IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Estimating underlying articulatory targets of Thai vowels by using deep learning based on generating synthetic samples from a 3D vocal tract model and data augmentation

LAPTHAWAN, T.[1], PROM-ON, S.[1], BIRKHOLZ, P.[2], XU, Y.[3]
[1]Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
[2]Institute of Acoustics and Speech Communication, TU Dresden, Germany
[3]Division of Psychology & Language Sciences, University College London, London

Corresponding author: Prom-on, S. (e-mail: santitham.pro@kmutt.ac.th).

ABSTRACT Representation learning is one of the fundamental issues in modeling articulatory-based speech synthesis using target-driven models. This paper proposes a computational strategy for learning underlying articulatory targets from a 3D articulatory speech synthesis model using a bi-directional long short-term memory recurrent neural network based on a small set of representative seed samples. From a seeding set, a larger training set was generated that provided richer contextual variations for the model to learn. The deep learning model for acoustic-to-target mapping was then trained to model the inverse relation of the articulation process. This method allows the trained model to map the given acoustic data onto the articulatory target parameters which can then be used to identify the distribution based on linguistic contexts. The model was evaluated based on its effectiveness in mapping acoustics to articulation, and the perceptual accuracy of speech reproduced from the estimated articulation. The results indicate that the model can accurately imitate speech with a high degree of phonemic precision.

INDEX TERMS Acoustic-to-articulatory inversion, deep learning, articulatory model, articulatory target acquisition

I. INTRODUCTION

In speech production, speakers convey messages to listeners in acoustic form by moving multiple articulators in specific patterns. These movement patterns are learned and are utilized regularly in everyday communications. By observing speech and visible articulations, a child gradually learns to speak with minimal specific instructions on articulatory movements [1]. Later, a child can mimic other speech by observing only a few samples and then practicing producing them. This learning phenomenon suggests that one of the key components in an early stage of language learning is the ability to recognize the potential articulatory movements and test them by producing similar instances and improving their correctness. Understanding this learning process will provide a critical answer to a question on how speech production learning should be represented. This will provide a framework for creating a better learning algorithm for speech synthesis systems that can automatically learn from observations and interactions, and other speech related applications [2]–[9].

One way to address this issue is to use a corpus-based analysis-by-synthesis method that learns underlying articulatory targets through an iterative exploration of candidate targets by comparing synthesized and original signals and using them to adjust the targets [10]. This modeling method simulates iterative learning of articulatory movements by the speaker where the speaker iteratively synthesizes the speech and then generalizes to estimate the movement pattern. While this approach allows the computational learning process to generate the targets, a sizable corpus with enough contextual variations is required to cover all the possible utterances. Also, it is computationally complex because of the large time

complexity of optimization. Moreover, it can only estimate a single utterance at a time. While this strategy is analogous to mimicking, it still does not address how a trained speaker can recognize or estimate the movement patterns of newly introduced speech utterances immediately after perceiving a few samples.

To address the latter issue, the learning process should be able to quickly estimate the articulatory targets once a few speech examples are received. This can be done by using an acoustic-to-articulatory inversion model [11] where the model learns the mapping from the acoustic to articulatory trajectories. After learning, this model can be used to quickly recognize the plausible articulatory trajectories and assimilate either acoustic or articulatory differences as an interactive learning strategy. However, the problem is complicated, as the solutions are non-linear, non-unique, and ambiguous [12], due to characteristics of speaker vocal tract shape [13], environmental noise [14], coarticulation [15], and speaking rate [16]. Different methods have been proposed to address the problem of learning the association between acoustic and articulation [17]–[19]. The most recent advances were methods that use deep learning models, which have achieved low error rates [20]–[26].

Deep learning [27] is a learning method by using an artificial neural network with the gradient-based optimization to approximate a complex function [28]. To accurately approximate the complex function, a large amount of learning samples are required. However, the process of gathering data is expensive and not always possible. The common strategy to increase the quality of the learning process without the need to acquire additional data is data augmentation, i.e., generating additional samples based on existing data [29]–[31]. Common strategies for a speech data augmentation are vocal tract length perturbation [32], speech perturbation [33], pitch-shifting [34], speech rate modification [35], speech's feature masking [36], and data synthesis [37] using a generative model. To utilize this strategy in target learning for articulatory synthesis, the augmentation should reflect the variabilities in speech production.

In recent developments of articulatory targets estimation, an analysis-by-synthesis approach using a distal learning strategy has been used [10], [38]. Conceptualizing learning by imitation, gradient descent optimization, and swarm optimization were used as a learning strategy to acquire the target articulation. A three-dimensional articulatory synthesis model [39] was used to generate the speech signal from different parameter sets. The results suggest that the optimizer can imitate single vowel utterances. Further improvement of the optimization process using genetic algorithms and long short-term memory (LSTM) neural networks has also been developed which showed promising results [40], [41].

Two kinds of articulatory spaces have been studied for acoustic-to-articulatory inversion methods: 1) the actual articulatory spaces using electromagnetic articulography [25], [26], and magnetic resonance imaging [24], and 2) a theoretical human articulatory space of a two-dimensional [42], [43] and a three-dimensional vocal tract model [44], [45]. Of these two, the theoretical space represented by a vocal tract model is more accessible for understanding speech production and has been much studied in recent years. One example is speech imitation via acoustic-to-articulatory inversion on a two-dimensional vocal tract synthesizer using distal learning [42] and chain metrics [43]. The results of these studies show that the differences between synthesized speech and human speech pose a major constraint on the modelling process. To improve the synthesis quality, a data generating method called babbling generator was proposed that uses an HMM to estimate realistic articulatory trajectories [46]. Further studies have used VocalTractLab, a three-dimensional vocal tract model [39] to improve the naturalness of the synthesized speech, and reinforcement learning using a reward function as a learning strategy [44] on a preset vowel samples, although there were no quantitative assessments of the synthesis quality. In addition, the learning process is not end-to-end, and the human was involved to select optimized tokens during the vowel's refinements process. There was also a proposal for implementing an imitation algorithm based on Echo State Network to refine synthetic syllables generated by VocalTractLab with preset gestural scores [45]. However, the results showed that the improvement due to the refinements was limited, trading off the 12% deteriorated intelligibility with only 40% improvement while others are not improved. Therefore, much more work is needed to improve the intelligibility of the synthesized speech from the model, and to close the generalization gap, i.e., the difference between the ability to re-synthesize speech with a high intelligibility of an utterance from learning samples and an utterance from unseen speakers.

This study proposes a speech acquisition strategy for learning the underlying articulatory targets that can generate synthetic Thai vowels using a three-dimensional vocal tract model. The strategy uses deep learning to directly map an acoustic vowel representation to an underlying articulatory target, and then uses the VocalTractLab, an articulatory synthesizer with a three-dimensional vocal tract model, as a forward function to reproduce speech utterances from the retrieved representations. The training samples were monosyllabic and disyllabic vowel-only utterances that were synthesized by interpolating and augmenting from a few observed samples. The quality of the re-synthesized speech by the model was evaluated using a perceptual recognition test, where the model re-synthesizes a vowel speech from the unseen human vowel utterance.

II. METHOD

A. OVERVIEW

Our proposed underlying articulatory target acquisition strategy applies deep learning to model the acoustic-to-articulatory targets mapping and a three-dimensional vocal tract model to both generate learning samples and re-synthesize speech from the estimated articulatory targets from the model, as illustrated in Figure 1. The deep neural

network maps the observed target speech to the underlying articulatory target. Next, the three-dimensional vocal tract model, designed by Birkholz et al. [47], maps estimated articulatory targets into motor commands to reproduce the speech. In the acoustic domain, the target is a surface acoustical pattern that a speaker aims to produce, while in the articulatory domain, the underlying target is a set of parameters used to control the models of the vocal tract and the vocal folds. This corresponds to the motor commands in speech production. The deep learning model analogous an auditory-to-motor mapping, which is a neural connection in the cortex of the human brain. To extract multiple representations from an observed sample, the data generator module was designed, generating a large speech and articulatory target corpus by using interpolation and augmentation.

The model was evaluated by comparing the reproduced speech with the observed target speech vowels (as part of disyllabic utterances). Two observed target speech vowel samples were used, 1) an observed pair of a speech and an articulatory target acquired from the VocalTractLab application, and 2) an observed disyllabic speech signal recorded from 12 native Thai speakers. To determine the effectiveness of the proposed learning strategy, a listening test, which included 25 native Thai participants, was used to recognize a disyllabic Thai vowel utterance which re-synthesized from the recorded speech of the native Thai speaker.

B. VOCALTRACTLAB

The VocalTractLab 2.2 (VTL), an articulatory speech synthesizer, is the core speech production model [48]. The application provided a 3D vocal tract model and a predefined articulatory parameter of German vowels and some of German consonants. The 3D vocal tract model used in the VocalTractLab was developed based on the volumetric MRI of a German native male speaker [49]. The VocalTractLab can synthesize a full range of speech sounds based on the set of articulatory parameters. For the aero-acoustic simulation, the 3D vocal tract shape is mapped to an enhanced area function and its equivalent transmission-line circuit representation, which is numerically simulated in the time domain [50]. The target approximation model implemented in VocalTractLab simulates continuous articulatory trajectories [51]. The model simulates dynamic articulation movements in the same way that tones in Thai have been successfully simulated [52]. The VocalTractLab generates speech waveforms with a sampling rate of 22.05 kHz with 16 bits resolution.

The shape of the vocal tract is controlled by 23 articulatory parameters, as shown in Table 1. These parameters control jaw angle, velum shape, velopharyngeal port, lips, tongue, and additional constraint parameters. The minimum and maximum parameter ranges are restricted as a soft constraint to prevent abnormal anatomic shapes.

Dynamic articulatory movements generated by VocalTractLab are controlled by gestural scores, which specify the underlying targets of each articulator in terms of their geometrical shapes and positions in a specified temporal speech interval. The motor commands for the individual articulators are then calculated based on the target approximation model [51], [53]. Beside articulatory parameters, pitch targets of the produced speech are also defined in terms of underlying targets [51].

TABLE 1. Articulatory parameters of the VocalTractLab 2.2

Description	Symbol	Min.	Max.	Unit
Horizontal hyoid position	HX	0.0	1.0	
Vertical hyoid position	HY	-6.0	-3.5	cm
Horizontal jaw displacement	JX	-0.5	0.0	cm
Jaw angle	JA	-0.7	0.0	deg
Lip protrusion	LP	-1.0	1.0	
Vertical lip distance	LD	-2.0	4.0	cm
Velum shape	VS	0.0	1.0	
Velic opening	VO	-0.1	1.0	
Tongue body center X	TCX	-3.0	4.0	cm
Tongue body center Y	TCY	-3.0	1.0	cm
Tongue tip X	TTX	1.5	5.5	cm
Tongue tip Y	TTY	-3.0	2.5	cm
Tongue blade X	TBX	-3.0	4.0	cm
Tongue blade Y	TBY	-3.0	5.0	cm
Tongue root X	TRX	-4.0	2.0	cm
Tongue root Y	TRY	-6.0	0.0	cm
Tongue side elevation 1	TS1	-1.4	1.4	cm
Tongue side elevation 2	TS2	-1.4	1.4	cm
Tongue side elevation 3	TS3	-1.4	1.4	cm
Tongue side elevation 4	TS4	-1.4	1.4	cm
Min area tongue back region	MS1	0.0	0.3	cm2
Min area tongue tip region	MS2	0.0	0.3	cm2
Min area lip region	MS3	0.0	0.3	cm2

C. DATA GENERATOR

Data generator module was designed to generate a high variation of the speech from a few observed samples. The acoustics were generated from the VTL, using 1) articulatory parameters, 2) speaker's vocal tract model, and 3) gestural score. The high degree of freedom of the three-dimensional vocal tract model resulted in many-to-many mapping between articulatory parameters and acoustics. The boundary of the interpolation function was defined from a predefined articulatory target of German vowels from the VTL. The linear interpolation of the articulation is defined as follows:

$$R = uP + (1 - u)Q, \quad u \in (0.6, 1) \tag{1}$$

P and Q are a vector consisting of 23 articulatory parameters from a randomly selected predefined articulatory target of German vowels. R is a generated articulatory target vector. u is an interpolation parameter indicating the interpolating range from P to Q. The interpolating range was constrained around P, which prevents oversampling of the central part of the vowel space produced from average articulations between P and Q.

Similarly, the speakers' vocal tract model was constructed using linear interpolation between the existing adult and a child vocal tract model from VTL, where the child vocal tract was transformed from an adult vocal tract model [54]. The function is defined as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3166922, IEEE Access

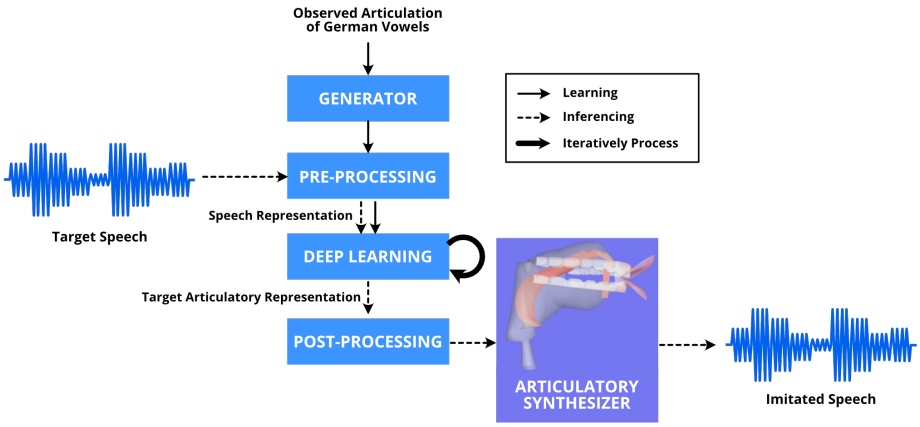


FIGURE 1. The overview of the proposed underlying target articulatory acquisition strategy

$$V_{intpl} = jV_{adult} + (1-j)V_{child}, \quad j \in (-0.3, 0.3) \quad (2)$$

The terms V_{intpl} , V_{adult} , and V_{child} are vectors of anatomical parameters of the interpolated, adult, and child speakers, respectively. The j is an interpolation factor, where the range of j was perceptually selected to ensure the naturalness of the synthesized speech.

Generated articulatory targets were then scaled using min-max articulation range of the new interpolated speaker vocal tract model, defined as follows:

$$\hat{y}_{jk} = \frac{y_k - \min(\hat{Y}_{jk})}{\max(\hat{Y}_{jk}) - \min(\hat{Y}_{jk})} \quad (3)$$

The term y_k is an articulatory k, where $k \in (1, 23)$. $\hat{Y}_{jk}$ is a generated target articulatory parameter k of the interpolated speaker vocal tract model j.

To simulate an articulatory movement, the gesture score was generated where gestures that related to the production of the vowel utterance were randomly selected from a distribution of a possible dynamic movement, while gestures related to the consonants (lip, tongue, and velic gesture) were left blank. The glottal shape gesture was fixed to modal phonation. The syllable duration of both monosyllabic and disyllabic vowel-only utterances was randomly selected from the uniform distribution between 0.5 and 1.5 seconds. For the disyllabic utterance, the transition between the first and second syllable was randomly selected, the uniform distribution between -20% and 20% from the half of the total duration. The time constant was uniformly sampled from a range of $C \in [0.015,\ 0.020]$ seconds. The glottal pressure was uniformly sampled from a range of $G \in [9000,\ 12000]$ dPa. This range of parameter values was chosen based on a perceptual evaluation of the intelligibility of the synthetic speech without distortion.

The speech was resampled to a 16 kHz with 16-bit resolution. The disyllabic vowel data along with its corresponding generated underlying articulatory target data were split into the first and second syllables. The split point was a midpoint of the disyllabic vowel speech sequence regardless of the transition time. These split syllables were treated as individual data. The amplitude of the speech was normalized into a range between -1 and 1. The learning samples were augmented into multiple representations per sample.

The speech augmentation methods included: 1) random noise injection, 2) volume perturbation, 3) pitch shifting, and 4) feature masking. The vocal tract length perturbation was excluded because it produces the same effect as in the speaker simulation method. All parameters of the augmentation function were perceptually selected to prevent a loss of intelligibility and a speech distortion from over-augmentation. In random noise injection, a sequence of noise $A(t)$ was generated at random from a continuous amplitude range of $A \in (0.001, 0.01)$. Given $X(t)$ is an normalized speech signal with an amplitude between -1 and 1 at time t, the noise injection is defined as follows:

$$X_{aug}(t) = X(t) + A(t)X(t) \quad (4)$$

For volume perturbation, a perturbation factor α was randomly selected from a continuous range $\alpha \in (1.5,\ 3)$, defined as follows:

$$X_{aug}(t) = \alpha X(t) \quad (5)$$

The pitch shifting augmentation was based on the PSOLA algorithm [34] implemented in the Librosa Python package [55]. The shifting factor β was randomly selected from a continuous range of $\beta \in (-1.0, 4)$. The range of α and β were perceptually selected to ensure the naturalness of the synthesized speech. The feature masking method was based on SpecAug [e1] where the masking was applied in 1) one of a random mel-frequency cepstral coefficient, and 2) a random segment of a speech feature time frame where the masking length was set to 10.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3166922, IEEE Access

D. PRE-PROCESSING METHOD

The speech signal was represented as Mel-frequency cepstral coefficient (MFCCs) [56] with 13 cepstral coefficient and an additional velocity and acceleration resulting in a total of 39 features per time frame. The spectrum was computed using a Hanning window with a window length of 32 ms and a frame step of 10 ms, then applying a filter bank frequency followed by the discrete cosine transformation to decorrelate a filter bank frequency. Normalization was applied using cepstral mean and variant normalization (CMVN) [57], which was a feature compensation method using the z-scoring and scaling per coefficient. The mean and variance were inferred from their distribution. The CMVN is defined as follows:

$$X_{ij}[c] = \frac{Xij[c] - \overline{X[c]}}{SD(X[c])} \tag{6}$$

The mean $X[c]$ and standard deviation $SD(X[c])$ of c^{th} coefficients are defined as follows:

$$X[c] = \frac{1}{N}\sum_{k=0}^{k=N}\sum_{m=0}^{m=J} X_{km}[c] \tag{7}$$

$$SD(X[c]) = \sqrt{\frac{\sum_{k=0}^{k=N}\sum_{m=0}^{m=J} X_{km}[c] - \overline{X[c]}}{N-1}} \tag{8}$$

The $X_{ij}[c]$ is a i^{th} th input feature of c^{th} coefficient at a frame j^{th} . N is the total number of samples. J is the total number of timeframes.

The target articulatory parameters were min-max scaled into a range between 0 and 1, using equation 3, where the min and max value was inferred from the training distribution. The JX, VO, TRX, TRY, MS1, MS2, and MS3 were excluded from the model estimation and set to a constant that is appropriate for all vowels. The TRX and TRY parameters were automatically calculated in VTL based on the tongue body, TCX, and TCY. The parameter VO controlling the velic opening was fixed for a closed velo-pharyngeal port. The JX, MS1, MS2, and MS3 were defined as a zero constant because its boundary is very close to zero and have little effect on a vowel articulation.

E. DEEP LEARNING

A bidirectional LSTM recurrent neural network (BiLSTM) [58], [59] was used as the deep learning model architecture. The BiLSTM was composed of five LSTM layers with 128 hidden units each with a backward recurrent direction. The output layer was a fully connected layer that mapped the extracted feature representation from BiLSTM to the articulatory representation. The dropout [60] with a 50% drop rate was applied. The simple multiple linear regression without any feature extraction layer was used as a baseline. Both BiLSTM and baseline take MFCC features as an input, and estimate 23 articulatory target parameters of the input as an output.

The model was trained by supervised learning using the gradient-based optimization, Adam optimizer [61], minimizing a mean square error (MSE) between estimated articulatory targets and generated articulatory targets, defined as follows:

$$MSE = \frac{\sum_{k=1}^{k=N}\sum_{j=1}^{j=M}(y_{kj} - \hat{y}_{kj})^2}{NM} \tag{9}$$

The y_{kj} and $\hat{y}_{kj}$ are a target underlying articulatory targets (labels) and estimated underlying articulatory target values of a parameter j at data-point k from the model. M is the total number of parameters. N is the total number of data in a mini-batch. The learning rate used in the optimization was 0.0001 , the batch size was 64 . Hyperparameters of the Adam optimizer were 0.9 and 0.999 for β_1 and β_2 respectively. The weight was initialized using the Kaiman initialization method [62]. Models were trained with 150 epochs with the early stopping mechanism monitoring the loss computed from the development set to prevent the overfitting problem [63].

F. POST-PROCESSING METHOD

The articulatory targets estimated by the deep neural network were inverted with min-max rescaling, using the same min and max parameter from the training distribution. Then, the parameters JX, WC, TRX, TRY, MS1, MS2, and MS3 were added, where JX and WC were 0.0., and MS1 to MS3 were -0.05.. These settings were based on a distribution of the predefined vowels in VTL. TRX and TRY were imputed using the equation from the VocalTractLab synthesizer, defined as follows:

$$TRX = 0.938TCY - 5.1100 \tag{10}$$

$$TRY = 0.831TCX - 3.0300 \tag{11}$$

G. PREDEFINED VOWEL DATASET

To test the model generalization on unseen speeches from the known speaker, and to evaluate the model in the articulatory domain, observed samples which is a predefined German vowel from the VocalTractLab were excluded from the training dataset and used as a model evaluation dataset instead. The predefined German vowel includes /a:/, /i:/, /u:/, /e:/, /E:/, /o:/, /@:/, /œ:/, /O:/, /A:/, /ø:/, and /U:/. These vowels were composed into both monosyllabic predefined vowel dataset and disyllabic predefined vowel dataset, where the later result in 144 disyllabic vowel utterances.

H. RECORDED THAI VOWEL DATASET

The speech material consisted of disyllabic Thai vowel-only utterances produced by 6 male and 6 female native Thai speakers with no reported speech and hearing disorders. The audio signals were recorded in a room without noticeable environment noise and some noticeable reverberation. The sampling rate of the recordings was 44.1kHz with a 16 bits resolution. The utterances consisted of the nine Thai vowels /a:/, /i:/, /u:/, /e:/, /ɛ:/, /ɯ:/, /ɤ:/, /o:/, and /ɔ:/ composed into

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3166922, IEEE Access

disyllabic utterances, which resulted in a total of 81 disyllabic vowel utterances per recorded set (one set per speaker). Thus, a total of 12 x 81 = 972 disyllabic vowel utterances were recorded. Seven additional sets were recorded by another native Thai speaker, which then were used as additional data to train the speech recognition model, as will be described later. Some transformations were performed by hand prior to data processing, which are 1) resampling, 2) amplitude rescaling, and 3) syllabic transition marking. Resampling was performed to reduce the speech sample rate from 44.1 kHz to 16 kHz. The amplitude of the recorded audio was scaled to match the distribution of a training synthetic speech by multiplication with a constant factor. The transition time between syllables in the disyllabic vowel utterances was manually marked based on visual inspection of the waveform.

I. DESIGN OF EXPERIMENT

The model was trained on two synthetic training sets from the generator, which were a monosyllabic vowel utterance and a disyllabic vowel utterance. To prevent overfitting from training the model too long, the dataset was split into a training, validating, and testing dataset, where the validating dataset was used for performance monitoring during training, and the testing dataset was used for a final performance test. After the training, the model performance was evaluated both in the articulatory domain and acoustic domains using the predefined vowel dataset. Lastly, the model was evaluated on the acoustic domain using the recorded disyllabic Thai vowel samples. Since the acoustic from different speakers cannot be directly compared, the model performance was evaluated in a phonemic domain using a listening test, described in the following section.

The effect of the proposed generator, e.g., speaker simulation and data augmentation during the data generating process, was studied in four experiments each using a different dataset: 1) the originally proposed dataset; 2) the dataset without speaker simulation; 3) the dataset without data augmentation; 4) the dataset without both speaker simulation and speech augmentation. The training on all four models used the same experimental setting. The performance of the training was evaluated by resynthesizing the recorded disyllabic Thai vowel samples, and then measuring the performance in terms of phoneme recognition accuracy using a speech recognition model.

J. EVALUATION METRICS

This study evaluated models in the articulatory domain and the acoustic domain using both visual and numerical assessment. In the articulatory domain, the root means square error (RMSE) and R-squared (R^2) were used to measure the error between the observed underlying articulatory target and estimated underlying articulatory target by the model. For the acoustic domain, F1, F2, F3 formant errors between the target speech and the reproduced speech by the model were measured using a mean absolute percentage formant error. Formants were extracted using the Praat script [64]. The mean absolute percentage formant error is defined as follows:

$$\text{MAPE} = \frac{\sum(|\frac{F_i - F_a}{F_a}|100)}{N} \tag{12}$$

where F_i is a formant of an imitated speech. F_a a is a formant of a target speech. N is a total amount of a formant sample in the speech data. For the phonemic domain, the precision metric was used as a numerical score. The precision is defined as follows:

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} - \text{FalsePositive}} \tag{13}$$

The TruePositive is a number of correct predictions. The FalsePositive is a number of incorrect predictions, where the actual target is negative.

K. LISTENING TEST

The listening test was conducted with 25 native Thai listeners who participated in this experiment. 14 listeners are female, and others are male. The listener's age is distributed around 23 to 27 years old. The listener was asked to identify the phoneme of the given set of disyllabic vowel utterances reproduced by the proposed model. These utterances were composed of vowels as described in Subsection II-H consisting of 81 utterances. The utterances were presented to the participants in a random order.

L. RECOGNITION TEST USING SPEECH RECOGNITION

To measure the effect of the proposed generator, the recognition test using a speech recognition model was used to evaluate the intelligibility of a reproduced speech in a phonemic domain. The speech recognition model was trained on the recorded disyllabic Thai vowel speech. This test assumes that if the reproduced speech was intelligible enough, the speech recognition trained from Thai vowels should be able to identify phonemes correctly. The model architecture was defined as a shallow LSTM recurrent network consisting of two LSTM layers and 64 hidden units per layer. The output layer was a fully-connected layer with nine units representing the phonemic target class. The model was trained by supervised classification learning, where the MFCCs representation was used as a speech feature and its phonemic representation was used as a learning target. The cross-entropy loss was used as an objective function to train this model, defined as follows:

$$L(\hat{z}, z) = -\sum_{c=1}^{C} z\log(\hat{z}) \tag{14}$$

$$\text{Softmax}(a) = \frac{e^{a_c}}{\sum_{d=1}^{C} e^{a_d}} \tag{15}$$

where $\hat{z}$ is a predicted probability produced by the $Softmax(a)$ function. C is the total number of classes. a is an activation from the previous layer. z is a ground truth of a predicted class. The permutation test and bootstrapping subsampling method were used to ensure the model's fitness.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3166922, IEEE Access

M. DATA VISUALIZATION

A uniform manifold approximation and projection (UMAP) [65] was used to visualize the cluster of phonemes from the speech and articulatory targets in a two-dimensional space. UMAP reduces dimensions of the input by constructing high-dimensional topological space from the input using simplicial complexes, and then using an optimization to find a topology of a lower dimension that is similar to the initial high-dimensional topology. UMAP preserved both local structure and global structure, meaning that similar data are clustered together, and similar categories of data are shown close to each other.

III. RESULTS

A. DATA EXPLORATION ANALYSIS

Figure 2 shows the plot of articulatory targets from the training dataset where articulatory parameters were projected into a two dimensional space using UMAP. As can be seen, there is a clear separation of the clusters of the phoneme groups, indicating that the training samples are well-defined. The bottom left clusters are articulations that have a wide open lip. The top right clusters are articulations with a closed lip. The articulations toward the top right have a rounded lip and a raised tongue back, and vice versa.

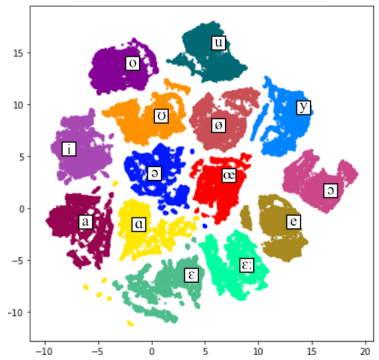


FIGURE 2. The underlyinh articulatory targets space using UMAP

Figure 3 shows the visualization of the phoneme groups of the recorded speech from the MFCC features using UMAP. As can be seen, /i:/, /e:/, /ɛ:/, /a:/, and /ɔ:/ are clearly separated from others. Both /u:/, /o:/, and /ɤ:/, /ɯ:/ were clustered together but still separable. /ɔ:/ and /u: o:/ formed one big cluster, representing the vowels with rounded lips. /i:/, /e:/. /ɤ:/ and /ɯ:/ form another big cluster, representing speech produced with wide to median opened mouth without lip rounding. Speaker variations present in the speech samples may have caused the overlaps between the clusters. The sources of these variations are likely to be: 1) individual speakers; 2) gender differences; 3) and phoneme contexts.

B. MODEL PERFORMANCE ON SYNTHETIC DATASET

Table 2 shows the result of the model performance evaluated on the synthetic samples. The error in the articulatory domain

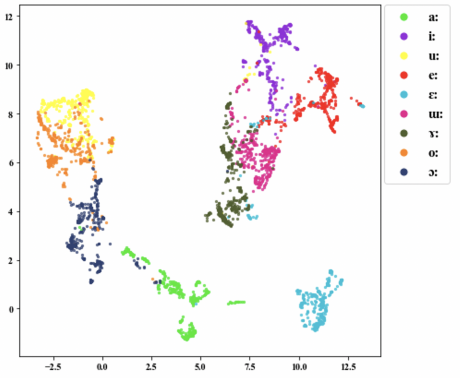


FIGURE 3. MFCCs feature plot of a recorded Thai vowels dataset using UMAP

shows that the BiLSTM performed better than the baseline on both monosyllabic and disyllabic vowel utterances. Next, the models were evaluated on the predefined vowel dataset. The results are shown in Table 3. The BiLSTM achieved small RMSEs for both monosyllabic and disyllabic vowel utterances, indicating that the model can estimate the articulation of the unseen speech sample from a known speaker. To further analyze the model performance, the mean absolute percentage formant error with a 95% confidence interval was used to measure the resynthesizing error between the target speech from predefined vowels and the re-synthesized speech from the model in the acoustic domain. As shown in Table 4, the BiLSTM achieved a low percent error rate, indicating that the model can accurately reproduce the unseen target speech from a known speaker.

TABLE 2. The model performance evaluation results on learning samples

Monosyllabic vowels utterance				
	Train Samples		Test Samples	
Model	RMSE	R2	RMSE	R2
baseline	0.3441	0.5117	0.3348	0.5344
BiLSTM	0.1172	0.9440	0.1385	0.9202
Disyllabic vowels utterance				
	Train Samples		Test Samples	
Model	RMSE	R2	RMSE	R2
baseline	0.3575	0.4795	0.3462	0.5048
BiLSTM	0.1220	0.9395	0.1428	0.9147

TABLE 3. The model performance evaluation results on predefined vowels on an articulatory domain

	Monosyllabic vowels		Disyllabic vowels	
Model	RMSE	R2	RMSE	R2
baseline	0.3857	0.5923	0.4043	0.5381
BiLSTM	0.1968	0.8564	0.1434	0.9258

The model performance of each estimated articulatory parameter on a predefined monosyllabic vowel is shown in Table 5. The result shows that the BiLSTM was weak in estimating TTX and TS1 parameters. While TS1 has little effect on the produced speech, TTX, the tongue tip, may cause a slight error when reproducing a speech.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3166922, IEEE Access

TABLE 4. The formant MAPE results on predefined vowels

Monosyllabic vowels utterance			
Model	F1 % Error	F2 % Error	F3 % Error
baseline	77.38 ± 16.03	151.50 ± 27.09	161.87 ± 31.03
BiLSTM	7.90 ± 0.62	7.60 ± 0.55	14.21 ± 1.90
Disyllabic vowels utterance			
Model	F1 % Error	F2 % Error	F3 % Error
baseline	119.23 ± 4.55	171.34 ± 5.99	226.90 ± 8.22
BiLSTM	2.73 ± 0.04	7.74 ± 0.19	8.38 ± 0.15

TABLE 5. The RMSE and R^2 of the estimated underlying articulatory targets on predefined monosyllabic vowels

Articulatory Parameter	RMSE	R^2	Articulatory Parameter	RMSE	R^2
HX	0.153	0.950	TTX	0.385	0.645
HY	0.260	0.767	TTY	0.281	0.775
JA	0.124	0.931	TBX	0.144	0.923
LP	0.236	0.853	TBY	0.292	0.756
LD	0.193	0.936	TS1	0.473	0.514
VS	0.171	0.945	TS2	0.173	0.882
VO	0.243	0.913	TS3	0.180	0.905
TCX	0.138	0.958	TS4	0.094	0.949
TCY	0.141	0.957			

The model performance of each resynthesize vowel in an acoustic domain measured by absolute percentage formant error is shown in Table 6. The re-synthesized speech of the phoneme /i:/, /u:/, /o:/, and /ø:/ from the BiLSTM had a higher F1 error compared to other vowels. Using a t-statistical test, F1, F2, and F3 have p-values larger than 0.1, as shown in Table 7. Thus, the null hypothesis of having similar speech does hold. Therefore, these errors did not cause the target speech and re-synthesized speech to be significantly different.

TABLE 6. Absolute percentage formant error result on predefined monosyllabic vowels

Phonetic	F1 %Error	F2 %Error	F3 %Error
/a:/	1.41	1.64	4.38
/i:/	15.47	1.79	5.72
/u:/	17.38	4.19	1.09
/e:/	5.93	0.06	1.24
/ɛ:/	8.20	5.18	1.32
/o:/	12.62	2.82	2.67
/ə:/	0.67	3.17	0.41
/œ:/	8.81	1.32	0.64
/ɔ:/	2.16	5.72	1.77
/ɑ:/	0.54	0.17	0.50
/ø:/	14.70	2.80	2.81
/ʊ:/	6.86	2.80	0.67

TABLE 7. Statistical comparison between the target speech and the re-synthesized speech of a monosyllabic vowel

	F1	F2	F3
t-test	0.151	-0.059	-0.088
p-value	0.882	0.954	0.930

Table 8 shows the model performance of each estimated articulatory parameter on predefined disyllabic vowels. The model estimated articulation of a disyllabic vowel better than that of a monosyllabic vowel, where most of the estimated articulatory parameters have lower RMSE and higher R2. Table 9 shows the average F1, F2, and F3 formant error between target and the re-synthesized disyllabic vowel utterance, where the first and the second syllables of disyllabic vowel utterance were measured separately. From the result, the model noticeably has a high F2 error on the vowel /u:/ for both halves of the disyllabic vowel. Using the t-statistical test to test the difference between target and re-synthesized disyllabic vowel speech, both speeches were not significantly different where the p-value for F1, F2, and F3 of both parts are more than 0.1, as shown in Table **??**.

TABLE 8. The RMSE and R^2 of the estimated underlying articulatory targets on predefined disyllabic vowels

Articulatory Parameter	RMSE	R^2	Articulatory Parameter	RMSE	R^2
HX	0.152	0.950	TTX	0.259	0.840
HY	0.158	0.914	TTY	0.192	0.895
JA	0.109	0.946	TBX	0.136	0.931
LP	0.178	0.916	TBY	0.166	0.922
LD	0.139	0.966	TS1	0.290	0.817
VS	0.165	0.948	TS2	0.156	0.904
VO	0.176	0.954	TS3	0.148	0.936
TCX	0.083	0.985	TS4	0.109	0.931
TCY	0.091	0.982			

TABLE 9. The formant MAPE results on predefined disyllabic vowels

Label	First syllable			Second syllable		
	F1 %Error	F2 %Error	F3 %Error	F1 %Error	F2 %Error	F3 %Error
/a:/	1.79	2.11	2.45	1.66	1.94	2.10
/i:/	3.99	0.69	1.40	2.63	0.41	1.41
/u:/	2.95	15.46	1.15	4.27	11.10	1.58
/e:/	2.30	0.15	0.32	1.75	0.36	1.51
/ɛ:/	3.10	1.47	1.84	3.91	3.05	1.81
/o:/	3.85	2.31	0.75	4.28	2.38	1.13
/ə:/	0.55	0.49	0.37	0.92	1.06	0.17
/œ:/	2.34	6.49	1.74	3.24	5.62	2.21
/ɔ:/	1.67	4.87	0.40	1.22	1.38	0.75
/ɑ:/	0.37	1.10	0.91	0.98	2.40	0.44
/ø:/	5.59	0.47	2.08	5.37	0.81	1.44
/ʊ:/	1.49	4.70	1.39	5.40	7.98	1.35

TABLE 10. Statistical comparison between the target and the reproduced speech formant of a disyllabic vowel utterance

	First syllable			Second syllable		
	F1	F2	F3	F1	F2	F3
t-test	-0.027	0.029	-0.169	0.023	0.080	-0.248
p-value	0.978	0.977	0.867	0.981	0.937	0.806

Figure 4 shows the estimated articulation from the model when reproducing monosyllabic vowels (top) and disyllabic vowels (bottom) visualized from the VocalTractLab. These estimated articulations from the model were behaved according to the international phonetic alphabet chart (IPA) [66], where /a:/ had tongue towards front and month open, /i:/ had a tongue towards the front and mouth slightly close, and /u:/ had tongue towards back and mouth slightly close.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3166922, IEEE Access

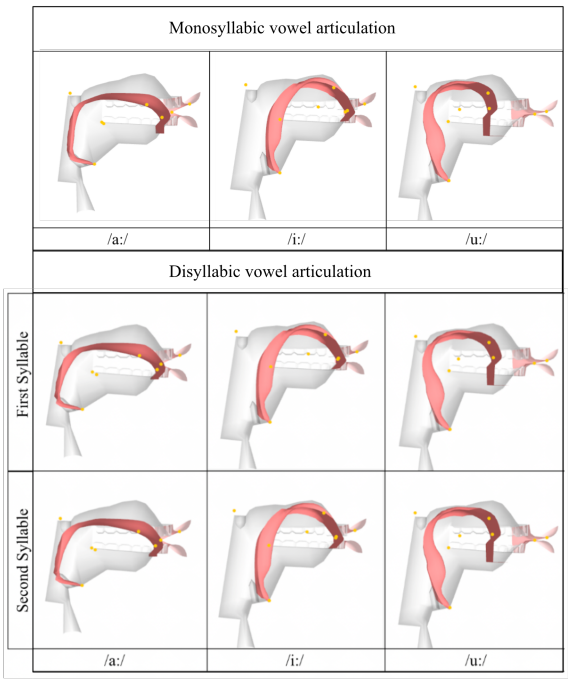


FIGURE 4. Average disyllabic vowel articulation according to the IPA chart

C. MODEL PERFORMANCE ON THE RECORDED THAI VOWEL DATASET

Table 11 shows the mean absolute percentage formant error between the target disyllabic Thai vowel speech and the re-synthesized speech from estimated underlying articulatory targets by the model. Figure 5 shows the similar arrangement of formants between target and re-synthesized speech. The result in Table 12 from the t-statistical test also shows that average F1 and F2 frequencies were not statistically significantly different, where the p-value for F1, F2, and F3 of both first and second syllable of a disyllabic vowel are larger than 0.1, as shown in . While the t-statistical test shows that average F3 was statistically significantly different, only F1 and F2 are enough to identify vowels [67]. As shown in both numerically and visually, while it is not statistically different, the error seems large. However, the formant from the different speakers cannot be directly compared because it is affected by the shape of the vocal tract. Therefore, formants of the re-synthesized speech were compared with the empirical formant [68] range. Figure 6 illustrates that most of the average formants are within the empirical formant range, indicating that the model accurately re-synthesized target speeches which were recorded by the actual human.

Figure 7 shows the comparison result between the spectrograms of the target speech recorded by the Thai speaker and re-synthesized speech by the model. The red contour shows the speech formants. Visually, the F1 and F2 formants of both target and re-synthesized speech signals are comparable, while F3 in some speeches were different, e.g., /ɛːɔː/ and

TABLE 11. The formant MAPE results on recorded disyllabic Thai vowel data

Label	First syllable			Second syllable		
	F1 %Error	F2 %Error	F3 %Error	F1 %Error	F2 %Error	F3 %Error
/aː/	25.45	15.20	17.05	26.86	17.35	17.06
/iː/	18.78	27.56	14.71	31.11	32.97	14.57
/uː/	16.82	32.54	10.96	21.79	36.91	13.96
/eː/	19.56	21.36	15.99	22.53	21.47	15.76
/ɛː/	18.48	28.97	16.13	24.56	26.75	17.15
/ɯː/	12.79	12.73	16.06	17.55	15.72	16.28
/ɣː/	14.05	11.58	15.98	21.42	13.20	16.23
/oː/	21.36	25.64	12.25	23.65	22.90	12.16
/ɔː/	15.54	20.82	15.09	23.32	21.51	15.69

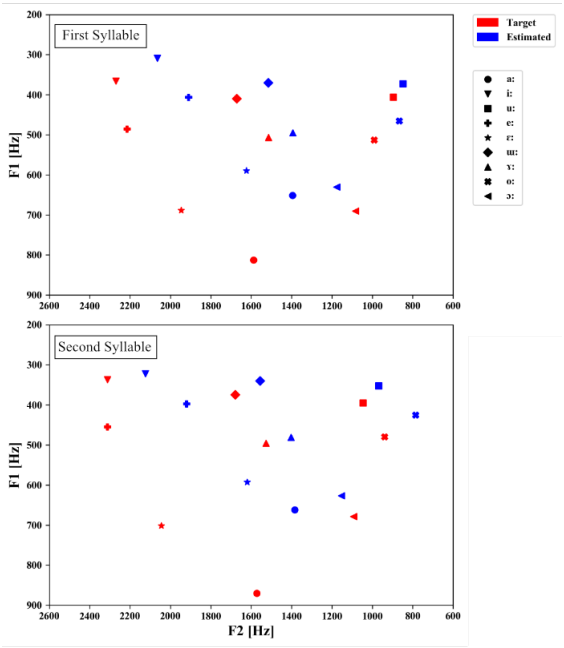


FIGURE 5. Average formant plot comparing recorded Thai vowel and a reproduced speech

/ɯːoː/. Figure 8 shows the group of estimated underlying articulatory targets from the model where the color represents the phoneme of the target speech. The articulatory parameters were projected into a two-dimensional space using the UMAP. Articulations of the same phoneme were clustered together. However, while there was a clear distinction between some groups, some mixture between phonemes were presented.

Figure 9 shows articulations of an estimated underlying articulatory target from the recorded disyllabic Thai vowel. The visualization is shown as an estimation of the first and

TABLE 12. Statistical comparison between the recorded and the reproduced speech of a disyllabic vowel utterance

	First syllable			Second syllable		
	F1	F2	F3	F1	F2	F3
t-test	-0.718	-0.506	-5.253	-0.847	-0.708	-3.765
p-value	0.483	0.620	0.000	0.410	0.489	0.002

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3166922, IEEE Access

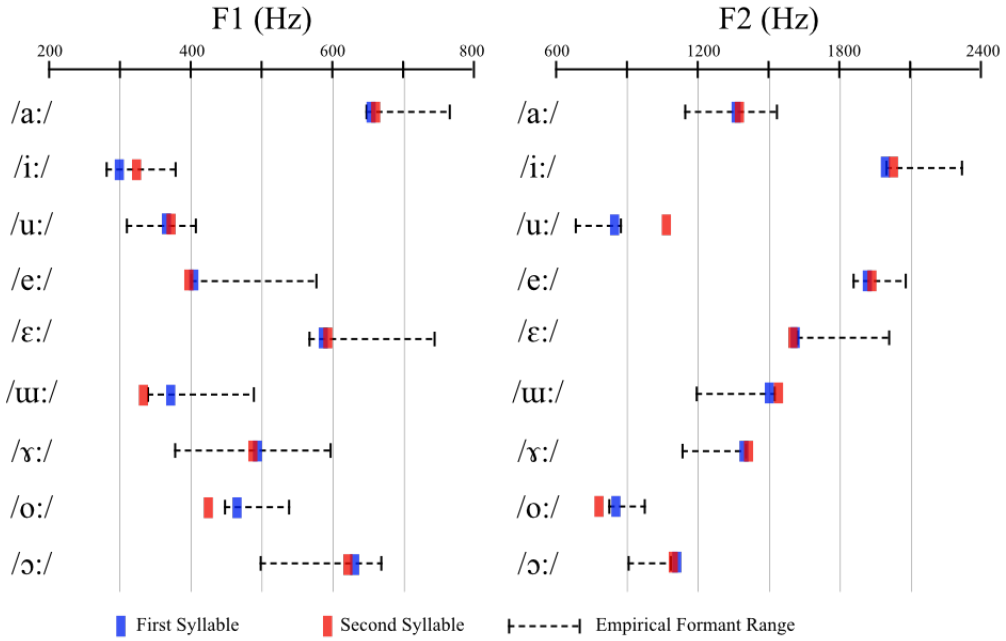


FIGURE 6. F1 and F2 Comparison between reproduced speech and the empirical formant range

the second syllable from the disyllabic vowel. The position of a tongue and the mouth of each vowel were according to the Thai phonetic chart [69]. The top right shows the tongue towards the back and mouth close, and the bottom left shows the articulation with the tongue forward and mouth open. Figure 10 shows a lip's shape of the /a:/, /i:/, and /u:/ vowel to visualize the roundedness. The lip's shape of /u:/ was rounded, while /a:/ and /i:/ are unrounded.

Figure 12 shows the phoneme recognition rate, where the participants were asked to identify the Thai phoneme of the re-synthesized speech from the target recorded Thai speech by the model. The high recognition rate meant that the re-synthesized speech by the model was intelligible enough to be classified correctly. The result shows that most of the vowels were correctly identified by participants. Overall, the model achieved more than 80% classification accuracy for re-synthesizing disyllabic vowel utterances, except for /a:/ with the recognition rate slightly below 80% and /ɔ:/ with a recognition rate below 40%. The only difference in the articulation of /a:/ and /ɔ:/ is the rounding of the lip where the articulation for /ɔ:/ is rounder than /a:/. Figure 11 shows the table where each row is a vowel that the model tried to reproduce, and each column is the vowel that was recognized by the participants. The diagonal represents the correct recognition. Most of the incorrect identifications were from a pair of vowels that having similar phoneme, which are /a:/ and /ɔ:/.

D. EFFECTS OF THE DATA GENERATOR

The effect of the proposed data generator module was evaluated by measuring the model target speech re-synthesizing performance on recorded disyllabic Thai vowel dataset. The result shown in Table 13 indicates that applying both speaker simulation and augmentation methods improved the model performance. The result clearly shows that the performance significantly dropped without both speaker simulation and data augmentation.

TABLE 13. The model performance trained on different dataset evalutated using recognition test

Model trained with different speech dataset	Avg. Precision
With purposed data generation	0.826
dataset without speaker simulation	0.714
dataset without data augmentation	0.683
dataset without both speaker simulation and data augmentation	0.587

IV. DISCUSSION

Results showed in both visually and numerically indicate that the deep learning model can estimate the underlying articulatory targets from monosyllabic and disyllabic vowel utterances by both known and unknown speakers. The model was trained with synthetic samples generated by VocalTract-Lab combining with both interpolation and data augmentation methods using a few observed samples as a seed. The model evaluation result on the dataset with label articulatory information showed that the model could learn the acoustic and the underlying articulatory targets relationship from the synthetic training dataset, and it can generalize this knowledge to the unseen vowel utterances speaked by the same speaker from the synthetic dataset. The model evaluation result on the recorded Thai vowel speech dataset, where the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3166922, IEEE Access

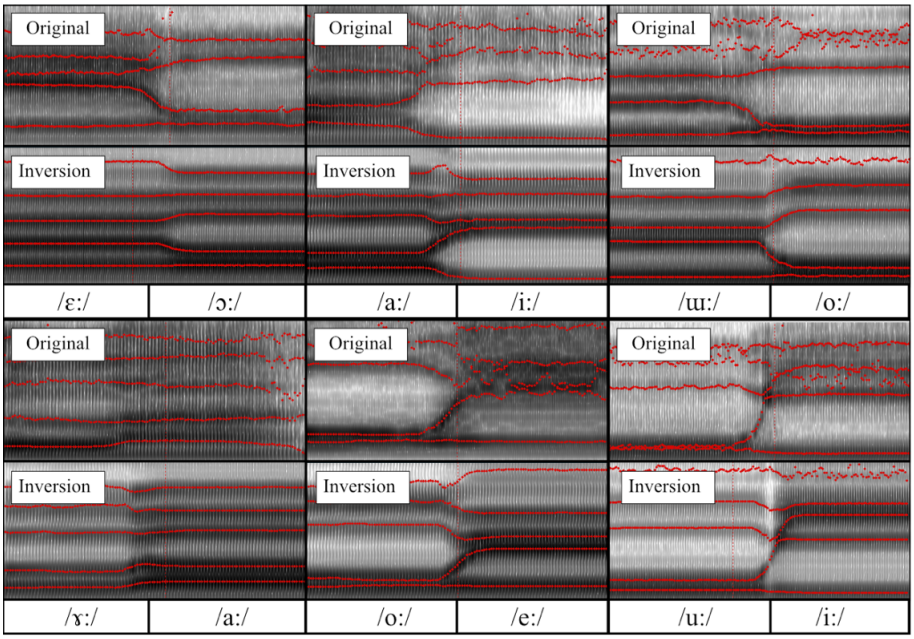


FIGURE 7. Spectrogram comparing recorded Thai vowel and a reproduced speech. The red line is a formant contour, and the blue line is a pitch contour

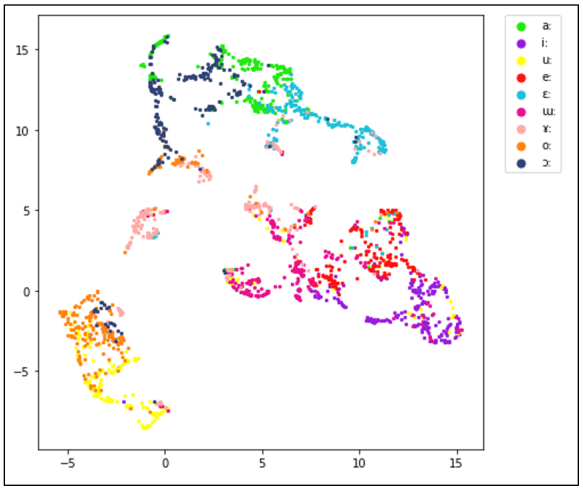


FIGURE 8. The group of estimated underlying articulatory targets from the model visualized using UMAP

model was trained on the synthetic samples, showed that the model could generalize the acoustic to the underlying articulatory targets relationship to unseen utterances speaked by unseen native Thai speakers as well. The variations from the speaker simulation and data augmentation increased the model's knowledge about the acoustic-to-underlying articulatory targets relationship, leading to the better estimation of the underlying articulatory targets. The data augmentation method affected the model performance more than speaker simulation. This is because the speaker simulation only increased the variation in terms of speaker characteristic, where the vocal tract area was interpolated from the same vocal tract model thus the overall shape of the vocal tract did not change much, while the data augmentation augmented pitch, volumn, and added random noise which reduce the model overfitting. The result from the listening test showed that the proposed strategy provides a nearly perceptually accurate mapping between the Thai vowel speech with the German vocal tract configurations provided in VocalTractLab, indicating the ability to decouple the articulatory mechanisms from the linguistic information. Thus, the generalization of the strategy and the learned targets should provide evidence that this learning strategy can also provide a potential means of target learning besides a general supervised learning approach.

The advantage of the proposed strategy over the previous study [10] is that the model performed well although its training was based on only a few observed samples. While only trained with synthetic samples, the model can estimate the underlying articulatory targets from the recorded Thai speech from various native Thai speakers, which are unknown to the model, with actual minimum background noise. Therefore, it shows a promising result to apply these underlying articulatory targets acquisition strategies on the real-world application.

Further improvement of this model is highly needed. First, the model assumed that the perceptual segmentation of the speech was learned before the learning of speech production. Second, the effect consonants were not studied. Future work should explore how humans segment speech, e.g., how many syllables are in the target speech, and then learn to imitate those signals. The attention model for machine learning has proved very useful for speech recognition [70], [71], where a speech utterance was translated into a sequence of characters. Therefore, the attention model could also be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3166922, IEEE Access

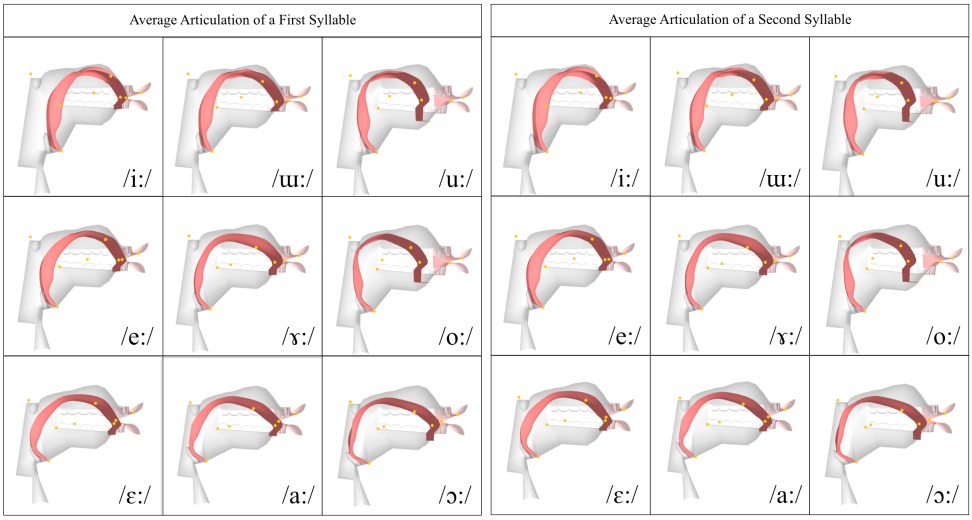


FIGURE 9. Average reproduced vowel articulation fsplit into a first syllable (left) and a second syllable (right) of a disyllabic vowel utterance

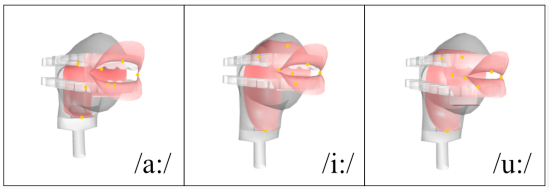


FIGURE 10. The front view of an average reproduced vowel articulation of /a:/, /i:/, and /u:/

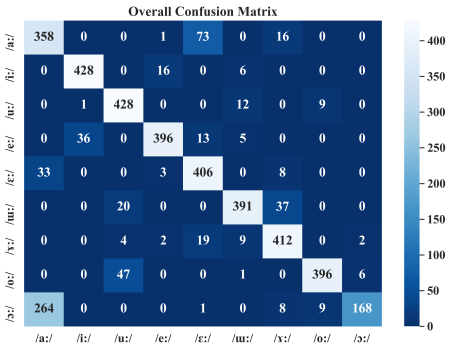


FIGURE 11. Confusion matrix result between recorded Thai vowel and a reproduced vowel from the model

applied to estimate a set of underlying articulatory targets from the given speech utterance. Next, the method could be modified to be able to estimate articulatory targets related to consonants. In addition, to improve estimation performance and generalization, self-supervised methods [72] are worth exploring on how the speech is featured without any explicit target.

V. CONCLUSION

This study explored the estimation of underlying articulatory targets by learning the mapping between acoustic and the underlying articulatory targets of Thai vowels using a bidirectional long short-term memory recurrent neural network. The VocalTractLab was used as a generative model to generate acoustic data from articulatory parameters, and the deep learning approach was used to model the acoustic-to-articulatory relationship. Using a few data points as representative of Thai vowels, the speech data augmentation and a speaker simulation method allowed us to extract more information from the data and improve the estimation of the underlying articulatory targets. The results demonstrated that the proposed strategy was able to accurately reproduce speech from a given target utterance from unseen Thai speakers. Thus, the model represents an effective strategy for rapid mapping of acoustic data to articulatory target parameters.

The caveats of this study are: 1) the proposed method required a predefined syllable segmentation of the input speech, and 2) this study excluded consonants. Therefore, the recommended improvements from this study are to include the estimation of underlying articulatory targets of consonant-vowel utterances, and to explore methods which can directly estimate the sequence of speech syllables without a need for the predefined speech segmentation.

REFERENCES

[1] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 54, no. 4, pp. 1070—-1082, 1973.

[2] A. Noiray, A. Popescu, H. Killmer, E. Rubertus, S. Krüger, and L. Hintermeier, "Spoken language development and the challenge of skill integration," *Frontiers in Psychology*, vol. 10, no. 1, p. 2777, 2019.

[3] H. A. Lim, "Effect of "developmental speech and language training through music," *Journal of music therapy*, vol. 47, no. 1, pp. 2–26, 2010.

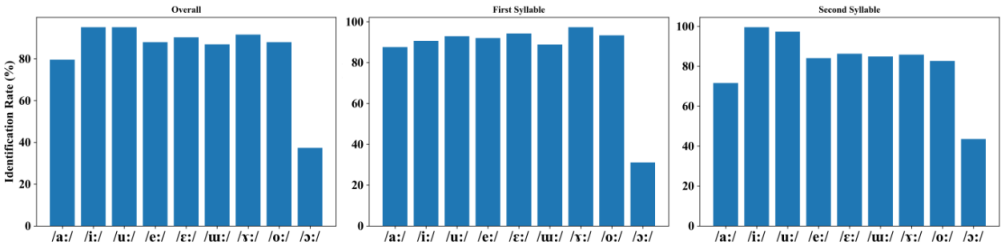


FIGURE 12. Overall identifiable phonetic rate from the perception test (left) and the identification rate of a first syllable (middle) and second syllable (right)

[4] G. Sivaraman, C. Espy-Wilson, and M. Wieling, "Analysis of acoustic-to-articulatory speech inversion across different accents and languages," *Proc. Interspeech 2017*, pp. 974–978, 2017.

[5] A. Zolnay, R. Schluter, and H. Ney, "Acoustic feature combination for robust speech recognition," *In Acoustics, Speech, and Signal Processing*, vol. 1, no. I, p. 457, 2005.

[6] Z. Ling, K. Richmond, J. Yamagishi, and R. Wang, "Integrating articulatory features into hmm-based parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[7] J. Zhao, H. Yuan, W. Leung, H. Meng, J. Liu, and S. Xia, "Audiovisual synthesis of exaggerated speech for corrective feedback in computer-assisted pronunciation training," *In Acoustics, Speech and Signal Processing*, pp. 8218–8222, 2013.

[8] K. Leung, M. Mak, and S. Kung, "Applying articulatory features to telephone-based speaker verification," *In Acoustics, Speech and Signal Processing*, vol. 1, pp. 1–85, 2004.

[9] G. Hofer and K. Richmond, "Comparison of hmm and tmd methods for lip synchronization," *Proc. Interspeech 2010*, p. 454–457, 2010.

[10] S. Prom-on, P. Birkholz, and Y. Xu, "Identifying underlying articulatory targets of thai vowels from acoustic data based on an analysis-by-synthesis approach," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 1, p. 23, 2014.

[11] A. Toutios and K. Margaritis, "Rough guide to the acoustic-to-articulatory inversion of speech," *Hellenic European Conference of Computer Mathematics and its Applications*, pp. 746–753, 2003.

[12] C. Qin and C.-P. MA., "An empirical investigation of the non-uniqueness in the acoustic-to-articulatory mapping," *International Speech Communication Association*, pp. 74–77, 2007.

[13] A. Afshan and P. Ghosh, "Improved subject-independent acoustic-to-articulatory inversion," *Speech Communication*, vol. 66, pp. 1–16, 2007.

[14] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.

[15] J. Ohala, "Coarticulation and phonology," *Language and speech*, vol. 36, p. 2–3, 1993.

[16] A. Illa and P. Ghosh, "The impact of speaking rate on acoustic-to-articulatory inversion," *Computer Speech and Language*, vol. 59, pp. 75–90, 2020.

[17] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[18] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, 2004.

[19] T. Toda, A. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," *In Eighth International Conference on Spoken Language Processing*, pp. 1129–1132, 2004.

[20] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," *In Acoustics, Speech and Signal Processing*, pp. 4450–4454, 2015.

[21] P. Zhu, L. Xie, and Y. Chen, "Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings," *In Sixteenth Annual Conference of the International Speech Communication Association*, pp. 2192–2196, 2015.

[22] H. Li, J. Tao, M. Yang, and B. Liu, "Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings," *In Acoustics, Speech and Signal Processing*, pp. 4854–4858, 2015.

[23] A. Illa and P. K. Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion," *In Acoustics, Speech and Signal Processing*, p. 5931–5935, 2019.

[24] H. Li, J. Tao, M. Yang, and B. Liu, "Estimate articulatory mri series from acoustic signal using deep architecture," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 4854–4858, 2015.

[25] B. Uria, S. Renals, and K. Richmond, "A deep neural network for acoustic-articulatory speech inversion," *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, p. 1, 2011.

[26] P. Tobing, H. Kameoka, and T. Toda, "Deep acoustic-to-articulatory inversion mapping with latent trajectory modeling," *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, p. 1274 – 1277, 2017.

[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[28] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[29] I. Rebai, Y. BenAyed, W. Mahdi, and J. P. Lorré, "Improving speech recognition using data augmentation and acoustic model fusion," *Procedia Computer Science*, vol. 112, pp. 316–322, 2017.

[30] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," *In Sixteenth Annual Conference of the International Speech Communication Association*, vol. 112, pp. 3586–3589, 2015.

[31] T. S. Nguyen, S. Stüker, J. Niehues, and A. Waibel, "Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation," *International Conference on Acoustics, Speech and Signal Processing*, pp. 7689–7693, 2020.

[32] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013, p. 21.

[33] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019, 2613-2617*, 2019.

[34] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," *IEEE International Conference on Acoustics, Speech, And Signal Processing*, vol. 11, pp. 2015–2018, 1986.

[35] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," *IEEE International Conference on Acoustics, Speech, And Signal Processing*, vol. 2, pp. 554–557, 1993.

[36] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[37] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu. (2019) Speech recognition with augmented synthesized speech.

[38] S. Fairee, B. Sirinaovakul, and S. Prom-on, "Acoustic-to-articulatory inversion using particle swarm optimization," *In Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, pp. 1–6, 2015.

[39] P. B. D. Jackèl and B. J. Kroger, "Construction and control of a three-dimensional vocal tract model," *In Acoustics, Speech and Signal Processing ICASSP Proceedings*, vol. 1, pp. 1–1, 2006.

[40] Y. Gao, S. Stone, and P. Birkholz, "Articulatory copy synthesis based on a genetic algorithm." in *INTERSPEECH*, 2019, pp. 3770–3774.

[41] G. Yingming, P. Steiner, and P. Birkholz, "Articulatory copy synthesis using long-short term memory networks," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pp. 52–59, 2020.

[42] I. Howard and M. Huckvale, "Training a vocal tract synthesiser to imitate speech using distal supervised learning," *SpeCom: 10th International Conference on Speech and Computer*, vol. 2, pp. 159–162, 2005.

[43] S. Panchapagesan and A. Alwan, "A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the maeda articulatory model," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. 2144–2162, 2011.

[44] I. Howard and M. Huckvale, "Modelling vowel acquisition using the birkholz synthesizer," *Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pp. 304–311, 2019.

[45] A. K. Philippsen, R. R. Felix, and W. Britta, "Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model," in *4th International Conference on Development and Learning and on Epigenetic Robotics*. IEEE, 2014, pp. 195–200.

[46] I. Howard and M. Huckvale, "Learning to control an articulatory synthesizer by imitating real speech," *ZAS Papers in Linguistics*, vol. 40, pp. 63–78, 2013.

[47] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS one*, vol. 8, no. 4, p. e60603, 2013.

[48] ——, "Vocaltractlab: Towards high-quality articulatory speech synthesis [computer program]."

[49] P. Birkholz and B. J. Kröger, "Vocal tract model adaptation using magnetic resonance imaging," in *7th International Seminar on Speech Production (ISSP'06)*, 2006, pp. 493–500.

[50] P. Birkholz, D. Jackèl, and B. J. Kroger, "Simulation of losses due to turbulence in the time-varying vocal system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1218–1226, 2007.

[51] S. Prom-On, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in mandarin and english as a process of target approximation," *The journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 405–424, 2009.

[52] S. Prom-on, "Pitch target analysis of thai tones using quantitative target approximation model and unsupervised clustering," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[53] P. Birkholz, B. J. Kroger, and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant–vowel sequences," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1422–1433, 2010.

[54] P. Birkholz and B. J. Kröger, "Simulation of vocal tract growth for articulatory speech synthesis," in *Proceedings of the 16th international congress of phonetic sciences*, 2007, pp. 377–380.

[55] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[56] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[57] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[58] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[59] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[60] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[63] Y. Yao, L. Rosasco, and A. Caponnetto, "On early stopping in gradient descent learning," *Constructive Approximation*, vol. 26, no. 2, pp. 289–315, 2007.

[64] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[65] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[66] I. P. Association, I. P. A. Staff *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[67] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970, no. 2.

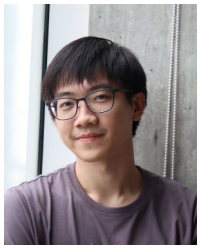
[68] A. S. Abramson, "The tones of central thai: Some perceptual experiments," *Studies in Thai linguistics in honor of William J. Gedney*, vol. 1, p. 16, 1975.

[69] K. Tingsabadh and A. Abhramson, "Thai sound: An acoustic study," *Journal of the International Phonetic Association*, vol. 22, no. 1, pp. 24–48, 1993.

[70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[71] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[72] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," *arXiv preprint arXiv:1906.12340*, 2019.



THANAT LAPTHAWAN received a B.Eng. in Computer Engineering from King Mongkut's University of Technology Thonburi in 2017. He is currently a postgraduate student at the Department of Computer Engineering, King Mongkut's University of Technology Thonburi. His research interests in artificial intelligence related to speech, natural language, and image.

SANTITHAM PROM-ON (Member, IEEE) received the B.Eng. degree in computer engineering and the Ph.D. degree in electrical and computer engineering from the King Mongkut's University of Technology Thonburi, Thailand. He is currently an Assistant Professor with the Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi.

PETER BIRKHOLZ received the Diploma in computer science and the Ph.D. degree (with distinction) in signal processing from the Institute for Computer Science, University of Rostock, Rostock, Germany, in 2002 and 2005, respectively. Currently, he is the chair of Speech Technology and Cognitive Systems, Technische Universitt Dresden Fakultt Elektrotechnik and Informationstechnik.

YI XU received the Ph.D. degree in Linguistics from the University of Connecticut, United States in 1993. He is currently a Professor of Speech Sciences at the Department of Speech, Hearing and Phonetic Sciences, Division of Psychology and Language Sciences, University College London.