Multimodal Data Fusion of Electromyography and Acoustic Signals for Thai Syllable Recognition

Nida Sae Jong, Alba García Seco de Herrera, *Member*, *IEEE*, and Pornchai Phukpattaranont, *Member*, *IEEE*

***Abstract*— Speech disorders such as dysarthria are common and frequent after suffering a stroke. Speech rehabilitation performed by a speech-language pathologist is needed to improve and recover. However, in Thailand, there is a shortage of speech-language pathologists. In this paper, we present a syllable recognition system, which can be deployable in a speech rehabilitation system to provide support to the limited speech-language pathologists available. The proposed system is based on a multimodal fusion of acoustic signal and surface electromyography (sEMG) collected from facial muscles. Multimodal data fusion is studied to improve signal collection under noisy situations while reducing the number of electrodes needed. The signals are simultaneously collected while articulating 12 Thai syllables designed for rehabilitation exercises. Several features are extracted from sEMG signals and five channels are studied. The best combination of features and channels is chosen to be fused with the mel-frequency cepstral coefficients extracted from the acoustic signal. The feature vector from each signal source is projected by spectral regression extreme learning machine and concatenated. Data from seven healthy subjects were collected for evaluation purposes. Results show that the multimodal fusion outperforms the use of a single signal source achieving up to $\sim 98\%$ of accuracy. In other words, an accuracy improvement up to 5% can be achieved when using the proposed multimodal fusion. Moreover, its low standard deviations in classification accuracy compared to those from the unimodal fusion indicate the improvement in the robustness of the syllable recognition.**

***Index Terms*— Acoustic signal, electromyography, feature-level fusion, multimodal fusion, speech recognition.**

Manuscript received March 23, 2020; revised September 15, 2020; accepted October 18, 2020. Date of publication Xxx xx, 2020; date of current version Xxx xx, 2020. This work was supported in part by the Office of the Higher Education Commission, Ministry of Education, Thailand, under Grant No. 006/2558. *(Corresponding author: Pornchai Phukpattaranont.)*

Nida Sae Jong is with the Department of Electrical Engineering, Prince of Songkla University, Hat Yai, 90110 Thailand, and with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, UK, and also with the Department of Electrical Engineering, Faculty of Engineering, Princess of Naradhiwas University, Narathiwat, 96000, Thailand (e-mail: nida.s@pnu.ac.th)

Alba García Seco de Herrera is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, UK (e-mail: alba.garcia@essex.ac.uk)

Pornchai Phukpattaranont is with the Department of Electrical Engineering, Faculty of Engineering, Prince of Songkla University, Hat Yai, Songkhla, 90110, Thailand (e-mail: pornchai.p@psu.ac.th)

I. Introduction

Stroke is one of the most frequent causes of death and disability [1]. Around 0.4% of people over the age of 45 years in the United States, Europe, and Australia have a first stroke each year [2]. Similar to other parts of the world, stroke in Asia has a high incidence and a high mortality rate [1]. In particular, in Thailand the age- and sex-standardized disability-adjusted-life-years (DALYs) lost was 1,108 per 100,000 people in 2010 [1]. One of the possible stroke effects is a speech disorder, i.e. dysarthria, which is caused by a disturbance in neuromuscular control of speech production including lips, tongue, and vocal folds [3]. In clinical practice, a speech-language pathologist (SLP) treats patients by improving articulation, focusing on their way of pronouncing clear phonemes [4]. However, in Thailand, the number of neurologists [5] and SLPs specialized in speech ability recovering [6] is limited, especially in rural areas. An intelligent system based on speech recognition will help because it may substitute the SLPs in providing assistance and feedback to the patients.

The majority of speech recognition methods used with dysarthric patients is based on the analysis of acoustic signals [6], [7]. However, noisy environments are not suitable for this method because the acoustic signal is easily contaminated with ambient noise [8]. Hence, surface electromyography (sEMG) based speech recognition has been extensively proposed, especially in silent and non-audible environments [9]. However, some challenges may be encountered while using sEMG for speech recognition. Examples include: 1) difficulty measuring sEMG signals on a flaccid dysarthric speaker; 2) interference affecting the signal quality and recognition performance, such as motion artifact, power line noise or electrocardiographic artifact [10]; and 3) a large number of electrodes leading to an increase in the price level and inconveniences of use.

Speech recognition systems have been developed in other languages [11]–[13] but there is limited literature available for the Thai language. Pothirat *et al.* [14] classified five oral movement activities used to improve oral motor for speech production using six channels of sEMG signals recorded from the face and the neck. The average classification accuracy achieved was 91.3%. In previous work [15], we proposed a syllable recognition system using five channels of sEMG

signals to classify nine Thai syllables. The average accuracy obtained from healthy volunteers was 94.5%. In this work, we extend and continue the development of a speech recognition system obtaining improved accuracy while reducing the number of electrodes, which will be more comfortable and easier to use for patients. To achieve this, we propose a multimodal data fusion of sEMG and acoustic signals. Moreover, in this work, more Thai syllables (12) are used, which are more appropriate for the deployment in a speech rehabilitation system for dysarthric patients. The three additional syllables are suggested by the SLP to use at the beginning of speech rehabilitation because these syllables represent the most difficult activities for non-speech oral treatment in dysarthria.

The rest of the paper is organized as follows. Section II brings current related work in the domain including electrode position, feature calculation, and data fusion. Section III gives a description of the data acquisition system and experiment. Section IV describes our proposed framework on the syllable recognition system based on multimodal data fusion. The results are presented and discussed in Section V and Section VI, respectively. Finally, conclusions are given in Section VII.

II. Related work

This section brings some related work that serves as background and motivation for the presented research. It reviews three key aspects of this work: electrode position, feature extraction, and data fusion.

A. Electrode Position

Surface EMG signals from the facial muscles can be applied in a variety of applications. Lapaki *et al.* [16] applied a new skin attachment technique to produce a small electrode with the outer dimension of 4-mm diameter. Results from a group of 11 professional trumpeters showed that the sEMG signals from seven different perioral muscles recorded with the new electrodes were similar to those obtained with intramuscular fine-wire electrodes. Schumann *et al.* [17] recorded the sEMG signals of the facial muscles while performing 29 facial movements of high clinical relevance to characterize essential muscle activity. Also, the sEMG signals recorded in the facial muscles were applied in a speech recognition system to classify Spanish [11] and Thai [15] syllables, English words [12], and English phonemes [13].

The electrode position in the facial muscles affects the accuracy and reliability of sEMG based speech recognition system. The motor units can be considered the sEMG signal sources of a muscle of interest. We place the electrodes over the skin to detect the electrical activity from the motor units when the muscle of interest underneath contracts. The electrode-skin interface is modeled as a spatial low-pass filter, which reduces the amplitude and frequency content of the detected signal [18]. Moreover, the amplitude and frequency content of the detected signal decrease when the distance between the electrodes and the muscle increases [19], [20]. We enhance the signal-to-noise ratio of the signal by placing the electrodes close to the muscle of interest [20], thereby improving the accuracy and reliability of the sEMG based speech recognition system. The electrode positions placed on facial muscles for sEMG recording included the following muscles: the anterior belly of the digastric and the depressor anguli oris [11]–[13], [15], the levator anguli oris [12], [13], [15], the zygomaticus major [11], [15], and the mentalis [15].

In addition to the electrodes positioned on the muscles, the data acquisition system must be connected with either a ground or a reference electrode. Two common electrode configurations used for measuring sEMG signals are monopolar and bipolar [21]. Normally, in the bipolar configuration, the ground electrode is placed on a location where there is no muscular activity during the measurements, such as the forehead, the back of the neck, and the wrist. When the monopolar configuration is used, the heart should not be placed between the reference and the electrode to minimize the contamination of the electrocardiography signal in the measurement [21].

B. Feature Extraction

Feature extraction is an essential process to reduce redundant data and highlight relevant data. Features used in recognition of sEMG signals from the facial muscles can be determined based on their amplitude values, frequency contents, and statistical values. The popular amplitude based features include root-mean-squared value [11]–[14], mean absolute value, zero crossing [11], [14], [15], waveform length [14], [15], and slope sign change [14]. While frequency based features commonly extracted are Fast Fourier transform coefficients [11] and mean frequency [15], statistical based features are kurtosis [11], [15] and skewness [15].

The mel-frequency cepstral coefficients (MFCCs) are popular features for acoustic signals [12], [13] since the mel filterbank is designed to mimic the human hearing perception. In other words, the variation in the low-frequency range is more important than that in the high-frequency range. Therefore, its bandwidth is wider when the center frequency of the filter increases. The MFCCs were also successfully applied with the sEMG signals in the speech recognition system. For example, the MFCCs with 12 and 13 coefficients were used as the features for the sEMG signals in [12], [13], and [11], respectively.

C. Data Fusion

Data fusion comprises two main approaches based on the nature of the data: unimodal and multimodal fusions [22]. The unimodal fusion combines data from the same source, whereas the multimodal fusion combines data from different sources.

In the unimodal fusion, features from multiple channels of sEMG signals are combined in the syllable recognition system [11], [15]. Lopez-Larraz *et al.* [11] concatenated the feature vectors from 8 sEMG-combined channels. The length of feature vectors resulting from this unimodal fusion was 328 (41 features/channel $\times$ 8 channels). Thirty Spanish syllables were classified using the Adaptive Boosting (AdaBoost) algorithm with a decision tree. The AdaBoost algorithm is an algorithm that can combine many weak classifiers into a strong classifier using a weighted sum to improve classification

accuracy. A weak classifier is the classifier, which gives an accuracy slightly greater than 50%. Average accuracy of 70% was reported by using data from three subjects. In previous work [15], the feature vectors from 5 sEMG channels were combined when the number of features from each channel was 6. As a result, the feature vectors with 30 in length were used for the recognition of 9 Thai syllables.

Regarding the multimodal fusion, Chan *et al.* [12] and Scheme *et al.* [13] combined 5 channels of sEMG with an acoustic signal in decision-level fusion for recognizing words and phonemes, respectively. A voting process was performed using the results from classification. The method was based on a mathematical framework of evidence theory called plausibility method [23]. Results showed that the classification accuracy from the data fusion using plausibility method was better than that from the acoustic or sEMG signals only.

III. Data Acquisition

Fig. 1 shows an overview of the data acquisition system. Following previous work [15], five channels of electrodes were used for acquiring sEMG signals. They were placed on: 1) the zygomaticus major (CH1); 2) the levator anguli oris (CH2); 3) the depressor anguli oris (CH3); 4) the mentalis (CH4); and 5) the anterior belly of the digastrics (CH5). While the sEMG signals from CH2, CH3, CH4, and CH5 were recorded with the monopolar configuration, the sEMG signals from CH1 were recorded with the bipolar configuration. The reference electrode was placed on the earlobe and the ground electrode was placed on the left wrist. Small disc-shaped sEMG electrodes (5 mm diameter, Ag/AgCl) and shielded cables were connected to a commercial sEMG measurement system for recording sEMG signals. The sEMG signals were digitized at a sampling frequency of 1024 Hz.

To collect the acoustic signal, a wired headset microphone was connected to the computer as shown in Fig. 1. A voice recorder system controlled with LabVIEW[1] was used to acquire the acoustic signal at a sampling frequency of 20 kHz. When the voice recorder system started to record the acoustic signal, a trigger signal was generated and sent to the commercial sEMG measurement system to start the sEMG signal acquisition. As a result, both sEMG and acoustic signals were synchronously acquired. The sEMG and acoustic signals were recorded for 4 seconds including preparing, articulating, and ending times.

Seven healthy subjects with no speech impediment or disorders (four males and three females; age $20-22$ years; height $160-180$ cm; weight $46-75$ kg) participated in the experiments. The experiments were carried out at the Department of Electrical Engineering, Faculty of Engineering, Prince of Songkla University.

Each participant articulated 12 isolated syllables from the Thai language. Three vowels consisting of "a", "i" and "u" were chosen because they represented the maximum muscle contraction in the opening of the mouth, the smiling, and the pursing of the lips, respectively, which were needed for non-speech oral motor treatment [24]. The consonants were divided into three groups according to the place of articulation: velar ("k"); alveolar ("n"); and bilabial ("m"). Hence, the training syllable consisted of the vowels by themselves in addition to the combination of the consonants and the vowels as shown in Table I.

TABLE I
SET OF THAI TRAINING SYLLABLES USED FOR SPEECH THERAPY. THE FIRST COLUMN CONTAINS THREE VOWELS. THE LAST THREE COLUMNS CONSIST OF THAI SYLLABLES ORGANIZED BY THEIR PLACE OF ARTICULATION.

Vowels	Velar	Alveolar	Bilabial
/a/	/ka/	/na/	/ma/
/i/	/ki/	/ni/	/mi/
/u/	/ku/	/nu/	/mu/

To improve the articulation of flaccid dysarthria speakers, repetitive practice was required as part of the treatment protocol. Therefore, the participants repeated each syllable five times. While the total number of sEMG signals from each participant was 300 (12 syllables $\times$ 5 channels $\times$ 5 trials), the total number of acoustic signals from each participant was 60 (12 syllables $\times$ 1 channel $\times$ 5 trials).

IV. Proposed System

Fig. 2 shows the framework of the proposed syllable recognition system based on multimodal data fusion. A number in parenthesis indicates the length of a feature vector. There are four processing steps. The feature calculation step generates feature vectors from sEMG and acoustic signals of each syllable, whose lengths are 40 and 18, respectively. The feature projection step reduces the length of feature vectors from sEMG and acoustic signals to 11. Then, the feature fusion step concatenates the two feature vectors. The classification step predicts the syllable from the fusion feature vector. Details of each processing step are as follows.

A. Feature Extraction

In this section, we present the methods used to calculate features from sEMG and acoustic signals in this paper.

1) Features from the sEMG signals: Five features were calculated from sEMG signals consisting of mean absolute value (MAV), waveform length (WL), zero crossing (ZC), slope sign change (SSC), and the fourth-order autoregressive (AR) coefficient. The feature selection process based on the sequential forward floating selection technique chose them from eight popular features. While six of them, i.e. MAV, WL, ZC, mean frequency, skewness, and kurtosis, were used in the recognition system with the sEMG signals from facial muscles as details given in Section II-B [11]–[15], SSC and AR were used in the recognition system with the sEMG signals from lower [25] and upper [26] limb muscles, respectively. Details of each feature calculation are as follows.

MAV: is the average of the absolute values of sEMG signal amplitudes in a sampled segment, which can be defined as [15]:

$$\mathrm{MAV} = \frac{1}{N}\sum_{i=1}^{N}|x_i|, \tag{1}$$

[1]http://www.ni.com/en-gb/shop/labview.html

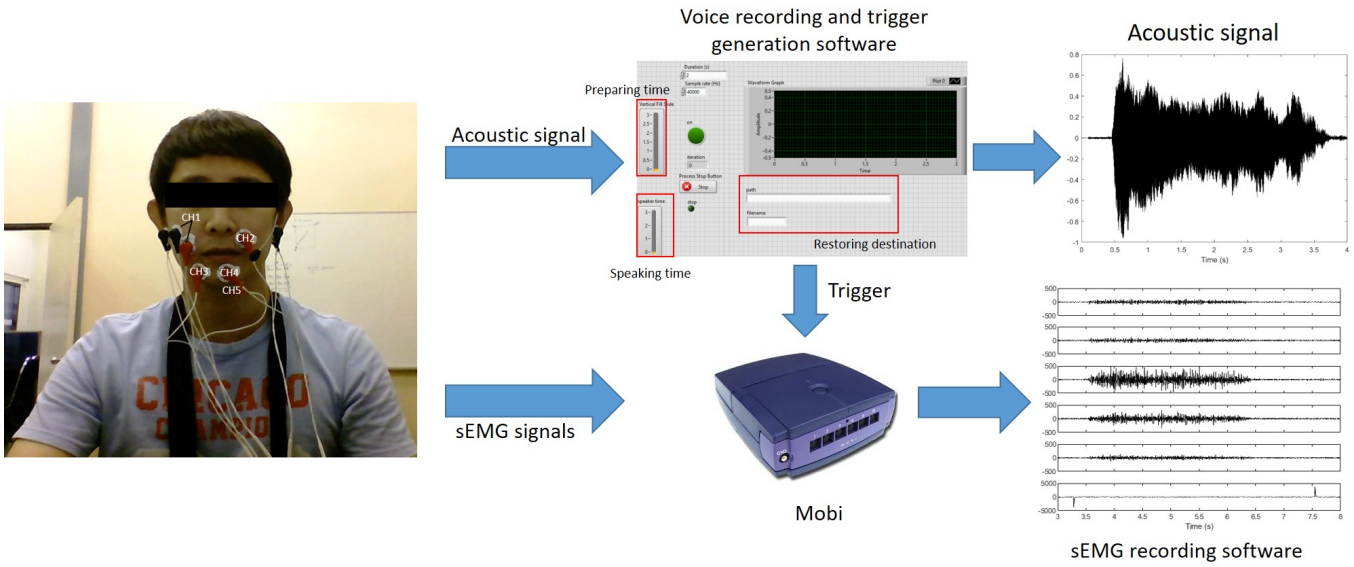

Fig. 1. Data acquisition system of sEMG and acoustic signals. After the subject articulates, the acoustic signal is recorded by a voice recording system while the sEMG signals are acquired by a commercial sEMG measurement system.

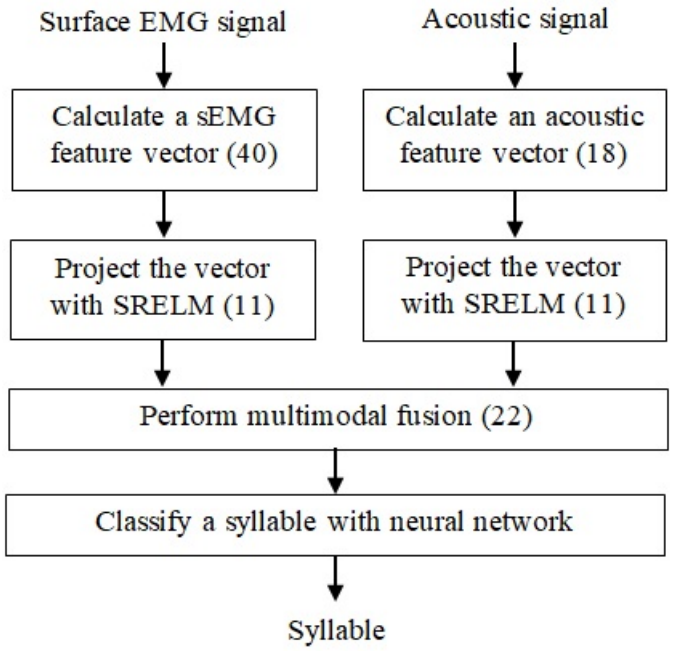

Fig. 2. Framework of the proposed syllable recognition system based on multimodal data fusion between *n* sEMG channels and an acoustic signal. Note that *n* is equal to 5 in this figure and a number in parenthesis indicates the length of a feature vector.

where x_i represents the amplitude of sEMG signal at sample i and N denotes the total number of sEMG samples under calculation.

WL: is a measure of the complexity of the sEMG signal. It is defined as the cumulative length of the sEMG waveform over a time segment. It can be calculated by [15]:

$$\mathrm{WL} = \sum_{i=1}^{N-1} |x_{i+1} - x_i|. \tag{2}$$

ZC: is a measure of the frequency information of the sEMG signal. It represents the number of times that the amplitude value of the sEMG signal crosses the zero amplitude level. A threshold condition T is implemented to avoid low voltage fluctuations or background noises. ZC is defined as [15]:

$$\mathrm{ZC} = \sum_{i=1}^{N-1} [f(x_i \times x_{i+1}) \text{ and } |x_i - x_{i+1}| \geq T], \quad \text{where } f(x) = \begin{cases} 1, & \text{if } x < 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

SSC: can be used as the supplementary information of the signal frequency. It can be expressed as [25]:

$$\mathrm{SSC} = \sum_{i=2}^{N-1} [s\{(x_i - x_{i+1})(x_i - x_{i-1})\}], \quad \text{where } s(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

AR: estimates the present value using a linear combination of the previous observation value x_{i-p} and white noise w_p, which is given by [26]:

$$\mathrm{AR} = \sum_{p=0}^{P-1} a_p x_{i-p} + w_p, \tag{5}$$

where P is the order of the AR model, which is 4 in this paper. As a result, the number of AR coefficients is also 4.

After the sEMG signals are acquired, they are applied with a bandpass filter with cut-off frequencies of 20 and 450 Hz for noise removal. The passband of 20 and 450 Hz is used for the sEMG signal because it matches with the bandwidth of sEMG signals between 20 and 450 Hz. Next, for each syllable (2.4 milliseconds (ms) in length), the filtered sEMG signal is segmented into 18 frames with 250 ms in length and

50% overlap. Five features consisting of 8 values (1 MAV, 1 WL, 1 ZC, 1 SSC, and 4 ARs) are computed for each frame. Consequently, for each syllable, there are 18 feature vectors per sEMG channel where the length of each feature vector is 8.

2) Features from the acoustic signal: We extract 18 order MFCCs feature vector from the acoustic signal. Details of the feature vector calculation are as follows:

1) Segment 2.4 s of the acoustic signal from each syllable.
2) Eliminate noise in the recorded acoustic signal using a lowpass filter with a cut-off frequency of 5 kHz.
3) Apply the pre-emphasis filtering with a pre-emphasis coefficient of 0.97 to the filtered signal from step 2). The pre-emphasis filter is a highpass filter, which can be implemented as $y[n] = x[n] - 0.97x[n-1]$.
4) Divide the signal from step 3) into 18 frames using the frame size 250 ms and the frame shift 125 ms.
5) Apply Hamming windowing on each frame from step 4) to attenuate discontinuities at the frame edges.
6) Compute magnitude spectrum of each frame from step 5) based on fast Fourier transform (FFT).
7) Design filterbank with 20 uniformly spaced triangular filters on the mel-scale between 15.99 and 2363.50 or on the frequencies between 10 and 5000 Hz, which cover the dominant signal energy from the syllables. When we divide the mel-scale between 15.99 and 2363.50 into 20 uniform intervals. The bandwidth for each interval is (2363.50-15.99)/20 = 117.38. Then, the lower and upper cutoff frequencies of 20 uniformly spaced triangular filters on the mel-scale are defined. For example, the lower and upper cutoff frequencies of the first triangular filter in the mel-scale are 15.99 and 133.37, respectively. The mel-scale is converted into the frequency in Hertz for filter design. The mapping between the frequency (f) in Hertz to the mel-scale is given by Young *et al.* [27]:
$$\text{Mel}(f) = 1127ln(1 + \frac{f}{700}) \quad (6)$$
As a result, the lower (15.99) and upper (133.37) cutoff frequencies in the mel-scale are converted into frequencies 10 Hz and 88 Hz, respectively.
8) Multiply FFT magnitude coefficients from each frame by the corresponding filter gain and accumulate the results. Therefore, twenty energies in each band were obtained.
9) Convert the natural logarithm of 20 energy bands to 18 order MFCCs through the Discrete Cosine Transform.
10) To reduce the problem on significant variation in amplitude between the low-order and the high-order MFCCs, apply liftering with a cepstral sine lifter parameter of 22 to the MFCCs from step 9) so that they have similar amplitudes.

Finally, the feature vector is obtained. For each syllable, there are 18 feature vectors from the acoustic signal where the length of each feature vector is 18.

B. Feature Projection

This section describes the details of feature projection applied to the features obtained from both sEMG and acoustic signals. Before feature projection, the features from both sEMG and acoustic signals are normalized to keep their values in a range of -1 to 1. Subsequently, a feature projection method called spectral regression extreme learning machine (SRELM) [15] is applied to the normalized features. SRELM not only increases the relevant data but also reduces the data dimension.

SRELM consists of three layers, the input, hidden, and output layers. The feature vector before projection is an input of the SRELM input layer. The number of nodes in the input layer depends on the type of signal and the number of channels. For example, for the 5 sEMG-combined channels, the number of nodes in the input layer is 40 because the length of the feature vector is 8 for each sEMG channel. The number of hidden nodes and alpha are two parameters used for optimizing SRELM performance. We vary the number of hidden nodes from 100 to 1500 with an increment of 100 and alpha from 1 to 20 with an increment of 1 to evaluate the optimal parameters. The feature vector after projection is an output of the SRELM output layer. The number of nodes in the output layer is the total number of syllables under classification minus one, which are 11 in this work.

C. Feature Fusion

As described in Section II-C, previous publications have shown that the unimodal and multimodal fusions can improve speech recognition performance [12], [13], [28], [29]. We perform both unimodal and multimodal fusions in this paper. The results from the unimodal fusion are used in performance comparisons with those from the multimodal fusion.

The *unimodal fusion* is performed with the features determined from all possible sEMG-combined channels. The groups of combined channels comprise 2 channels (10 subgroups), 3 channels (10 subgroups), 4 channels (5 subgroups), and 5 channels. Therefore, the lengths of feature vectors from 2, 3, 4, and 5 sEMG-combined channels are 16, 24, 32 and 40, respectively. Subsequently, the feature vectors from sEMG signals are projected by SRELM. After SRELM applications, the lengths of feature vectors from 2, 3, 4, and 5 sEMG-combined channels are reduced from 16, 24, 32, and 40 to 11. Hence, these features are ready to be classified.

In the *multimodal fusion*, both feature vectors from sEMG and acoustic signals are projected by SRELM before a feature concatenation. For the sEMG signals, the feature vector is obtained using the method described for unimodal fusion. Therefore, the length of the feature vector is 11. Also, the feature vector from the acoustic signal is projected by SRELM. The length of the feature vector from the acoustic signal after SRELM applications (FACO) is also reduced from 18 to 11. For the multimodal fusion, the sEMG feature vector after SRELM projection is concatenated with the FACO resulting in a new feature vector with a length of 22. This concatenated feature vector is used as an input of a classifier.

D. Classification

The feature vectors described in Section IV-C are classified using a feed-forward neural network. The structure of feed-forward neural network consists of an input layer, a hidden

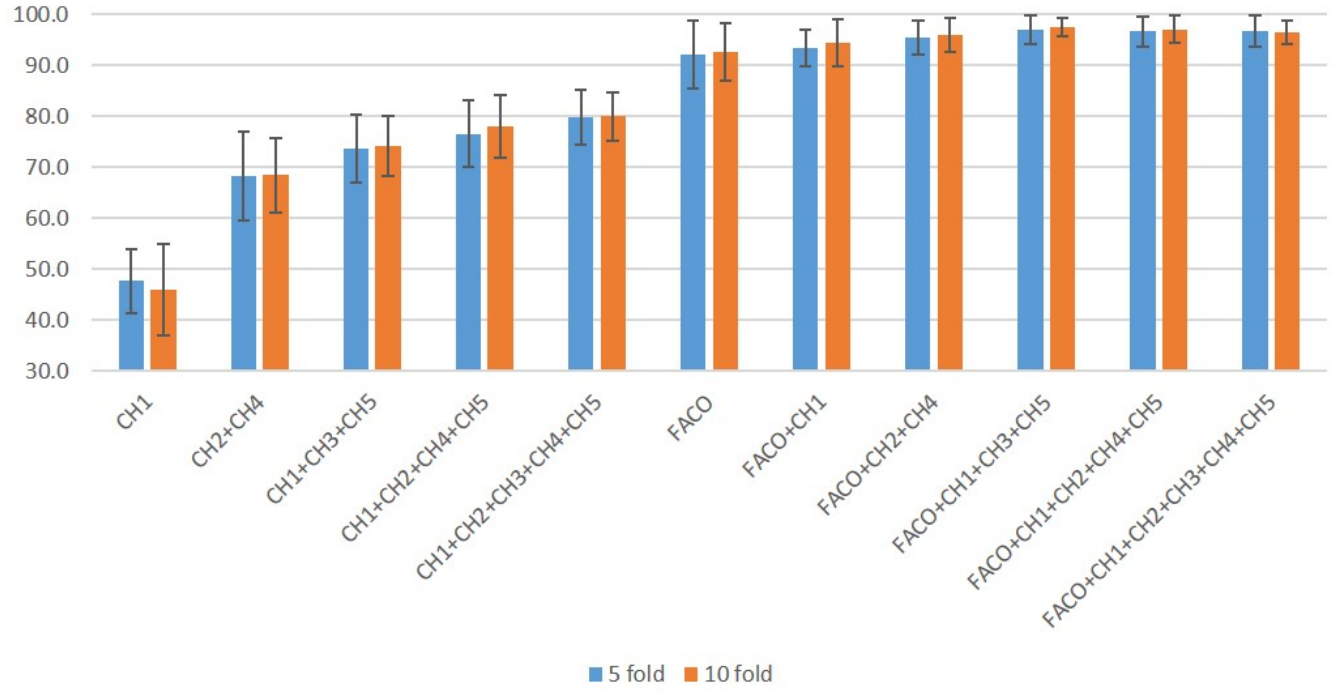

Fig. 3. Comparison of classification accuracies obtained by 5-fold and 10-fold cross validation from unimodal and multimodal fusions when the classifier is neural network.

layer, and an output layer. The number of nodes in the input layer is the length of feature vector obtained from Section IV-C, which is 11 for unimodal fusion and 22 for multimodal fusion. For the number of nodes in the hidden layer, we vary the nodes from 1 to 15 to evaluate the optimal value. The number of nodes in the hidden layer that gives the maximum accuracy is selected, which is 10 in this paper. When the number of nodes in the hidden layer increases from 11 to 15, the accuracy is not significantly different from 10 nodes. The number of nodes in the output layer is 12, which is the number of syllables under classification. The transfer function for the hidden and output layers is a hyperbolic tangent sigmoid.

After classification, the evaluation is performed per each subject by dividing the data into two sets containing training and testing data. The classification is performed using 5-fold cross validation. In other words, the total data for each subject is divided into 5 subsets. In the first iteration, the first 4 subsets are used as the training data. The last subset is used as the testing data. The accuracy from the first iteration is kept. In the second iteration, while the first subset is used as the testing data, the other 4 subsets are used as the training data. The accuracy from the second iteration is obtained. Five iterations are performed until each subset is used as the testing data. The classification performance of each subject is determined by the average accuracy of the 5 iterations. Moreover, the results from 10-fold cross validation are determined for comparisons.

V. Results

A. Performance Evaluation

Fig. 3 shows mean (MEAN) and standard deviation (SD) of the accuracy of seven subjects from the classification described in Section IV-D. The first five groups of bar graphs show the results when applying the unimodal fusion to only sEMG signals. The label in x-axis shows the channels, which give the best classification accuracy in each subgroup of sEMG-combined channels. For example, while the best single sEMG channel is CH1, the best 2 sEMG-combined channels are CH2 and CH4. The overall accuracy of the 7 subjects approximately increases from 47% to 80% when the number of sEMG-combined channels increases from 1 to 5. The sixth group of bar graphs in Fig. 3 shows the results from the unimodal using the acoustic signal only (FACO). The MEAN value from 7 subjects is 92%, which is higher than that from the 5 sEMG-combined channels at 80%. When the multimodal fusion is employed by adding FACO to the single sEMG-combined channels (CH1), the MEAN value significantly improve to 94% as shown in the seventh group of bar graphs. Similarly, by adding FACO to the 2, 3, 4, and 5 sEMG-combined channels, the MEAN values significantly improve to 96%, 97%, 97%, and 97%, respectively as shown in the last four groups of bar graphs of Fig. 3.

When applying multimodal fusion, the SD values in the last five groups of bar graphs of Fig. 3 decrease compared to the SD values from the unimodal fusion. These results show that the multimodal fusion can reduce the variation of classification accuracy. The decrease in SD leads to the enhancement in the robustness of the syllable recognition system.

B. Performance Comparison

Fig. 4 shows comparison of accuracies from five classifiers obtained by 10-fold cross validation from multimodal fusion. We compare the classification accuracy from neural network

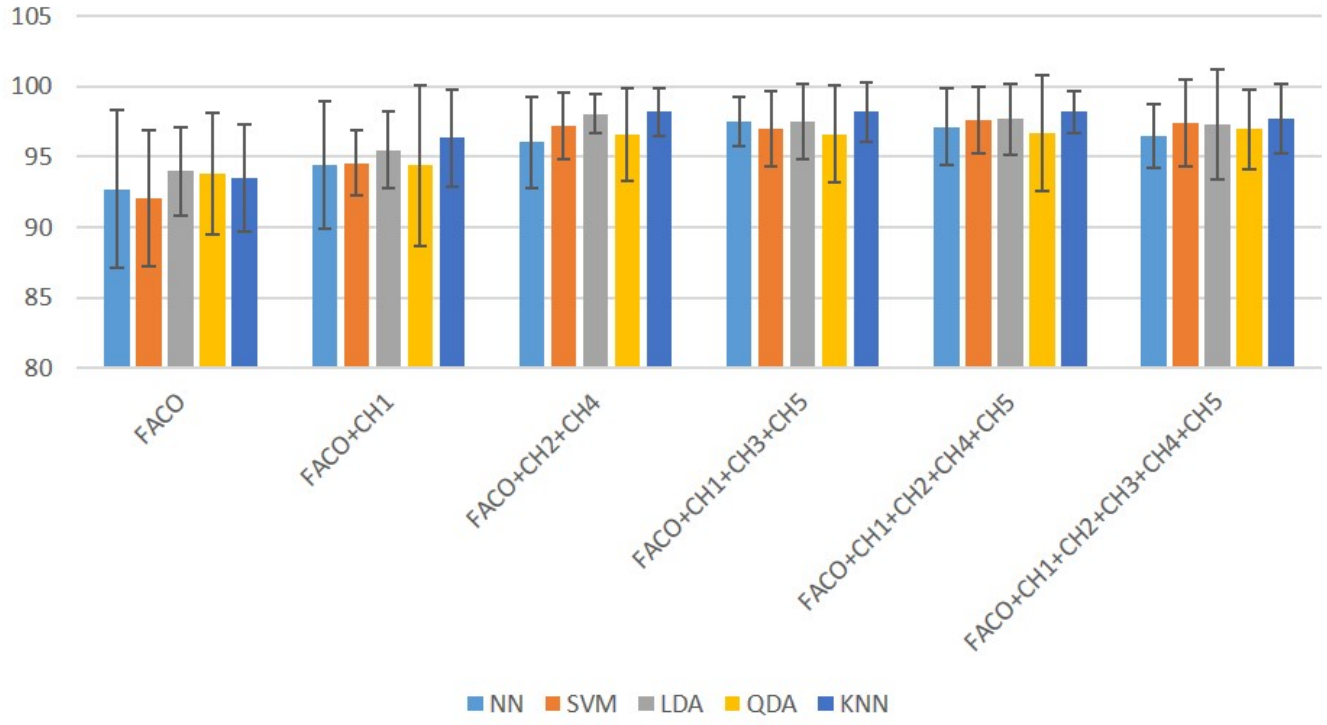

Fig. 4. Comparison of accuracies from five classifiers obtained by 10-fold cross validation from multimodal fusion. NN: neural network, SVM: support vector machine, LDA: linear discriminant analysis, QDA: quadratic discriminant analysis, and KNN: k-nearest neighbor.

(NN) classifier with other four popular classifiers, i.e., support vector machine (SVM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and k-nearest neighbor (KNN). Whereas there are no specified parameters for LDA and QDA, the optimal parameters of SVM are determined using a coarse grid-search method. We use SVM with the radial basis function kernel. The parameter ranges of the cost parameter C and kernel parameter γ are $[10^{-3} - 10^{3}]$. For KNN, the Euclidean distance is used and the number of the nearest neighbors k is optimized using a coarse grid-search method. The range of k is $[1 - 5]$.

Classification accuracies from all classifiers increase about 3-5% when the sEMG and acoustic signals are processed using multimodal fusion. The optimum number of sEMG channels is two. The increase in more than 2 sEMG channels does not increase significant classification accuracy. Among all five classifiers, KNN gives the best classification accuracy at about 98% when the 2, 3, 4, and 5 combined sEMG channel and FACO are obtained. However, two sEMG-combined channels are suggested because the cost of electrodes can be reduced and the electrode placement is easily attached.

VI. Discussion

The results shown in Fig. 3 show that the increase in the number of sEMG channels from one to five can enhance classification accuracy from 47% to 80%. However, it also increases the costs as well as making rehabilitation more difficult to implement by a patient. It is also noticeable that the increase in the MEAN values is not proportional to the number of sEMG-combined channels. In other words, when the number of sEMG-combined channels increases from 1 to 2, the MEAN value increases by 20%. However, when the number of sEMG-combined channels increases from 2 to 3, 3 to 4, and 4 to 5, the MEAN values increase only 6%, 3%, and 3%, respectively.

The maximum accuracy from the multimodal fusion between one acoustic signal and five sEMG channels is 98% for classification of twelve syllables as shown in Fig. 3. It is superior to previous results from the unimodal fusion, such as accuracy at 91% and 94% for classification of five oral activities from six sEMG channels [14] and classification of nine Thai syllables from five sEMG channels [15], respectively. In [13], the multimodal fusion between one acoustic signal and five sEMG channels is used to classify ten words based on phonemes. Accuracies from the multimodal fusion with knowledge of acoustic SNR at 0 and 17.5 dB are 95% and 99%, respectively. These results suggest that the inclusion of the knowledge of SNR into our proposed multimodal fusion may enhance classification accuracy.

Another advantage of multimodal fusion is that it can increase accuracy while the number of sEMG channels reduces. For example, as shown in Fig. 3, We can see that the MEAN value of FACO+CH2+CH4 from the multimodal fusion at 98% is significantly greater than that from the 5 sEMG-combined channels at 80% while the number of electrodes is reduced by 60% resulting in cost-saving and ease of use.

Fig. 5 shows scatter plots between the first two elements of the projected feature vectors from 12 syllables. The result shows that the first two elements of the projected feature vectors from CH2+CH4 in Fig. 5(a) and CH1+CH3+CH5 in

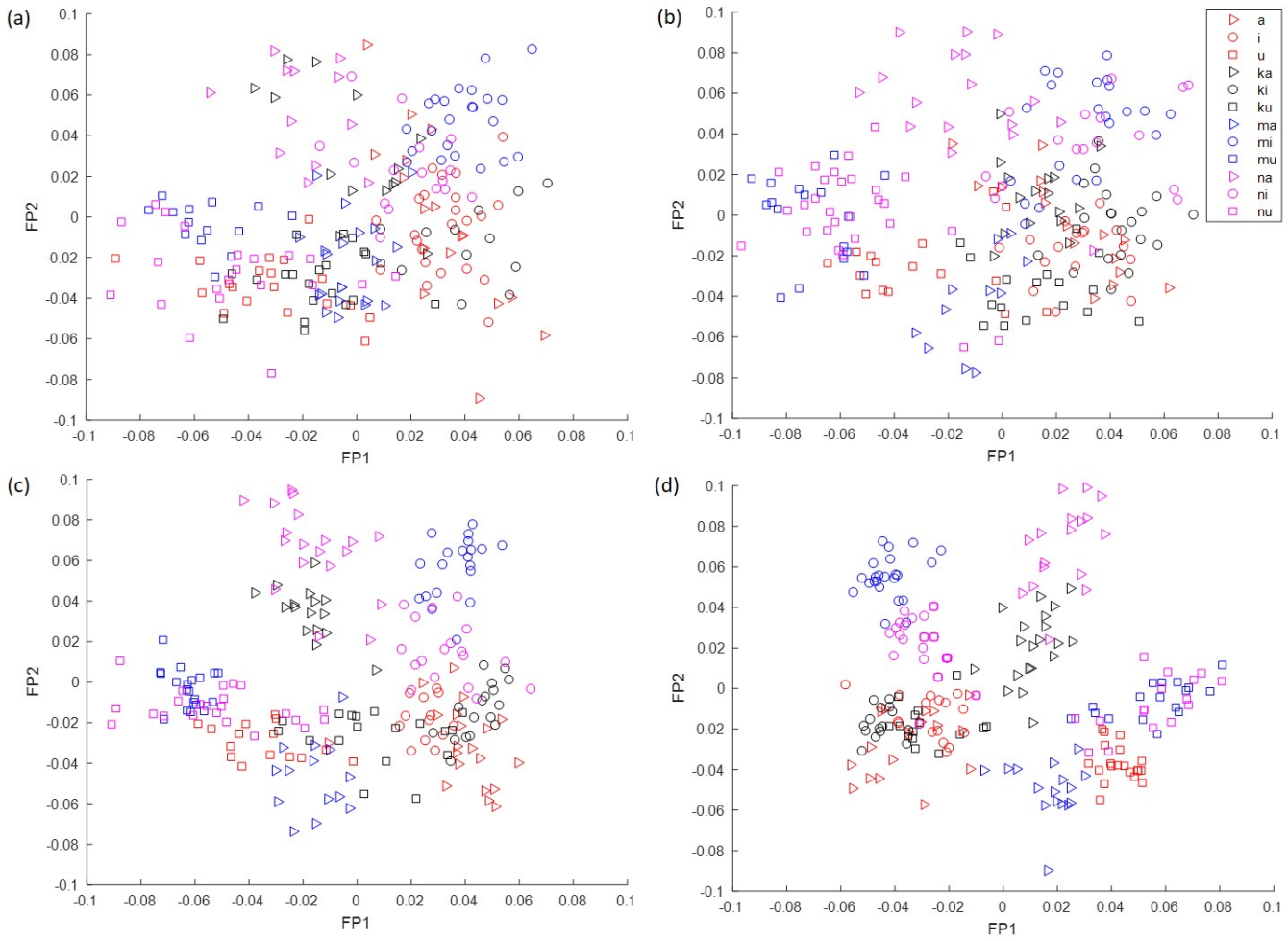

Fig. 5. Scatter plots of the first two elements of the projected feature vector from (a) CH2+CH4; (b) CH1+CH3+CH5; (c) CH1+CH2+CH3+CH4+CH5; and (d) FACO. Twelve markers represent the first two elements from 12 syllables. On the one hand, the scatter plots from CH2+CH4 and CH1+CH3+CH5 are quite overlapped corresponding to average accuracy at 68% and 74%, respectively. On the other hand, the scatter plot from CH1+CH2+CH3+CH4+CH5 shows a better degree of separation corresponding to average accuracy at 80%. However, the scatter plot from FACO shows the best degree of separation corresponding to average accuracy at 92%.

Fig. 5(b) are slightly overlapped, corresponding to the MEAN values at 68% and 74% shown in the bar graphs from Fig. 3, respectively. Fig. 5(c) shows the first two elements of the projected feature vectors from CH1+CH2+CH3+CH4+CH5. Better separation in scatter plot can be seen and supports the corresponding MEAN values at 80% shown in the bar graphs from Fig. 3. Fig. 5(d) shows the first two elements of the projected feature vectors from FACO. The best separation in scatter plot can be seen and supports the corresponding MEAN values at 92% shown in the bar graphs from Fig. 3. However, The scatter plot from FACO shows a different pattern of distribution. In other words, there is no redundancy between them resulting in higher accuracy when performing multimodal fusion.

To gain a clearer insight into the scatter plots in Fig. 5, a circle is drawn for each syllable cluster in Fig. 6. While the center of a circle is a cluster centroid, the radius of a circle is an average Euclidean distance from the centroid to all the points in the cluster. A degree of separation among syllables in the circle plots from Fig. 6(a)-(d) coincides with a degree of cluster separation in the scatter plots from Fig. 5(a)-(d).

As a quantitative measurement, we determine a separation index for each scatter plot with its corresponding circle plot. The separation index is defined as an average inter-cluster distance divided by an average cluster size, where the average inter-cluster distance is the average distance from all possible pairwise Euclidean distances between pairs of 12 centroids ($0.5 \times 12 \times 11 = 66$) and the average cluster size is the average of 12 radii. The average inter-cluster distance is high when different clusters are well separated, whereas the average cluster size is low when all the points in the same cluster are compact. A higher separation index indicates a better degree of syllable cluster separation, which is desirable.

Table II shows that the average inter-cluster distance increases and the average cluster size decreases when the circle plots have a better degree of separation. As a result, the separation indexes from CH2+CH4, CH1+CH3+CH5, CH1+CH2+CH3+CH4+CH5, and FACO show an upward trend at 2.56, 3.30, 3.96, and 4.54, respectively, which match with the MEAN accuracy values at 68%, 74%, 80% and 92% shown in the bar graphs from Fig. 3.

The advantages of the proposed multimodal fusion are the

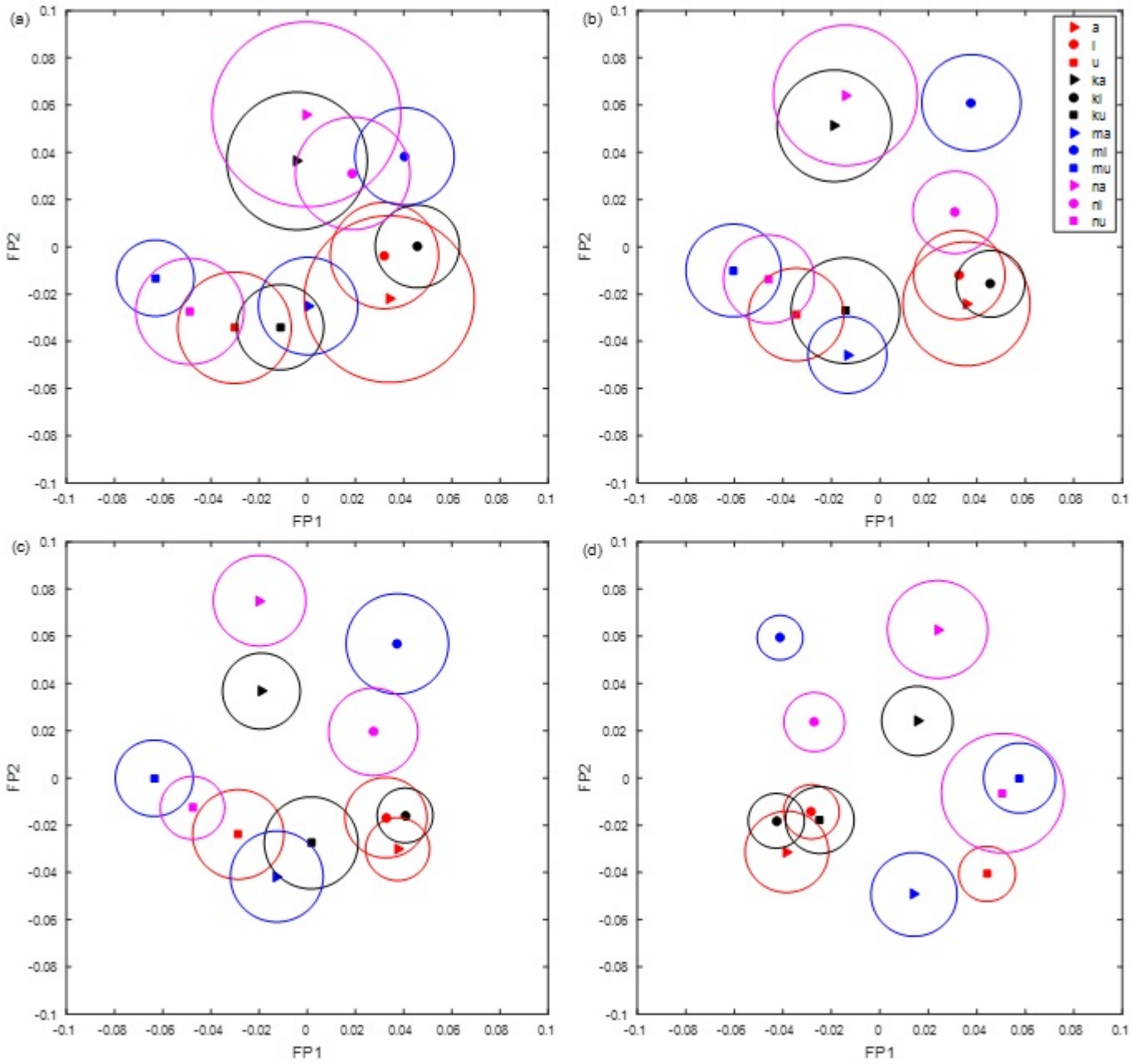

Fig. 6. Circle plots representing the syllable clusters determined based on the scatter plots shown in Fig. 5: (a) CH2+CH4; (b) CH1+CH3+CH5; (c) CH1+CH2+CH3+CH4+CH5; and (d) FACO. The circle plots from CH2+CH4, CH1+CH3+CH5, CH1+CH2+CH3+CH4+CH5, and FACO show a degree of separation corresponding to the average accuracy at 68%, 74%, 80%, and 92%, respectively.

TABLE II
PERFORMANCE COMPARISON OF THREE METRICS DETERMINED BASED ON THE CIRCLE PLOTS FROM FIG. 6.

Metric	CH2+CH4	CH1+CH3+CH5	CH1+...+CH5	FACO
AIC	0.0616	0.0680	0.0674	0.0685
ACS	0.0241	0.0206	0.0170	0.0151
SI	2.56	3.30	3.96	4.54

Notes: AIC: Average inter-cluster distance, ACS: Average cluster size, SI: Separation index.

increase in classification accuracy and the enhancement in the robustness compared to the traditional unimodal fusion. However, further research development on integrating sEMG and acoustic data acquisition systems into a portable device should be carried out for user convenience.

VII. CONCLUSION

This paper proposes a recognition system for classifying 12 Thai syllables, which are used for rehabilitation in the dysarthric patient. The knowledge from the syllable recognition system studied with a healthy subject in this paper will be used as a reference and used to develop a speech rehabilitation system for dysarthric patients in the future. The objective is to replace the SLPs with the speech rehabilitation system in

giving feedback to the dysarthric patients when they perform their rehabilitation at home.

In this paper, unimodal and multimodal fusions are employed to combine sEMG and acoustic signals. While five channels of sEMG signals are acquired from facial muscles and neck muscles, a single channel of the acoustic signal is simultaneously recorded when a subject articulates a syllable. Results show that applying the proposed multimodal fusion on sEMG and acoustic signals outperforms the results using just one source even when applying unimodal fusion. Moreover, the standard deviation when using the multimodal fusion is lower than those from unimodal fusion. These results indicate that the presented multimodal fusion improves not only the accuracy but also the robustness of the speech recognition system. Using multimodal fusion, the number of electrodes for sEMG signal acquisition can be reduced while keeping the recognition accuracy. Hence, the proposed system can be deployed in a speech rehabilitation system for dysarthric patients. Soon, it will be implemented and tested in Hat Yai Hospital, Songkhla, Thailand. The results will be reported in the near future.

Acknowledgment

The authors would like to thank Ms. Duangmon Vongjandaeng from the Department of Otolaryngology, Hat Yai Hospital, Songkhla, Thailand, for her suggestions on the syllables selected based on their use for the rehabilitation of dysarthric speakers and her training on syllable pronunciation for the volunteers.

References

[1] N. Venketasubramanian, B. W. Yoon, and J. C. Navarro, "Stroke epidemiology in South, East, and South-East Asia: A review," *J. Stroke*, vol. 19, no. 3, pp. 286–294, Sep. 2017.

[2] B. H. Dobkin, "Rehabilitation after stroke," *New England J. Med.*, vol. 352, no. 16, pp. 1677–1684, Apr. 2005.

[3] A. B. Kaina, J.-P. Hosom, X. Niu, J. P. van Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Commun.*, vol. 49, no. 9, pp. 743–759, Apr. 2007.

[4] R. Palmer and P. Enderby, "Methods of speech therapy treatment for stable dysarthria: A review," *Adv. Speech Language Pathol.*, vol. 9, no. 2, pp. 140–153, Jul. 2007.

[5] N. C. Suwanwela, "Stroke epidemiology in Thailand," *J. Stroke*, vol. 16, no. 1, pp. 1–7, Jan. 2014.

[6] P. Kayasith and T. Theeramunkong, "Speech confusion index (ϕ): A confusion-based speech quality indicator and recognition rate prediction for dysarthria," *Comput. Math. with Appl.*, vol. 58, no. 8, pp. 1534–1549, Jun. 2009.

[7] M. J. Kim, Y. Kim, and H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 4, pp. 694–704, Apr. 2015.

[8] A. B. Dobrucki, P. Pruchnicki, P. Plaskota, P. Staroniewicz, S. Brachmaski, and M. Walczyski, "Silent speech recognition by surface electromyography," in *New Trends and Developments in Metrology*, L. Cocco, Ed. Rijeka: IntechOpen, 2016, ch. 4.

[9] N. Srisuwan, P. Phukpattaranont, and C. Limsakul, "Comparison of feature evaluation criteria for speech recognition based on electromyography," *Med. Biol. Eng. Comput.*, vol. 56, no. 6, pp. 1041–1051, Nov. 2018.

[10] R. H. Chowdhury, M. Reaz, M. Ali, A. A. Bakar, K. Chellappan, and T. G. Chang, "Surface electromyography signal processing and classification techniques," *Sensors*, vol. 13, no. 9, pp. 12 431–12 466, Sep. 2013.

[11] E. Lopez-Larraz, O. M. Mozos, J. M. Antelis, and J. Minguez, "Syllable-based speech recognition using EMG," in *Proc. 32nd Annu. Int. Conf. Eng. Med. Biol. Soc.*, Buenos Aires, Argentina, 2010, pp. 4699–4702.

[12] A. D. C. Chan, K. Englehart, B. Hudgins, and D. F. Lovely, "Multiexpert automatic speech recognition using acoustic and myoelectric signals," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 4, pp. 676–685, Apr. 2006.

[13] E. J. Scheme, B. Hudgins, and P. A. Parker, "Myoelectric signal classification for phoneme-based speech recognition," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 4, pp. 694–699, Apr. 2007.

[14] T. Pothirat, S. Chatpun, P. Phukpattaranont, and D. Vongjandaeng, "The optimum myography feature for oral muscle movements," in *Proc. 6th Biomed. Eng. Int. Conf.*, pp. 1–5.

[15] N. Sae Jong and P. Phukpattaranont, "A speech recognition system based on electromyography for the rehabilitation of dysarthric patients: A Thai syllable study," *Biocybern. Biomed. Eng.*, vol. 39, no. 1, pp. 234–245, 2019.

[16] B. G. Lapatki, D. F. Stegeman, and I. E. Jonas, "A surface EMG electrode for the simultaneous observation of multiple facial muscles," *J. Neurosci. Methods*, vol. 123, no. 2, pp. 117–128, Mar. 2003.

[17] N. P. Schumann, K. Bongers, O. Guntinas-Lichius, and H. C. Scholle, "Facial muscle activation patterns in healthy male humans: A multi-channel surface EMG study," *J. Neurosci. Methods*, vol. 187, no. 1, pp. 120–128, Mar. 2010.

[18] R. Merletti, A. Botter, A. Troiano, E. Merlo, and M. A. Minetto, "Technology and instrumentation for detection and conditioning of the surface electromyographic signal: State of the art," *Clin. Biomech.*, vol. 24, no. 2, pp. 122–134, 2009.

[19] N. J. O'Dwyer, P. T. Quinn, B. E. Guitar, G. Andrews, and P. D. Neilson, "Procedures for verification of electrode placement in EMG studies of orofacial and mandibular muscles," *J. Speech Lang. Hear. Res.*, vol. 24, no. 2, pp. 273–288, 1981.

[20] M. Hakonen, H. Piitulainen, and A. Visala, "Current state of digital signal processing in myoelectric interfaces and related applications," *Biomed. Signal Process. Control*, vol. 18, pp. 334–359, 2015.

[21] C. E. Stepp, "Surface electromyography for speech and swallowing systems: Measurement, analysis, and interpretation," *J. Speech Language Hear. Res.*, vol. 55, pp. 1232–1246, Aug. 2012.

[22] A. Lumini and L. Nanni, "Overview of the combination of biometric matchers," *Inf. Fusion*, vol. 33, pp. 71–85, Jan. 2017.

[23] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Stat.*, vol. 38, no. 2, pp. 325–339, Apr. 1967.

[24] D. M. Ruscello, "Nonspeech oral motor treatment issues related to children with developmental speech sound disorders," *Language Speech Hear. Ser.*, vol. 39, no. 3, pp. 380–391, Jul. 2008.

[25] Q. Ai, Y. Zhang, W. Qi, Q. Liu, and K. Chen, "Research on lower limb motion recognition based on fusion of sEMG and accelerometer signals," *Symmetry*, vol. 9, no. 8, p. 147, Aug. 2017.

[26] A. H. Al-Timemy, G. Bugmann, J. Escudero, and N. Outram, "Classification of finger movements for the dexterous hand prosthesis control with surface electromyography," *IEEE J. Biomed. Health Inform.*, vol. 17, no. 3, pp. 608–618, May 2013.

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.

[28] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session independent non-audible speech recognition using surface electromyography," in *Proc. IEEE Workshop Autom. Speech Recog. Understanding*, San Juan, Puerto Rico, 2005, pp. 331–336.

[29] Y. Deng, R. Patel, J. T. Heaton, G. Colby, L. D. Gilmore, J. Cabrera, S. H. Roy, C. J. D. Luca, and G. S. Meltzner, "Disordered speech recognition using acoustic and sEMG signals," in *Proc. INTER-SPEECH*, 2009, pp. 644–647.