applied sciences

MDPI

Article

An Acoustic Feature-Based Deep Learning Model for Automatic Thai Vowel Pronunciation Recognition

Niyada Rukwong and Sunee Pongpinigpinyo *

Department of Computing, Faculty of Science, Silpakorn University, Nakhon Pathom 73000, Thailand; rukwong_n@silpakorn.edu

* Correspondence: pongpinigpinyo_s@su.ac.th

Abstract: For Thai vowel pronunciation, it is very important to know that when mispronunciation occurs, the meanings of words change completely. Thus, effective and standardized practice is essential to pronouncing words correctly as a native speaker. Since the COVID-19 pandemic, online learning has become increasingly popular. For example, an online pronunciation application system was introduced that has virtual teachers and an intelligent process of evaluating students that is similar to standardized training by a teacher in a real classroom. This research presents an online automatic computer-assisted pronunciation training (CAPT) using deep learning to recognize Thai vowels in speech. The automatic CAPT is developed to solve the inadequacy of instruction specialists and the complex vowel teaching process. It is a unique system that develops computer techniques integrated with linguistic theory. The deep learning model is the most significant part of recognizing vowels pronounced for the automatic CAPT. The major challenge in Thai vowel recognition is the correct identification of Thai vowels when spoken in real-world situations. A convolutional neural network (CNN), a deep learning model, is applied and developed in the classification of pronounced Thai vowels. A new dataset for Thai vowels was designed, collected, and examined by linguists. The result of an optimal CNN model with Mel spectrogram (MS) achieves the highest accuracy of 98.61%, compared with Mel frequency cepstral coefficients (MFCC) with the baseline long short-term memory (LSTM) model and MS with the baseline LSTM model have an accuracy of 94.44% and 90.00% respectively.

Keywords: computer-assisted pronunciation training; convolutional neural networks; Thai vowels; speech recognition; mel spectrogram; mel frequency cepstral coefficients

check for updates

Citation: Rukwong, N.; Pongpinigpinyo, S. An Acoustic Feature-Based Deep Learning Model for Automatic Thai Vowel Pronunciation Recognition. *Appl. Sci.* **2022**, *12*, 6595. https://doi.org/10.3390/app12136595

Academic Editor: Douglas O'Shaughnessy

Received: 3 May 2022
Accepted: 21 June 2022
Published: 29 June 2022

1. Introduction

Speech is the most basic form of human communication, it is simple and convenient to use. Language instruction is now allocated to different groups of people to meet different needs of targets. Language learning, generally, can be divided into first language learning, second language learning, or third language learning. Moreover, language therapy is also available for people with disabilities or disorders. Therefore, there are various groups of language learners with different objectives. For example, (1) according to government policy, students learn their mother tongue and a second language in schools; (2) to achieve career advancement, a third language is needed for the working-age group; and (3) in some cases, patients undergoing laryngeal surgery need to practice speaking after treatment. Those who practice language teaching, therefore, can be divided into three main groups: language teachers, linguists, and language therapists.

1.1. Motivation to Solve Pronunciation Problem of Thai Vowels

Pronunciation correction is an essential goal of language learning. Correct and clear pronunciation is crucial for ensuring that listeners correctly comprehend the meaning of certain words. Good pronunciation makes communication more effective [1]. Proper language practice can improve pronunciation skills and helps learners pronounce languages more accurately. Generally, language teachers, specialists, linguists, speech therapists, and native speakers are the people who explain and teach pronunciation to language learners using the listening method.

However, because of diverse circumstances between teachers and learners, inaccuracies in judging the appropriateness of pronunciation through hearing are conceivable. The quantity of teachers available is insufficient to meet the needs of all students. A teacher's job is to listen to and correct the mispronunciations of many students throughout the day. As a result, weariness may cause them to lose focus on major information. Furthermore, learners can only perfect their pronunciation in the classroom.

Because of the aforementioned limitations, coupled with the situation of the COVID-19 pandemic at present, online learning has become necessary. To solve these problems, an automated system for the intelligent assessment of pronunciation practice is an alternative choice. It can improve the efficiency of the learning process for both learners and teachers because learners are able to study at any location and at any time. Therefore, this research collected recordings of Thai vowels in various environments, for instance at schools, cafeterias, parks, classrooms, bedrooms, and homes, as well as various native speakers in real-world situations, to train the model.

With the advancement of information technology and increased computing capabilities, deep learning is increasingly applied for recognition tasks. Deep learning is effective in learning and classification, [2] such as handwritten character recognition and speech recognition. Deep learning is an artificial intelligence (AI) mathematical technique for classification that depends on data using a multilayered neural network. Computer-assisted language learning (CALL) and computer-assisted pronunciation training (CAPT) [3–8] have gained much attention in the field of language teaching and training. CALL and CAPT systems are widely used to improve language learning and teaching methods. CALL and CAPT systems can recognize speech using automatic speech recognition (ASR), which implemented deep learning structures. The deep learning model is applied to increase the accuracy of Thai vowel pronunciation classification in an automated system.

1.2. Contributions in Automatic Thai Vowels Pronunciation Recognition

In order to accurately recognize pronunciations, there are many tasks that use specific techniques or special tools to analyze them, for example, developing a 3D pronunciation learning system in Chinese [9], using ultrasound with the pronunciation of consonants [10], developing an application that helps to analyze tones in Thai [11], and applying the Praat program for analyzing phonetics to demonstrate the pronunciation of vowels [12–17].

To increase the accuracy of pronunciation, linguists use the acoustic phonetics method for phonetic analysis. Praat [18] is a popular tool for acoustic phonetics analysis. It is used to display two formant frequencies, which are Formant1 (F1) and Formant2 (F2). In order to display whether the pronunciation of the vowel is correct or incorrect, the location of the tongue must be represented with a specific graph. This is a complex process that requires specialists. The tongue position analysis process by a linguist starts with (1) recording and playing the vowel sound with the Praat program; (2) checking and selecting the range of the vowel sound in the syllable to measure; (3) compiling the average of the F1 and F2 of the vowels that were chosen; (4) in Microsoft Excel, recording the average values of F1 and F2; (5) using Microsoft Excel, Python, or R, creating a vowel graph of native speakers and learners; (6) comparing the graphs of native speakers and learners; and (7) presenting and explaining the incorrect pronunciation from the results to the learners and correcting them. The method for determining pronunciation accuracy with graphs is a complex, time-consuming, and nonreal-time process. The user, moreover, must be an expert or someone with programming knowledge. Therefore, there are few specialists in the field today.

So far, there are a few research works conducted to classify Thai vowels pronunciation using an intelligent system. From the literature review, there is no study on automatic

CAPT with Thai vowels using AI with a deep learning structure. Therefore, this research aims to design and develop an automatic CAPT system using a deep learning structure for Thai vowel speech recognition. This system is developed for solving the problems of practicing the pronunciation of Thai vowels for (1) nonnative learners or nonstandard Thai speakers and (2) people with pronunciation disabilities; (3) for solving the shortage of specialists in teaching Thai vowels pronunciation; (4) for solving the original process, which is complicated and time-consuming and does not present results in real time; and (5) inventing a new tool for learning languages online that is appropriate for the current situation.

A deep learning structure of CAPT for Thai vowels speech is designed to recognize the 18 basic Thai vowels. The pronunciations of the 18 Thai vowels are difficult. Some phonemes have similar characteristics. Therefore, nonnative learners cannot distinguish them and require assistance from an expert. In this research, a convolutional neural network (CNN) is trained over the dataset for Thai vowels speech classification. This model is created to train a computer to recognize vowels like an expert who can identify learners' vowel pronunciation. The existing Thai audio corpus is not suitable for the objectives of this research. Therefore, to obtain theoretically qualitative vowel data in linguistic principles, a new dataset is designed and collected. The dataset used for training this model consists of voices collected from various dimensions in real-life contexts, such as gender, age, accent, environment, and noise. The major contributions to this work are as follows:

1. This research proposes a noisy dataset that is collected from standard Thai speakers from various dimensions. The dataset is designed, collected, and examined by a linguist.
2. This research proposes a method for Thai vowel speech recognition based on a convolutional neural network (CNN), which is one of the most well-known deep neural networks (DNNs). The CNN model is applied to automatic speech recognition on the automatic CAPT for Thai vowels. The optimal CNN model is utilized to learn the spectral characteristics of 18 Thai vowel classes.
3. This research generates two different acoustic feature inputs for CNN and long short-term memory (LSTM) models. They are MS and MFCC acoustic features from the raw speech waveform to learn deep multimodal features.
4. This research proposes automatic CAPT for Thai vowel speech that can display the learners' vowel pronunciation results. If the pronunciation is incorrect, then it will suggest the correct practice with text, real video, and 3D video.

The automatic CAPT for Thai vowel speech uses a deep learning structure. It is a new system that develops computer techniques integrated with linguistic theory. The system can be used to guide learners such as the voice impaired, nonnative learners, and nonstandard Thai speakers. This system, therefore, allows learners to practice vowel pronunciation in real time, similar to having an expert, Thai teachers, and linguists provide assistance on the correctness of pronouncing vowels.

The outcome of this work benefits the innovation of advanced systems, for example, the classification of tones, words, phrases, and sentences to facilitate learning standard Thai pronunciation.

The remainder of the paper is organized as follows. In Section 2, the background is described, related works are presented in Section 3, and materials and methods are shown in Section 4. Section 5 describes the results of the experiments, and conclusions are displayed in Section 6. In Section 7, the discussion is presented. Finally, the automatic computer-assisted pronunciation training for Thai vowels is presented. The definitions of variables and acronyms are shown in Appendix A (Table A1) after the last section.

2. Background

There are different levels to language learning, starting from practicing pronunciation of phonemes and words to conversations and storytelling. Phonemes are the smallest units of a language and are fundamental units for advanced language learning. Different types of phonemes differ from language to language such as vowels, consonants, or tones.

2.1. Vowel's Pronunciation Problems

In general, children learn their mother tongue from family members and their communities. At school age, they are trained by teachers based on the curriculum that has been approved by the Ministry of Education. Past works show that children with speech organ problems from diseases such as autism or from birth defects encounter first language learning problems. Children with speech sound disorders (SSD) struggle to produce vowel sounds in comparison with normal children [19], while speakers with Down syndrome (DS) have difficulty pronouncing corner vowels (/α/, /æ/, /i/, /u/) [20].

People communicate worldwide with different groups for a variety of reasons: trade, transport, medical aid, and education. Accordingly, foreign languages are learned as second languages and third languages. The differences in phonemes between one's mother tongue and the new language are essential factors in pronunciation practice. Previous studies [21–25] discovered that learners have difficulty with the pronunciation of vowels when learning second and third languages.

This research focuses on vowel sounds. Vowels are the core of syllables (nucleus) and are a significant part of speech [26,27]. Vowels, at least, are distinguished by three characteristics: the nature and position of the tongue, the duration, and the rhythm of speech or melody. Vowels are important phonemes that affect meaning and language behavior. Vowels can be trained in perception (listening) and production (speaking) with a phonetic approach. A vowel pronunciation test cannot be used as an articulation test like consonants [28]. Vowel pronunciation is more problematic than consonant pronunciation [29]. The practice of short–long vowel pronunciation is a problem for learners for whom in their mother tongue, the duration of the vowels does not affect the meaning of words [30]. Practicing vowels that do not exist in the mother tongue is more difficult than practicing consonants since the movement of the tongue heavily affects the pronunciation of a vowel. Since learners cannot know where exactly their tongue is, those who are practicing pronunciation should have an expert to guide them and teach them the correct pronunciation. The most popular description of vowel sounds involves describing the position of the high–low level, the front–back of the tongue, the area of the tongue, and the appearance of the lips [27]. However, this description is still difficult to understand and can often confuse learners.

2.2. Thai Vowels and Pronunciation Problems

The Thai language is one of the most difficult languages in the world to learn due to its challenging pronunciation [31]. Thai is a tonal language: one syllable consists of a first consonant, vowel, final consonant, and tone [32]. The forms of the syllables are: (1) Consonant + Vowel (CV) such as [ปา /pa:/ 'to hurl'] or [ดู / du:/ 'to look'], (2) Consonant + Vowel + Consonant (CVC) such as [บาน /ba:n/ 'to bloom'] or [กิน /kin/ 'to eat']. A diphthong can be written as C(C) such as [ปลา /pla:/ 'fish'] or [คลาน /khla:n/ 'to crawl'] [33,34]. Basic Thai vowels are divided into two types: short vowels and long vowels. V is a short vowel, such as [ติ /ti/ 'to censure'] or [ดุ /du/ 'to scold'], and V(V) is a long vowel, such as [ตี /ti:/ 'to hit'] or [ดู /du:/ 'to watch'] [35].

Each language has a different number of vowel sounds, for example, 20 sounds in English or 5 sounds in Japanese. There are 32 Thai vowel sounds, which are represented by 21 letters. This difference in the number of vowels can be a problem when learning Thai as a second or third language. Thai has short–long pairs of vowels that are often a problem for those whose mother tongue does not have a difference in duration such as [สด /sot/ 'fresh'] and [โสด /so:t/ 'unmarried'], or [ขุด /khut/ 'to dig'] and [ขูด /khu:t/ 'to scrape'] [36]. If the learner makes a mispronunciation, the meanings of the words will change.

3. Related Works

This section presents a study of speech recognition tasks that uses a deep learning structure, Mel spectrogram (MS) and Mel frequency cepstral coefficients (MFCCs), and the activation function.

3.1. Speech Recognition Tasks

As deep learning architectures, CNNs have been applied in speech recognition. The CNN model was applied to ASR [37] with various strategies such as pooling and weight sharing, which improved the experimental results. The CNN architecture was applied to classify 14 phrases in a small-footprint keyword spotting task [37] and achieved 27% to 44% improvement in the false rejection rate. The best CNN architecture and strategies were employed in large-scale speech tasks [38] and in noise–robust speech recognition [39]. The three large vocabularies continuous speech recognition (LVCSR) tasks were a 50 h and a 400 h news broadcast and a 300 h switchboard. The large-scale speech tasks obtained a word error rate (WER) of 12% to 14% relative improvement. In Aurora4, noise-robust speech recognition obtained 8.81% WER and 10.0% relative reduction over the traditional CNN on the augmented multiparty interaction meeting transcription task. Two techniques that are used to improve the performance and increase the robustness of CNN acoustic models [40] were autoregressive moving average (ARMA) spectrogram features and channel dropout. The channel dropout technique obtained 16% WER with ARMA features and 20% WER with FBANK features over the baseline CNN. The Urdu numerals 0–9 were classified into 10 classes [41]; Urdu is the national language of Pakistan and spoke in parts of India. The public dataset consisted of 25,518 sound samples that were collected from 740 speakers. A CNN for audio digit classification with Mel spectrogram received 97.53%. A phonetic posteriorgram (PPG) speech feature with CNN was applied in speech command control-based recognition [42]. The dataset was created by 3 cerebral palsy (CP) patients who spoke 19 Mandarin commands 10 times each. The data augmentation method was applied to receive 103 (100 corruption with noise data and 3 time-domain variances). The 19 Mandarin commands were close, up, down, previous, next, in, out, left, right, home, one, two, three, four, five, six, seven, eight, and nine. These commands were developed for controlling web browser applications using speech. The results presented that the CNN with PPG achieved 93.49% accuracy. The CNN–PPG system was better than the CNN with MFCC and ASR-based systems, which obtained 65.67% and 89.59%, respectively. For vowel speech recognition, the Thai vowel speech recognition task [43] used the convolutional neural network of the Thai Simple Vowel model with MFCCs. The noisy Thai vowel dataset used in the task was collected from 50 informants. The dataset consisted of 1800 vowel sound files. The output consisted of 18 classes. The results obtained 90.00% and 88.89% accuracy in the female and male datasets, respectively. The CNN was also applied to the vowels in Javanese, one of the languages spoken in Indonesia [44,45]. Mel frequency spectral coefficients (MFSCs) were applied to separate the features. The dataset consisted of 250 middle vowel sound files recorded by a Javanese speaker, and the output consisted of 5 classes. The results achieved 94% and 99.6% accuracy, respectively.

3.2. Acoustic Features

Mel spectrograms (MS) were converted from the raw speech signal (16 kHz) and then applied to the speech command recognition (SCR) task [46]. MS images with the feature size of 125 × 80 × 1 were used as acoustic features. The light interior search network (LIS-Net) model was applied to the SCR task using the Google Speech Command dataset. The experiment showed a significant improvement in accuracy on all 12, 20, and 35 commands with a fast prediction time and a small total number of parameters. The results of the LIS-Net model achieved 97% accuracy. The model consists of the input layer, followed by the light interior search block (LIS-Block) and a classification block. Each LIS-Block was composed of stacks of several LIS-Cores enclosed by two convolutional layers, followed by the batch normalization (BN) and activation layers. The Adam optimizer was used in the model. Yao et al. developed a structure combining three classifiers, namely, a DNN, a CNN, and recurrent neural network (RNN). A corpus recorded by 10 actors and containing 5 sessions was used [47]. The model was used to recognize four emotions, namely, angry, happy, neutral, and sad. At the frame level, low-level descriptors (LLDs) were transferred to the RNN to obtain the LLD–RNN model. At the segment level, MS was transferred to a CNN to obtain the MS–CNN model. At the utterance level, the outputs of high-level statistical functions (HSFs) were transferred to DNN to obtain the HSF–DNN model. A multitask learning strategy was applied in the three models to obtain generalized features. The classification of discrete emotional categories and the regression of continuous emotional attributes were simultaneously performed. Finally, a confidence-based fusion strategy was used to combine diverse classifiers in recognizing different emotional states. The model with the fusion strategy achieved a weighted accuracy of 57.1% and an unweighted accuracy of 58.3%. Speech emotion classification [48] with two datasets (i.e., Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Emotional Tagged Corpus on Lakorn (EMOLA)) was categorized into four classes, namely, anger, happiness, neutral, and sadness. The experimental results showed that each emotion used different features. The Mel frequency cepstral coefficient (MFCC) with zero-crossing rate worked well for the anger and happiness emotional classes with 81.95% and 69.86% accuracy, respectively. Recently, MFCC has also operated as an input in neonatal bowel sound detection [49]. The CNN and a Laplace hidden semi-Markov model (HSMM) were presented. The abdominal sounds from 49 newborn samples in the tertiary neonatal intensive care unit (NICU) were applied. The peristalsis (P) and no peristalsis (NP) classes were used. The totals of P and NP were 14,410 and 1991, respectively. The results were presented with accuracy and area under curve (AUC) scores being 89.81% and 83.96%, respectively.

3.3. Activation Function

The 1D and 2D CNN combined long short-term memory (LSTM) models were applied to the speech emotion recognition task [50] and used raw speech and log-Mel spectrograms as inputs. These models shared a similar architecture that consisted of four local feature learning blocks (LFLBs). Each block contained one convolutional layer, one BN layer, one exponential linear unit (ELU) layer, one max pooling layer, and one LSTM layer. The results showed that the 2D CNN LSTM model outperforms the traditional approaches, i.e., deep belief network and CNN. The structure of the 2D CNN LSTM model included four LFLBs, one LSTM layer, and one fully connected layer. The Softmax classifier was applied to the top layer. The log Mel spectrogram with 251 frames and 128 Mel frequency bins was used as acoustic features. The experimental results showed that overall, the 2D CNN LSTM model outperformed the 1D CNN LSTM model. The ELU method [51] was used to achieve fast learning and high accuracy in DNNs and to solve the vanishing gradient problem. The ELUs yielded negative values that pushed the mean unit activations closer to 0, which differs from that of rectified linear units (ReLUs). ELUs reduced the gap between normal and unit natural gradients, which led to active and fast learning. On different vision datasets, the results show that ELUs significantly outperform other activation functions. The model with ELUs performed better than the model with ReLUs trained with BN.

Those previous works inspired the classification of Thai vowel speech recognition. The Thai vowel speech recognition process is the most important part of the automatic CAPT for Thai vowel speech. From the literature review, it was found that there was not a study on automatic CAPT in Thai vowel recognition using CNN as a deep learning model. Thai vowel speech recognition is used to recognize vowels pronounced by the speaker in a real environment. CNN is applied to the Thai vowel speech recognition model. MS is used for audio input features in the model. This model improves performance by using ELU activation function.

4. Materials and Methods

Materials and methods for deep learning to recognize Thai vowels in speech are explained. The details are presented as follows: Section 4.1. describes the design, collection, preparation, and exploration of the dataset. Spectrogram conversion is shown in Section 4.2. Section 4.3 presents classification by convolutional neural network model. The final section shows performance evaluation.

4.1. Dataset

This section explains the steps of how the dataset was designed, collected, prepared, and explored. Every step was verified by the linguists to ensure the accuracy and quality of the data.

4.1.1. Dataset Design

Currently, no public datasets for Thai vowels are available for this research objective. Most of them were collected unsystematically and unstandardized. Therefore, the dataset preparation was designed for the objective of this study. The linguist listed a set of Thai words to be recorded based on linguistic theory. Every word has the same characteristics: the same consonant, the same tone, and the same final consonant, differing only in the vowels.

4.1.2. Dataset Collection

The speech dataset was collected from Thai speakers who speak the central standard dialect, which is considered the official Thai language. The speech dataset was collected from native speakers in real-world environments, for example, schools, cafeterias, parks, classrooms, bedrooms, and homes. The environments consisted of noises at 30–40 dB SNR (signal-to-noise ratio) such as vehicles, music, wind, and animal sounds (dogs and birds). Therefore, the available data are categorized as the "noisy Thai vowels" dataset. The sounds in the dataset were recorded from 50 standard Thai native speakers (25 males and 25 females between the ages of 20 and 25), fully validated by the linguist according to a good sampling selection process. The dataset contains 44,100 Hz of standard speech data that were recorded from a mobile phone.

4.1.3. Dataset Preparation

After recording the dataset, all the audio files were then sent to linguists for revision. The sound with each speaker completing each vowel was selected. Afterwards, the linguist cut the sound file with Praat [12]. When all the audio files were cut, each file was reaudited for recheck. A linguist and a native speaker who listened to and validated the sounds defined the criteria for good-quality sound files. Vowel sounds were distinct and unambiguous in the high-quality sound recordings. Both linguists and native speakers classified each sound file by vowel sound and chose the file if the sound file was of good quality. Conversely, if the sound file had bad quality, the linguist would remove it and replace it with a new one. To achieve the best quality of the sound files, the linguist rechecked all the selected sound files 5 times (1800 sound files each time).

At the end of this step, the total number of good quality vowels obtained was 1800 sound files, which were divided into 900 male (18 vowels × 25 males × spoken twice) and 900 female (18 vowels × 25 females × spoken twice) voices. The 18 vowels in the dataset can be classified into 9 short vowels and 9 long vowels. Table 1 presents each phonetic label that represents a class.

Table 1. Thai Simple Vowels in the International Phonetic Alphabet (IPA) adapted with permission from Refs. [33,34] Copyright 2015, IEEE.

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

The ambiguous speech audio files were removed from the dataset to ensure the best-quality input. The dataset was a combination of the female and the male voice data. The dataset was divided into training and testing sets by K-fold cross-validation, in this case k = 5.

4.1.4. Exploration of Data

After the preparation of the Thai vowel dataset, the data for each vowel sound were explored. Table 2 presents the durations of the Thai vowels in males and females: maximum (max), minimum (min), and average (avg).

Table 2. The durations in the Thai vowel dataset.

Durations of Each Thai Vowel						
Thai Vowels	Male Sounds (Seconds)			Female Sounds (Seconds)		
	Max	Min	Avg	Max	Min	Avg
a:	0.649	0.200	0.391	0.791	0.177	0.394
i:	0.616	0.161	0.357	0.667	0.175	0.395
ɯ:	0.580	0.159	0.356	0.658	0.170	0.392
u:	0.601	0.234	0.383	0.874	0.121	0.470
e:	0.501	0.145	0.331	0.714	0.213	0.410
ɛ:	0.672	0.230	0.378	0.801	0.152	0.447
o:	0.531	0.155	0.333	0.759	0.208	0.410
ɔ:	0.646	0.179	0.365	0.821	0.182	0.391
ɤ:	0.559	0.173	0.352	0.937	0.210	0.425
a	0.181	0.059	0.120	0.707	0.065	0.140
i	0.170	0.072	0.112	0.268	0.057	0.133
ɯ	0.295	0.074	0.124	0.493	0.076	0.176
u	0.332	0.061	0.140	0.538	0.062	0.189
e	0.175	0.082	0.116	0.464	0.066	0.144
ɛ	0.353	0.072	0.129	0.542	0.085	0.158
o	0.373	0.067	0.131	0.657	0.071	0.164
ɔ	0.337	0.069	0.122	0.757	0.078	0.157
ɤ	0.206	0.080	0.125	0.576	0.075	0.156

The pronunciation of each individual vowel varies individually as well as by gender and age. Even if the same speaker speaks twice, the sound is still different. The values for the duration of 18 Thai vowels are presented in Table 2. For male sounds, the max duration is 0.672 s in /ɛ:/, and the min duration is 0.059 s in /a/. The average maximum and minimum durations are 0.391 s in /a:/ and 0.112 s in /i/, respectively. For female sounds, the max duration is 0.937 s in /ɤ:/, and the min duration is 0.057 s in /i/. The average maximum and minimum durations are 0.470 s in /u:/ and 0.133 s in /i/, respectively. The pronunciation of /i/ for both males and females have a very short duration.

Figure 1 shows the maximum, minimum, and average duration of male and female vowel sounds. The top three symbols represent the durations of the male sounds, and the lower three symbols represent the durations of the female sounds. The greatest duration of vowel sounds is found in females; the female duration is longer than the male in all vowels. The minimum duration in each vowel for males and females is approximately the same, but the average duration of all vowels in females is longer than males.

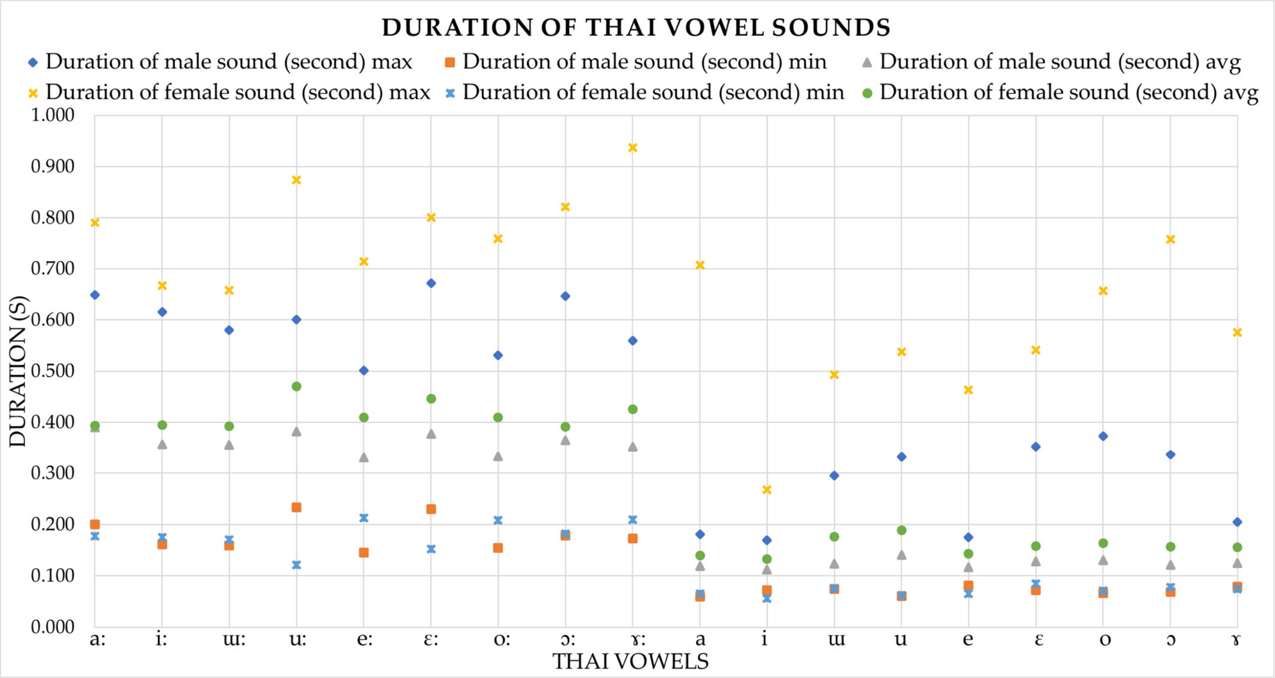


Figure 1. Durations of Thai vowel sounds.

Figure 2 shows the average durations of Thai vowels. The first nine bars on the x-axis are long vowels. and the last nine bars on the x-axis are short vowels. The y-axis represents the duration of Thai vowel sounds, ranging from 0.0 to 0.45 s. It can be seen that the long vowels' average duration of pronunciation is longer than that of the short vowels. The long vowels average is 0.3–0.4 s, whereas the short vowels average at 0.1–0.2 s. It is apparent that the vowels are different lengths.

Based on an exploration of vowel wave files that have been examined by a linguist, certain features were identified. For instance, while speaking, the characteristics of each vowel sound are different because the pronunciation of each vowel varies in many aspects such as gender, style of speaking, loudness, duration, age, and environment. Vowel sound files should therefore be preprocessed to obtain the appropriate form of information for the next stage.

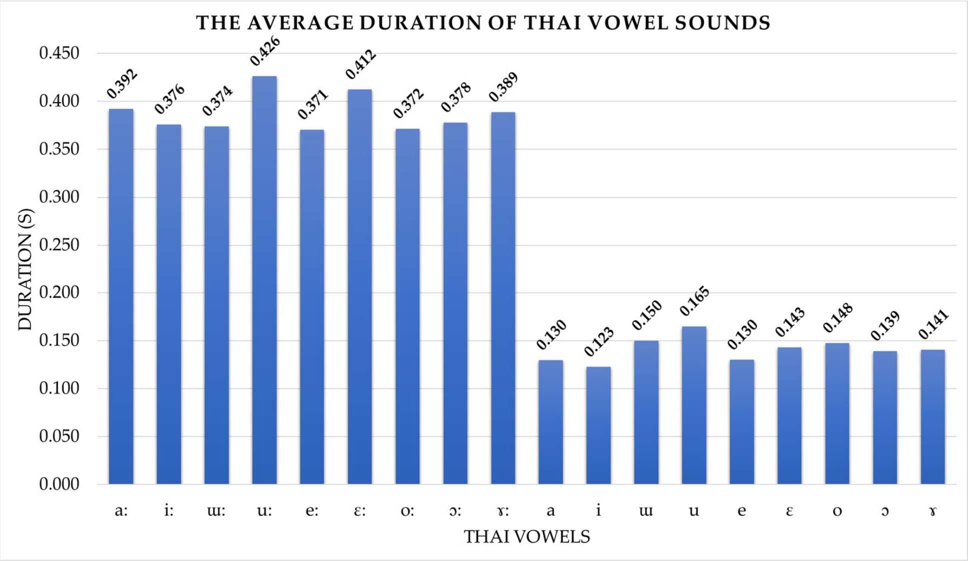


Figure 2. The average durations of Thai vowel sounds.

4.2. Spectrogram Conversion

The raw audio signals were converted to the waveform and then transformed into spectrograms of various sizes to find the appropriate acoustic feature inputs. The 2D images of a spectrogram consist of one axis of time and one axis of frequency, represented by the sequences of the spectra. For audio analysis, spectral representations retain more information than traditional hand-crafted features. The spectrograms have fewer dimensions than the raw audio [52].

4.2.1. Preprocessing

Proper preprocessing of input data is one of the keys to good feature representation. The Thai vowel speech signals were preprocessed using the LibROSA [53] library in Python, a package for audio and music analysis. In the preprocessing step, the monophonic speech signal was downsampled from 44,100 Hz to 16,000 Hz sampling rate. The audio analysis method used small frames of the signal that were spaced by a hop length. The length of the window was 2048 samples (approximately 128 ms), and the hop size was 512 samples (approximately 32 ms). The speech signal was converted into a time-frequency representation based upon the short-time Fourier transform (STFT). The discrete-time STFT was calculated using the fast Fourier transform algorithm (FFT). The mathematical representation of STFT [54] is written as:

$$X[m,\omega] = \sum_{n=-\infty}^{\infty} x[n]w[n-m]exp^{-j\omega n} \tag{1}$$

In Equation (1), $x[n]$ denotes the sequence of a discretized time-domain signal to be transformed, $w[n]$ denotes the window function, m denotes the time index, ω denotes the frequency, and $X[m,\omega]$ denotes the STFT of the time-domain sequence.

4.2.2. Feature Extraction

The squared magnitude (power spectrum) of the STFT is the linear-scaled spectrogram. For the MS, the linear frequency scale of the spectrogram is scaled to the Mel scale using overlapping triangular filters. The Mel scale provides the linear scale for the human auditory system [55] and is defined as follows:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

where m denotes Mels and f denotes Hertz as explained by Equation (2). Figure 3 presents an example of Thai vowel speech that is converted to MS acoustic features.

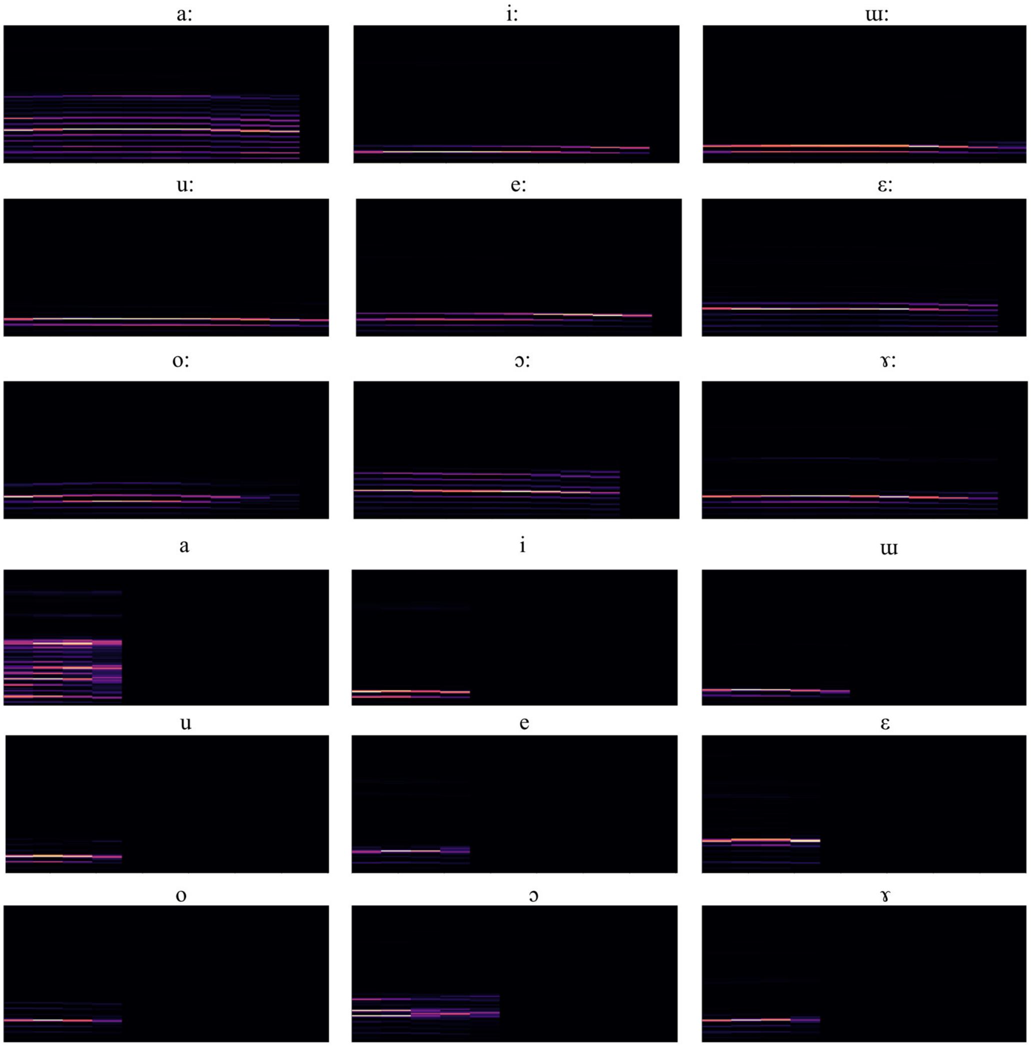


Figure 3. MS acoustic features of Thai vowels.

For the MFCC, the logarithm of MS is converted using the discrete cosine transform (DCT). The result of the conversion is called the Mel frequency cepstrum coefficient (MFCC). Figure 4 presents an example of Thai vowel speech that is converted to MFCC acoustic features. In this research, the default acoustic features are 40 Mel bands that are utilized for MS acoustic features, 40 MFCC for MFCC acoustic features, and 11 contextual vectors [39,43]. The acoustic features were used as input features for passing data into the deep learning architectures in the classification process.

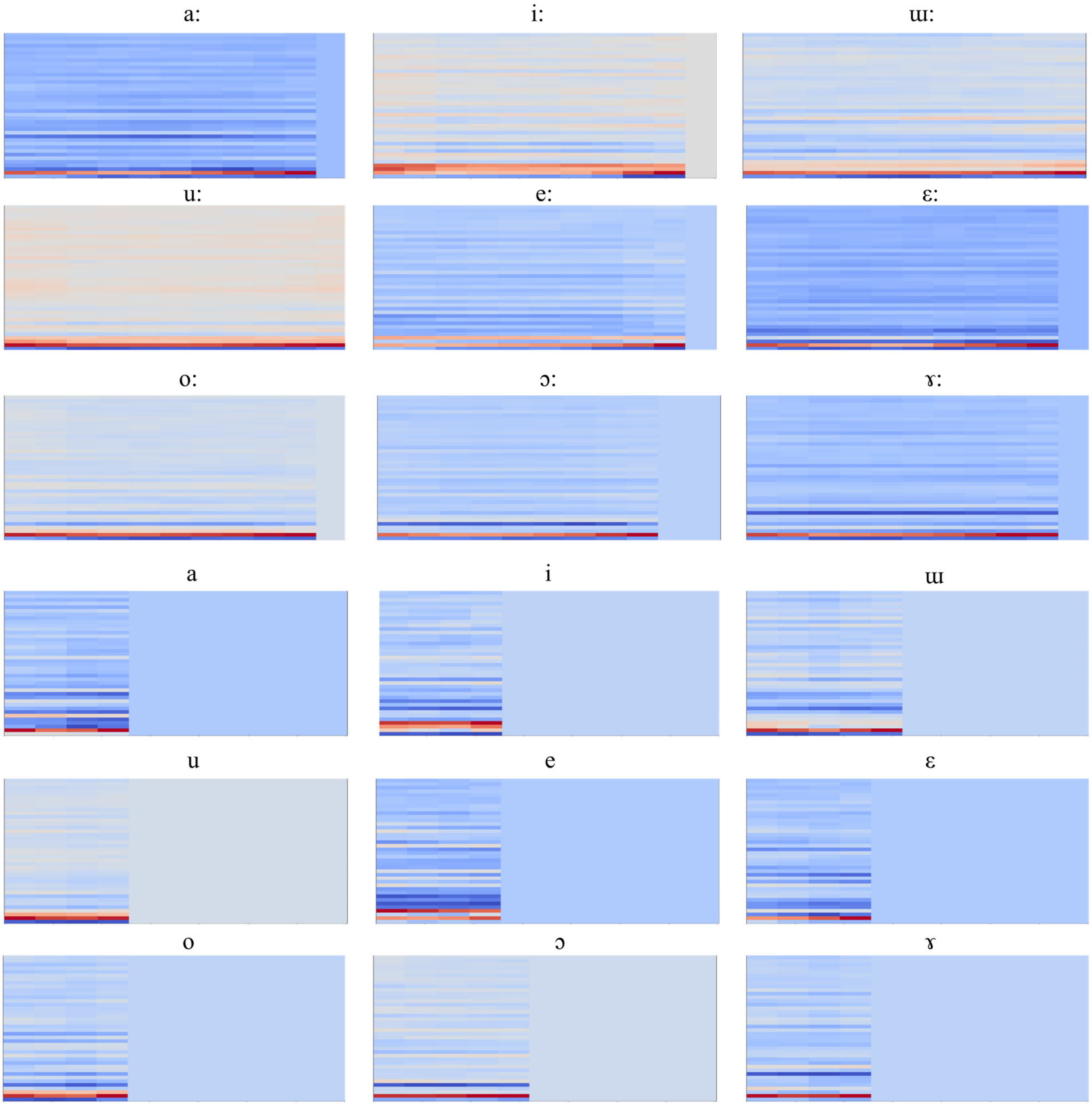


Figure 4. MFCC acoustic features of Thai vowels.

4.3. The Classification by Convolutional Neural Networks Model

The learners practiced by speaking Thai vowels. The speaker's voice was preprocessed to convert the sound wave to time-frequency input data. Then, the input data were extracted into input features and passed to the Thai vowel speech recognition model, which is the deep learning architecture. The model is the most important part of recognizing vowels the speaker pronounces in the automatic CAPT for Thai vowel speech. After that, the vowel classification output was sent to the comparison stage. The vowel obtained from the model's recognition was then compared with the vowels the learner selected to check whether they matched. If they matched, the model indicated that the learner's pronunciation was correct.

Convolutional neural networks (CNNs) are one of the deep learning architectures. CNNs have been applied to not only computer vision but also speech recognition. A CNN is an architecture that is a combination of the feature extractor and the classifier [56]. The CNN model typically consists of convolutional, pooling, normalization, and fully connected layers [57], and the convolutional layers are used to extract the local features from the input data. The classification stage was employed to classify the class labels in the fully connected layers. After feature extraction, the Thai vowel speech signal was converted into MS or MFCC acoustic features vectors and was passed to the CNN model to classify Thai vowels speech.

4.3.1. Baseline Structure

In the baseline CNN model, the researcher started with shallow convolutional neural networks which consisted of two convolutional layers and one pooling layer. ReLU activation function [58], Adam optimizer, and a batch size of 32 were used in this model. The first convolutional layer consisted of 32 filters (2 × 2), the ReLU activation function, and max pooling (2 × 2). The second convolutional layer consisted of 64 filters (2 × 2) and ReLU activation function but without the pooling layer. The fully connected layer consisted of 64 hidden units. In the final layer, the Softmax activation function was used for classification.

The LSTM layer is designed for modeling the time signal and learning the long-term contextual dependencies from the sequences [50] including the baseline LSTM model and the reshaped dimension of the input layer. A dropout of 0.35 is used after the input layer. Then, the output passes through the LSTM layers. The first and second LSTM layers consist of 512 units, hyperbolic tangent activation function (tanh), and dropout (0.35). Finally, the Adam optimizer and Softmax activation function are applied to the classification.

The LSTM has a special mechanism to control the flow of information using four components. Cells with a self-recurrent connection, an input gate, a forget gate, and an output gate are used as components in the LSTM layer. The LSTM unit is updated at every timestep *t* as explained by Equations (3)–(7) [50]:

$$i_t = \sigma(W_i y_t^{l-1} + U_i y_{t-1}^{l} + b_i) \tag{3}$$

$$f_t = \sigma(W_f y_t^{l-1} + U_f y_{t-1}^{l} + b_f) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_c y_t^{l-1} + U_c y_{t-1}^{l} + b_c) \tag{5}$$

$$o_t = \sigma(W_o y_t^{l-1} + U_o y_{t-1}^{l} + b_o) \tag{6}$$

$$y_t^l = o_t \tanh c_t \tag{7}$$

where y_t^{l-1} denotes the input of the LSTM unit, y_t^l denotes the output of the LSTM unit; i_t, f_t, and o_t denote gate vectors; c_t denotes the LSTM unit state; W, U, and b denote the parameter matrices and vectors; σ denotes the sigmoid function; and $\tan h$ denotes hyperbolic tangents. In Equations (3)–(7), the superscript $l-1$ and l denote the indices of the input and output features. The subscript i, f, o, and c in Equations (3)–(6) denote the input gate, forget gate, output gate, and cell, respectively.

4.3.2. The Optimized CNN Model

To improve the model, the baseline model was updated by adding convolutional layers, max pooling, padding, and dropout strategies. Padding strategies [39] preserve the size of the feature maps and make more improvements. Pooling is a significant concept in CNN architectures that reduces spectral variability in input features [38]. Dropout strategies [58] reduce the overfitting problem. Hyperparameter configuration was applied to the models. The Adam optimizer increased the converging rate and provided a performance increase. The initial learning rate was 0.001. The batch size was 32, and the epoch was 500. The dataset was divided into training and testing sets by K-fold cross-validation (k-fold = 5). Each convolutional layer applied convolution filters to the acoustic features input followed by a nonlinear activation function. A kernel of size 2 was defined for each convolutional layer. The number of filters was set to 32, 64, or 128 in different convolutional layers. The numbers of batch sizes were 32, 64, and 128, and the different dropout values were 20%, 25%, 30%, 35%, 40%, 45%, and 50%.

In the optimal convolutional neural networks model for Thai vowel recognition, increasing the convolutional layer and the filter size; adding max pooling; and using padding, dropout strategies, and appropriate activation functions could improve accuracy. The CNN model was trained to learn in a trial-and-error way and maintained by increasing the testing accuracy in small steps, which allowed testing accuracy to improve to 90.0%. This research proposed an optimal CNN model for Thai vowels consisting of three convolutional layers, two max pooling layers, one flatten layer, and two fully connected layers. This architecture is shown in Figure 5. The details of the model are as follows:

- The first convolutional layer consists of 128 filters (2 × 2), an ELU activation function, max pooling (2 × 2), and dropout (0.35).
- The second convolutional layer consists of 64 filters (2 × 2), the ELU activation function, max pooling (2 × 2), and dropout (0.35).
- The third convolutional layer consists of 128 filters (2 × 2), a ReLU activation function, and dropout (0.35) but no pooling layer.
- Finally, the fully connected layer consists of 64 hidden units with dropout (0.35), and Softmax activation function is used for the classification in the last layer.

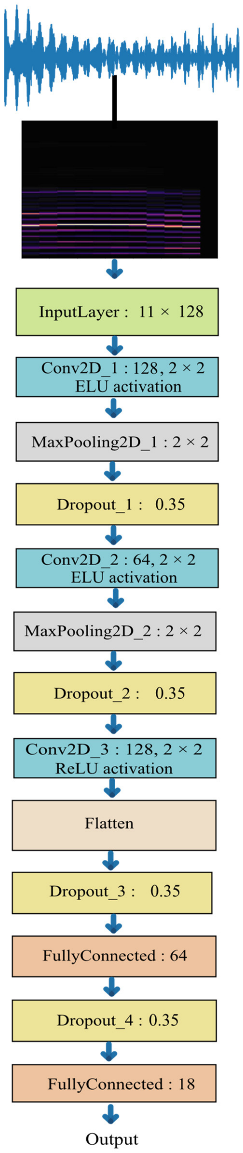


Figure 5. The architecture of the optimized CNN model for Thai vowels.

The MS or MFCC acoustic feature inputs were formatted as images and consisted of rows, columns, and one channel (#frequencies, #times, 1) for feeding into this model. The acoustic features vector was determined to have different input nodes in the 2-dimensional (2D) convolutional layers. The 2D convolution layer is a layer that extracts significant patterns from the input. The purpose was to create a feature map with convolution filters, and it used nonlinear activation functions. The input of the 2D convolution layer in Equation (8) is $x(i,j)$, and the result $y(i,j)$ can be obtained by convolving the input $x(i,j)$ with the convolution filter or kernel $w(i,j)$ [50], defined as follows:

$$y(i,j) = x(i,j) * w(i,j) \tag{8}$$

When features into nonlinear activation functions, the output of the convolution layer is defined as follows:

$$y_i^l = \left(\sum^{j} y_j^{l-1} * w_{ij}^l + b_i^l\right) \tag{9}$$

In Equation (9), where y_i^l denotes the *i-th* output feature at the *l-th* layer. y_j^{l-1} denotes the *j-th* input feature at the *(l-1)-th* layer. w_{ij}^l denotes convolution filter between the *i-th* and *j-th* feature. b_i^l denotes the *i-th* bias at the *l-th* layer. $(\cdot)$ denotes activation functions.

Currently, ReLUs are well-known activation functions [59]. Relatively, ReLUs are implemented in deep learning. The function of ReLUs is defined as follows:

$$O(z) = \max(0, z) \tag{10}$$

where z denotes the input, ϕ (z) is 0 when z is less than 0, and ϕ (z) is equal to z when z is greater than or equal to 0. Thereby, the range is between 0 and ∞. The ReLUs convert the negative values into 0 and maintains the positive values.

The ELUs solve the vanishing gradient problem [51,59]. ELUs yield negative values that push the mean unit activations closer to 0. The ELU activation function in Equation (11) is defined as follows:

$$O(z) = \begin{cases} z & , \ z > 0 \\ \alpha(exp^z - 1), & z \leq 0 \end{cases} \tag{11}$$

where α denotes the ELUs' hyperparameter that controls the value to which an ELU saturates the negative inputs.

After the convolution and activation function layers, the acoustic features were passed into the max pooling layer. The goal of the max pooling layer was to reduce the resolution of the feature maps. The outputs at the last 2D convolutional layers were fed into the flatten layer and passed to the fully connected layer. The fully connected layer combined the output with the final layer for classification. In the final layer, the Softmax function was utilized for multiclass classification, and the Softmax outputs gave the probabilities for the input data. The model classification results represented the 18 classes of Thai vowel speech. In this research, the classification results contained 18 classes, which consisted of 9 short vowels and 9 long vowels. They were set as the Thai simple vowels grouping. Finally, only one class was selected after the classification step.

4.3.3. Implementation Details

In the experiments, the models were implemented using Python on the Keras framework. TensorFlow was used for the backend. This work evaluated the model using Google Colaboratory [60] with Intel®® Xeon®® CPU @ 2.20 GHz and Nvidia Tesla P100 GPU.

In this research, to create 128 × 11 × 1 MS image-like (frequencies × times × channel) files to be used as appropriate acoustic features, the parameters of the spectrogram were operated. The length of the window was 2048. The hop length between frames in the sample was 512. The input shape of the audio channel and the audio sampling rate were 1 and 16,000, respectively. The number of Mel bands was 128. The maximum frequency of the MS was 8000.

4.4. Performance Evaluation

The evaluation method has generally been applied to various recognition tasks. The method for measuring efficiency in this research uses accuracy, precision, recall, and F1 score, which are defined as follows:

$$\text{Accuracy} = \frac{(\text{tp} + \text{tn}) \times 100}{(\text{tp} + \text{tn} + \text{fp} + \text{fn})} \tag{12}$$

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{13}$$

$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{14}$$

$$\text{F1 score} = \frac{2\ (\text{Precision} \ \times \ \text{Recall})}{\text{Precision} + \text{Recall}} \tag{15}$$

In Equations (12)–(15), tp denotes true positive, fp denotes false positive, tn denotes true negative, and fn denotes false negative for each of the class labels.

5. Results

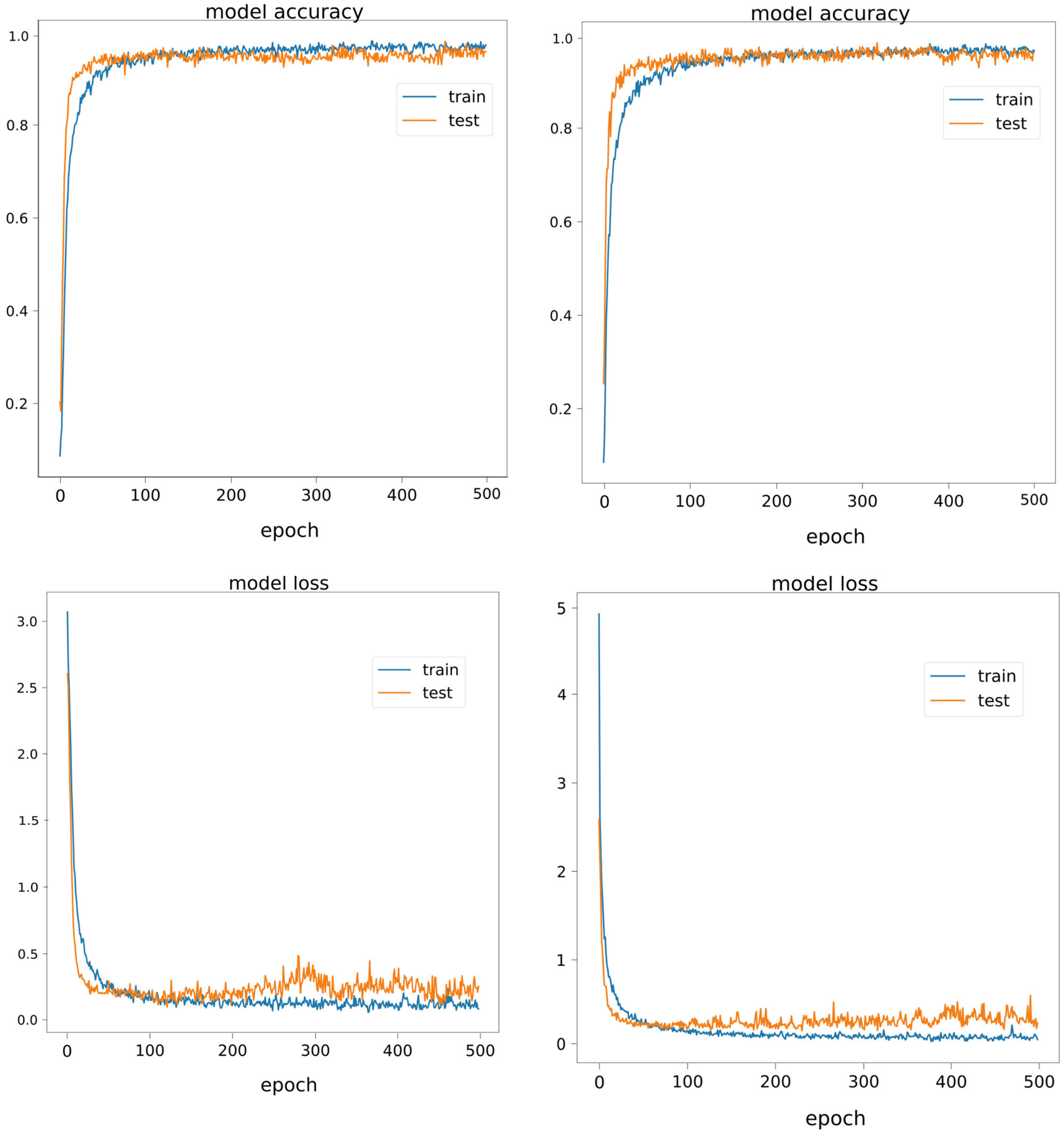
This experiment utilizes combinations of two acoustic feature inputs and three model settings as follows: (1) the MFCC acoustic features combined with a baseline CNN model, (2) the MS acoustic features combined with the baseline CNN model, (3) the MFCC acoustic features combined with the baseline LSTM model, (4) the MS acoustic features combined with the baseline LSTM model, (5) the MFCC acoustic features combined with the optimized model, and (6) the MS acoustic features combined with the optimized model.

Table 3 presents the different experimental results. The parameters are set with 32 batch sizes and 500 epochs. The MS acoustic features combined with the baseline CNN model achieves the lowest accuracy at 88.89%. In the third and fourth experiments, the MFCC and MS acoustic features on the baseline LSTM model achieve low accuracy, 94.44% and 90.00%, respectively. LSTM layers are beneficial for learning long-term contextual dependencies from long sequences. Here, in contrast, LSTM is used on the Thai vowel dataset which is one-syllable words, not long sentences. Therefore, LSTM is not outstanding in this task. The optimized CNN model combined with MS acoustic features achieved improved accuracy of 98.61%. The result of the optimized CNN model shows that the MS acoustic features perform the best for Thai vowel classification.

Table 3. The results from the different experiments (k-fold = 5).

No.	Experiment Settings	Accuracy (%)	Error of the Model (Loss)
1.	MFCC + baseline CNN model	93.33	0.60
2.	MS + baseline CNN model	88.89	0.65
3.	MFCC + baseline LSTM model	94.44	0.25
4.	MS + baseline LSTM model	90.00	0.45
5.	MFCC + optimized CNN model	98.06	0.12
6.	MS + optimized CNN model	98.61	0.18

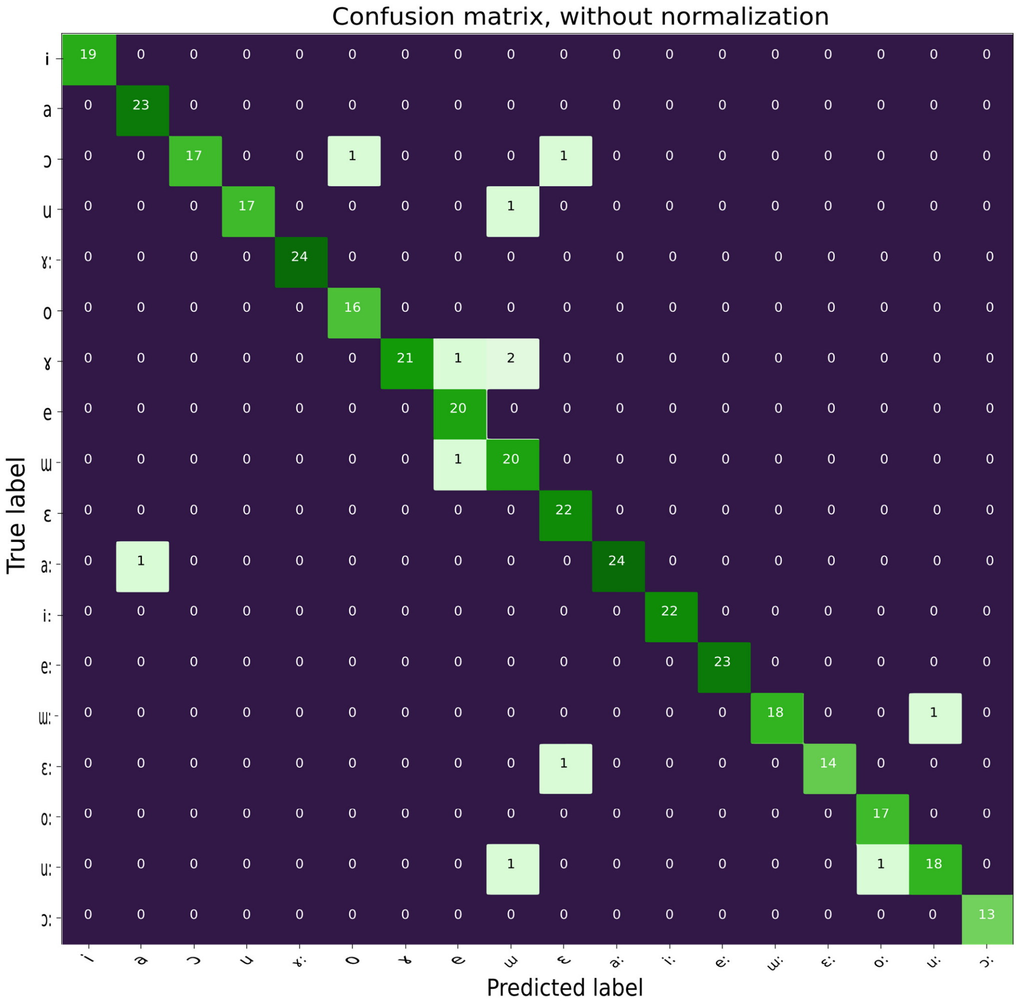
The line graphs of the accuracy and loss of the optimized CNN models combined with MFCC or MS acoustic features are presented in Figure 6a,b. The visualization shows the line graph that compares the accuracy and loss of the training and testing models from 0 to 500 epochs. The optimized MS acoustic features combined with the optimized CNN model outperform the optimized CNN combined with MFCC and achieve the best accuracy of 98.61% as shown in Figure 6b. The loss values are presented in Table 3. The baseline CNN model with the MFCC or MS acoustic features obtains higher loss values and has more overfitting problems than the other models. The optimized CNN model combined with the MFCC or MS acoustic features can reduce the overfitting problems as illustrated in Figure 6a,b.



(**a**) (**b**)

Figure 6. The accuracy and loss of the optimized CNN model. (**a**) MFCC + the optimized CNN model. (**b**) MS + the optimized CNN model.

For the error analysis, the confusion matrix of the optimized CNN model for Thai vowel recognition is shown in Figure 7. For the details of misclassification, 10 of the 18 classes have a 0% error rate. The 'เออะ' /ɤ/ vowel is the class that has three mispredictions. In the confusion matrix, the most perplexing Thai vowel pairs are ('เออะ' /ɤ/) and ('อื' /ɯ/). These sounds are similar, which can be explained by linguistic theory as they share the same characteristics. The (‘เออะ’ /ɤ/) and (‘อึ’ /ɯ/) pronunciations both use the back part of the tongue [61]. As a result, the Thai vowel recognition model may be confused.



	i	a	ɔ	u	ɤ:	o	ɤ	e	ɯ	ɛ	a:	i:	e:	ɯ:	ɛ:	o:	u:	ɔ:
i	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
a	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ɔ	0	0	17	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
u	0	0	0	17	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ɤ:	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0
o	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0
ɤ	0	0	0	0	0	0	21	1	2	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
ɯ	0	0	0	0	0	0	0	1	20	0	0	0	0	0	0	0	0	0
ɛ	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0
a:	0	1	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0
i:	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0
e:	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0
ɯ:	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	1	0
ɛ:	0	0	0	0	0	0	0	0	0	1	0	0	0	0	14	0	0	0
o:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0
u:	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	18	0
ɔ:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13

Figure 7. The confusion matrix of the MS acoustic features combined with the optimized model. The dark green shades represent the correct prediction. The light green shades represent the incorrect prediction.

Table 4 shows the precision, recall, and F1 scores of the CNN model for classifying each Thai vowel. The lowest F1 score on the dataset is 0.89 for (‘อึ’ /ɯ/). The F1 score results are relevant to the confusion matrix. On the other hand, the highest F1 score (i.e., 1.00) on the dataset can be seen for (‘อิ’ /i/), (‘เออ’ /ɤ:/), (‘อี’ /i:/), (‘เอ’ /e:/), and (‘ออ’ /ɔ:/).

Based on error analysis and the evaluation of the CNN, an accuracy of more than 95% was achieved. This optimized CNN model was implemented on the automatic computer-assisted pronunciation training (CAPT). In this experiment, Thai vowel sounds were classified using the CAPT in real situations with unseen data. The recognized vowel results from the system were compared with those perceived by a linguist and a native speaker. The unseen dataset was received from 4 users (2 males and 2 females). All of them were 16 to 30 years old. Each user practiced 18 vowels and spoke 3 times. The total unseen dataset comprised 216 sound files (18 vowels × 4 users × 3 times). In the unseen data, the recognized results for the system of 22 vowel sounds (10.19%) do not match the vowel sounds perceived by a linguist and a native speaker, which are presented in Table 5. A total of 194 vowel sounds (89.81%) match the linguist's and the native speaker's perceptions.

Table 4. The precision, recall, and F1 scores of the CNN model.

Thai Vowels	Dataset		
	Precision	**Recall**	**F1 Score**
i	1.00	1.00	1.00
a	0.96	1.00	0.98
ɔ	1.00	0.89	0.94
u	1.00	0.94	0.97
ɤ:	1.00	1.00	1.00
o	0.94	1.00	0.97
ɤ	1.00	0.88	0.93
e	0.91	1.00	0.95
ɯ	0.83	0.95	0.89
ɛ	0.92	1.00	0.96
a:	1.00	0.96	0.98
i:	1.00	1.00	1.00
e:	1.00	1.00	1.00
ɯ:	1.00	0.95	0.97
ɛ:	1.00	0.93	0.97
o:	0.94	1.00	0.97
u:	0.95	0.90	0.92
ɔ:	1.00	1.00	1.00

Table 5. The unseen CAPT data in real situations that do not match the linguist's and the native speaker's perceptions.

Vowel		Perceived Vowel by	
Practiced Pronunciation	**System Recognition**	**Linguist**	**Native Speaker**
ɛ:	a:	ɛ:	ɛ:
ɛ:	ɔ:	ɛ:	ɛ:
ɛ:	ɔ:	ɛ:	ɛ:
ɛ:	ɛ	ɛ:	ɛ:
ɛ:	ɤ:	ɛ:	ɛ:
e:	ɤ:	e:	e:
e:	o:	e:	e:
e:	ɤ:	e:	e:
i:	e:	i:	i:
i:	e:	i:	i:
ɤ:	o:	ɤ:	ɤ:
ɤ:	ɯ:	ɤ:	ɤ:
ɯ:	ɔ:	ɯ:	ɯ:
ɯ:	ɔ:	ɯ:	ɯ:
ɯ:	ɤ:	ɯ:	ɯ:
ɯ:	i:	ɯ:	ɯ:
i	e	i	i
o:	ɛ:	o:	o:
o:	o	o:	o:
u	ɯ	u	u
u:	o:	u:	u:
u:	i:	u:	u:

Table 6 shows the system’s most often predicted pairings of vowels that do not match those of a linguist or a native speaker, which are (‘แอ’/ɛ:/) and (‘ออ’/ɔ:/), (‘เอ’/e:/), and (‘เออ’ /ɤ:/), (‘อี’/i:/) and (‘เอ’/ e:/), and (‘อือ’/ɯ:/) and (‘ออ’/ɔ:/). Each of them has a two-time mismatched pronunciation frequency. These can be explained in linguistic theory with the fact that the mispronounced pairs are related to the similar tongue positions, which are front–back and high–low.

Table 6. The frequency of incorrectly predicted pairs for Thai vowels.

Vowel		Frequency
Practiced Pronunciation	**System Recognition**	
ɛ:	a:	1
ɛ:	ɔ:	2
ɛ:	ɛ	1
ɛ:	ɤ:	1
e:	ɤ:	2
e:	o:	1
i:	e:	2
ɤ:	o:	1
ɤ:	ɯ:	1
ɯ:	ɔ:	2
ɯ:	ɤ:	1
ɯ:	i:	1
i	e	1
o:	ɛ:	1
o:	o	1
u	ɯ	1
u:	o:	1
u:	i:	1

6. Conclusions

Vowels are the core of syllables (nucleus) and are an important part of speech. Vowels are produced in the oral cavity depending on the tongue’s position. Thai vowel pronunciation practice is difficult for nonnative speakers to easily understand by themselves. Experts are required to provide advice. However, today, there is often an inadequacy of instructional specialists. To solve these problems, technology for pronunciation practice should be implemented. This research presents the appropriate acoustic features and an optimal CNN model for noisy Thai vowel speech recognition that is applied in an automatic CAPT system.

The CAPT system is developed to be used in daily life learning activities that can be practiced anywhere and anytime. Therefore, the noisy Thai vowels dataset is collected from native speakers in real-world environments with variations in dimensions such as gender, age, accent, environment, and noise. The dataset, moreover, is designed, collected, and verified by a linguist based on linguistic theory. The 2D-CNN model combined with MS acoustic features improves performance in Thai vowel speech recognition. The model achieves a significant increase in accuracy of 98.61% over the baseline model by employing various strategies and hyperparameter tuning. Finally, the model is implemented on the CAPT system in a realistic situation. The input data received from learners are considered to be invisible data. The recognized vowel resulting from the CAPT system is compared with perceived vowel sounds by a linguist and a native speaker, and it achieves an accuracy of 89.81%. The extraction of vowel acoustic features that apply to MS combined with

CNN provides distinctive acoustic features for Thai vowel speech recognition. This model can distinguish vowel sounds even though the data have various noise, ages, accents, environments, and physical characteristics (i.e., female vs. male voices).

The automatic CAPT system uses the optimal CNN model combined with appropriate MS acoustic features for Thai vowel speech recognition. It can solve problems such as lack of expertise, complexity, time consumption, and lack of real-time feedback. The contribution of this work is that its findings are beneficial for stakeholders who are interested in developing assistive Thai vowel recognition systems or similar pronunciation systems. This work enables researchers to produce learning applications by following similar operations. Moreover, it can help guide and assist voice-impaired, nonnatives, or nonstandard Thai dialect learners, allowing learners to practice vowel pronunciation in real time, anywhere, and anytime, comparable with having an expert, Thai teacher, or linguist advise on pronunciation at all times. For future works, the authors will focus on novel methods of Thai vowel speech recognition in other aspects of Thai words to improve CAPT system.

7. Discussion

This research uses AI with a deep learning model to automate computer-assisted pronunciation training (CAPT) with Thai vowels. The recognition model for the standard 18 Thai vowels is an essential function for automatic speech recognition. The model is applied to recognize learners pronouncing vowel sounds. To be effective in recognition, proper data and structure are required for training the model.

The existing Thai audio corpus cannot be used for the objectives of this research. Therefore, to obtain theoretically qualitative vowel data in linguistic principles, a new dataset was designed, collected, and verified based on linguistic theory by a linguist. The dataset was gathered from native speakers in real-world scenarios, so the CAPT system may be used everywhere and at any time. The dataset consisted of noise at 30–40 dB SNR such as vehicles from the road, people talking, wind at the park, and animal sounds. Therefore, the available data are categorized as the “noisy Thai vowels” dataset. The sounds were recorded with a mobile phone by 50 Thai native speakers. The total number of vowels were 1800 sound files divided into 18 classes.

The CNN model in this research can reduce the frequency variation of the acoustic features and extract appropriate features. This research preserves the sizes of the feature maps with padding strategies as in [39]. The CNN model reduces the spectral variability in the input features with max pooling as done in [38] and decreases the overfitting problem with the dropout strategies as in [58]. The ELU adopted in the CNN model provides outstanding results. It corresponds to those reported by [51,59]. In [50], ELU was applied to each LFLB. CNN-LSTM networks were constructed to learn local and global emotion-related features from speech and log-Mel spectrograms. The 2D CNN_LSTM network that used the ELU activation function achieved recognition accuracies of 95.33% and 95.89% in Berlin EmoDB in speaker-dependent and speaker-independent experiments. The ELU facilitates faster CNN learning and led to higher classification accuracies. The benefits of those strategies improve the performance of the Thai vowels CNN model.

Given the acoustic features of the model, the time and frequency extension techniques in [39,43] can be used for this CNN model. The results indicate that the optimum MS acoustic features are 11 × 128 (time × frequencies). The results of this study differ from those in [43], which applied the CNN model with MFCC acoustic features. Two datasets (female and male) were used in that work. The suitable acoustic feature sizes for the two datasets were different. They were 11 × 40 and 11 × 64, respectively. The most accurate rates of the CNN model were 90.00% and 88.89% for female and male voices, respectively. Another difference is that in this study, the dataset is a combination of female and male voices. This study also differs from [49], which used MFCC acoustic features with CNN and Laplace HSMM for neonatal bowel sounds. The sequence of 24 MFCCs long was applied to the input to feed into the 1D CNN model. The two classes were peristalsis and no peristalsis. The four convolutional layers were contained in that work. The results had accuracies of 89.81% and 83.96% AUC, respectively. In those works, MFCC used a logarithmic frequency scale and DCT, while in this study, MS uses a linear frequency scale. Therefore, different types of sounds and acoustic features lead to different sizes of acoustic feature settings in terms of frequencies and model structure.

According to the results, the optimized CNN model with Mel spectrogram provides outstanding performance with 98.61% accuracy, which is higher than MFCC with the baseline LSTM model and MS with the baseline LSTM model, which had accuracies of 94.44% and 90.00%, respectively. LSTM layers are beneficial for learning long-term contextual dependencies from long sequences. In contrast, LSTM is used on the Thai vowel dataset, which is one-syllable words mand not long sentences; therefore, LSTM is not effective for this task. MS is used in conjunction with CNN, and it can distinguish vowel sounds, although the aggregate dataset is more complex due to many dimensions, such as various noises, ages, accents, environments, and physical characteristics (i.e., female vs. male voices). In the same way [46], MS was applied to the speech command recognition (SCR) task and achieved good performance. MS images with a feature size of 125 × 80 × 1 were used as acoustic features. The light interior search network (LIS-Net) model was applied to the SCR task using the Google Speech Command dataset. The results of the LIS-Net model achieved 97% accuracy.

In this research, the vowel speech dataset used is similar to those used in [44,45] that used a vowel speech dataset for classification with CNN. The classification results were 94% and 99.6% accurate, respectively. This research differs from [44,45], which focus on Javanese vowel sounds. The dataset in [44,45] did not describe how to design, collect, and verify data under a linguistic theory. That dataset was recorded from only one speaker. Only 250 Javanese middle vowels sound files were collected, which were divided into 5 classes (/e/, /ε, /↔/, /o/, and /ɔ/). The dataset in this research work has more diverse dimensions than [44,45]: the number of speakers, the total audio files, the variety of classes, and diverse sound environments. The recognition model also provides satisfactory results. Vowel pronunciation was applied to classical Arabic phonemes [52], which also differs from this research; in that study, the data consisted of 28 consonants associated with 3 short vowels, and a CNN was used to categorize 84 classes. A total of 6229 recorded items in the dataset were documented online from 85 speakers (81 native and 4 nonnative Arabic speakers). The model obtained an accuracy of 95.77%. Thai vowel recognition using deep learning is compared with vowel recognition tasks using deep learning in Table 7.

Table 7. Vowels recognition using deep learning tasks.

Vowels	Methods and Results
Javanese vowels [44,45]	A CNN model with MFSC is applied to recognize 5 classes of Javanese vowels. The dataset consisted of 250 middle vowels sound files recorded by a Javanese speaker. The results achieved are 94% and 99.6% accuracy.
Thai vowels [43]	A CNN of the Thai Simple Vowel model with MFCCs is applied. Two datasets are used, female and male. They are collected from 50 informants. Each dataset contains 900 sound files. The output consists of 18 classes. The results show 90.00% and 88.89% accuracy in the female and male datasets, respectively.
Arabic short vowels [52]	Arabic short vowels using the CNN model are applied to classical Arabic phonemes. The model categorizes 84 classes from 28 consonants associated with 3 short vowels. The online dataset is recorded from 85 speakers producing 6229 records. The model archives an accuracy of 95.77%.
Thai vowels [This work]	A CNN model with MS achieves an accuracy of 98.61%. The model classifies 18 classes. The dataset consists of 1800 vowel sound files recorded from 50 native speakers. The dataset is designed, collected, and examined by a linguist and a native speaker.

In many works, models are not deployed in real-world situations. As a result, when they are applied to systems, the true impacts cannot be determined. To examine robustness, the CNN model is implemented on the CAPT system in a real situation. The input data received from learners are considered invisible data. When compared with the linguist's and the native speaker's perceptions of the vowel sounds, the CAPT system's vowel detection was 89.81% accurate. This indicates that utilizing a variety of data dimensions and designing, collecting, and verifying data are very helpful for creating high-quality input data for speech recognition models.

The automatic CAPT uses the CNN model for Thai vowel speech recognition. It solves problems such as lack of expertise, complexity, time-consuming processes, and nonreal time processes. Deep learning is applied to an ASR recognition model to recognize the learner's pronunciation. The system uses raw speech input data and uses MS acoustic features along with CNN to extract the distinctive features of the vowels and classify them. After that, the vowel derived from the classification is compared with the vowel selected by the learner. If the comparison matches, the learner's pronunciation is correct.

To increase the preciseness of pronunciation, linguists use the acoustic phonetics method for phonetic analysis. In general, those research works necessitate the hand-crafted extraction of formant1 (F1) and formant2 (F2), which presents significant challenges due to the large amount of data and speaker acoustic variation. This work differs from previous studies that used Praat with phonetics principles in the analysis of pronunciation [12–17]. In general practice, linguists usually employ the F1 and F2 using Praat. Then the F1 and F2 of native and nonnative speakers are plotted using Microsoft Excel, Python, or R programming language. After that, the graphs are compared to find the differences between the native speakers and the learner; differences mean the pronunciation is incorrect. These normal steps are a multistep rather than real-time approach, and they require someone with expertise in Praat or someone who can program in R or Python language.

Those traditional methods by linguists may be complex for others, whereas this research provides a very uncomplicated method. It works in real time and can solve those problems. Moreover, the technique can be utilized for developing CNN models in a similar domain by applying various strategies, optimizing the neural networks, and adjusting the hyperparameters for improved performance in learning assist systems in the future.

8. The Automatic Computer-Assisted Pronunciation Training for Thai Vowel Speech

The automatic CAPT for Thai vowel speech is shown in Figure 8. The details of this system are as follows:

(1) The user selects a Thai vowel that they want to practice pronouncing, such as /a/.

(2) Then the system goes to the /a/ pronunciation practice page. The page contains three types of vowel description:

  - A video of the 3D /a/ sound pronunciation that shows the movement of the tongue.
  - The vowel descriptions including details such as vowel duration: short or long, mouth shape: round or not round, and tongue characteristics: low, mid, or high.
  - A video of a straight-faced person showing the position of the mouth as the vowel is pronounced.

(3) When users practice pronunciation, they can press the microphone icon and speak (the speech signal is an analog audio signal).

(4) The signal is digitalized and converted into the time-frequency domain during the preprocessing step.

(5) Afterward, features are extracted during the extraction step, and their values characterize the phonemes in the acoustic features.

(6) Then, the acoustic features are utilized in the classification process. In this process, a deep learning architecture for the recognition of the Thai vowel speech is applied. The class result contains 18 classes (9 short vowels and 9 long vowels like Thai simple vowels grouping). After the classification step, only one class is selected as the output for the vowel classification result.

(7) Finally, a comparison between the vowel selected by the user and the vowel classification result is presented. This procedure compares the vowel chosen by the user for pronunciation practice with the result from the recognition of the Thai vowels.

(8) If the vowel selected by the user (e.g., /a/) and the vowel classification result (e.g., /a/) are the same, then the pronunciation result will display a message "Your pronunciation is correct".

(9) In contrast, if the vowel selected by the user and the vowel classification result are not the same, e.g., the user selects /a/ but the vowel classification result is /o/, the pronunciation result will display a message "Your pronunciation is incorrect" and "You pronounced /o/". In addition, messages will show a 3D video of the pronunciation and text describing how to pronounce it. The user can go back to the vowel pronunciation practice page to retry by clicking the bottom arrow.

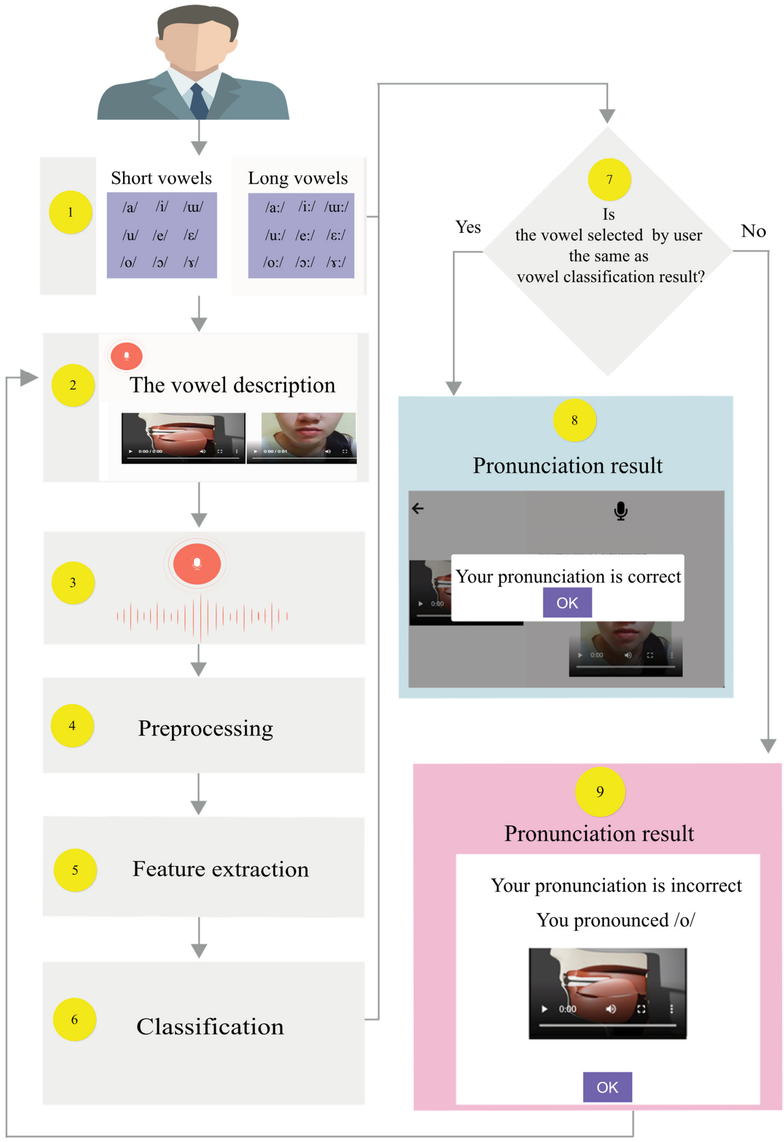


Figure 8. The automatic computer-assisted pronunciation training for Thai vowel speech.

Author Contributions: Conceptualization, N.R. and S.P.; Data curation, N.R.; Formal analysis, N.R. and S.P.; Investigation, N.R. and S.P.; Methodology, N.R. and S.P.; Project administration, S.P.; Resources, N.R.; Software, N.R.; Supervision, S.P.; Validation, N.R. and S.P.; Visualization, N.R.; Writing—original draft, N.R.; Writing—review & editing, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Variables and Acronyms

Table A1. The definitions of variables and acronyms.

Variables and Acronyms	Meaning
CAPT	Computer-Assisted Pronunciation Training
CNN	Convolutional Neural Network
MS	Mel Spectrogram
MFCC	Mel Frequency Cepstral Coefficients
LSTM	Long Short-Term Memory
AI	Artificial Intelligence
CALL	Computer-Assisted Language Learning
ASR	Automatic Speech Recognition
Praat	The program for analyzing phonetics
F1	Formant1
F2	Formant2
DNN	Deep Neural Network
SSD	Speech Sound Disorder
DS	Down syndrome
CV	Consonant + Vowel
CVC	Consonant + Vowel + Consonant
C(C)	A diphthong consonant
LVCSR	The large vocabularies continuous speech recognition tasks
WER	Word Error Rate
ARMA	Autoregressive Moving Average spectrogram features
PPG	Phonetic Posteriorgram
CP	Cerebral Palsy
MFSC	Mel frequency spectral coefficients
SCR	Speech Command Recognition
LIS-Net	The Light Interior Search Network
LIS-Block	The Light Interior Search Block
BN	Batch normalization
RNN	Recurrent Neural Network
IEMOCAP	The Interactive Emotional Dyadic Motion Capture
LLDs	Low-Level Descriptors
HSFs	High-Level Statistical Functions
EMOLA	Emotional Tagged Corpus on Lakorn
ELU	Exponential Linear Unit
ReLUs	Rectified Linear Units
SNR	Signal to Noise Ratio
LibROSA	The library in Python, which is a package for audio and music analysis.
STFT	Short-Time Fourier Transform
FFT	Fast Fourier Transform algorithm
DCT	Discrete Cosine Transform
tanh	Hyperbolic Tangent Activation Function

References

1. Anderson-Hsieh, J.; Koehler, K. The effect of foreign accent and speaking rate on native speaker comprehension. *Lang. Learn.* **1988**, *38*, 561–613. [CrossRef]
2. Lauzon, F.Q. An introduction to deep learning. In Proceedings of the 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), Montreal, QC, Canada, 2–5 July 2012; pp. 1438–1439.
3. Fu, J.; Chiba, Y.; Nose, T.; Ito, A. Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Commun.* **2020**, *116*, 86–97. [CrossRef]
4. Ferrer, L.; Bratt, H.; Richey, C.; Franco, H.; Abrash, V.; Precoda, K. Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Commun.* **2015**, *69*, 31–45. [CrossRef]
5. Short, G.; Hirose, K.; Kondo, M.; Minematsu, N. Automatic recognition of Japanese vowel length accounting for speaking rate and motivated by perception analysis. *Speech Commun.* **2015**, *73*, 47–63. [CrossRef]
6. Gamper, J.; Knapp, J. A review of intelligent CALL systems. *Comput. Assist. Lang. Learn.* **2002**, *15*, 329–342. [CrossRef]
7. Martens, W.L.; Wang, R. Applying adaptive recognition of the learner's vowel space to English pronunciation training of native speakers of Japanese. *SHS Web Conf.* **2021**, *102*, 01004. [CrossRef]
8. Rogerson-Revell, P.M. Computer-Assisted Pronunciation Training (CAPT): Current issues and future directions. *RELC J.* **2021**, *52*, 189–205. [CrossRef]
9. Peng, X.L.; Chen, H.; Wang, L.; Wang, H.A. Evaluating a 3-D virtual talking head on pronunciation learning. *Int. J. Hum. Comput. St.* **2018**, *109*, 26–40. [CrossRef]
10. Tabain, M.; Beare, R. An ultrasound study of coronal places of articulation in Central Arrernte: Apicals, laminals and rhotics. *J. Phon.* **2018**, *66*, 63–81. [CrossRef]
11. Teeranon, P. Thai tones in chinese students after using the tone application and their attitudes. *J. Lang. Linguist. Stud.* **2020**, *16*, 1680–1697. [CrossRef]
12. Boersma, P.; Van Heuven, V. Speak and unspeak with PRAAT. *Glot Int.* **2001**, *5*, 341–347.
13. Ling, L.; Wei, H. A research on guangzhou dialect's negative transfer on british english pronunciation by speech analyzer software Praat and ear recognition method. In Proceedings of the 2021 2nd International Conference on Computers, Information Processing and Advanced Education, Ottawa, ON, Canada, 25–27 May 2021; pp. 1123–1132.
14. Intajamornrak, C. Variation and change of the phrae pwo karen vowels and tones induced by language contact with the Tai Languages. *Manusya J. Humanit.* **2012**, *15*, 1–20. [CrossRef]
15. Georgiou, G.P. Discrimination of L2 Greek vowel contrasts: Evidence from learners with arabic L1 background. *Speech Commun.* **2018**, *102*, 68–77. [CrossRef]
16. Liu, H.; Liang, J.; van Heuven, V.J.; Heeringa, W. Vowels and tones as acoustic cues in Chinese subregional dialect identification. *Speech Commun.* **2020**, *123*, 59–69. [CrossRef]
17. Nimz, K. Vowel perception and production of late Turkish learners of L2 German. In Proceedings of the ICPhS, Hong Kong, China, 17–21 August 2011; pp. 1494–1497.
18. Boersma, P.; Weenink, D. Praat: Doing phonetics by computer. *Glot Int.* **2001**, *5*, 341–347. [CrossRef]
19. Roepke, E.; Brosseau-Lapré, F. Vowel errors produced by preschool-age children on a single-word test of articulation. *Clin. Linguist. Phon.* **2021**, *35*, 1161–1183. [CrossRef]
20. Carl, M.; Kent, R.D.; Levy, E.S.; Whalen, D. Vowel acoustics and speech intelligibility in young adults with down syndrome. *J. Speech Lang. Hear. Res.* **2020**, *63*, 674–687. [CrossRef]
21. Lee, S.; Cho, M.-H. The impact of L2-learning experience and target dialect on predicting English vowel identification using Korean vowel categories. *J. Phon.* **2020**, *82*, 100983. [CrossRef]
22. Lu, Y.-A.; Lee-Kim, S.-I. The effect of linguistic experience on perceived vowel duration: Evidence from Taiwan Mandarin speakers. *J. Phon.* **2021**, *86*, 101049. [CrossRef]
23. Ghaffarvand Mokari, P.; Werner, S. Perceptual assimilation predicts acquisition of foreign language sounds: The case of Azerbaijani learners' production and perception of Standard Southern British English vowels. *Lingua* **2017**, *185*, 81–95. [CrossRef]
24. Kartushina, N.; Martin, C.D. Third-language learning affects bilinguals' production in both their native languages: A longitudinal study of dynamic changes in L1, L2 and L3 vowel production. *J. Phon.* **2019**, *77*, 100920. [CrossRef]
25. Sahatsathatsana, S. Pronunciation problems of Thai students learning english phonetics: A case study at Kalasin University. *J. Educ.* **2017**, *11*, 67–84.
26. Noss, R.B. *Thai Reference Grammar*; Foreign Service Institute, Department of State: Washington, DC, USA, 1964.
27. Ladefoged, P.; Johnson, K. *A Course in Phonetics*; Cengage Learning: Boston, MA, USA, 2005.
28. Kent, R.D.; Rountrey, C. What acoustic studies tell us about vowels in developing and disordered speech. *Am. J. Speech-Lang. Pathol.* **2020**, *29*, 1749–1778. [CrossRef]
29. Evans, B.G.; Alshangiti, W. The perception and production of British English vowels and consonants by Arabic learners of English. *J. Phon.* **2018**, *68*, 15–31. [CrossRef]
30. Rallo Fabra, L.; Romero, J. Native Catalan learners' perception and production of English vowels. *J. Phon.* **2012**, *40*, 491–508. [CrossRef]
31. Catron, E. The Hardest Languages in the World to Learn. Available online: https://bestlifeonline.com/most-difficult-languages/ (accessed on 27 February 2022).

32. Kanokphara, S. Syllable structure based phonetic units for context-dependent continuous Thai speech recognition. In Proceedings of the Eighth European Conference on Speech Communication and Technology, Geneva, Switzerland, 1–4 September 2003.
33. Jeerapradit, L.; Suchato, A.; Punyabukkana, P. HMM-based Thai singing voice synthesis system. In Proceedings of the 2018 22nd International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, 21–24 November 2018; pp. 1–4.
34. Aunkaew, S.; Karnjanadecha, M.; Wutiwiwatchai, C. 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), Hatyai, Thailand, 22–24 July 2015; pp. 69–73.
35. Munthuli, A.; Tantibundhit, C.; Onsuwan, C.; Kosawat, K.; Wutiwiwatchai, C. 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), Glasgow, UK, 10–14 August 2015; pp. 3–7.
36. Abramson, A.S.; Reo, N. Distinctive vowel length: Duration vs. spectrum in Thai. *J. Phon.* **1990**, *18*, 79–92. [CrossRef]
37. Sainath, T.N.; Parada, C. Convolutional Neural Networks for small-footprint keyword spotting. In Proceedings of the Interspeech, Dresden, Germany, 6–10 September 2015; pp. 1478–1482.
38. Sainath, T.N.; Kingsbury, B.; Saon, G.; Soltau, H.; Mohamed, A.R.; Dahl, G.; Ramabhadran, B. Deep Convolutional Neural Networks for large-scale speech tasks. *Neural. Netw.* **2015**, *64*, 39–48. [CrossRef]
39. Qian, Y.; Woodland, P.C. Very deep Convolutional Neural Networks for robust speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2263–2276. [CrossRef]
40. Kovács, G.; Tóth, L.; Van Compernolle, D.; Ganapathy, S. Increasing the robustness of CNN acoustic models using autoregressive moving average spectrogram features and channel dropout. *Pattern Recognit. Lett.* **2017**, *100*, 44–50. [CrossRef]
41. Aiman, A.; Shen, Y.; Bendechache, M.; Inayat, I.; Kumar, T. AUDD: Audio urdu digits dataset for automatic audio urdu digit recognition. *Appl. Sci.* **2021**, *11*, 8842.
42. Lin, Y.-Y.; Zheng, W.-Z.; Chu, W.C.; Han, J.-Y.; Hung, Y.-H.; Ho, G.-M.; Chang, C.-Y.; Lai, Y.-H. A speech command control-based recognition system for dysarthric patients based on deep learning technology. *Appl. Sci.* **2021**, *11*, 2477. [CrossRef]
43. Rukwong, N.; Pongpinigpinyo, S. Thai vowels speech recognition using Convolutional Neural Networks. In Proceedings of the 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Chiang Mai, Thailand, 30 October–1 November 2019.
44. Dewa, C.K. Javanese vowels sound classification with Convolutional Neural Network. In Proceedings of the 2016 International Seminar on Intelligent Technology and Its Applications (ISITIA), Lombok, Indonesia, 28–30 July 2016; pp. 123–128.
45. Dewa, C.K.; Afiahayati. Suitable CNN weight initialization and activation function for Javanese vowels classification. *Procedia Comput. Sci.* **2018**, *144*, 124–132. [CrossRef]
46. Anh, N.T.; Hu, Y.J.; He, Q.H.; Tran, T.N.L.; Hoang, T.K.D.; Guang, C. LIS-Net: An end-to-end light interior search network for speech command recognition. *Comput. Speech Lang.* **2021**, *65*, 101131. [CrossRef]
47. Yao, Z.W.; Wang, Z.H.; Liu, W.H.; Liu, Y.Q.; Pan, J.H. Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. *Speech Commun.* **2020**, *120*, 11–19. [CrossRef]
48. Sukhummek, P.; Kasuriya, S.; Theeramunkong, T.; Wutiwiwatchai, C.; Kunieda, H. Feature selection experiments on emotional speech classification. In Proceedings of the 2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Hua Hin, Thailand, 24–27 June 2015; pp. 1–4.
49. Sitaula, C.; He, J.; Priyadarshi, A.; Tracy, M.; Kavehei, O.; Hinder, M.; Withana, A.; McEwan, A.; Marzbanrad, F. Neonatal bowel sound detection using Convolutional Neural Network and Laplace Hidden Semi-Markov Model. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *20*, 1853–1864. [CrossRef]
50. Zhao, J.F.; Mao, X.; Chen, L.J. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323. [CrossRef]
51. Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). In Proceedings of the ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016.
52. Asif, A.; Mukhtar, H.; Alqadheeb, F.; Ahmad, H.F.; Alhumam, A. An approach for pronunciation classification of classical Arabic phonemes using deep learning. *Appl. Sci.* **2022**, *12*, 238. [CrossRef]
53. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. Librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; pp. 18–25.
54. Papadimitriou, I.; Vafeiadis, A.; Lalas, A.; Votis, K.; Tzovaras, D. Audio-based event detection at different SNR settings using two-dimensional spectrogram magnitude representations. *Electronics* **2020**, *9*, 1593. [CrossRef]
55. Thornton, B. Audio Recognition Using Mel Spectrograms and Convolution Neural Networks. 2019. Available online: http://noiselab.ucsd.edu/ECE228_2019/Reports/Report38.pdf (accessed on 27 February 2022).
56. Han, Y.; Kim, J.; Lee, K. Deep Convolutional Neural Networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 208–221. [CrossRef]
57. Demir, F.; Turkoglu, M.; Aslan, M.; Sengur, A. A new pyramidal concatenated CNN approach for environmental sound classification. *Appl. Acoust.* **2020**, *170*, 107520. [CrossRef]
58. Dahl, G.E.; Sainath, T.N.; Hinton, G.E. Improving deep neural networks for LVCSR using rectified linear units and dropout. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings, Vancouver, BC, Canada, 26–31 May 2013; pp. 8609–8613.

59. Gu, J.X.; Wang, Z.H.; Kuen, J.; Ma, L.Y.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.X.; Wang, G.; Cai, J.F.; et al. Recent advances in Convolutional Neural Networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
60. Carneiro, T.; Da Nobrega, R.V.M.; Nepomuceno, T.; Bian, G.B.; De Albuquerque, V.H.C.; Reboucas, P.P. Performance analysis of Google Colaboratory as a tool for accelerating deep learning applications. *IEEE Access* **2018**, *6*, 61677–61685. [CrossRef]
61. Slayden, G. Central Thai Phonology. 2009. Available online: http://www.thai-language.com/resources/slayden-thai-phonology.pdf (accessed on 27 February 2022).