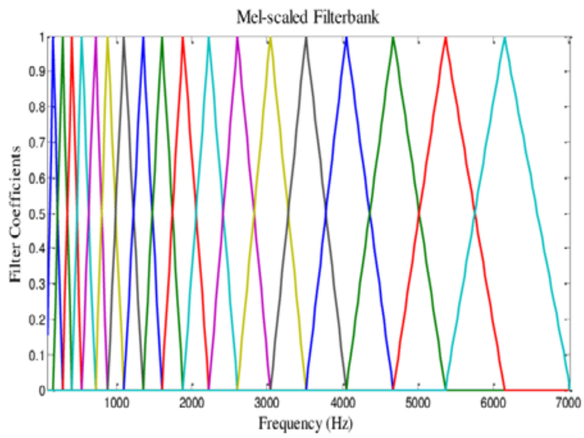

Figure 4. A Mel-scaled Filterbank [19]

The Mel spectrum of the magnitude spectrum $X(k)$ is calculated by multiplying the magnitude spectrum by each of the triangular Mel weighting filters following (2).

$$s(m) = \sum_{k=0}^{N-1} [|X(k)|^2 H_m(k)] \quad ; 0 \le m \le M-1 \tag{2}$$

when M is the total number of the triangular Mel weighting filters. $H_m(k)$ is the weight given to the k energy spectrum bin participation to the m output band.

3.3. Dynamic Time Warping (DTW)

In this stage, the DTW algorithm is used to compare the features of a word calculated in the previous step and the features of another word by calculating the least distance between two features of the word. If the distance is close to zero, the terms are similar, and if the distance is large, the terms are not similar. DTW is a technique to find an optimal alignment between two arrays or time series of different lengths [20]. The extent of matching between two arrays or time series is measured in terms of the distance factor. Euclidean distance is used to measure the distance between two features of a word. The DTW distance between two voice samples is shown in Figure 5.

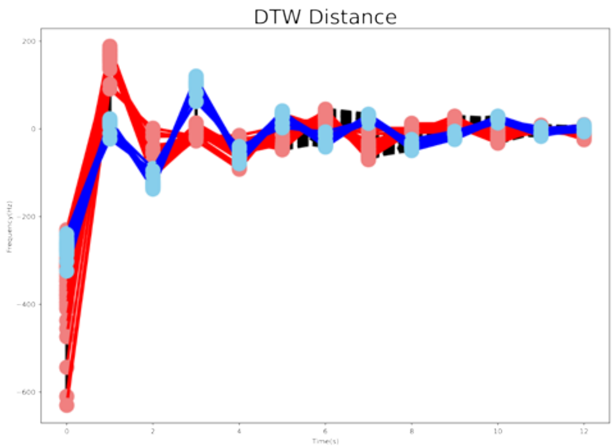

Figure 5. Dynamic Time Warping distance between two voice samples

4. EXPERIMENTAL RESULTS

The MFCC features of the Ling Six Sounds are shown in Figure 6 - 11. Figure 6 shows the frequency content of the "ah" sound, the frequencies higher than 100 Hz are distributed in the second MFCC coefficients and the frequencies between -200 to -400 Hz are distributed in the first MFCC coefficients. Figure 7 shows the frequency content of the "ee" sound, the frequencies higher than 100 Hz are distributed in the 4th MFCC coefficients and the frequencies between -200 to -400 Hz are distributed in the first MFCC coefficients. Figure 8 shows the frequency content of the "mm" sound, the frequencies higher than 100 Hz are distributed in the second MFCC coefficients and the frequencies lower than -200 Hz are distributed in the first MFCC coefficients. Figure 9 shows the frequency content of the "oo" sound, the frequencies higher than 100 Hz are distributed in the second MFCC coefficients and the frequencies lower than -200 Hz are distributed in the first MFCC coefficients. Figure 10 shows the frequency content of the "sh" sound, the frequencies higher than 100 Hz are distributed in the 4th MFCC coefficients and the frequencies between -200 to -300 Hz are distributed in the first MFCC coefficients. Figure 11 shows the frequency content of the "ss" sound, the frequencies higher than 100 Hz are distributed in the 3rd and 5th MFCC coefficients, and the frequencies lower than -200 Hz are distributed in the first MFCC coefficients. The DTW distances of comparisons between the same words and the different words are shown in Tables I and II. The DTW distance closer to 0 showed that the two words are similar, but the larger the distance, the less correlated the two words are. The results were that the same words had a significantly smaller distance value than the different words. From Tables I, a comparison between the same words is given the DTW distances below 3000. From Tables II, a comparison between the different words is given the DTW distances over 4000.

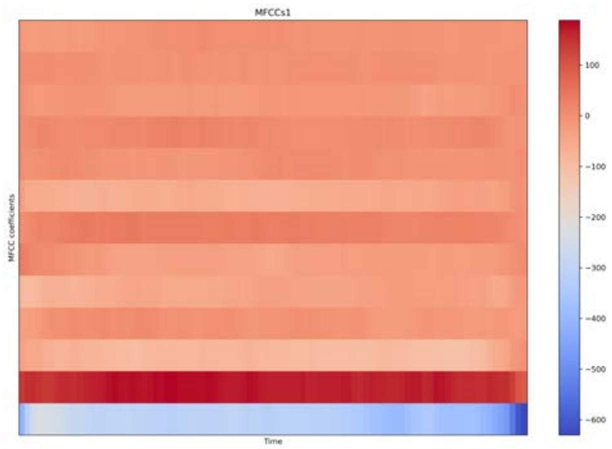

Figure 6. The Mel Frequency Cepstral Coefficients features of the "ah" sound

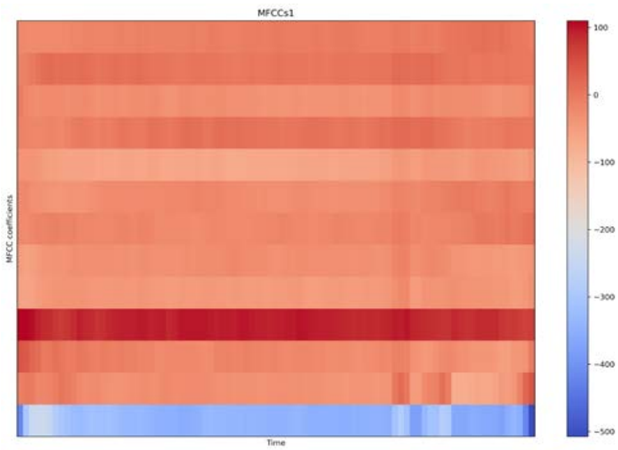

Figure 7. The Mel Frequency Cepstral Coefficients features of the "ee" sound

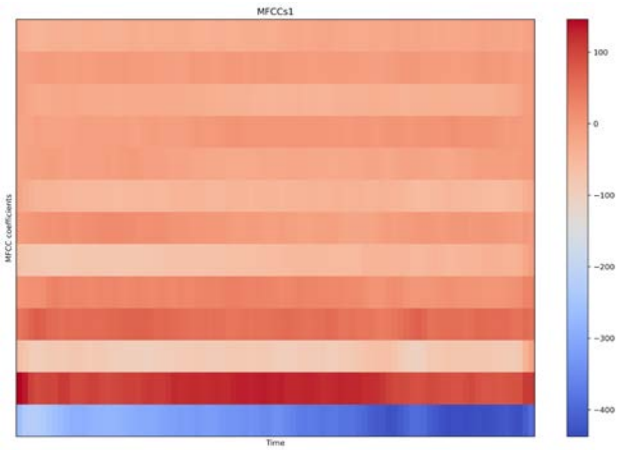

Figure 8. The Mel Frequency Cepstral Coefficients features of the “mm” sound

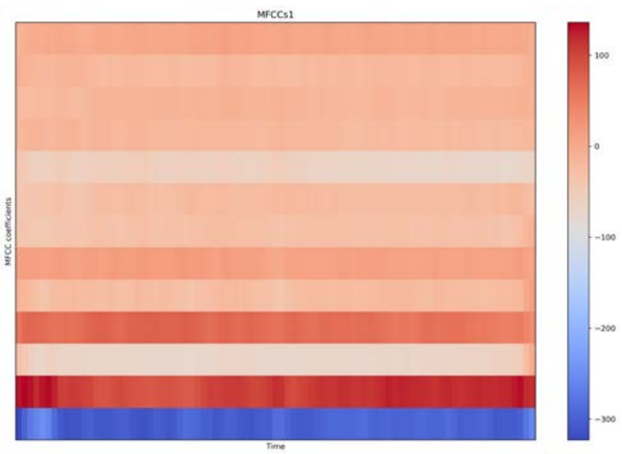

Figure 9. The Mel Frequency Cepstral Coefficients features of the “oo” sound

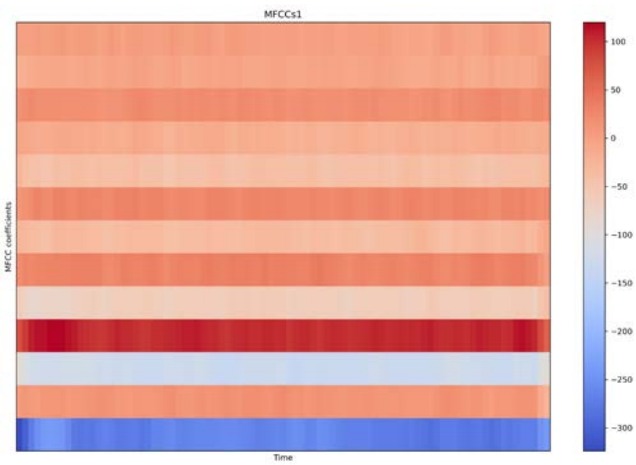

Figure 10. The Mel Frequency Cepstral Coefficients features of the “sh” sound

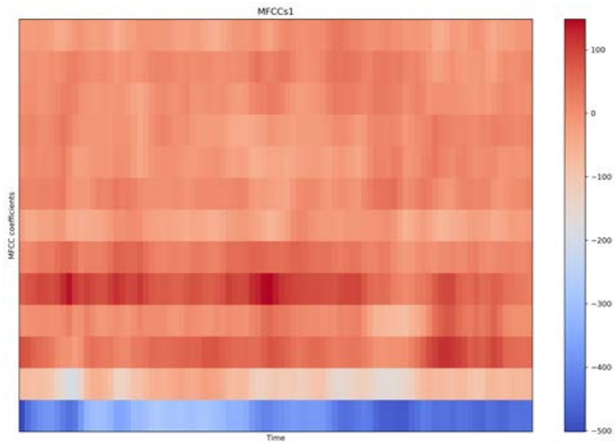

Figure 11. The Mel Frequency Cepstral Coefficients features of the “ss” sound

TABLE I. COMPARISONS BETWEEN THE SAME WORDS

Similar words	Average of DTW distance
ah	2721.39
ee	2857.97
mm	2992.34
oo	2319.23
sh	1904.13
ss	2472.79

TABLE II. COMPARISONS BETWEEN DIFFERENT WORDS

Word 1	Word 2	Average DTW distance
ah	ee	5612.80
ah	mm	4528.52
ah	oo	4440.71
ah	sh	5858.34
ah	ss	7510.71
ee	mm	4960.79
ee	oo	4776.36
ee	sh	4827.32
ee	ss	5674.33
mm	oo	3465.50
mm	sh	4893.58

mm	ss	6379.13
oo	sh	4848.19
oo	ss	6914.31
sh	ss	6434.43

5. DISCUSSION AND CONCLUSION

This paper presents an analysis of the Ling Six Sounds of the Thai people with a cochlear implant. The Mel-frequency Cepstral Coefficient (MFCC) was used as a feature extraction. Dynamic Time Warping (DTW) was used to compare the similarity of each MFCC value of the Ling Six Sounds. The results from Tables I and II were that the Ling Six Sounds with MFCC features can be effectively used in a speech-rehabilitation system because the significant-differences index of the DTW distance involving similar words was below 3000, and was over 4000 when involving different words. The difference between this research and other research [7-10, 14-16] was the database in this research is the words used to rehabilitation the communication of a cochlear implant person, and the audio data is collected from the people with cochlear implants. People with cochlear implants have communication disorders due to they have never heard for a long time.

The limitations in this research were this research was not a system that can separate words, it was only a study of feature extraction of sound to prove that the Ling Six Sounds can be used in a speech-rehabilitation system and the words used in this research just a very basic words for the rehabilitation of communication.

In the next step of our research, a speech-rehabilitation system will be developed for Thai people who have a cochlear implant to help them recover their communication ability. More words for the rehabilitation of communication from Speech-Language Pathologist will be added and using neural network for separate words in the speech-rehabilitation system. Successful research in the future will greatly reduce the rehabilitation of communication time of the cochlear implant person because the patient will be able to practice speaking at home more effectively.

ACKNOWLEDGMENT

This research was supported by the National Research Council of Thailand (NRCT). The researcher would like to give special thanks to the Department of Speech & Language Pathology, Department of Audiometry, HRH Princess Maha Chakri Sirindhorn Medical Center (MSMC), Faculty of Medicine, Srinakharinwirot University, which provided a great opportunity to record the Lin Six Sounds of our subjects.

REFERENCES

[1] S. Scollie *et al.*, "Stimuli and Normative Data for Detection of Ling-6 Sounds in Hearing Level," *American journal of audiology,* vol. 21, 07/30 2012, doi: 10.1044/1059-0889(2012/12-0020).

[2] C. I. A. (Thailand). "Recovery guide." https://www.cochlearassociationth.org/recovery-guide (accessed 12 Sep, 2020).

[3] A. Bionics. "The Ling Six Sound Check." https://advancedbionics.com/content/dam/advancedbionics/Documents/libraries/Tools-for-Toddlers/tools-for-parents/The-Ling-Six-Sound-Check.pdf (accessed 9 Sep, 2020).

[4] C. E. Ltd. "Ling Sounds." https://www.cochlear.com/8a1630d4-b089-4e06-8160-b1b24b7ccf3e/general_rehabilitationresources_earlyintervention_chitchat_handout2ling+sounds_1.0_en_119kb.pdf?MOD=AJPERES&CONVERT_TO=url&CACHEID=8a1630d4-b089-4e06-8160-b1b24b7ccf3e (accessed 12 Sep, 2020).

[5] N. Kanjanamongkol. "Ling 6 sound." https://hearingaidsbestprice.com/2501/ (accessed 2 Sep, 2020).

[6] N. Teakle. "Rehab At Home: What Is The Ling Six Sound Test?" https://blog.medel.com/all-about-the-ling-six-sound-test/ (accessed 5 Nov, 2020).

[7] P. Pisit, N. Kriengkrai, and P. Supachai, "Thai speech recognition using Double filter banks for basic voice commanding," in *2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering*, 24-26 Aug. 2010 2010, vol. 6, pp. 33-36, doi: 10.1109/CMCE.2010.5609930. [Online]. Available: https://ieeexplore.ieee.org/document/5609930/

[8] M. Bhadragiri Jagan and N. Ramesh Babu, "Speech recognition using MFCC and DTW," in *2014 International Conference on Advances in Electrical Engineering (ICAEE)*, 9-11 Jan. 2014 2014, pp. 1-4, doi: 10.1109/ICAEE.2014.6838564.

[9] K. Srijiranon and N. Eiamkanitchat, "Thai speech recognition using Neuro-fuzzy system," in *2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 24-27 June 2015 2015, pp. 1-6, doi: 10.1109/ECTICon.2015.7207075. [Online]. Available: https://ieeexplore.ieee.org/document/7207075/

[10] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, 14-15 Jan. 2016 2016, pp. 493-497.

[11] D. Hardt and K. Fellbaum, *Spectral subtraction and RASTA-filtering in text-dependent HMM-based speaker verification*. 1997, pp. 867-870 vol.2.

[12] S. Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot," 2011.

[13] E. Mansour, M. Sayed, A. M. Moselhy, and A. A. Abdelnaiem, "LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification," 2013.

[14] M. A. Imtiaz and G. Raja, "Isolated word Automatic Speech Recognition (ASR) System using MFCC,

DTW & KNN," in *2016 Asia Pacific Conference on Multimedia and Broadcasting (APMediaCast)*, 17-19 Nov. 2016 2016, pp. 106-110.

[15] E. R. Swedia, A. B. Mutiara, M. Subali, and Ernastuti, "Deep Learning Long-Short Term Memory (LSTM) for Indonesian Speech Digit Recognition using LPC and MFCC Feature," in *2018 Third International Conference on Informatics and Computing (ICIC)*, 17-18 Oct. 2018 2018, pp. 1-5.

[16] S. Tantisatirapong, C. Prasoproek, and M. Phothisonothai, "Comparison of Feature Extraction for Accent Dependent Thai Speech Recognition System," in *2018 IEEE Seventh International Conference on Communications and Electronics (ICCE)*, 18-20 July 2018 2018, pp. 322-325, doi: 10.1109/CCE.2018.8465705. [Online]. Available: https://ieeexplore.ieee.org/document/8465705/

[17] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980, doi: 10.1109/TASSP.1980.1163420.

[18] X. Huang, A. Acero, and H.-W. Hon, "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development," 01/01 2001.

[19] K. Rao and M. k e, *Speech Recognition Using Articulatory and Excitation Source Features*. 2017.

[20] M. Müller, "Dynamic time warping," *Information Retrieval for Music and Motion*, vol. 2, pp. 69-84, 01/01 2007, doi: 10.1007/978-3-540-74048-3_4.

Kanut Puengsiricharoen was born in Bangkok, Thailand in 1997. He received the B.Eng. degrees in biomedical engineering from Srinakharinwirot University, Thailand, in 2020. Currently, he is working toward M.Eng. degree in biomedical engineering from Srinakharinwirot University, Thailand. His current research interests include biomedical signal and biomedical instrumentation.

Direk Sueaseenak was born in Lampang, Thailand, in 1980. He received the B.Eng. degree in electrical engineering from Srinakharinwirot University, Thailand, in 2005. He received the M.Eng. degree in biomedical electronics and the D.Eng. degree in electrical engineering, both from King Mongkut's Institute of Technology Ladkrabang(KMITL), Bangkok, Thailand, in 2007 and 2013, respectively.

After working as a biomedical instrumentation technician at HRH Princess Maha Chakri Sirindhorn Medical Center(MSMC), Faculty of Medicine, Srinakharinwirot University, in 2015, he joined the Department of Biomedical Engineering, Faculty of Engineering, Srinakharinwirot University, where he is currently an Assistance Professor. His current research interests include biomedical electronics and signals, biomedical instrumentation, and biomedical innovation.

Jarinratn Sirirattawan was born in Bangkok, Thailand, in 1970. She received the B.SC. degrees in food science and nutrition from Srinakharinwirot University, Bangkok, Thailand, in 1993. She received the medical degrees (MD.) from Chulalongkorn University, Bangkok, Thailand, in 2001. She received the Diploma of the Thai Board of Otorhinolaryngology-Head and Neck Surgery from Bhumibol Adulyadej Hospital RTAF, Bangkok, Thailand, in 2008. She received the research fellow (Doctor Med.) in inner ear disease and visiting fellow: Sinus surgery in head and neck surgery from Johannes Gutenberg University, Mainz, Germany, in 2008 and 2011, respectively.

She worked in the HRH Princess Maha Chakri Sirindhorn Medical Center(MSMC), Faculty of Medicine, Srinakharinwirot University, she is currently a Deputy Dean in the Faculty of Medicine. Her current research interests include Otolaryngologist and medical innovation and application in a hospital.