RS-MSConvNet: A Novel End-to-End Pathological Voice Detection Model

WONGSATHON PATHONSUWAN1, KHOMDET PHAPATANABURI2, PRAWIT BUAYAI3, TALIT JUMPHOO1, PATIKORN ANCHUEN4, MONTHIPPA UTHANSAKUL1, (Member, IEEE), AND PEERAPONG UTHANSAKUL1, (Member, IEEE)

1School of Telecommunication Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand
2Department of Telecommunication Engineering, Faculty of Engineering and Technology, Rajamangala University of Technology Isan (RMUTI), Nakhon Ratchasima 30000, Thailand
3Graduate Faculty of Interdisciplinary Research, University of Yamanashi, Kofu 400-8511, Japan
4Navaminda Kasatriyadhiraj Royal Air Force Academy, Bangkok 10220, Thailand

Corresponding authors: Khomdet Phapatanaburi (khomdet.ph@rmuti.ac.th) and Peerapong Uthansakul (uthansakul@g.sut.ac.th)

This work was supported in part by the Suranaree University of Technology (SUT), in part by the Thailand Science Research and Innovation (TRSI), and in part by the National Science Research and Innovation Fund (NSRF) through NRIIS under Grant 160361.

ABSTRACT Recent studies have reported the success of multi-scale convolution neural network (MSConvNet) model for many classification applications due to its powerful ability of exploring multi-scale convolution block to extract multi-scale representations to make a detection. However, a new design based on MSConvNet for pathological voice detection has not been explored. In this paper, we propose RS-MSConvNet, a novel end-to-end MSConvNet model using raw speech for pathological voice detection. The main contribution of the proposed RS-MSConvNet method is to exploit the multi-scale convolution block, followed by spatial-temporal feature block, and fully connected layer as classification. In addition, to further improve accuracy performance, we propose a novel hybrid detection model by integrating the feature extraction ability of the RS-MSConvNet model and the classifier of support vector machine (SVM) method, called RS-MSConvNet-SVM model. The effectiveness of our proposed models is investigated using the TORGO database. The experimental results reveal that the RS-MSConvNet model outperforms other baseline methods in the speaker-independent task. Moreover and as compared to the RS-MSConvNet-SVM model, a further improved accuracy is obtained using the RS-MSConvNet-SVM model. These outcomes exhibit that our proposed models are useful for pathological voice detection.

INDEX TERMS Pathological voice detection, end-to-end architecture, multi-scale convolution, spatial-temporal feature, hybrid model.

I. INTRODUCTION

Pathological voice detection is a technique of determining pathological voice or healthy voice from a provided utterance signal. It plays an important role in voice healthcare systems [1] such as voice clinics [2] and telemonitoring application [3], [4], [5] because the detection of changed speech is a diagnostic tool to identify the onset of disabling physical symptoms [6], where the results are exploited to screen patients at risk of having certain diseases. Moreover, the pathological voice detection is an essential pre-processing step for automatic speaker recognition for dysphonic voice assessment [7] and dysarthric speech recognition [8]. In this study, we focus on a pathological voice detection, which is a subject area of the pattern recognition task in the field of biomedical and health informatics.

Typical pathological voice detection system can be divided into two groups: traditional pipeline system [9] and modern end-to-end system [10]. In the earlier studies [11], the systems usually consist of the front-end feature extraction and the back-end classifier. Based on traditional pipeline systems, the handcrafted design feature extraction converts speech signal into parametric representation while the back-end classifier learn feature representation for predicting
pathological/healthy voice class. For modern end-to-end systems that can extract features without using the handcrafted design feature extraction, a deep learning-based classifier used for predicting target classes is learned using a raw speech or its spectrogram. A brief survey based on existing traditional pipeline systems and modern end-to-end systems approaches for pathological voice detection is reviewed below.

Based on the traditional pipeline approach, most existing studies in pathological voice detection have focused on exploring effective hand-crafted design feature extraction with effective classifiers. Researchers have introduced various feature extraction methods for pathological voice detection. Mel-frequency cepstral coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), Linear Prediction Coefficients (LPC), Multi-Dimensional Voice Program (MDVP)-based features were proposed in [12]. Harmonics-to-noise ratio [13], Jitter [13], Shimmer [14], Kullback-Leibler divergence (KLD) histogram [15] and KLD higher amplitude suppression spectrum [15] were proposed for pathological voice detection. Autocorrelation and entropy features in different frequency regions were proposed in [16].

In addition to using the above mentioned individual features, the openSMILE set and Glottal source set-based fusion feature were introduced in order to combine acoustic features with statistical function sets or combine frequency-domain glottal and time-domain glottal feature sets with statistical function sets, respectively. Moreover, the combination of the openSMILE set and Glottal source set-based feature fusion was introduced [17] to fuse two merits based on different features. For the classifier, support vector machine (SVM) have been utilized as popular classifier in most previous studies [18], [19], [20], [21] because it can provide promising result for pathological voice detection. In addition to using SVM, researchers have applied various classifiers such as artificial neural networks [22], [23], linear discriminant analysis [24], Gaussian Mixture Model [25], and decision trees [26]. In all these traditional pipeline approaches, the ability of detecting pathological speech from healthy speech is strongly dependent on the effectiveness of effective handcrafted design feature extraction. This suggests that the detection performance requires expert knowledge in speech processing to devise relevant features.

Regarding pathological voice detection using end-to-end systems, previous works [11], [27], [28] have shown that they do not require expert feature engineering because deep learning models can be trained using either raw speech signal or its spectrum. For example, the combinations of convolutional neural network and multilayer perceptron (CNN-MLP)/long short-term memory (CNN-LSTM) using raw speech signal were proposed in [17]. The results showed that the CNN-MLP and CNN-LSTM could provide good results for pathological voice detection. However, using raw speech signal with any modification was not efficient enough as an input for training the end-to-end model under small training data. To further improve the end-to-end CNN-MLP and CNN-LSTM, the authors of [11] proposed to use glottal flow signal to replace raw speech signal as the input. The results showed that end-to-end CNN-LSTM and CNN-LSTM using glottal flow signal performed better than conventional CNN-MLP and CNN-LSTM using raw speech signal. Even though the end-to-end CNN-MLP and CNN-LSTM using either raw speech or glottal source signals could provide encouraging results, there is still an open research subject to design a new end-to-end model for pathological voice detection.

In this paper, a new end-to-end multi-scale convolution neural network architecture using raw speech, RS-MSConvNet is proposed for pathological voice detection. The main idea of the proposed architecture is to exploit multi-scale convolution block to scale the input information into different scaled representation, followed by spatial-temporal feature block and fully connected (FC) layer as a classifier block. In addition, to further improve detection accuracy, we propose a hybrid of RS-MSConvNet and SVM (RS-MSConvNet-SVM) models. Here, SVM classifier was explored to learn the automatically extracted features derived from fully trained RS-MSConvNet model. RS-MSConvNet-SVM model provide promising results for speaker-independent pathological voice detection.

The contributions of this article can be summarized as follows:

1) A novel end-to-end model architecture, RS-MSConvNet is proposed to learn raw speech. The proposed RS-MSConvNet architecture which is end-to-end does not require expert knowledge in feature engineering.

2) We investigate our model on TORGO dataset. Here, the proposed RS-MSConvNet model performs comparably to other baseline systems in a speaker-independent approach.

3) The RS-MSConvNet is also modified using SVM method to replace FC layer as classification to learn the automatically extracted features derived from fully trained RS-MSConvNet model. The modified RS-MSConvNet is referred to as a hybrid RS-MSConvNet-SVM model architecture. The RS-MSConvNet-SVM model provides improved accuracy result, compared to the RS-MSConvNet classifier.

The rest of this paper is organized as follows: Our proposed methods are introduced in Section II. Section III describes pathological voice detection setup including the details of the database, network training, baseline method, and experimental evaluation. In Section IV, the results and discussions are presented. Section V presents our conclusion.

II. PROPOSED METHOD

A. RS-MSConvNet

The proposed RS-MSConvNet framework for pathological voice detection consists of pre-processing block, multi-scale
convolution block, spatial-temporal feature extraction block, and classifier block as shown by the flowchart in Fig. 1. The configuration of the proposed RS-MSConvNet model is summarized in Table 1.

### TABLE 1. Configuration of RS-MSConvNet architecture, where \((N, T)\) are the dimension of input representation and \(k\) denotes the order of layer in block \(b\).

| Block | Layer | Kernel | Output | Activation | Parameters |
|-------|-------|--------|--------|------------|------------|
| b     | Input | \((2, 2)\) stride = 2 | \((1, N, T)\) | Linear | 5 |
|       | Conv2D | \((2, 2)\) stride = 2 | \((1, N/2, T)\) | Linear | 5 |
|       | Conv2D | \((2, 2)\) stride = 2 | \((1, N/4, T)\) | Linear | 5 |
|       | Conv2D | \((2, 2)\) stride = 2 | \((1, N/8, T)\) | Linear | 5 |
|       | Conv2D | \((2, 2)\) stride = 2 | \((1, N^{16}/16, T)\) | Linear | 5 |
| c     | Input | \(32 \times (\frac{N}{2T}, 1)\) | \((1, N/2, T)\) | Relu | 416 |
|       | Activation | | \((32, 1, T)\) | | |
|       | Conv2D | \(16 \times (1, 4)\) | \((16, 1, T/2)\) | Relu | 2064 |
|       | Activation | | \((16, 1, T/4)\) | | |
|       | Activation | | \((16, 1, T/8)\) | Dropout | |
|       | Global average pooling | | 16 | | |
| d     | Input | | 48 | Relu | |
|       | FC | | 128 | Linear | |
|       | FC | | 2 | | |
|       | Classifier | | 1 | Log Softmax | |

1) **PRE-PROCESSING BLOCK**

This subsection describes how to form input data for training the RS-MSConvNet model. Pre-emphasis is initially employed to compensate the high-frequency component of the input speech signal. Next, the framing operations is used. In this paper, a 20 ms frame length and 10 ms frameshift of raw speech signals were divided into several speech frames and then a Hamming window is applied to enhance the harmonics and smooth the edges of the framed speech signals. Finally, a 2 dimension (2D) input data for training the proposed RS-MSConvNet model is formed by stacking the speech frames.

2) **MULTI-SCALE CONVOLUTION BLOCK**

Motivated by [29], [30], and [31], the feature pyramid networks based on Multi-scale convolution block has been proven to be effective feature technique in the built-up areas detection using synthetic aperture radar images [32] and electroencephalography seizure detection [33] because Multi-scale convolution block can extract multi-scale semantics information and make a more precise prediction by mean...
of gathering more robust semantics information of scaled features. In this paper, Multi-scale convolution block was implemented to extract the 2D-input data into multi-scale features, making the designed detection model able to learn objects across a large range of scales. Each block was designed to extract the input fixed information in half the scale of the previous layer’s resolution. The block could automatically learn the weight to effectively distinguish the valuable level features while reducing the signal to half.

In this block, the input data are shaped as \((C, N, T)\) where \(C\), \(N\) and \(T\) are defined as the number of channel, number of frames, and number of samples in each frame, respectively. Next, as seen in block (b), each 2D-convolution layer performs a convolution to reduce the input size with the same kernel size of \((2, 2)\), a stride of 2, and no padding. By this configuration, the number of rows (frames) and columns (its samples) from \(k\) convolution layer become half of \(k - 1\) convolution layer providing the output size from the \(k\) layer to be \((1, \frac{N}{2^k}, \frac{T}{2^k})\). Here, outputs from the second to forth layers are used as the input for the next block for further extraction of spatial and temporal representation based on different field of views.

3) SPATIAL-TEMPORAL FEATURE EXTRACTION BLOCK

In this block, the objective is to extract spatial-temporal features from each output’s scale of the Multi-scale convolution block. Here, two regular 2D-convolution layers are used to capture three last scaled output of the block. For the first 2D-convolution layers, three kernel sizes of \((\frac{N}{2}, 1)\), \((\frac{N}{4}, 1)\), \((\frac{N}{8}, 1)\) with a stride of 2, no padding, and output channels of 32 are used to capture the outputs from the second layer, third layer and forth layer of previous phase, respectively. For the second 2D-convolution layers, the same kernel size of \((1,4)\) with a stride of 2, no padding, and channel outputs of 16 are used to extract the different output of the first layers. Finally, global average pooling is applied to the output passed from the two regular 2D-convolution layers. By this configuration, a total of 48 features (16 representations per each scale) are obtained to be the input for next block.

4) FULLY CONNECTED LAYER BLOCK

After two convolutional layers and global average pooling, the spatial feature module \(s’\) outputs of different scales are augmented and fed to FC layers. Log softmax is used as our last layer for predicting binary classes. Based on Log softmax, the logarithm of the prediction probability of binary classes is computed as follows:

\[
\hat{y}(x_i) = \log\left(\frac{\exp(x_i)}{\sum_j \exp(x_j)}\right),
\]

where \(x_i\) is the input vector with the \(i^{th}\) element and \(j\) is the number of classes (possible outcomes).

To calculate classification loss, cross entropy is implemented in this study. This loss function calculates the similarity between the label and predicted probability values as follows:

\[
\mathcal{L}_{CE}(y, \hat{y}) = -\sum_{k=1}^{\text{class}} y_k \log \hat{y}_k.
\]

where \(\hat{y}\) and \(y\) are the predicted probability and the true label, respectively.

B. RS-MSConvNet-SVM

SVM-based method has been proven to be effective for both classification and regression problems, so it has been utilized for many applications [34] such as face recognition [35], electroencephalography seizure detection [36], and automatic emotion speech recognition [37]. Moreover, most previous studies have used SVM-based method as baseline classifier for pathological voice detection because it can deal with two-class classification problem. In this paper, the SVM-based method is applied to learn automatically optimized features based on the RS-MSConvNet model.

Motivated by [38] which integrated CNN as a trainable feature representation and SVM as a classifier, a hybrid CNN and SVM (CNN-SVM) provided better accuracy results than CNN model for tumor detection. This was attributed to the fact that the developed model combined the advantages of CNN and SVM models. Similarly, hybrid RS-MSConvNet and SVM (RS-MSConvNet-SVM) as shown in Fig. 2 is proposed by using SVM as the classification to replace the FC layers after the RS-MSConvNet classifier was fully trained as shown in Fig. 2 (a). In this paper, to construct the SVM in a hybrid model, we adopt radial basis function (RBF) and determined penalty parameter \(\mathcal{C}\) and the optimal kernel parameter \(\gamma\) by investigating the validation data on the hybrid model learned by training data. Both training and validation data are explained in next section.

The implementation process of the RS-MSConvNet-SVM model is shown in Fig. 2 (b) and can be summarized as follows:

1) For the training process, the samples of training set were fed to RS-MSConvNet model.

2) After the RS-MSConvNet classifier was fully trained, the corresponding feature information could be automatically extracted for each input map.

3) The FC layers were replaced with SVM-based classifier to learn the automatically extracted feature vectors derived from the fully trained RS-MSConvNet classifier.

4) For the test process, the samples of test set were fed to the fully trained RS-MSConvNet classifier to obtain the automatically extracted features as the test feature representation.

5) The test feature data was fed to the well-trained SVM for predicting healthy or pathological class.

III. EXPERIMENTAL SETUP

A. DATABASE

TORG0 [39] and UA-Speech [40] are commonly used databases for pathological voice detection. In this paper, TORG0 database is used to investigate our RS-MSConvNet
and RS-MSConvNet-SVM models. The main reason for using this database is that the database is more challenging than UA-Speech database due to the limited data, which makes the end-to-end models based on the TORGO database difficult to achieve better accuracy than the end-to-end models based on the UA-Speech database as seen in [17]. Moreover, the results are also directly compared with the experimental settings as in [17]. The publicly available TORGO corpus was produced by three females (F01, F03, F04) with dysarthria, three healthy females, five males (M01, M02, M03, M04, M05) with dysarthria, and five healthy males (MC01, MC02, MC03, MC04). In this database, participants without dysarthria recorded approximately 900 utterances on average while participants with dysarthria recorded approximately 400 utterances on average. Further details of the TORGO database can be seen in [39]. All speech utterances were sampled at 16 kHz. In this study, since a substantial amount of silence was contained in the TORGO database, it need to be removed for training/testing classification model.

FIGURE 2. Implementation process of the RS-MSConvNET-SVM model.

TABLE 2. Details of the TORGO database.

| Training  | Validation | Testing    |
|-----------|------------|------------|
| MC03, MC04 | MC02, FC01, FC03 | MC01, M01, M03 |
| FC02, M02, M05, F01, F03 | M01, M03 | M04, F04 |

TABLE 3. Model parameters of the RS-MSConvNet and RS-MSConvNet-SVM models.

| Parameters          | RS-MSConvNet | RS-MSConvNet-SVM |
|---------------------|--------------|------------------|
| Batch Size          | 256          | 256              |
| Learning Rate       | 0.0001       | 0.0001           |
| Dropout Rate        | 0.5          | 0.5              |
| Epoch               | 1000         | 1000             |
| SVM (C)             | N/A          | 1                |

1,491 pathological utterances with 3.5 hr), validation subset (944 healthy and 795 pathological utterances with 2 hr), and testing subset (2,087 healthy and 861 pathological utterances with 3 hr). Table. 2 summarizes all three subsets of TORGO database used for our experiments.

B. NETWORK TRAINING

In this paper, we used the PyTorch v1.10.1 framework to build the proposed method. NVIDIA RTX3090 with 24 GB memory was used to train the networks. Adam optimizer was exploited to optimize the loss function in each iteration of training process. The model parameters for training the RS-MSConvNet and RS-MSConvNet-SVM models are listed in Table. 3.

C. BASELINE SYSTEMS

Based on the same database and training/testing condition, the effectiveness of the proposed RS-MSConvNet
and RS-MSConvNet-SVM methods are compared with the results of five baseline system groups: OpenSMILE+SVM, Glottal+SVM, OpenSMILE-Glottal+SVM, and conventional end-to-end methods, modified end-to-end methods using glottal flow signal.

1) OpenSMILE+SVM methods: two acoustic feature extraction sets obtained using the OpenSMILE toolkit were used as the input feature information for the classifier. First OpenSMILE s' acoustic feature sets (OpenSMILE-1) with a total of 384 dimensions ((16 dimensions of the chosen acoustic features + 16 Δ dimensions) × 12 statistical functions) and second OpenSMILE s' acoustic feature sets (OpenSMILE-1) with a total of 6552 dimensions (56 dimensions of the chosen acoustic features + 56 Δ dimensions) × 39 statistical functions) were used as the input for the SVM classifier. The lists of OpenSMILE-1 set and OpenSMILE-2 set with its statistical information are summarized in Table. 4. For this baseline system, the SVM-based classifiers using the OpenSMILE-1 and OpenSMILE-2 sets are referred to as OpenSMILE-1+SVM and OpenSMILE-2+SVM, respectively.

2) Glottal+SVM methods: two glottal feature sets were used as the input feature for the SVM-based classifier. First glottal feature set (Glottal-1) with 192 feature vectors ((12 dimensions of the chosen acoustic features + 12 Δ dimensions) × 8 statistical functions) was obtained by capturing glottal flow signal using several time-and frequency-domain feature extraction method as listed in Table. 5. For second glottal feature set (Glottal-2), the principal component analysis (PCA) with 30 principal component weights was applied to the normalized Glottal-1 to calculate 480-dimension features ((30 dimensions of the chosen acoustic features + 30 Δ dimensions) × 8 statistical functions). For this baseline system, the SVM-based classifiers with the Glottal-1 and Glottal-2 sets are referred to as Glottal-1+SVM and Glottal-2+SVM, respectively.

3) OpenSMILE-Glottal+SVM methods: To take the advantages of two types of feature extraction sets mentioned above, the OpenSMILE-1/OpenSMILE-2 and Glottal-1/Glottal-2 sets were joined as the input for further improving the SVM-based classifier. Here, SVM using joint OpenSMILE-1 and Glottal-1 sets, joint OpenSMILE-1 and Glottal-2 sets, joint OpenSMILE-2 and Glottal-1 sets, and joint OpenSMILE-2 and Glottal-2 sets were referred to as OpenSMILE-1/Glottal-1+SVM, OpenSMILE-1/Glottal-2+SVM, OpenSMILE-2/Glottal-1+SVM, and OpenSMILE-2/Glottal-2+SVM, respectively.

4) Conventional end-to-end methods: the CNN-MLP and CNN-LSTM methods were used as baseline end-to-end method using raw speech.

5) Modified End-to-end methods using glottal flow: In similar way as conventional end-to-end methods, the CNN-MLP and CNN-LSTM methods were also used. Unlike conventional end-to-end methods, the glottal flow signals were used to replace raw speech signals as the input for CNN-MLP and CNN-LSTM-based classifier.

Further details of five baseline methods compared with our proposed methods can be seen in [17].

D. EXPERIMENTAL EVALUATION

In order to investigate the effectiveness of our proposed methods, three common evaluation criteria suggested in [42] are used: classification accuracy, sensitivity, and specificity. Accuracy is computed as:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}.
\]

where TP and TN are the true pathological voice and true healthy voice where the fully trained network correctly predicts the pathological voice and healthy voice classes. FP and FN are the true pathological voice and true healthy voice which are incorrectly classified.

The sensitivity and specificity are calculated as follows:

\[
\text{Sensitivity} = \frac{TP}{TP + FN}.
\]

\[
\text{Specificity} = \frac{TN}{TN + FP}.
\]

IV. RESULTS AND DISCUSSIONS

A. RESULTS ON RS-MSConvNet

This subsection reports the performance of RS-MSConvNet. The following conclusions based on varying the configuration of parameters can be drawn:

- Since the fixed-length segments has an effect on the performance of end-to-end network, it is important to find out suitable fixed-length segments. In this paper, different fixed-length segments of 240-ms, 250-ms, 500-ms, 1 s and 3 s were first investigated to find out the optimal fixed-length segment. Table. 6 reports the results of RS-MSConvNet model using different fixed-length segments.

It can be seen that the fixed-length segment of 500 ms provided the best performance compared with the others. The reasons were that the network could not sufficiently learn the overly short data of raw speech signals (240 or 250 ms), which might have not contained enough pathological information while a fixed-length speech segment longer than 500 ms could not be exploited due to the short durations of some of the vowels as summarized in [47]. Moreover, the fixed-length segment longer than 500 ms might have led to overly small training data making the fully trained network incompetent to detect pathological speech from healthy speech. The results indicate that the fixed-length segment of 500 ms (49 × 320 pixels) was the most suitable for the RS-MSConvNet model.

- Many trials, which is not reported in this section (data not shown), were conducted by adding/reducing the convolution layers and changing the parameters in Multi-scale convolution block and spatial-temporal feature extraction block but they did not achieve better results. Moreover, batch normalization was also applied to spatial-temporal feature extraction block and it could not improve the detection performance. This outcome means that changing Multi-scale
convolution block, spatial-temporal feature extraction block, and the parameter was unsuitable for our RS-MSConvNet method.

• Finally, since the number of FC layers has an effect on the detection performance of pathological voice, the numbers of FC layers was varied from 1 to 5 layers. Table. 3 shows the comparison among different FC layers. It was found that the detection performances decreased while using more than two layers. This is because one FC layer is suitable for the classification based on two classes and the limited training data suggested in [43] and [44]. This suggested that using one FC layer as classification was suitable for our RS-ConvNet model.

To visualize discriminating information using the scaled feature representation for pathological voice detection, a healthy voice and a pathological voice signal were chosen to be fed into a fully trained RS-ConvNet model. The output representation derived from second convolution layer to fourth layers were then compared to show discriminating feature information of healthy and pathological voices. In this paper, the representation images are displayed using the matplotlib function based on bilinear interpolation method [45]. Here, since the fixed-length segment of 500 ms which provided the best results mentioned above was used as the input for RS-MSConvNet, the output sizes of second layer, third layer, fourth layer are $10 \times 80$ pixels, $5 \times 40$ pixels, and $2 \times 20$ pixels, respectively. Fig. 4 shows a comparison of feature representation between healthy voice and a pathological voice signals with a similar amplitude signature. We can observe from Fig. 4 that the convolution layers provided different representation between healthy voice and pathological voice signals with a similar amplitude signature. This indicated that the proposed Multi-scale scale convolution block could give discriminative features for the pathological voice detection.

Next, to observe discriminating ability using the spatial-temporal features derived from the trained RS-MSConvNet

| Feature sets   | Acoustic features                                                                 | Statistical functionals                                                                 |
|---------------|----------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|
| openSMILE-1   | zero-crossing rate, RMS-energy, pitch, MFCCs (12 coefficients), voicing probability | min (or max) value and its relative position, range, median, kurtosis, skewness, standard deviation, 2 linear regression coeff. and quadratic error |
| openSMILE-2   | log-energy, zero-crossing rate, pitch, MFCCs (13 coefficients), Mel-spectrum (26 coefficients), jitter, shimmer, spectral flux, voicing probability, roll-off points, spectral centroid, position of spectral minimum and maximum | min (or max) value and its relative position, range, median, kurtosis, skewness, standard deviation, 2 linear regression coeff., linear and quadratic errors, 3 quartiles, 2 percentiles (95% & 98%), 3 inter-quartile errors, number of peaks, mean of peaks, mean distance between peaks, geometric, arithmetic and quadratic means |

| Time-domain glottal features | Frequency-domain glottal features | Statistical functionals |
|------------------------------|-----------------------------------|-------------------------|
| Amplitude quotient, Closing quotient, Speed quotient computed from the primary glottal opening, Speed quotient, computed from the secondary glottal opening, Normalized amplitude quotient, Open quotient, extracted from the LF model, Open quotient obtained from the primary glottal opening | Difference between the lowest two glottal harmonics, Harmonic richness factor, and Parabolic spectrum parameter | Skewness, Standard deviation, Kurtosis, Maximum, Minimum, Median, Mean, and Range |

| Input speech length | Accuracy(%) | Sensitivity(%) | Specificity(%) |
|--------------------|-------------|---------------|---------------|
| 240 ms             | 79.44       | 72.36         | 82.37         |
| 250 ms             | 82.19       | 71.89         | 86.44         |
| 500 ms             | 86.46       | 83.03         | 87.88         |
| 1s                 | 82.87       | 82.22         | 83.13         |
| 3s                 | 80.08       | 70.15         | 87.01         |
model for detecting pathological voices, the t-distributed stochastic neighbor embedding (t-SNE) [46] which is a commonly used method for dimensionality reduction was exploited to consider the distributions between healthy voice and pathological voice categories. Here, 200 pathological and 200 healthy voice samples were selected to show the distributions of the two classes based on the t-SNE analysis. Fig. 5 shows visual distribution of the spatial-temporal feature derived from the trained RS-MSConvNet model. As seen in Fig. 5 (a), the data distributions of the different classes using raw speech signals without any feature extraction were significantly overlapped. This caused difficulty in distinguishing the different voices. By comparing Fig. 5 (a) with (b), it can be seen that the data distribution of the proposed spatial-temporal feature performed better than using raw speech signals without any processing method because it provided clear contours and small inter-class distances. This suggested that the spatial-temporal feature based on the RS-MSConvNet could be useful for pathological voice detection.

B. RESULTS ON RS-MSConvNet-SVM

This subsection presents the results of RS-MSConvNet-SVM. Because the $\gamma$ value directly affects the SVM-based detection performance, it is important to find out the optimal $\gamma$. Here, the $\gamma$ values varied from 0.1 to 30 by the step size of 0.1, with the optimal $\gamma$ at 0.1 which provided the highest accuracy by investigating the validation set on the hybrid model trained by training set. Therefore, the fully trained RS-MSConvNet model using the optimal $\gamma$ at 0.1 was used to evaluate the testing set because the decision ranked by the distance between the RS-MSConvNet based-automatically extracted features and the hyperplane of the trained SVM model achieved the highest number of correct predictions. Fig. 6 shows the result of the RS-MSConvNet-SVM model compared with the results of the RS-MSConvNet model.

As seen in Fig. 6, improved accuracy performance was obtained using the hybrid model. The accuracy was improved from the RS-MSConvNet with 86.46% to the RS-MSConvNet-SVM with 87.61%. This can be attributed to the decision ranked by the distance between the RS-MSConvNet based-automatically extracted features and the hyperplane of the trained SVM model performing higher specificity compared with the RS-MSConvNet classifier, which led to directly improving the detection accuracy. This result indicated that the RS-MSConvNet-SVM seems useful for detecting pathological voice from healthy voice.

C. COMPARISON WITH BASELINE SYSTEMS

In this subsection, the performances of our proposed methods are compared to those of some known systems. As mentioned in the introduction section, some systems may not be discussed due to the experiments being based on speaker-dependent approach and different database from our experiments. Here, the results based only on the TORGO database for a speaker-independent approach and different database from our experiments, were compared. Table. 7 shows the results of some known systems compared to our proposed methods.
FIGURE 4. Comparison representation of raw speech and the outputs from second convolution layer to fourth convolution layer. Both are from patient #01 speaking “Zero”.

FIGURE 5. Visual distributions of the different features based on t-SNE. (a) raw speech signals, (b) spatio-temporal features.

TABLE 7. Comparison with baseline systems on the TORGO database.

| Systems (results in [17]) | Feature sets          | Classifier | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|---------------------------|-----------------------|------------|--------------|-----------------|-----------------|
| Conventional pipeline     | openSMILE-1           | SVM        | 78.24        | 72.94           | 83.54           |
|                           | openSMILE-2           | SVM        | 80.62        | 73.73           | 87.52           |
|                           | Glottal-1             | SVM        | 67.17        | 71.22           | 63.12           |
|                           | Glottal-2             | SVM        | 66.93        | 71.35           | 62.17           |
|                           | openSMILE-1 + Glottal-1| SVM     | 79.62        | 73.21           | 86.03           |
|                           | openSMILE-2 + Glottal-1| SVM     | **82.12**    | 79.02           | 85.22           |
|                           | openSMILE-1 + Glottal-2| SVM     | 80.63        | 72.59           | **88.68**       |
|                           | openSMILE-2 + Glottal-2| SVM     | 81.35        | 76.83           | 85.87           |
| Conventional end-to-end   | Raw speech            | CNN-MLP    | **78.83**    | 82.85           | **76.24**       |
|                           | Glottal flow          | CNN-MLP    | **81.12**    | **85.88**       | 75.26           |
|                           | Raw speech            | CNN-LSTM   | 71.15        | 78.45           | 66.17           |
|                           | Glottal flow          | CNN-LSTM   | 75.41        | 81.32           | 69.68           |
| Our proposed              | Raw speech            | RS-MSConvNet-PC | **86.46** | **83.04**       | **87.88**       |
|                           | Raw speech            | RS-MSConvNet-SVM (γ = 0.1) | **87.61** | **78.86**       | **91.23**       |

As seen in Table 7, the results obtained with the RS-MSConvNet and RS-MSConvNe models outperformed all known systems in terms of accuracy and specificity performances. For the specificity result, it was observed the end to end system approach (CNN-MLP) using glottal flow information performed better than the proposed methods because the glottal flow signal gave better discriminative information than raw speech signal as summarized in [47].
making the detection of pathological voice more specific. However, accuracy and specificity of the CNN-MLP using glottal flow information was worse than those of the proposed systems. This indicated that the proposed methods could give more reliable classification performance without requiring expert knowledge in pre-processing for computing alternative efficient signal to replace raw speech signal.

V. CONCLUSION
In this paper, we proposed a new RS-MSConvNet architecture for pathological voice detection. The main contribution of the proposed RS-MSConvNet method is to use Multi-scale convolution neural network, followed by spatial-temporal feature, and FC layer as classification. In addition, we proposed a hybrid model by integrating RS-MSConvNet as trainable feature presentation and support vector machine (SVM) as a classifier and referred to it as RS-MSConvNet-SVM model. The performances of our proposed models were evaluated using the TORG0 database. From the experimental results, it was observed that the RS-MSConvNet gave the discriminating feature information between healthy and pathological voice via the t-SNE method and provided an accuracy of 86.46 %, which outperformed other baseline systems. In addition, improved accuracy performance was obtained using RS-MSConvNet-SVM model. The accuracy was improved from the RS-MSConvNet with 86.46 % to the RS-MSConvNet with 87.61 %. The results indicated that our proposed RS-MSConvNet and RS-MSConvNet-SVM approaches could be useful for pathological voice detection.

In the future, the effectiveness of using the attention mechanism will be explored to further improve our proposed RS-MSConvNet and RS-MSConvNet-SVM approaches. We will also use Glottal flow signal [48] to replace raw speech signal as the input of our proposed methods.

REFERENCES
[1] Y. Wu, C. Zhou, Z. Fan, D. Wu, X. Zhang, and Z. Tao, “Investigation and evaluation of glottal flow waveform for voice pathology detection,” IEEE Access, vol. 9, pp. 30–44, 2021.
[2] C. Watters, B. Miller, M. Kelly, V. Burnay, Y. Karagama, and E. Chevretton, “Virtual voice clinics in the COVID-19 era: Have they been helpful?” Eur. Arch. Oto-Rhino-Laryngology, vol. 278, no. 10, pp. 4113–4118, Oct. 2021.
[3] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, “Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease,” IEEE Trans. Biomed. Eng., vol. 56, no. 4, pp. 1010–1022, Apr. 2009.
[4] C. G. Goetz, G. T. Stebbins, D. Wolff, W. DeLeeuw, H. Bronte-Stewart, R. Ebble, M. Hallett, J. Nutt, L. Ramig, T. Sanger, A. D. Wu, P. H. Kraus, L. M. Blasucci, E. A. Shamim, K. D. Sethi, J. Spielman, K. Kubota, A. S. Grove, E. Dishman, and C. B. Taylor, “Testing objective measures of motor impairment in early Parkinson’s disease: Feasibility study of an at-home testing device,” Movement Disorders, vol. 24, no. 4, pp. 551–556, Mar. 2009.
[5] P. Klumpp, T. Janu, T. Arias-Vergara, J. C. Vásquez-Correia, J. R. Orozco-Arroyave, and E. Nóth, “Apkinson—A mobile monitoring solution for Parkinson’s disease,” in Proc. Interspeech, Aug. 2017, pp. 1839–1843.
[6] A. E. Aronson and D. M. Bless, Clinical Voice Disorders. New York, NY, USA: Springer, 2009.
[7] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio, “Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia),” in Proc. Interspeech, Sep. 2005, pp. 149–152.
[8] Y. Lin, L. Wang, S. Li, J. Dang, and C. Ding, “Staged knowledge distillation for End-to-End dysarthric speech recognition and speech attributes transcription,” in Proc. Interspeech, Oct. 2020, pp. 4791–4795.
[9] N. P. Narendra, B. Schuller, and P. Alku, “The detection of Parkinson’s disease from speech using source voice information,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 29, pp. 1925–1993, 2021.

[10] M. Alhussein and G. Muhammad, “Voice pathology detection using deep learning on mobile healthcare framework,” IEEE Access, vol. 6, pp. 41034–41041, 2018.

[11] P. Harar, J. B. Alonso-Hernandez, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, “Voice pathology detection using deep learning: A preliminary study,” in Proc. Int. Conf. Workshop Bioinspired Intellect. (IWOBI), Jul. 2017, pp. 1–4.

[12] T. A. Mesallah, M. Farahat, K. H. Malki, M. Alsulaiman, Z. Ali, A. Al-Nasheri, and G. Muhammad, “Development of the Arabic voice pathology database and its evaluation by using speech features and machine learning algorithms,” J. Healthcare Eng., vol. 2017, pp. 1–13, Feb. 2017.

[13] B. Sabir, F. Rouda, Y. Khazri, B. Touri, and M. Moussetaud, “Improved algorithm for pathological and normal voices identification,” Int. J. Elect. Comput. Eng., vol. 7, no. 1, pp. 238–243, 2017.

[14] E. Vaciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, Z. Kons, A. Satt, and R. Hooij, “Fusion of voice signal information for detection of mild laryngeal pathology,” Appl. Soft Comput., vol. 18, pp. 91–103, May 2014.

[15] R. A. Barreira and L. L. Ling, “Kullback–Leibler divergence and sample skewness for pathological voice quality assessment,” Biomed. Signal Process. Control, vol. 57, Mar. 2020, Art. no. 101697.

[16] A. Al-Nasheri, “Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions,” IEEE Access, vol. 6, pp. 6961–6974, 2018.

[17] N. P. Narendra and P. Alku, “Glottal source information for pathological voice detection,” IEEE Access, vol. 8, pp. 67745–67755, 2020.

[18] M. K. Arjmandi and M. Pooyan, “An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine,” Biomed. Signal Process. Control, vol. 7, no. 1, pp. 3–19, Jan. 2012.

[19] V. Sellam and J. Jagadeesan, “Classification of normal and pathological voice using SVM and RFENN,” J. Signal Inf. Process., vol. 5, no. 1, pp. 1–7, 2014.

[20] Z. Ali, “Voice pathology detection based on the modified voice contour and SVM,” Biol. Inspired Cognit. Archae., vol. 15, pp. 10–18, Jan. 2016.

[21] S.-H. Fang, Y. Tsao, M.-J. Hsiao, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, “Detection of pathological voice using cepstrum vectors: A deep learning approach,” J. Voice, vol. 33, no. 5, pp. 634–641, Sep. 2019.

[22] R. T. Ritchings, M. McGillon, and C. J. Moore, “Pathological voice quality assessment using artificial neural networks,” Med. Eng. Phys., vol. 24, nos. 7–8, pp. 561–566, Sep. 2002.

[23] J. P. Teixeira, P. O. Fernandes, and N. Alves, “Vocal acoustic analysis—Classification of dysphonic voices with artificial neural networks,” Proc. Comput. Sci., vol. 121, pp. 19–26, Dec. 2017.

[24] P. Gómez-Vilda, R. Fernández-Baillo, V. Rodellar-Biarge, V. N. Lluis, J. P. Teixeira, P. O. Fernandes, and N. Alves, “Vocal acoustic analysis—Classification of dysphonic voices with artificial neural networks,” Proc. Comput. Sci., vol. 121, pp. 19–26, Dec. 2017.

[25] P. Harar, J. B. Alonso-Hernandez, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, “Voice pathology detection using deep learning: A preliminary study,” in Proc. Int. Conf. Workshop Bioinspired Intell. (IWOBI), Jul. 2017, pp. 1–4.

[26] B. Sabir, F. Rouda, Y. Khazri, B. Touri, and M. Moussetaud, “Improved algorithm for pathological and normal voices identification,” Int. J. Elect. Comput. Eng., vol. 7, no. 1, pp. 238–243, 2017.

[27] E. Vaciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, Z. Kons, A. Satt, and R. Hooij, “Fusion of voice signal information for detection of mild laryngeal pathology,” Appl. Soft Comput., vol. 18, pp. 91–103, May 2014.

[28] R. A. Barreira and L. L. Ling, “Kullback–Leibler divergence and sample skewness for pathological voice quality assessment,” Biomed. Signal Process. Control, vol. 57, Mar. 2020, Art. no. 101697.

[29] A. Al-Nasheri, “Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions,” IEEE Access, vol. 6, pp. 6961–6974, 2018.

[30] N. P. Narendra and P. Alku, “Glottal source information for pathological voice detection,” IEEE Access, vol. 8, pp. 67745–67755, 2020.

[31] M. K. Arjmandi and M. Pooyan, “An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine,” Biomed. Signal Process. Control, vol. 7, no. 1, pp. 3–19, Jan. 2012.

[32] V. Sellam and J. Jagadeesan, “Classification of normal and pathological voice using SVM and RFENN,” J. Signal Inf. Process., vol. 5, no. 1, pp. 1–7, 2014.

[33] Z. Ali, “Voice pathology detection based on the modified voice contour and SVM,” Biol. Inspired Cognit. Archae., vol. 15, pp. 10–18, Jan. 2016.

[34] S.-H. Fang, Y. Tsao, M.-J. Hsiao, Y.-J. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, “Detection of pathological voice using cepstrum vectors: A deep learning approach,” J. Voice, vol. 33, no. 5, pp. 634–641, Sep. 2019.

[35] R. T. Ritchings, M. McGillon, and C. J. Moore, “Pathological voice quality assessment using artificial neural networks,” Med. Eng. Phys., vol. 24, nos. 7–8, pp. 561–566, Sep. 2002.

[36] J. P. Teixeira, P. O. Fernandes, and N. Alves, “Vocal acoustic analysis—Classification of dysphonic voices with artificial neural networks,” Proc. Comput. Sci., vol. 121, pp. 19–26, Dec. 2017.

[37] P. Gómez-Vilda, R. Fernández-Baillo, V. Rodellar-Biarge, V. N. Lluis, A. Álvarez-Marquina, L. M. Mazaira-Fernández, R. Martínez-OLLalla, and J. I. Godino-Llorente, “Glottal source biometrical signature for voice pathology detection,” Speech Commun., vol. 51, no. 9, pp. 759–781, Sep. 2009.

[38] I. M. M. El Emary, M. Fezari, and F. Amara, “Towards a comparative database of dysarthric articulation,” in Proc. Int. Conf. Int. J. Commun. Technol. Electron., vol. 18, no. 8, pp. 5547–5557, Aug. 2022.

[39] K. Phapatanaburi, L. Wang, Z. Oo, W. Li, S. Nakagawa, and M. Iwashashi, “Noise robust voice activity detection using joint phase and magnitude based feature enhancement,” J. Ambient Intell. Humanized Comput., vol. 8, no. 6, pp. 845–859, Nov. 2017.

[40] L. Wang, K. Phapatanaburi, Z. Go, S. Nakagawa, M. Iwashashi, and J. Dong, “Phase aware deep neural network for noise robust voice activity detection,” in Proc. ICME, 2017, pp. 1087–1092.

[41] J. D. Hunter, “Matplotlib: A 2D graphics environment,” Comput. Sci. Eng., vol. 9, no. 3, pp. 90–95, May–Jun. 2007, doi: 10.1109/MCSE.2007.55.

[42] L. Van der Maaten and G. Hinton, “Visualizing data using T-SNE,” J. Mach. Learn. Res., vol. 8, no. 11, p. 728, 2008.

[43] S. Tirronen, S. R. Kadiri, and P. Alku, “The effect of the MFCC frame size and window length on the detection of built-up areas in high-resolution SAR images,” in Proc. ICASSP, Toronto, ON, Canada, Jun. 2021, pp. 7328–7332.

[44] K. Phapatanaburi, W. Pathonsuwan, L. Wang, P. Anchuen, T. Jumphoo, P. Buayai, M. Uthansakul, and P. Uthansakul, “Whispered speech detection using glottal flow-based features,” Symmetry, vol. 14, no. 4, pp. 777, Apr. 2022.

WONGSAVONG PATHONSUWAN received the B.E. degree in telecommunication engineering and the M.E. degree in telecommunication and computer engineering from the Suan Sunandha Rajabhat University, Thailand, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree in telecommunication and computer. His research interests include wireless communication, artificial intelligence (AI), machine learning (ML), and artificial neural networks (ANN).
KHOMDET PHAPATANABURI received the B.E. degree in electronic and telecommunication engineering and the M.E. degree in electrical engineering from the Rajamangala University of Technology Thanyaburi (RMUTT), Thailand, in 2010 and 2012, respectively, and the Dr.Eng. degree in information science and control engineering from the Nagaoka University of Technology (NUT), Japan, in 2017.

In 2018, he joined the Department of Telecommunication Engineering, Rajamangala University of Technology Isan, as a Lecturer, and became an Assistant Professor, in 2020. His main research interest includes audio and brainwave classification. During his study in Japan, he received the Mombukagakusho (MEXT) Scholarship, from 2014 to 2017. He is a Reviewer for several international journals including IEEE SIGNAL PROCESSING LETTERS, Computer Speech and Language, APSIPA Transactions on Signal and Information Processing, and IEEE ACCESS.

PRAWIT BUAYAI received the B.E. and M.E. degrees in computer engineering from Khon Kaen University (KKU), Thailand, in 2013 and 2016, respectively, and the Ph.D. degree in computer engineering from the University of Yamanashi, Japan, in 2022.

He is currently a Specially Appointed Assistant Professor with the Department of Computer Science and Engineering, University of Yamanashi, and an AI and the IoT Engineer at Mirapro Company Ltd., Japan. His main research interest includes the application of artificial intelligence technology to improve accessibility and productivity in the agriculture and manufacturing.

TALIT JUMPHOO received the B.E. degree in telecommunication and electronic engineering and the Ph.D. degree in telecommunication engineering from the Suranaree University of Technology, Thailand, in 2014 and 2022, respectively. He is currently a Postdoctoral Researcher at the School of Telecommunication Engineering, Institute of Engineering, Suranaree University of Technology. His research interests included biosignal processing, biomedical devices, brainwave classification, and applied machine learning.

PATIKORN ANCHUEN received the B.E. degree in telecommunication engineering from the King Mongkut’s Institute of Technology Ladkrabang, Thailand, in 1997, and the M.E. degree in electrical engineering from Chulalongkorn University, Thailand, in 1999, and the Ph.D. degree in information technology and electrical engineering from The University of Queensland, Australia, in 2007.

She is currently an Associate Professor with the Telecommunication School, Suranaree University of Technology. Her research interests include wideband/narrowband smart antennas, automatic switch beam antenna, DOA finder, microwave components, application of smart antenna, and advance wireless communications. She received the Second Prize of Young Scientist Award from 16th International Conference on Microwaves, Radar, and Wireless Communications, Poland, in 2006.

PEERAPONG UTHANSKUL (Member, IEEE) received the B.E. and M.E. degrees in electrical engineering from Chulalongkorn University, Bangkok, Thailand, in 1996 and 1998, respectively, and the Ph.D. degree in information technology and electrical engineering from The University of Queensland, Brisbane, QLD, Australia, in 2007.

From 1998 to 2001, he was employed as a telecommunication engineer at one of the leading telecommunication companies in Thailand. He is currently working as an Associate Professor and the Dean of the Research Department, Suranaree University of Technology, Nakhon Ratchasima, Thailand. He has more than 100 research publications and the author/coauthor of various books related to MIMO technologies. His research interests include green communications, wave propagation modeling, MIMO, massive MIMO, brain wave engineering, OFDM and advanced wireless communications, wireless sensor networks, embedded systems, the Internet of Things, and network security. He has won various national awards from the Government of Thailand for his contributions and motivation in the field of science and technology. Furthermore, he is the Editor of Suranaree Journal of Science and Technology and other leading Thailand journals related to science and technology.