Thai Vowels Speech Recognition using Convolutional Neural Networks

Niyada Rukwong
Department of Computing, Faculty of Science
Silpakorn University
Nakhon Pathom, Thailand
rukwong_n@su.ac.th

Sunee Pongpinigpinyo*
Department of Computing, Faculty of Science
Silpakorn University
Nakhon Pathom, Thailand
pongpinigpinyo_s@su.ac.th

***Abstract*— The vowel is considered as the core of syllable in each word. This paper aims to present noisy Thai vowels speech recognition by using Convolutional Neural Network (CNN). The noisy Thai vowels dataset is the speech of Thai vowels in real-world situations. The sound is collected in a real environment from several areas which consist of many types of noise at 30 - 40 dB SNR (Signal to Noise Ratio). It constrains 16 kHz speech data is recorded from a mobile phone. The vowel speeches are separated into 2 groups: male's voice and female's voice from 25 male and 25 female speakers. In this research, it constrains 18 classes (9 short vowels and 9 long vowels). Mel Frequency Cepstral Coefficients (MFCCs) are used for feature extraction. The most accuracy rate of the CNN_Thai Simple Vowel (CNN_TSV) model is 90.00% and 88.89% on female and male voices respectively. The comparison results of CNN_TSV model with other models such as Multilayer Perceptron (MLP) and Support Vector Machines (SVM) show that the CNN_TSV model is the most effective for both female and male voices. This research can be used as one of an alternative model to apply for Computer-assisted language learning (CALL) in the future development direction.**

Keywords—Convolutional Neural Networks, CNNs, Vowels, Thai vowels, Classification, Speech Recognition.

I. INTRODUCTION

Vowels always make problems than consonants for second language learners [1] whose native language do not have a difference in the duration of vowels [2]. Moreover, some vowels are difficult to pronounce because the vowel is the type of sound, depending largely on very slight variations of tongue position. Speakers or learners cannot determine for themselves where their tongues are. There is no constriction which we can feel with any precision as we do in consonants. So vowels are most easily described in terms of auditory relationships, or terms of the position of the highest point of the tongue and the position of the lips [3]. To explain or to teach vowel pronunciation effectively, experts or linguists such as phonologists, speech therapist or experienced native speakers with special techniques and extra instruments [4],[5] are required.

Thai language is a tonal language which one syllable combines of Consonant + Vowel (CV) or Consonant + Vowel + Consonant (CVC) [6],[7]. C can be an initial consonant and a final consonant. C(C) is a consonantal cluster. Thai simple vowels are separated into short and long vowels. V is a short vowel. V(V) is a long vowel [8]. Each Thai syllable always starts with a consonant and a tone [9]. Basically, the vowel in each syllable is a nucleus of its syllable [10] and it is the most significant part of the speech event. There are 18 simple monophthongs in Thai vowel system [11] which are divided into 9 short and 9 long. The short and long pairs of vowels are quantitatively different (duration), but they are qualitatively similar (frequency). So, it can be seen that Thai vowels have complex pronunciation. There is no doubt that Thai vowels are much more very difficult for Thai beginner learners or non-native Thai speakers to pronounce Thai vowels properly and correctly.

In this paper, we propose the model of vowels speech recognition for Thai language by using Convolutional Neural Networks (CNNs) as it is one of the most popular deep neural networks. Recently, Convolutional Neural Networks (CNNs) have been applied for Automatic Speech Recognition (ASR) and showed higher performance. CNNs can reduce frequency variant in robustness models. There are a few studies in noise-robust Thai speech recognition, especially for Thai vowels. Since Thai vowel data set are not available for public usage, the data set has recently been collected from 50 Thai native speakers (25 male's voice and 25 female's voice). This paper focuses on apply CNN model in 18 Thai vowels speech recognition with real-world noise.

Remaining part of the paper is prepared as follows: Related work is shown in section II. Section III describes the dataset and the model architecture in our experiments. Section IV reports the results that we obtained. Finally, conclusions are presented.

II. RELATED WORK

There are many research works of vowels that have been proposed, which can be found in [12], [13], [14], [15], [16]. Recently, Convolutional Neural Networks (CNNs) model has been applied in not only computer vision but also speech recognition. In speech recognition tasks, the benefits of the CNNs model are applied in various works because CNNs can be used to reduce frequency/spectral variations. CNN for Automatic speech recognition (ASR) [17] used a variety of strategies, such as pooling and weight sharing, which was a technique used in CNN architecture, therefore results were improved. In a small-footprint keyword spotting (KWS) task [18], CNN architecture was used for 14 phrases classification ('answer call', 'decline call', 'email guests', 'fast forward', 'next playlist', 'next song', 'next track', 'pause music', 'pause this', 'play music', 'set clock', 'set time', 'start timer', and 'take note'). It provided 27 to 44% relative improvement in the false reject rate. Large-scale Speech Tasks [19] and Noise Robust Speech Recognition [20], researches proposed the best CNN architecture and strategies. In Large-scale Speech Tasks, achieved a word error rate (WER) of 12% to 14% relative improvement on 3 Large Vocabulary Continuous Speech Recognition (LVCSR) tasks. These tasks

* Corresponding author.

were a 50 and a 400-hour Broadcast News (BN) task and a 300-hour Switchboard (SWB) task respectively. Noise Robust Speech Recognition on Aurora4 reached 8.81% in WER. It also achieved 10.0% relative reduction over the traditional CNN on AMI meeting transcription task. The robustness of CNN acoustic models [21] was used with 2 techniques to increase performance; autoregressive moving average spectrogram features and channel dropout. Channel dropout method reached 16% in WER with ARMA features and 20% with FBANK features over the baseline CNN. Combination of models was used to increase the efficiency of speech recognition on LVCSR task [22]. Model architecture consisted of CNNs, LSTMs, and DNNs which was called CLDNN. The CLDNN achieved 4 to 6% relative reduction in WER over LSTM. Speech recognition tasks in Thai are shown in [23] using the neuro-fuzzy system with Thai 8 words such as forward, back, left, right recorded in a different noisy environment. It showed that each factor has different effects on recognition accuracy. Double Filter Banks for feature extraction and Euclidian distance for the recognition processes [24] were used with Thai basic voice commanding in various conditions from volunteers (9,000 speech) and achieved accuracy rate is about 96.3 %. Speech Classification was experimented on emotion [25] from 2 corpora which were Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Emotional Tagged Corpus on Lakorn (EMOLA). The emotion was classified into four categories including anger, happiness, neutral, and sadness, It found that each emotion used different features. MFCC with Zero Crossing Rate (ZCR) were good for anger and happiness emotion class with result of 81.95% and 69.86% accuracy respectively. CNN was applied to speech emotion classification [26]. The raw input speech signal was extracted by Convolutional Long Short-Term Memory Neural Network (ConvLSTM-RNN model) and it was classified by Support Vector Machines. It was experimented on IEMOCAP database. The result of SVM with Polynomial Kernel in 192 Phoneme archived accuracy rate at 65.13%. For tasks on vowels, CNN was applied to use with Javanese language which is a language of Indonesia [27], [28]. Mel-frequency spectral coefficients (MFSC) was used to extract the feature. Dataset consisted of 250 Javanese middle vowels sound files recorded by only one speaker. The output consisted of 5 classes. The result was 94% accuracy. In [29] applied the reduction of the order of Linear Predictive Coefficients (LPC). The reduced set of Critical Band Intensities (CBI) was selected. The optimization was used in short and long vowels classification and unmixed and mixed vowels recognition in Thai spoken language. The voices were collected from 6 speakers. Result of classifying frames for short and long unmixed vowels for the 3 male model, the 3 female model, and the 2 male-2 female model archived accuracy at 89.39, 89.83, and 87.67 respectively. The 1,134 samples were used for training in the male model and female model. The 1,512 voice samples for the mixed-gender model. Our research is inspired by many speech recognition tasks that used CNNs that we apply to noisy Thai vowels.

III. EXPERIMENTS

We design the experiments to determine the suitable parameters in the CNN architecture for noisy Thai vowels recognition. Our research compares some methods using various strategies such as Padding [20] which experimental results showed that padding in feature maps for very deep CNNs was important. It could save the size of feature maps and made more improvements. Dropout [30] could reduce the over-fitting problem. Dropout, ReLU, and DNN were applied on a 50-hour English Broadcast News task, results over a DNN with sigmoid, and a GMM/HMM system 4.2%, 14.4% relative improvement respectively. Batch Normalize [31] was used to support the convolution neural network for faster convergence in training. And increasing the number of convolution layers and hidden units are used to evaluate the performance of the model. Finally, our research compares the CNN model with the Multilayer Perceptron (MLP) model and the Support Vector Machines (SVM) model.

A. Dataset

Thai vowels data set is not available for public usage. Therefore, in this research, noisy Thai vowels dataset is the speech of Thai vowels in real-world situations. The sound is collected in a university environment from several areas which consist of many types of noise at 30 - 40 dB SNR (Signal to Noise Ratio) such as vehicles from the road, people talking in the canteen, music at the college of music, wind in the park, and animal sounds like dog and bird. It constrains 16 kHz speech data is recorded from a mobile phone.

The speech of male and female are separately recorded and collected because of the pitch. Following the principle of linguistics study, the pitch of male and female are different: male's pitch is low, but female's pitch is high. The vowel speeches are separated into 2 groups: male's voice and female's voice from 25 males and 25 females. All of them are 20-25 year-old standard Thai speakers. In this research, it constrains 18 classes (9 short vowels and 9 long vowels like Thai simple vowels grouping). Each speaker speaks 2 times each vowel, the total of the collected voice is 1,800 sound files of Thai vowels included of 900 male's files (18 vowels x 25 males x spoken twice) and 900 female's files (18 vowels x 25 females x spoken twice). The 80% of the total files in each group are used for training and the 20% for testing. The sound is collected in normal environments, thus there are many types of noise such as vehicles, people talking, music, wind, animals, and others. 16,000 Hz sampling rate is used for each record in datasets.

TABLE I. THAI SIMPLE VOWEL IN THE INTERNATIONAL PHONETIC ALPHABET (IPA) [6], [7].

Vowels			
Short		***Long***	
Thai letter	***Phonetic***	***Thai letter***	***Phonetic***
อะ	/a/	อา	/a:/
อิ	/i/	อี	/i:/
อึ	/ɯ/	อือ	/ɯ:/
อุ	/u/	อู	/u:/
เอะ	/e/	เอ	/e:/
แอะ	/ɛ/	แอ	/ɛ:/
โอะ	/o/	โอ	/o:/
เอาะ	/ɔ/	ออ	/ɔ:/
เออะ	/ɤ/	เออ	/ɤ:/

After the collecting of noisy Thai vowels speech dataset, a Thai linguist has the sound files cut and selected only vowel sound by using linguistic measurement with PRAAT tool. PRAAT is a computer program for analyzing, synthesizing, and manipulating speech developed by Paul Boersma and David Weenink [32]. It is a formidable research and teaching tool for phonetics that is commonly used by linguists in worldwide phonetic researches because it is probably the most comprehensive toolbox, and it is certainly the most affordable with the top-quality graphic representations of speech.

B. Input features

Previous works, CNNs for speech recognition [19] have been defined input features with a size of #times x #frequencies = 11x40. In the research [20], the default input map size for the model was set to 11x40 as well, and researchers experimented with extending the time and the frequency. They received better results with the full-extension model (21x64) and achieved a WER of 9.8%.

In this research, the speech signal of Thai vowels is preprocessing by a package for audio and music analysis calls LibROSA library in python. For input feature extraction, librosa.feature.mfcc is used for extracted MFCC features: sampling rate of 16000, number of MFCCs to return are 40 and 64 is set in parameters.

To experiment with the appropriate input features, we initialize the default MFCC features to 11x40, and we extend both time and frequency to find the appropriate value in this research.

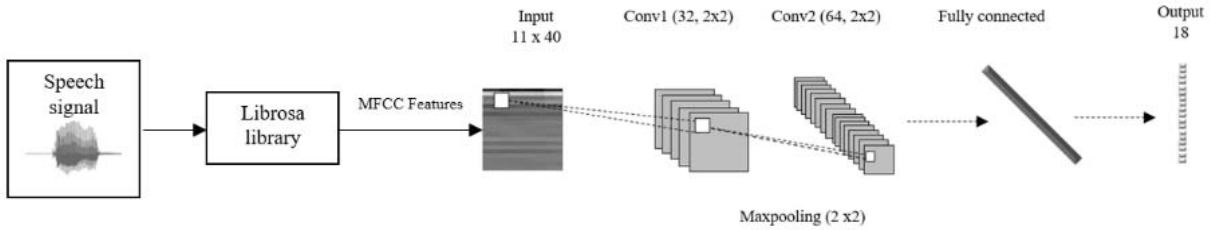

Fig. 1. Feature Extraction and Baseline CNN Architecture

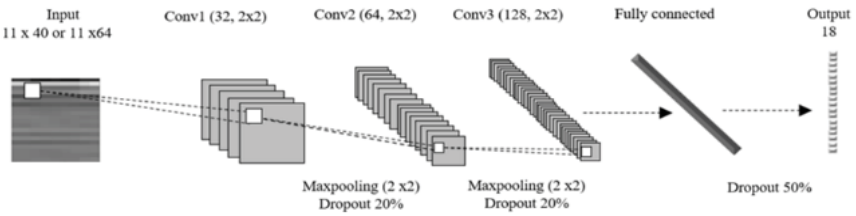

Fig. 2. CNN Architecture of Thai Simple Vowel (CNN_TSV)

C. Baseline structure

In our research, the baseline CNN structure consists of 2 convolutional layers. The first convolutional layer has 32 filters (2x2) followed by max-pooling (2x2), while the second convolutional has 64 filters (2x2) but no pooling layer. ReLU activation function [30] is used in this architecture as it has been used most widely in convolutional neural networks or deep learning and can reduce the calculation time. All of the pooling layers use filter (2x2) and stride of 2. Finally, the fully connected layer uses 64 hidden units and a Softmax activation function. Adam optimizer [11] is used in this model because Adam converges faster and provides high performance. The output classes are 18 classes. The Fig. 1. shows details of the baseline CNN model architecture are built for Thai vowels classification.

D. Model architecture

Our CNN architecture of Thai Simple Vowel in this paper is called the CNN_TSV model that derived from the benefits of experimental results section IV (A-E), and we use it to compare the MLP and SVM models in section IV (F). The CNN_TSV model consists of 3 convolutional layers. The first convolutional layer has 32 filters (2x2) followed by max-pooling (2x2) and dropout 20%, the second convolutional has 64 filters (2x2) followed by max-pooling layer and dropout as the first convolutional layer. While the third convolutional layer has 128 filters (2x2) and uses dropout 20% but no pooling layer. Finally, a fully connected layer uses 64 hidden units, a dropout 50%, and a Softmax activation function. This model uses padding strategy, Adam optimizer, batch normalize is 32. In the female, the appropriate input features are 11x40 and the male are 11x64. This architecture is shown in Fig. 2.

E. Implementation details

For our experiments, we implement models with python on Keras framework and backend is a TensorFlow. We evaluate our model on Windows 64 bit with Intel CORE i7 CPU, 8 GB memory and Nvidia GeForce GTX 1050 GPU.

IV. Results

This research aims to study the appropriate structure using CNN acoustic modeling for Thai vowels speech recognition. The experimental results are presented in the following tables.

A. Time and frequency extension

The literature review [20], our research has found that using our input features and extending time and frequency are useful for the model. To find the appropriate value for the nosy Thai vowels recognition task, this technique has experimented.

TABLE II. RESULTS OF TIME AND FREQUENCY EXTENSION

Input Features*	Accuracy (%)			
	No padding		*padding*	
	Female	*Male*	*Female*	*Male*
11x40	82.78	76.67	80.00	78.89
11x64	80.00	**80.00**	**80.56**	**80.56**
17x40	78.33	78.33	77.78	78.33
17x64	**83.89**	78.33	78.33	78.89
Avg.	81.25	78.33	79.17	79.17

*Input Features: #times x #frequencies

Table II above presents the experimental results, the appropriate input features for both male and female voices are 11x64. In contrast, female at 17x64 shows the best result at 83.89% .Using padding does not improve the performance (not used with any strategy).

B. Dropout

Based on our experimental results on computer vision that demonstrate effective experiments when using dropout and padding. This paper makes more experiments to achieve better performance by using dropout. Moreover, when padding has been used, the results are improved by 5 to 8 % over No padding.

TABLE III. RESULTS OF DROPOUT

Input Features*	Accuracy (%)			
	No padding		*padding*	
	Female	*Male*	*Female*	*Male*
11x40	**87.78**	83.89	87.22	84.44
11x64	87.22	85.56	**88.89**	**87.22**
17x40	85.56	85.56	86.11	83.89
17x64	87.22	**86.67**	86.67	**87.22**
Avg.	86.95	85.42	87.22	85.69

*Input Features: #times x #frequencies

Table III shows the better results by using input features at 11x40, 11x64 and 17x64, therefore these 3 input features are used in the next experiment.

C. Batch normalize

In this section, both with and without batch normalize strategies are tested. Values of batch normalize of 32, 64 and 128 are used respectively.

TABLE IV. RESULTS OF BATCH NORMALIZE (BN)

BN	Accuracy (%)					
	11x40		*11x64*		*17x64*	
	Female	*Male*	*Female*	*Male*	*Female*	*Male*
no	87.22	84.44	**88.89**	87.22	86.67	87.22
32	**88.89**	83.89	87.22	**88.89**	85.56	86.67
64	**88.89**	**86.67**	86.67	86.67	86.11	**88.33**
128	88.33	84.44	86.67	87.22	**87.78**	85.00
Avg.	**88.33**	84.86	87.36	**87.50**	86.53	86.81

Table IV shows the input features (11x40, 11x64 and 17x64) of female and male are taken to find the average value to consider the results. The experiment for the female voice, the appropriate input features are 11x40. Average accuracy at 88.33%. For the male voice, the appropriate input features are 11x64. Average accuracy is 87.50%. The appropriate batch normalize of female and male is 32 which gives 88.89% accuracy.

D. Number of the convolution layer

When extending the convolution layer from 2 to 3, Table V below shows the better results. Especially, the improvement is clear for the female that achieve 90.00%. Although for male voice, the results of increasing the convolution layer are not different, but we believe that adding more convolutional layers will give better results in future experiments.

TABLE V. RESULTS OF NUMBER OF CONVOLUTION LAYER

Number of convolution layer	Accuracy (%)	
	Female (11x40)	*Male (11x64)*
2 layers	88.89	88.89
3 layers	**90.00**	**88.89**

E. Number of hidden units

Our research compares the results of experiments with a different number of hidden units with 3 convolutional layers. Table V provides the effective results that we obtain for both female and male.

TABLE VI. RESULTS WITH DIFFERENT NUMBERS OF HIDDEN UNITS

Number of hidden units	Accuracy (%)	
	Female (11x40)	*Male (11x64)*
64 units	**90.00**	**88.89**
256 units	88.33	87.22
1024 units	86.67	86.11

Table VI, the results conclude that adding hidden units in the fully connected layer does not improve performance, therefore a 64 number of hidden units are used.

F. Comparison between CNN_TSV, MLP and SVM model (k-fold = 10)

The experimental represents comparing the CNN_TSV model with the MLP model and the SVM model. The CNN_TSV model is derived from experimental results section IV (A-E). The result from Fig. 3. shows that the appropriate epochs on the female voice are 500 epochs. The mean of accuracy is 84.86% and the standard deviation is +/- 4.14%. On male voice, the appropriate epochs are 1000 at 89.72% and the standard deviation is +/- 2.79%. The Multilayer Perceptron Classifier (MLP Classifier) is used on a baseline of the MLP model consists of a hidden layer of 256 units. The RELU activation is used in this model, Solver is Adam optimization, 32 batch size and the initial learning rate is 0.001. For the SVM model uses Support Vector Classification (SVC), linear is set to the kernel, and decision function of shape is one-vs-rest ('ovr'). All models use the same input features.

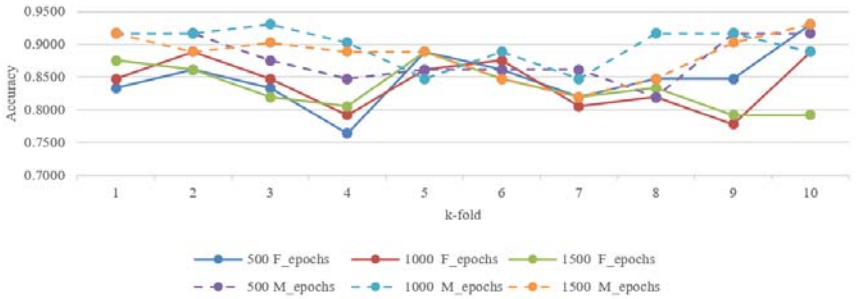

Fig. 3. The result of epochs (500, 1000, 1500) on the female/male voice.

TABLE VII. RESULTS WITH DIFFERENT METHODS

Methods	Mean Accuracy (%)	
	Female (11x40)	*Male (11x64)*
CNN_TSV	**84.86**	**89.72**
MLP	68.12	71.85
SVM	76.00	82.17

As described in Table VII, the results show the CNN_TSV model provides the highest mean accuracy and efficiency for both female and male, 84.86% and 89.72% mean accuracy respectively.

G. The confusion matrix, Precision, Recall, and F1-score of the CNN_TSV model

For error analysis, the confusion matrix of the CNN_TSV model on female and male voices are shown in Fig. 4. From the confusion matrix, the most confusing pair of Thai vowels on female voice are (‘โ’ /o:/ and ‘โะ’ /o/). Subordinate confusing pairs are (‘อือ’ /ɯ:/ and ‘เอ’ /ɤ:/), (‘เออ’ /ɤ:/ and ‘เออะ’ /ɤ/), (‘อิ’ /i/ and ‘อี’ /i:/), (‘อิ’ /i/ and ‘เอะ’ /e/. On male voice, the most confusing pairs of Thai vowels are (‘โ’ /o:/ and ‘โะ’ /o/), as the female voice, and (‘เออ’ /ɤ:/ and ‘เออะ’ /ɤ/). Subordinate confusing pair are (‘ออ’ /ɔ:/ and ‘เอาะ’/ɔ/. The experiment has found that ‘โ’ /o:/ and ‘โะ’ /o/ vowels are the most confusing pair of both genders. Moreover, the confusing pairs are short and long vowels which contrast in the duration.

Predicted class \ Actual class	/a:/	/i:/	/ɯ:/	/u:/	/e:/	/ɛ:/	/o:/	/ɔ:/	/ɤ:/	/a/	/i/	/ɯ/	/u/	/e/	/ɛ/	/o/	/ɔ/	/ɤ/
/a:/	**8**	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
/i:/	0	**15**	0	0	2	0	0	0	0	0	**2**	0	0	0	0	0	0	0
/ɯ:/	0	**1**	**5**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/u:/	0	0	0	**5**	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0
/e:/	0	0	0	0	**9**	0	0	0	0	0	0	0	0	0	0	0	0	0
/ɛ:/	0	0	0	0	0	**8**	0	0	0	0	0	0	0	0	0	0	0	0
/o:/	0	0	0	**1**	0	0	**8**	0	0	0	0	0	0	0	0	0	0	0
/ɔ:/	**1**	0	0	0	0	0	0	**9**	0	0	0	0	0	0	0	0	**1**	0
/ɤ:/	0	0	**2**	0	**1**	0	0	0	**10**	0	0	0	0	0	0	0	0	0
/a/	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	0	0	0
/i/	0	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	0	0
/ɯ/	0	0	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	0
/u/	0	0	0	0	0	0	0	0	0	0	**1**	0	**6**	0	**1**	0	0	0
/e/	0	0	0	0	**1**	0	0	0	0	0	**2**	0	0	**10**	0	0	0	0
/ɛ/	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**8**	0	0	0
/o/	0	0	0	0	0	0	**3**	0	0	0	0	0	0	0	0	**8**	0	0
/ɔ/	0	0	0	0	0	0	**1**	**1**	0	0	0	0	0	0	0	0	**11**	0
/ɤ/	0	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0	0	0	**7**

Predicted class \ Actual class	/a:/	/i:/	/ɯ:/	/u:/	/e:/	/ɛ:/	/o:/	/ɔ:/	/ɤ:/	/a/	/i/	/ɯ/	/u/	/e/	/ɛ/	/o/	/ɔ/	/ɤ/
/a:/	**7**	0	0	0	0	0	0	0	**1**	**1**	0	0	0	0	0	0	0	0
/i:/	0	**15**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/ɯ:/	0	0	**7**	0	0	0	0	0	0	0	**1**	**2**	0	0	0	0	0	0
/u:/	0	0	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/e:/	0	0	0	0	**13**	0	0	0	0	0	0	0	0	0	0	0	0	0
/ɛ:/	**1**	0	0	0	0	**8**	0	0	0	0	0	0	0	0	**2**	0	0	0
/o:/	0	0	0	0	0	0	**7**	0	0	0	0	0	0	0	0	**1**	0	0
/ɔ:/	0	0	0	0	0	0	0	**7**	0	0	0	0	0	0	0	0	0	0
/ɤ:/	0	0	0	0	0	0	0	0	**7**	0	0	0	0	0	0	0	0	**1**
/a/	**1**	0	0	0	0	0	0	0	0	**9**	0	0	0	0	0	0	0	0
/i/	0	**1**	0	0	0	0	0	0	0	0	**12**	0	0	0	0	0	0	0
/ɯ/	0	0	0	0	0	0	0	0	0	0	**1**	**7**	0	0	0	0	0	0
/u/	0	0	0	0	0	0	**1**	0	0	0	0	0	**6**	0	0	0	0	0
/e/	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0	0
/ɛ/	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**7**	0	0	0
/o/	0	0	0	0	0	0	**4**	0	0	0	0	0	0	0	0	**8**	0	0
/ɔ/	0	0	0	0	0	0	0	**3**	0	0	0	0	0	0	0	0	**12**	0
/ɤ/	0	0	0	0	0	0	0	0	**4**	0	0	0	0	0	0	0	0	**6**

Fig. 4. The confusion matrix of the CNN_TSV model on the female (top)/male(bottom).

TABLE VIII. PRECISION, RECALL, AND F1-SCORE OF THE CNN_TSV MODEL.

Thai Vowels	Female			Male		
	Precision	*Recall*	*F1-score*	*Precision*	*Recall*	*F1-score*
a:	0.89	0.89	0.89	0.78	0.78	0.78
i:	0.79	0.94	0.86	1.00	0.94	0.97
ɯ:	0.83	0.71	0.77	0.70	1.00	0.82
u:	0.83	0.83	0.83	1.00	1.00	1.00
e:	1.00	0.69	0.82	1.00	1.00	1.00
ɛ:	1.00	0.89	0.94	0.73	0.89	0.80
o:	0.89	0.67	**0.76**	0.88	0.58	**0.70**
ɔ:	0.82	0.90	0.86	1.00	0.70	0.82
ɤ:	0.77	0.83	0.80	0.88	0.58	**0.70**
a	1.00	0.90	0.95	0.90	0.90	0.90
i	1.00	0.64	0.78	0.92	0.86	0.89
ɯ	1.00	1.00	1.00	0.88	0.78	0.82
u	0.75	1.00	0.86	0.86	1.00	0.92
e	0.77	1.00	0.87	1.00	1.00	1.00
ɛ	0.89	0.89	0.89	0.88	0.78	0.82
o	0.73	0.89	0.80	0.67	0.89	0.76
ɔ	0.85	0.92	0.88	0.80	1.00	0.89
ɤ	0.78	1.00	0.88	0.60	0.86	0.71

Table VIII presents the precision, recall, and f1-score of the CNN_TSV model for classifying each vowel. The lowest of F1 score on the female voice is ‘โอ’ /o:/ (0.76) , and the male voice are ‘โอ’ /o:/ (0.70) and ‘เออ’ /ɤ:/ (0.70). The f1-score results are relevant to the confusion matrix. On the other hand, the highest of F1 score (1.00) on the female voice is ‘อึ’ /ɯ/, and the male voice are ‘อู’ /u:/, ‘เอ’ /e:/, and ‘เอะ’ /e/.

V. CONCLUSIONS

This research presents a nosy Thai vowels speech recognition task using a CNN acoustic model. Our research experiment on a newly collected noisy data set. This dataset consists of 25 female and 25 male voice records. The voices are grouped into 18 classes for each gender. The evaluation is done with time and frequency expansion. We have found that the appropriate input features are 11x40 for female voice and 11x64 for male voice. Padding and Dropout are used by 20%, it improves the performance by 5% - 8%. The appropriate value of Batch normalize strategy is at 32. Increasing the convolution layers to 3 gives a better to be performed in both groups. The appropriate of hidden units is 64. The most accuracy rate of CNN_TSV model is 90.00% and 88.89% on female and male voices respectively.

Finally, the comparison of the CNN_TSV model with MLP and SVM model that the CNN_TSV model is the most effective for both female and male voices. The mean accuracies reached are 84.86% and 89.72% respectively. The most confusing pair of both genders are ‘โอ’ /o:/ and ‘โอะ’ /o/ vowels. The highest of F1 score (1.00) on the female voice is ‘อึ’ /ɯ/, and the male voice are ‘อู’ /u:/, ‘เอ’ /e:/, and ‘เอะ’ /e/.

REFERENCES

[1] B. G. Evans and W. Alshangiti, “The perception and production of British English vowels and consonants by Arabic learners of English,” *J. Phon.*, vol. 68, pp. 15–31, 2018.

[2] L. Rallo Fabra and J. Romero, “Native Catalan learners’ perception and production of English vowels,” *J. Phon.*, vol. 40, no. 3, pp. 491–508, 2012.

[3] K. J. Peter Ladefoged, *A Course in Phonetics*, Sixth. Michael Rosenberg.

[4] X. Peng, H. Chen, L. Wang, and H. Wang, “Evaluating a 3-D virtual talking head on pronunciation learning,” *Int. J. Hum. Comput. Stud.*, vol. 109, no. August 2017, pp. 26–40, 2018.

[5] M. Tabain and R. Beare, “An ultrasound study of coronal places of articulation in Central Arrernte: Apicals, laminals and rhotics,” *J. Phon.*, vol. 66, pp. 63–81, 2018.

[6] L. Jeerapradit, A. Suchato, and P. Punyabukkana, “HMM-based Thai Singing Voice Synthesis System,” in *2018 22nd International Computer Science and Engineering Conference (ICSEC)*, 2019, pp. 1–4.

[7] S. Aunkaew, M. Karnjanadecha, and C. Wutiwiwatchai, “Constructing a phonetic transcribed text corpus for Southern Thai dialect Speech Recognition,” in *Proceedings of the 2015 12th International Joint Conference on Computer Science and Software Engineering, JCSSE 2015*, 2015, pp. 69–73.

[8] A. Munthuli, C. Tantibundhit, C. Onsuwan, K. Kosawat, and C. Wutiwiwatchai, “Frequency of occurrence of phonemes and syllables in Thai: Analysis of spoken and written corpora,” in *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*, 2015, pp. 3–7.

[9] K. Supphanat, “Syllable Structure Based Phonetic Units for Context-Dependent Continuous Thai Speech Recognition,” in *EUROSPEECH*, 2003, pp. 797–800.

[10] R. Grammar, *Thai Reference Grammar*. U.S. Government Printing Office, 1964.

[11] S. K. Gouda, S. Kanetkar, D. Harrison, and M. K. Warmuth, “Speech Recognition: Keyword Spotting Through Image Recognition,” 2018. [Online]. Available: http://arxiv.org/abs/1803.03759.

[12] Š. Šimáčková and V. J. Podlipský, “Production accuracy of L2 vowels: Phonological parsimony and phonetic flexibility,” *Res. Lang.*, vol. 16, no. 2, pp. 169–191, 2018.

[13] S. Sahatsathatsana, “Pronunciation Problems of Thai Students Learning English Phonetics: A Case Study at Kalasin University,” *J. Educ.*, vol. 11, no. 4, pp. 67–84, 2017.

[14] P. Ghaffarvand Mokari and S. Werner, “Perceptual assimilation predicts acquisition of foreign language sounds: The case of Azerbaijani learners’ production and perception of Standard Southern British English vowels,” *Lingua*, vol. 185, pp. 81–95, 2017.

[15] K. Mirzaei, H. Gowhary, A. Azizifar, and Z. Esmaeili, “Comparing the Phonological Performance of Kurdish and Persian EFL Learners in Pronunciation of English Vowels,” *Procedia - Soc. Behav. Sci.*, vol. 199, pp. 387–393, 2015.

[16] M. Navehebrahim, “An Investigation on Pronunciation of Language Learners of English in Persian Background: Deviation Forms from the Target Language Norms,” *Procedia - Soc. Behav. Sci.*, vol. 69, no. Iceepsy, pp. 518–525, 2012.

[17] S. Newatia and R. K. Aggarwal, “Convolutional Neural Network for ASR,” in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 638–642.

[18] T. N. Sainath and C. Parada, “Convolutional Neural Networks for Small-footprint Keyword Spotting,” in *Interspeech*, 2015, pp. 1478–1482.

[19] T. N. Sainath *et al.*, “Deep Convolutional Neural Networks for Large-scale Speech Tasks,” *Neural Networks*, vol. 64, pp. 39–48, 2015.

[20] Y. Qian and P. C. Woodland, “Very Deep Convolutional Neural Networks for Robust Speech Recognition,” *IEEE/ACM Trans. AUDIO, SPEECH, Lang. Process.*, vol. 24, no. 12, pp. 2263–2276, 2016.

[21] G. Kovács, L. Tóth, D. Van Compernolle, and S. Ganapathy, “Increasing the robustness of CNN acoustic models using autoregressive moving average spectrogram features and channel dropout,” *Pattern Recognit. Lett.*, vol. 100, pp. 44–50, 2017.

[22] T. N. Sainath, O. Vinyals, A. Senior, and N. York, “Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4580–4584.

[23] K. Srijiranon and N. Eiamkanitchat, “Thai speech recognition using Neuro-fuzzy system,” in *ECTI-CON 2015 - 2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2015, pp. 1–6.

[24] P. Phokharatkul, K. Nantanitikorn, and S. Phaiboon, “Thai speech recognition using Double filter banks for basic voice commanding,” in *2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering*, 2010, vol. 6, pp. 33–36.

[25] P. Sukhummek, S. Kasuriya, T. Theeramunkong, C. Wutiwiwatchai, and H. Kunieda, “Feature Selection Experiments on Emotional Speech Classification,” in *2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2015, pp. 1–4.

[26] N. Kurpukdee, T. Koriyama, and T. Kobayashi, “Speech Emotion Recognition using Convolutional Long Short-Term Memory Neural Network and Support Vector Machines,” 2017, no. December, pp. 1744–1749.

[27] C. K. Dewa and Afiahayati, “Suitable CNN Weight Initialization and Activation Function for Javanese Vowels Classification,” *Procedia Comput. Sci.*, vol. 144, pp. 124–132, 2018.

[28] C. K. Dewa, “Javanese vowels sound classification with convolutional neural network,” in *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2016, pp. 123–128.

[29] N. Suktangman, K. Khanthavivone, and K. Songwatana, “Optimizing vowel recognition in Thai spoken language using reduced LPC spectrum and reduced feature set of critical band intensities,” in *2006 International Symposium on Communications and Information Technologies, ISCIT*, 2006, no. 4, pp. 128–132.

[30] G. E. Dahl, T. N. Sainath, and G. E. Hinton, “Improving deep neural networks for LVCSR using rectified linear units and dropout,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8609–8613.

[31] L. Wenjie, G. Cheng, F. Ge, P. Zhang, and Y. Yan, “Investigation on the Combination of Batch Normalization and Dropout in BLSTM-based Acoustic Modeling for ASR,” in *Interspeech 2018*, 2018, vol. 2018, pp. 2888–2892.

[32] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program],” *Glot Int.*, vol. 5, no. 9–10, pp. 341–347, 2001.