IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 4, 2020, accepted August 12, 2020, date of publication August 18, 2020, date of current version August 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3017462

End-to-End Speech Emotion Recognition With Gender Information

TING-WEI SUN, (Graduate Student Member, IEEE)
Graduate Institute of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan
e-mail: willy@access.ee.ntu.edu.tw

ABSTRACT Many works have focused on speech emotion recognition algorithms. However, most rely on the proper selection of speech acoustic features. In this paper, we propose a novel emotion recognition algorithm that does not rely on any speech acoustic features and combines speaker gender information. We aim to benefit from the rich information from speech raw data, without any artificial intervention. In general, speech emotion recognition systems require manual selection of appropriate traditional acoustic features as classifier input for emotion recognition. Utilizing deep learning algorithms, and the network automatically select important information from raw speech signal for the classification layer to accomplish emotion recognition. It can prevent the omission of emotion information that cannot be direct mathematically modeled as a speech acoustic characteristic. We also add speaker gender information to the proposed algorithm to further improve recognition accuracy. The proposed algorithm combines a Residual Convolutional Neural Network (R-CNN) and a gender information block. The raw speech data is sent to these two blocks simultaneously. The R-CNN network obtains the necessary emotional information from the speech data and classifies the emotional category. The proposed algorithm is evaluated on three public databases with different language systems. Experimental results show that the proposed algorithm has 5.6%, 7.3%, and 1.5%, respectively accuracy improvements in Mandarin, English, and German compared with existing highest-accuracy algorithms. In order to verify the generalization of the proposed algorithm, we use FAU and eNTERFACE databases, in these two independent databases, the proposed algorithm can also achieve 85.8% and 71.1% accuracy, respectively.

INDEX TERMS Affective computing, speech emotion recognition, gender classifier, deep-learning, interpretability of deep-learning.

I. INTRODUCTION

Speech is the most direct and natural method of communication between humans, and even between human and machine. However, we still cannot achieve natural interaction between humans and machines because the current machines cannot sufficiently understand the emotional status of humans. This discrepancy has led to a new research field: speech emotion recognition [1]. In this paper, we focus on the algorithm analyzing speech characteristics and apply deep learning (DL) algorithm to recognize underlying emotions.

Generally speaking, according to the emotion conceptualization, speech emotion recognition systems can be divided into two categories: continuous-label emotion recognition [2], and discrete-label emotion recognition. Early research on discrete speech emotion focused mainly on selecting speech acoustic features that can represent different emotions. Thus, many short- and long-term acoustic features combining static mathematics formulas have been proposed. The most popular approach is to extract a large number of features based on mathematic models at the utterance level. Then apply a statistical machine learning algorithm, such as a Support Vector Machine (SVM) [3]. However, it remains unclear which features are more effective in representing speech emotion, and it is challenging to leverage feature sets in different language systems. Recently, with the increasing interest in deep learning algorithm, it has been commonly applied for the automatic learning of useful features from speech utterances. Prior research used a Deep Neural Network (DNN) on top of statistical features to improve accuracy compared with statistical learning algorithms [4]. Fig 1(a) shows the block diagram of the conventional speech emotion recognition algorithm based on statistical machine learning, where a speech feature extraction follows speech utterance.

The associate editor coordinating the review of this manuscript and approving it for publication was Alma Y. Alanis.

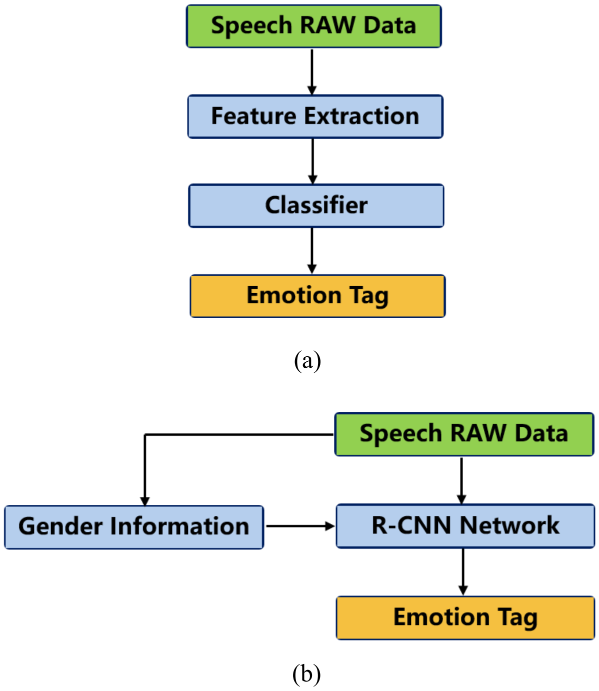


FIGURE 1. Block diagram of speech emotion recognition algorithms. (a) Conventional algorithm based on statistical machine learning. (b) The proposed algorithm.

These features could be based on either statistical acoustics or machine learning. A classifier algorithm is applied at the end to determine the emotion category. Although current existing speech emotion recognition algorithms achieve high recognition rates, two main issues still bother most researchers in speech emotion recognition. The first problem is how to balance the short-term characterization at the frame level and long-term aggregation at the utterance level [5]; that is, how to select proper speech features to represent corresponding emotion categories.

The second problem is that it is notoriously difficult to acquire a lot of useful labeled speech data since labeling data requires an expert's knowledge [6], which is prohibitively expensive and time-consuming in large quantities. In [5], it was found that in the same utterance, there are segments with strong emotional expression and segments without, such as a period of silence or friction sounds. Thus, a crucial gap exists in the ability to obtain reliably labeled speech utterances. Using raw speech data will reduce the effort in the traditional labeling method since we do not need to subdivision into every useful word of a speech utterance. The raw speech signal means the speech signal without any preprocessing, it theoretically contains the wealthiest information. Work on speech emotion sciences over recent decades has led to a proliferation of acoustic feature. Mostly based on established procedures in phonetics and speech sciences to measure different aspects of phonation and articulation in speech [12]. If we can use the raw speech signal without any acoustic feature selection, we can prevent any artificial intervention from omitting emotion information that cannot be mathematically modeled. Accordingly, raw speech data can be used to address both of the problems mentioned above.

Encouraged by the recent success of speech emotion recognition with deep learning algorithms and the idea of Convolutional Neural Networks (CNNs) [7], we propose a speech emotion recognition algorithm *that only requires the raw speech data for training.* Fig. 1(b) shows the block diagram of the proposed algorithm. The raw speech data is fed into both the CNN together with the gender information block. The CNN simultaneously obtains necessary information from the speech data and classifies the emotion category. This paper contributions are as follows:

1) A DISCRETE EMOTION RECOGNITION ALGORITHM WHICH COMBINES GENDER INFORMATION AND WITHOUT MANUAL FEATURE EXTRACTION

The proposed algorithm uses a CNN to endow with an ability to distill essential knowledge from the raw speech data. The learning algorithm can simultaneously ignore silence frames and other parts of the utterance that do not carry emotional content and focus on the more emotional regions of the speech signal. CNN learns the speech emotion information and predicts the emotion category in one joint block. Several studies have indicated the impact of gender information on the performance of speech emotion recognition [26], [27]. In the proposed algorithm, we consider this factor and develop a deep learning gender information algorithm, which improves the overall recognition accuracy rate. To our best knowledge, the proposed work is one of the early researches that uses only the raw speech signal and combines gender information to accomplish speech emotion recognition in a discrete speech emotion system.

2) RECORD-HIGH ACCURACY OF EMOTION RECOGNITION IN DIFFERENT LANGUAGE DATABASES

Experimental results show that we can achieve respective increases of 5.6%, 7.3%, and 1.5% in the accuracy of Chinese Mandarin, English, and German compared with existing highest-accuracy speech emotion recognition algorithms with different structures.

3) DISCUSS INTERPRETABILITY OF LEARNED FEATURES

Unlike traditional feature extraction algorithms that have well-defined physical meanings for each feature, deep learning algorithms cannot provide explanations for their predictions. Here, we demonstrate the interpretability of the proposed algorithm and how they predict the emotion categories. Based on the results of the Class Activation Map (CAM), we can understand how the network distinguishes emotions and allows us, humans, to judge whether they are reasonable.

The remainder of this paper is organized as follows: Section II describes related works focused on speech emotion recognition, and Section III details the proposed algorithm. Section IV demonstrates the experimental results on different language systems using three public emotional databases.

II. RELATED WORKS

Traditional speech emotion recognition can be divided into two steps. First performs feature extraction from the speech signal, and secondly applies a classifier to determine the emotional category. Some approaches add gender information to improve accuracy. With the rise of deep learning techniques, recent studies have proposed emerging end-to-end recognition structures. In this section, we review these related works.

A. FEATURE EXTRACTION AND EMOTION CLASSIFIER

Humans express emotion through communication, and a speech signal is the main container with all the necessary information to express emotion. Most studies in early speech emotion recognition research have focused on selecting the proper acoustic features. The nature of speech wave is a time-varying signal, where the speech signal contains emotion information, and the emotion clues are dispersed across both the time and frequency domains. Therefore, we cannot directly use the original speech signal, but must instead extract many short-term and long-term acoustic features. Then, we extract a large number of statistical features at each utterance level to represent the emotion information.

Generally speaking, we can use three categories of acoustic features to represent emotion information: prosody features and voice quality features, spectral features [8]. The prosody features can express the stress and intonation of the speaker. Loudness, intensity, duration, and pitch are most representative of speech prosody features [9]. Voice quality features can express the positive or negative emotions of the speaker [10]. The harmonics to noise ratio, pitch jitter, formats, glottal features [11], and spectral distribution are most representative of voice quality features. Spectral features are the original acoustic features considered in many speech applications; they can express ordinary emotion information from the view of human auditory perception. Mel-frequency cepstral coefficients, linear prediction cepstral coefficients, and log frequency power coefficients are the most characteristic spectral features, all of which are computed from the power spectrum of speech. In real applications of speech emotion recognition, we usually extract these three kinds of features, combine other features based on mathematical formulas such as Fourier parameters, and calculate a certain number of statistical models, such as mean and standard deviation. The resulting number array is called the low-level description of the emotional speech or statistical acoustic features and is used as the training data for the emotion classifier algorithm. The newly-developed Geneva minimalistic acoustic parameter set (GeMAPS) [12], which uses voice quality, prosody, and related spectral features, has performed well and provides a typical example of a speech emotion feature set.

Because the artificially selected features mentioned above are based on mathematical formulas and a static approach, they may not be distinctive enough to identify subjective emotions, mainly because humans' express emotions very subtly in speech signals. Recently, emerging deep learning techniques extract useful speech emotion information and features from speech signals. Even if the speech features have no physical meaning, they will theoretically improve the overall performance of the recognition system. This advantage is the original intention of the proposed work.

The purpose of the emotion classifier is to distinguish the emotion categories using the speech features. Almost all machine learning classifiers, such as K-Nearest-Neighbor (KNN) [13], have been considered for emotion recognition. Classifiers based on statistical recognition algorithms, such as Gaussian Mixture Model (GMM) [14], Hidden Markov Models (HMM) [15], and SVM, have been widely adopted for speech emotion recognition and have been shown to have excellent performance. Further, some works have integrated different classifiers to merge their advantages [16]. Recent studies have proposed deep learning classifiers, such as Long Short-Term Memory (LSTM), LSTM architectures are well-suited to framewise emotion recognition task [17].

B. END-TO-END SPEECH EMOTION RECOGNITION

The term ''end-to-end'' in the deep learning field means a complex learning system by applying gradient-based learning to the system as a whole [18]. That is, the network can directly convert the input signal into corresponding mapped output, bypassing the intermediate step in traditional algorithms, such as feature extraction in some emotion recognition systems. Therefore, we can consider the network a black box trained by the global objective function. The main idea is that the network automatically learns a representation of the raw input signal that better suits the task at hand, leading to improved performance. Prior research has applied end-to-end learning in speech emotion recognition for both continuous and discrete categories. For example, [19]–[21] all focus on continuous emotion recognition. Previous works considering continuous emotion recognition, share similar structures. The input of the raw speech signal passes through several convolution layers, followed by LSTM layers to capture the temporal structure. In [19], the authors proposed a solution to the problem of 'context-aware' emotional relevant feature extraction by combining CNN with LSTM in order to automatically learn the best representation of the speech signal directly from the raw time representation. In [20], the authors proposed a multimodal system that operates on the raw signal, to perform an end-to-end spontaneous emotion prediction task from speech and visual data. To consider the contextual information, LSTM was used. Besides, they study the gate activations of the recurrent layers in the speech modality and find cells that are highly correlated with prosodic features that were always assumed to cause arousal. In [21], the authors also proposed CNN, which extracts features from the raw signal and stacked on top of it a 2-layer LSTM, to consider the contextual information in the data. By contrast, [22]–[25] focus on discrete emotion recognition. The authors in [22] proposed an algorithm using CNN and LSTM to learn the emotion information from a two-dimensional spectrogram.

They use 13 Mel Frequency Cepstral Coefficient (MFCC) with 13 velocity and 13 acceleration components as spectrogram features. In [23], the authors directly used the image classification network in [7] to explore the speech spectrogram for different emotion categories. They proposed an acoustic feature representation, denoted as deep spectrum features, derived from feeding spectrograms through CNN and forming a feature vector from the activations of the last fully connected layer. In [24], the authors used a speech frequency spectrogram as an input signal, followed by convolution layers. They applied discriminant temporal pyramid matching to form a global utterance level feature representation and used SVM as the classifier. In [25], the authors proposed a three-dimensional neural network composed using CNN and LSTM to learn discriminative features for emotion recognition, where the Mel-spectrogram with deltas and delta-deltas were used as inputs, and the classifier is SoftMax.

All current end-to-end discrete speech emotion recognition algorithms use the speech spectrogram as an input. That is, although current algorithms are an end-to-end structure, the algorithms model the speech emotion recognition task as an image classification problem. We diverge from this approach. We prefer, as much as possible, to use an end-to-end deep learning algorithm to capture the information clues from emotional speech. Therefore, the input of the proposed algorithm is the raw speech data, rather than a spectrogram.

C. GENDER INFORMATION IN SPEECH EMOTION RECOGNITION

Many factors impact the performance of the speech emotion recognition system. However, the acoustic character differences between humans are the main factor because they lead to a severe problem with classifier algorithms. Speech is a time-varying signal generated according to the vocal tract. Thus, the difference in the shape of the vocal tracts causes different acoustic characters of speech signals. The shape of vocal tracts is quite different between human genders. The variance in emotional features may be very different in the same emotion category and very small between different emotion categories. For example, the pitch value difference is relatively small in male speech, but the value difference is quite significant between male and female speech, leading to difficulties in speech emotion recognition. Human gender is a factor that causes an average physiological difference and can increase the overall precision of the recognition system. Since male and female express feelings in various manners and have distinctive vocal systems. By including gender information, the preparation and testing information will be progressively reliable, and also the neural network will have one more piece to make the emotional vocal features of the two genders as a basic grouping. Therefore, it reduces the mutual influence of each other. Thus, we treat gender as an input feature in the proposed algorithm.

In past research of speech emotion recognition algorithms, the inclusion of speaker gender information has been shown to improve performance [26], [27]. In [26], the authors proposed a speech emotion recognition algorithm incorporating a gender classifier. The authors used pitch features to build up the gender recognition algorithm and aimed to provide a-priori information about the speaker's gender. They used a SVM as a classifier and gender information as input. The results showed that the gender-dependent system performed better than the gender-independent system. In [27], the authors proposed an emotion recognition algorithm based on gender information. They extracted gender features without any emotion information and emotion features without any gender information to build up emotion recognition. The gender-based denoising autoencoder is built using non-emotional speech to capture distinct information between the two genders. The hidden emotional representation is shared for the two genders in order to model more emotion specific characteristics and is used as features in a back-end classifier for emotion recognition. The experimental results also showed that this system performed better than the gender-independent system.

Table 1 reports the comparison between the proposed algorithm and current state-of-the-art speech emotion recognition algorithms. In discrete speech emotion recognition algorithms, the proposed work is the only one who does not need a feature extraction and to integrate gender information in a joint algorithm at the same time. Therefore, we proposed a concise and explicit algorithm, with lower design complexity and time consumption.

III. END-TO-END SPEECH EMOTION RECOGNITION

In this section, we detail the proposed speech emotion recognition algorithm. As shown in Fig. 2, the proposed algorithm combines a Residual Convolutional Neural Network (R-CNN) and a gender information block. The raw speech data is sent to these two blocks simultaneously. The R-CNN network obtains the necessary emotional information from the speech data and classifies the emotional category.

A. RAW SPEECH DATA

Unlike emotion recognition algorithms based on statistical machine learning, which extract a large number of mathematical features at each utterance level to use as input training data, we take advantage of deep learning and let the R-CNN block learn the speech emotional information and predict the emotion category in one joint block. Most acoustic speech features based on mathematical formulas selected by humans may not be discriminative enough to identify subjective emotions. Therefore, if we can use the raw speech signal without any feature selection, we prevent any artificial intervention from omitting emotional information.

B. RESIDUAL CONVOLUTIONAL NEURAL NETWORK

In this work, we have five different CNN blocks and two pooling strategies. Two extra convolution blocks perform the residual operation. In the bellowing paragraphs, we explain each block's working principle and mathematical formula.

TABLE 1. **Comparison between state-of-the-art works and proposed algorithm.**

	Feature Extraction	Gender Information	Emotion Label	Algorithm Structure
[8]	Yes (Multiple Features)	No	Discrete	Statistical Learning
[19]	No	No	Continuous	Deep Learning
[20]	No	No	Continuous	Deep Learning
[21]	No	No	Continuous	Deep Learning
[22]	Yes (Spectrogram)	No	Discrete	Deep Learning
[23]	Yes (Spectrogram)	No	Discrete	Deep Learning
[24]	Yes (Spectrogram)	No	Discrete	Deep Learning
[25]	Yes (Spectrogram)	No	Discrete	Deep Learning
[26]	Yes (Multiple Features)	Yes	Discrete	Statistical Learning
This work	**No**	**Yes**	**Discrete**	**Deep Learning**

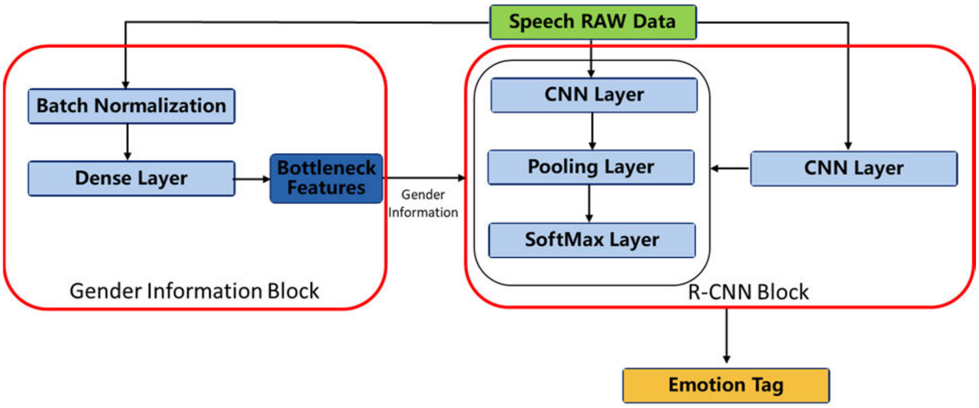


FIGURE 2. **Architecture of the proposed speech emotion recognition algorithm.**

1) CONVOLUTION LAYER

The convolution layer is the main fundamental block of a convolutional neural network. Each convolution layer is composed of a set of independent filters, which convolves, or combines, the previous layer's feature maps and then applies the activation function to form the output feature map. Each convolution layer extracts different local features at each local area of input data. Here, the filters perform dot products of the input of the previous convolution layers.

2) MAX POOLING LAYER

In the proposed algorithm, after a specific convolution layer, we use a max-pooling layer to down-sample the output data from the previous convolution layer. After max pooling, each local feature in each convolution area only has a single value. The max-pooling layer not only reduces the size of feature maps but also achieves spatial invariance for each feature map. The maximum value gives the output of the max-pooling layer over a particular window function.

3) GLOBAL AVERAGING POOLING LAYER

In the image classification task, AlexNet [7] and other similar CNNs, after the convolution and before the softmax layer, all need one fully-connected layer to integrate the outputs from previous convolution layers to yield the final feature representations for classification. A fully-connected layer has long been a standard component in the deep neural network.

Nevertheless, there is a significant issue with fully-connected layers: the feature size is too large, especially after the last convolution layer. A more significant feature size increases the computation loading of training and testing. On the other hand, larger feature sizes will also easily cause overfitting. In [28], the authors proposed a global averaging pooling layer to replace the fully-connected layer to address the above drawbacks. Fig. 3 shows the basic concept of the global averaging pooling operation. A fully-connected layer re-shapes each output channel of the convolution layer into a one-dimensional (1-D) vector, but a global averaging pooling

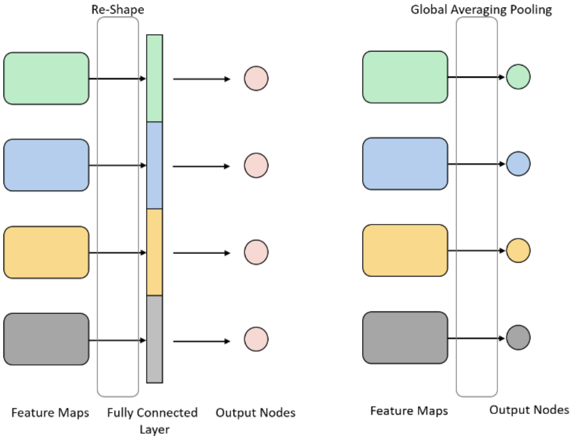


FIGURE 3. Fully-connected layer and global averaging layer.

layer averages all the output feature maps in each channel, meaning it outputs a single value instead of a 1-D vector. Because there is no parameter to optimize in the global average pooling, overfitting is avoided at this layer. Furthermore, global average pooling sums out the spatial information, so it is more robust to spatial translations of the input.

4) SOFTMAX LAYER

In order to complete the classification task in the deep neural network, the classification function is usually the softmax function, which specifies a discrete probability distribution for k classes. If we have $\mathbf{o}$ as the input of the softmax function, the softmax function is

$$\mathrm{p}_i = \frac{exp(\mathbf{o}_i)}{\sum_{i=0}^{K-1} exp(\mathbf{o}_i)}. \tag{1}$$

The final output classifier function is

$$\text{Category} = \underset{\mathrm{i}\in 1..\mathrm{K}}{\arg\max}\,(\mathrm{p}_i). \tag{2}$$

5) RESIDUAL CONNECTION

Residual learning [29], or a skip connection helps train the deep neural network. The residual connection can eliminate the vanishing and exploding gradient; the most important is the network degradation issue. Typically, we train a neural network by fitting a mapping relationship $\mathcal{L}(\mathbf{v})$ with the input signal $\mathbf{v}$. In the residual network, because of the residual connection, the mapping relationship becomes

$$\mathcal{T}(\mathbf{v}) = \mathcal{L}(\mathbf{v}) + \mathbf{v}, \tag{3}$$

where $\mathcal{T}(\mathbf{v})$ denotes the neural network output with the residual connection. Now, the original neural network $\mathcal{L}(\mathbf{v})$ learns the mapping $\mathcal{T}(\mathbf{v}) - \mathbf{v}$, called the residual mapping. If the network error becomes large in some layers, the residual network will turn into learn the input signal $\mathbf{v}$ directly. Thus, it is easier to optimize residual mapping than the original mapping.

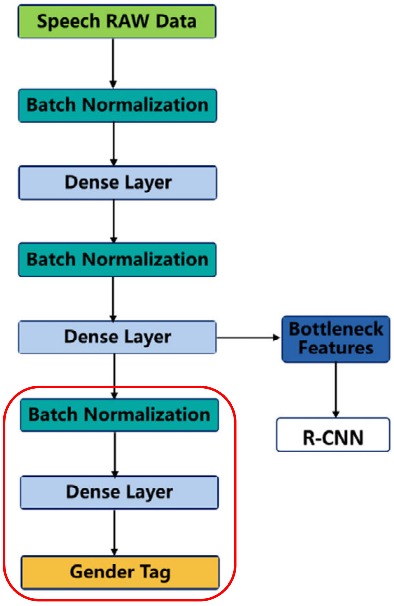


FIGURE 4. Detailed structure of the gender information network. The red mark part is only used for training.

C. GENDER INFORMATION NETWORK

Most current existing speech emotion recognition algorithms based on gender include gender information as a label in training or as a decision condition in the emotion recognition system. The proposed algorithm does not use the label directly but uses the gender information as an independent feature map to feed into CNN to train the whole recognition network. Fig. 4 illustrates the connections in the gender information network. Here, the dense layer is a regular layer in a neural network. Each node receives input from all the nodes in the previous layer, thus densely connected. The layer has a weight matrix, a bias vector, and the activations of previous layer.

1) BATCH NORMALIZATION

Batch normalization, which is a popular training technique for accelerating training for deep neural networks [30], can help improve the generalization and solve the internal covariate shift in deep neural training. The main idea is to use the mean and variance calculated from each batch to normalize the input of each hidden layer, and then linearly scale and shift the input after normalization. Mathematically, given a layer with output $\mathbf{z}$, batch normalization operation can be expressed as

$$BN\,(\mathbf{z}) = \gamma \frac{\mathbf{z} - \mathrm{E}[\mathbf{z}]}{(\mathrm{Var}[\mathbf{z}])^{1/2}} + \beta, \tag{4}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are the parameters to be learned.

2) BOTTLENECK LAYER

In the deep learning structure, if need to integrate two independent networks, and the dimension of the two networks

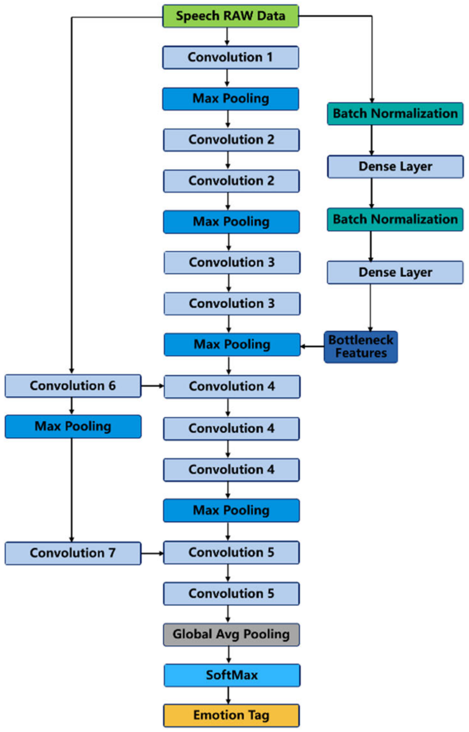


FIGURE 5. **Detailed structure of the speech emotion recognition network.**

are different then need the bottleneck layer. The bottleneck layer designed to smaller than the layers connect to it. The main purpose is to compress the feature map obtained by the forward layer and transfer it to the backward layer, like feature conversion between networks. In the proposed algorithm, the feature dimension of gender information is larger than the middle convolution layer. Therefore, we add a bottleneck layer to obtain a representation of gender information with reduced dimensionality and extract a new abstract representation of the gender information by taking the activations of the bottleneck layer.

D. TRAINING AND SETUP OF THE PROPOSED ALGORITHM

Fig. 5 shows the complete network of the proposed algorithm. In ordinary deep neural networks, the training and evaluation network are exactly the same. Nevertheless, we train the classifier and gender information network independently and combine them together into a final emotion classifier. We now detail the settings and working principles.

1) TRAINING THE GENDER INFORMATION NETWORK

In the training phase, we used the network shown in Fig. 4. We train gender information and emotion individually. First, we train the gender information network with binary output and cross-entropy loss. In the gender information network training, we select the Stochastic Gradient Descent (SGD) optimizer, which is commonly used in deep neural network training. The momentum factor is 0.9, and the weight decay value is 0.0005. Thus, the updated weight is calculated as

$$v_{j+1} = 0.9v_j - 0.0005\eta g_t, \tag{5}$$

$$w_{j+1} = w_j + v_{j+1}, \tag{6}$$

where v is the momentum value, η is the learning rate, j is the current iteration number, and g_t is the gradient of a given cost function. The entire network is updated by backpropagation.

2) TRAINING THE EMOTION CLASSIFIER NETWORK

After training the gender network, we use architecture presented in Fig. 5 to train the emotion classifier convolution network. Before training the classifier network, we disconnect the last two layers and the gender tag from the gender information network (the red mark part in Fig. 4). The weight and bias of the first four layers of gender information network are fixed. Therefore, gender information acts like independent feature maps to send to the emotion classifier network through the bottleneck layer. We train the emotional classifier network with Adam optimizer [31], which is a type of SGD optimizer that adaptively tunes the step size. Mathematically,

$$m_t = \mu m_{t-1}+(1-\mu)g_t, \tag{7}$$

$$n_t = r n_{t-1}+(1-r)g_t^2, \tag{8}$$

$$\widehat{m_t} = \frac{m_t}{1-\mu^t}, \tag{9}$$

$$\widehat{n_t} = \frac{n_t}{1-\nu^t}, \tag{10}$$

$$v_{j+1} = v_j - \eta\frac{\widehat{m_t}}{\sqrt{\widehat{n_t}+\epsilon}}, \tag{11}$$

where m_t and n_t are moving averages of the gradient and squared gradient, respectively, v is the momentum value, η is the learning rate. The values for μ, r, and ϵ are 0.9, 0.999, and 10^{-8}, respectively. We select softmax-cross-entropy as the loss function in the emotion classifier network. All the weights and bias in each layer are initialized without any pre-training. We use the initialization method in [32] to prevent gradient vanishing.

3) NETWORK SETUP

In the gender information network, the first dense layer node is 2400, the second dense layer is 1260, and the third dense layer is 610. The bottleneck features layer is a 1-D convolution with kernel size 1× 1. In the emotion classifier network, the first convolution layer is a 1-D convolution with kernel 1 × 80, the stride is 4, and channels 64. The second convolution layer is a 1-D convolution with kernel 1 × 3, the stride is 2, and channels are 64. The third convolution layer is a 1-D convolution with kernel 1 × 3, the stride is 2, and channels is 128. The fourth convolution layer is a 1-D convolution with kernel 1 × 3, the stride is 2, and channels is 256. The last convolution layer is a 1-D convolution with kernel 1 × 3, the stride is 2, and channels is 512. All the max-pooling layers are 1-D pooling with kernel size 8, and stride is 4.

IV. EXPERIMENTAL RESULTS

A. EMOTIONAL SPEECH DATABASES

First, we briefly introduce the databases used to evaluate the performance of the proposed algorithm. To confirm the proposed algorithm has generalization ability, we select three databases with different language systems.

1) DATASET IN CHINESE MANDARIN

CASIA [33], which was released by the Institute of Automation, Chinese Academy of Sciences, is composed of 9,600 wave files that represent six different emotional states: happiness, sadness, anger, surprise, fear, and neutral. Four professional native Chinese Mandarin speaking actors, two females, and two males simulated this set of emotions and produced 400 utterances in six classes of different emotions. Finally, ten people cross-checked the effectiveness of emotional utterances.

2) DATASET IN GERMAN

The Berlin Emotional Speech Database (EMODB) [34] was collected by the Institute of Communication Science at the Technical University of Berlin. Many researchers have used this database as a standard dataset for studying speech emotion recognition. EMODB comprises ten sentences that cover seven classes of emotion from everyday communication, namely, anger, fear, happiness, sadness, disgust, boredom, and neutral. EMODB contains 535 emotional utterances. Ten professional native German-speaking actors, five females and five males were asked to simulate these emotions. The emotions could be interpreted in all emotional contexts without semantic inconsistency. Finally, a human perception test with 20 other subjects was conducted to evaluate the quality of the recorded speech data.

3) DATASET IN ENGLISH

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [35] dataset was collected by the SAIL lab at USC. The corpus was organized into five sessions, each involving two actors in scripted scenarios or improvisations designed to elicit specific emotions. This database was recorded from 10 actors in dyadic sessions with markers on the face, head, and hands. The collected data were divided into small utterances of length between 3 and 15 seconds, which were labeled by evaluators. 3–4 assessors evaluated each utterance. The evaluation form contains ten emotions: neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excited, and other. The database contains approximately 12 hours of data.

B. EVALUATION METRICS

In the proposed work, we evaluate the performance by the Unweighted Averaged Recall (UAR), which is often used as the officially recommended measure for the speech emotion recognition challenge. UAR is the accuracy per class divided by the number of classes without consideration of instances per class. Thus, it better reflects the overall accuracy where there is unbalance among classes. Mathematically,

$$\text{UAR} = \frac{\textit{true positive}}{\textit{total positive}}, \tag{12}$$

where *true positive* is the number of correctly classified samples for a specified emotion category, and *total positive* is the actual number of utterances in the testing set with this emotion as ground truth. As suggested in [36] and many other previous emotion recognition works, test runs are carried out using the Leave-One-Speaker-Out (LOSO) or Leave-One-Speaker-Group-Out (LOSGO) strategies to ensure speaker independence, as required by most applications. In the case of fewer than ten speakers in one corpus, we apply the LOSO strategy. During each test loop, we leave one particular speaker out from dataset, and use the others as training data. We won't add that particular speaker into training. The speech data in each emotional database was normalized so that each vector dimension has a zero mean and unit variance; the size of the windows is 20ms, overlapping 10ms. All the speech signals were re-sampled into 16 kHz. In more detail, as for all the confusion matrices reported in this below section, the first row represents the recognized emotion while the first column contains the ground truth. For example, in Table 4, given anger as ground truth, the system predicts anger in 84.62% of the tests. Moreover, the mean value of the main diagonal of the matrix gives the average accuracy.

TABLE 2. Confusion Matrix (%) of the proposed algorithm with average UAR 84.6% in Chinese database CASIA. Each row presents the confusion of the ground truth emotion during prediction.

		True Classes					
		Ang	*Fear*	*Hap*	*Neu*	*Sad*	*Sur*
Predicted Classes	*Ang*	**94.74**	0	0	5.26	0	0
	Fear	0	**82.14**	3.58	0	10.71	3.58
	Hap	9.52	0	**76.19**	4.77	0	9.52
	Neu	0	0	7.69	**92.31**	0	0
	Sad	0	10	5	0	**85**	0
	Sur	7.7	0	15.38	0	0	**76.92**

Ang=Anger, Hap=Happy, Neu=Neutral, Sur=Surprise

C. EXPERIMENTAL RESULTS

Tables 2 to 4 list the confusion matrices resulting from the three different databases, respectively. Tables 7 to 9 compare the classification performance with previous state-of-the-art algorithms. Although the database is the same, the features, classifier type, and even algorithm structure are quite different. Thus, a performance comparison between works is not the best method to evaluate the performance of different algorithms. Nevertheless, it is still valuable to compare and analyze the results between the proposed and the other state-of-the-art methods to understand the progress and improvement of the proposed work. Here, we divide the state-of-the-art algorithms into three categories: algorithms based

TABLE 3. Confusion Matrix (%) of the proposed algorithm with average UAR 71.5% in English database IEMOCAP. Each row presents the confusion of the ground truth emotion during prediction.

		True Classes					
		Ang	*Exc*	*Fru*	*Hap*	*Neu*	*Sad*
Predicted Classes	*Ang*	**70.53**	22.08	0	0	0	7.39
	Exc	1.79	**69.64**	17.86	0	5.36	5.36
	Fru	0	21.28	**61.70**	0	17.02	0
	Hap	0	0	0	**72.5**	18.05	9.45
	Neu	0	7.69	13.19	1.10	**75.82**	2.20
	Sad	0	2.99	4.48	0	13.43	**79.10**

Ang=Anger, Exc=Excited, Fru=Frustrated, Hap=Happy, Neu=Neutral.

TABLE 4. Confusion Matrix (%) of the proposed algorithm with average UAR 90.3% in German database EMODB. Each row presents the confusion of the ground truth emotion during prediction.

		True Classes						
		Ang	*Bor*	*Dis*	*Anx*	*Hap*	*Sad*	*Neu*
Predicted Classes	*Ang*	**84.62**	0	0	0	15.38	0	0
	Bor	0	**100**	0	0	0	0	0
	Dis	0	0	**100**	0	0	0	0
	Anx	0	0	0	**77.78**	22.22	0	0
	Hap	10	0	0	0	**90**	0	0
	Sad	10	10	0	0	0	**80**	0
	Neu	0	0	0	0	0	0	**100**

Ang=Anger, Bor=Bored, Dis=Disgust, Anx=Anxiety, Hap=Happy, Neu=Neutral.

TABLE 5. Confusion Matrix of gender information.

	Male	Female
Male	**99.79%**	0.21%
Female	0.79%	**99.21%**

TABLE 6. Accuracy with and without gender information.

	CASIA	EMODB	IEMOCAP
With gender information	**84.6%**	**90.3%**	**71.5%**
Without gender information	81.4%	89.8%	69.9%

on statistical learning, algorithms based on deep learning, and algorithms based on gender information.

1) CONFUSION MATRIX IN THREE DATABASES

Tables 2 to 4 report the confusion matrix results for the three different databases. Most values in the EMODB confusion matrix are zero because the original emotion utterances were smaller than the other two databases, and the test samples may not be sufficient to cover all kinds of errors. The confusion matrix shows that it can distinguish different emotion categories with nearly uniform accuracy.

TABLE 7. Comparison of accuracy with statistical machine learning algorithm.

	CASIA	EMODB	IEMOCAP
This work	**84.6%**	**90.3%**	**71.5%**
[8]	79%	88.8%	–
[37]	–	85.1%	64.2%
[38]	–	80.1%	–
[39]	–	–	56.75%

TABLE 8. Comparison of accuracy with deep learning algorithm.

	CASIA	EMODB	IEMOCAP
This work	**84.6%**	**90.3%**	**71.5%**
[22]	–	88.01%	–
[24]	–	86.3%	–
[25]	–	82.82%	64.74%
[6]	–	63.6%	–

TABLE 9. Comparison of accuracy with algorithm based on gender information.

	CASIA	EMODB	IEMOCAP
This work	**84.6%**	**90.3%**	**71.5%**
[26]	–	81.5%	–

2) RESULTS WITH GENDER INFORMATION

Table 5 shows the confusion matrix of the single-gender information block. The speech characteristic between males and females are quite distinct. Therefore, the accuracy of gender information is very high.

Table 6 shows the accuracy comparison with and without gender information. In the experiment without gender information, we removed the bottleneck connection between gender information and the R-CNN block; that is, we trained the R-CNN block independently. The results show that gender information improves the overall performance in all three different language databases, thus confirming our assumption that gender information is a crucial factor in speech emotion recognition systems, regardless of the language system.

3) COMPARISON WITH STATISTICAL MACHINE LEARNING ALGORITHM

Table 7 presents four recent works with feature extraction. In [8], the authors proposed a new speech feature combined with an SVM classifier and evaluated it using the EMODB and CAISA databases. In [37], the authors proposed feature extraction in both vowel and non-vowel regions with extreme learning machine (ELM), which they evaluated with the EMODB and IEMOCAP databases. In [38], the authors proposed a new speech feature combined with an acoustic mask with a likelihood classifier, and they evaluated it using the EMODB database. In [39], the authors proposed a recognition system with both speech and lexical information with an SVM classifier. The proposed work preforms better with the three different language databases compared with the state-of-the-art works, regardless of the features or type of classifier. Notably, even if we remove the gender information from the proposed algorithm, the proposed work performs better with all three databases.

4) COMPARISON WITH DEEP LEARNING ALGORITHM

Table 8 lists four recent works based on deep learning algorithms, focusing solely on discrete emotion categories. In [22], the authors proposed an algorithm using a CNN and LSTM to learn the emotion information from a two-dimensional spectrogram, which they evaluated using the EMODB database. In [24], the authors proposed a CNN with temporal information and the SVM classifier, and they evaluated it with the EMODB database. In [25], the authors proposed an RNN combined with LSTM structure and a softmax layer as the emotion category output, and they evaluated it with both the EMODB and IEMOCAP databases. In [6], the authors proposed an autoencoder with a skip connection to directly classify the emotion category, and they evaluated using the EMODB database. The proposed work also outperforms these works for different language systems, regardless of a similar CNN or other deep learning structure. Even if we remove the gender information, the proposed algorithm still performs better in all three databases.

5) COMPARISON WITH ALGORITHM BASED ON GENDER INFORMATION

Table 9 shows a comparison of the proposed work with [26], which used gender information as a-priori information to design the algorithm, which was evaluated using the EMODB database. The authors used pitch features to build up the gender recognition algorithm, and they aimed to provide a-priori information about the speaker's gender. They used SVM as the classifier and gender information as an input. The proposed work performed significantly for the EMODB database.

6) COMPARISON BETWEEN RAW SPEECH AND SPECTROGRAM

Table 10 shows the accuracy comparison between different input signals. We train the proposed network with raw speech data and speech spectrogram individually. The speech signal is split into short frames with Hamming windows. Then, we calculate the power spectrum for each frame by using the discrete Fourier transform. The result in table 10 shows the accuracy with the raw speech signal is higher in both three databases, which validated our original intention. The emotion recognition algorithm benefits from the rich information of speech raw data.

TABLE 10. Comparison of accuracy between raw speech data and speech spectrogram.

	CASIA	EMODB	IEMOCAP
Raw speech	**84.6%**	**90.3%**	**71.5%**
Speech spectrogram	81.4%	88.2%	68.3%

TABLE 11. Confusion Matrix (%) of the proposed algorithm with average UAR 85.8% in German database FAU. Each row presents the confusion of the ground truth emotion during prediction.

		True Classes		
		Anger	*Neutral*	*Positive*
Predicted Classes	*Anger*	**85.8%**	7.9%	6.3%
	Neutral	5.7%	**87.4%**	6.9%
	Positive	8.6%	7.3%	**84.1%**

TABLE 12. Confusion Matrix (%) of the proposed algorithm with average UAR 71.1% in English database eNTERFACE. Each row presents the confusion of the ground truth emotion during prediction.

		True Classes					
		Ang	*Dis*	*Fear*	*Joy*	*Sad*	*Sur*
Predicted Classes	*Ang*	**78.61**	4.78	5.56	4.58	3.35	3.12
	Dis	5.35	**69.44**	6.25	9.21	5.58	4.17
	Fear	3.65	9.76	**65.65**	7.29	9.6	4.05
	Joy	8.98	5.31	5.79	**72.14**	5.25	2.53
	Sad	2.86	5.58	8.91	7.92	**65.49**	9.24
	Sur	4.85	3.63	5.83	4.91	5.66	**75.12**

Ang=Anger, Dis=Disgust, Sur=Surprise.

7) OVERALL RESULTS

The overall experimental results show that the proposed approach achieves accuracy improvements of 5.6%, 7.3%, and 1.5% for Mandarin, English, and German, respectively, compared with existing highest-accuracy speech emotion recognition algorithms with different structures. Note that the proposed algorithm still performs better than these state-of-the-art works even when we remove the gender information. Through these experiments with different language databases, the original intention is validated, i.e., we can use the raw speech signal without any feature selection to prevent artificial intervention from omitting emotional information.

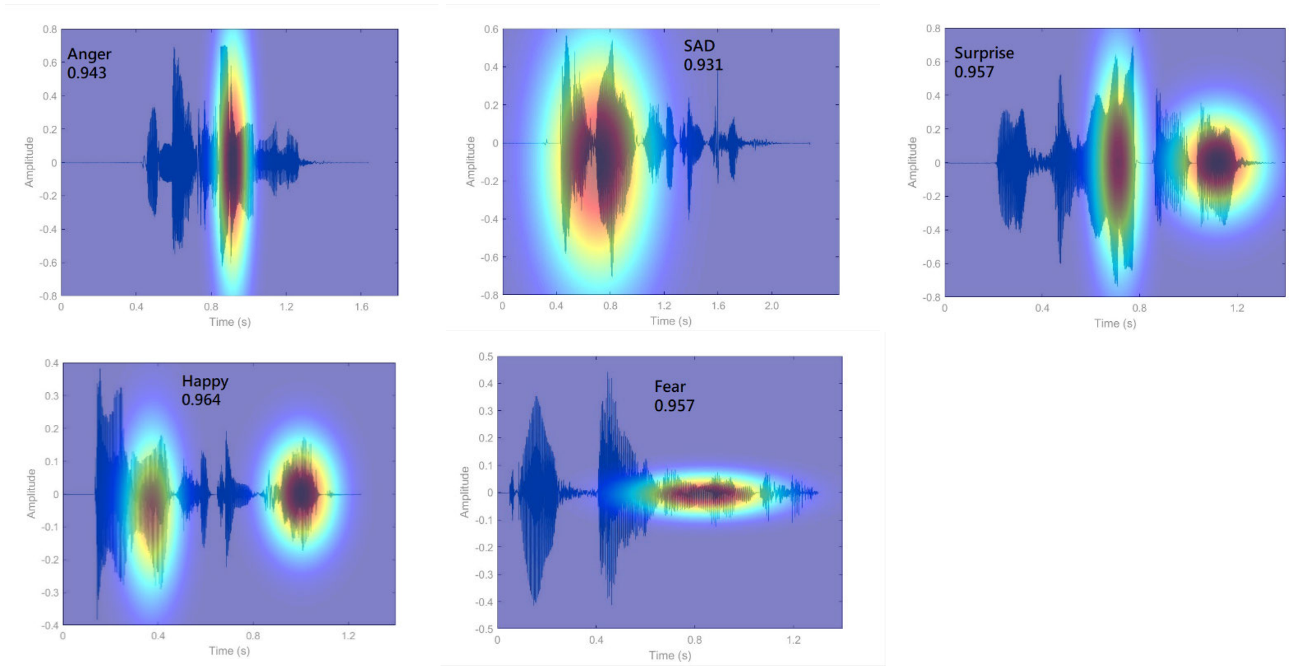


FIGURE 6. The CAMs of emotion recognition in CASIA database, with RAW speech signal. The maps highlight the discriminative image regions used for emotion classification.

TABLE 13. Accuracy based on different noise sources and level. Without train the network with noise sources.

	CASIA	EMODB	IEMOCAP
Without Noise	**84.6%**	**90.3%**	**71.5%**
Pink Noise SNR 20dB	82.1%	87.6%	70.1%
Crowd Noise SNR 20dB	81.3%	85.9%	69.6%
Footsteps Noise SNR 20dB	82.7%	83.5%	68.9%
Car Noise SNR 20dB	80.1%	82.5%	68.1%
Pink Noise SNR 10dB	80.3%	85.6%	66.4%
Crowd Noise SNR 10dB	65.6%	68.4%	55.1%
Footsteps Noise SNR 10dB	67.8%	69.6%	53.6%
Car Noise SNR 10dB	60.7%	60.1%	51.3%
Pink Noise SNR 0dB	71.7%	72.1%	53.7%
Crowd Noise SNR 0dB	45.0%	49.2%	41.4%
Footsteps Noise SNR 0dB	48.5%	58.1%	40.9%
Car Noise SNR 0dB	47.9%	51.5%	42.6%

D. GENERALIZATION OF PROPOSED ALGORITHM

The deep learning algorithm can learn models based on different numbers of training samples. The three databases we used to evaluate performance are relatively small compared to real-world data. The small training data set may lead to poor generalization of the model. Therefore, we use another database to verify the generalization. Table 11 shows the performance of the FAU database. FAU is a database of spontaneous emotion speech [40]. The database contains recordings of 51 German children, 21 males and 30 females between the ages of 10 – 13 years interacting with a pet robot. The database contains five different emotion classes: anger, emphatic, neutral, positive, and rest, and it contained approximately 9.2 hours of data. The children were told that

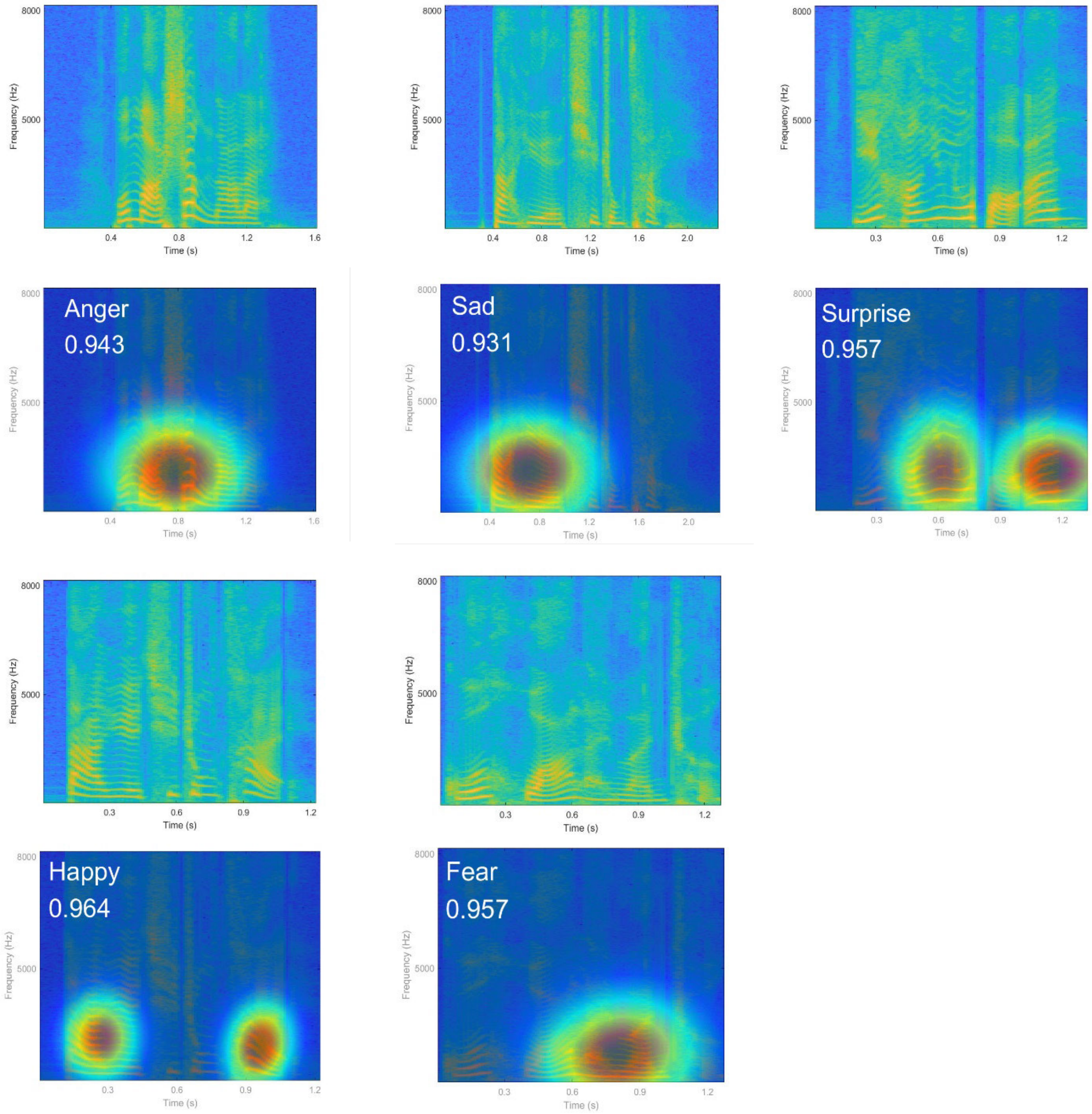


FIGURE 7. The CAMs of emotion recognition in CASIA database, with spectrogram speech signal. The maps highlight the discriminative image regions used for emotion classification.

the robot was responding to their voice commands regarding directions. However, the robot was controlled by a human operator to provoke strong emotional reactions from the children. Therefore, the FAU database can be considered as naturalistic emotional data. We use the FAU database as an independent test set; the training set used is the EMODB database. Because the emotion categories of EMODB and FAU are different, we only choose anger, neutral, and positive from the FAU database. The results show that the proposed algorithm still performs well even though the data do not exist in the training set, which validates the generalization ability. Since the FAU database is a children's emotional database, the pronunciation of children due to gender differences is very small. Here we also do an additional experiment that ignores the gender from the FAU database. The accuracy does not have a significant difference. It can be reasonably speculated that gender information is not very helpful in children.

In order to further verify the generalization, we retrained the IEMOCAP database and used all emotion labels. For the database used for evaluation, we chose eNTERFACE. The eNTERFACE [41] is an induced audio-visual emotion dataset with six basic emotions, *i.e.*, anger, disgust, fear, joy, sadness, and surprise. 42 subjects from 14 different nationalities are included. Each subject is asked to listen to six successive short stories, each of which is used to induce a particular emotion. Two experts are employed to evaluate whether the reaction unambiguously expresses the intended emotions. The speech utterances are pulled from video files

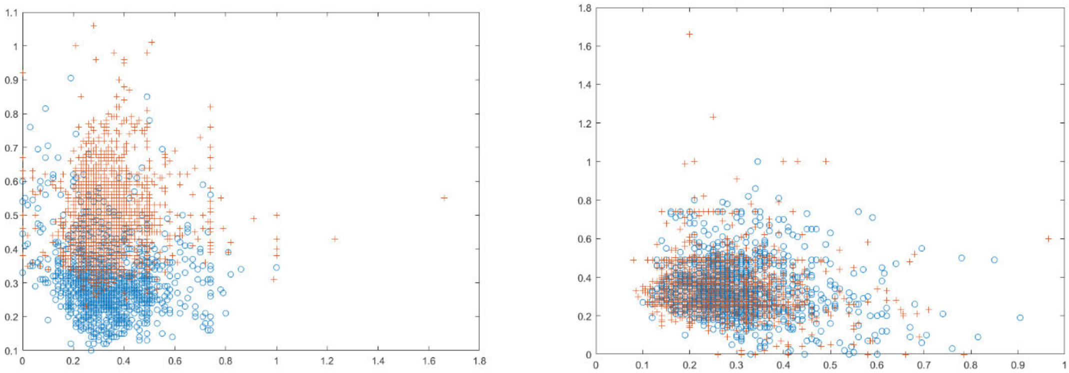


FIGURE 8. **(a) PCA data distribution of two emotion, with gender information. (b) PCA data distribution of two emotion, without gender information.**

of the subjects speaking in English. Here we did not train the network with the eNTERFACE database. Table 12 shows the results, even without training, the proposed algorithm can still achieve an average accuracy of 71% on this dataset. Fully illustrates the ability of the algorithm to generalization and realistic border non-acted emotion classes.

E. ALGORITHM PERFORMANCE UNDER NOISES CONDITIONS

The experimental results in section IV-C and D tested under noise-free condition. In order to make sure the performance will retain under noise. We evaluated the noise immunity performance here. We made the pink noise ourselves, and selected crowd, footsteps, car noise from [42]. The clean speech data is derived from the three databases we used to evaluate in section IV-C. All utterances from the training set of the databases were corrupted with the above mentioned four noise types at three levels of SNR, i.e., 20dB, 10dB, 0dB. Table 13 shows the accuracy under different SNR, and we did not train network with noise sources here. The accuracy shown in Table 13 represents the actual noise immunity. From the table, the proposed algorithm can still operate normally under SNR=20dB, no matter in stationary noise or non-stationary noise. Nevertheless, when the Signal to Noise Raito (SNR) becomes to worsen, the algorithm performance downgrades very quickly. When SNR=0dB, the proposed algorithm can hardly distinguish the emotion normally under the non-stationary noise environment. This result is not surprising; deep learning networks need to know what happens during the training phase. Adding noises only at the evaluation phase will not only bewilder the network but also destroy the speech signal at the same time. Table 14 presents the results of, we retrained the network on the same noise sources but only under SNR=20dB. As shown in the table, the overall performance has been improved. When SNR=20dB, which is the training condition, the accuracy is close to the first result in both four noises environment. When SNR=0dB, the accuracy improved by 20% to 30% under the non-stationary noise environment. The results imply the basic working principle about deep learning network – always let the network know what is happening in the training phase.

F. INTERPRETABILITY OF LEARNED FEATURES

The experimental results show that the proposed algorithm provided higher-accuracy predictions with a deep learning algorithm. Nevertheless, unlike traditional feature extraction algorithms that have well-defined physical meanings for each feature, deep learning algorithms cannot provide explanations for their predictions. Therefore, we try to provide some insight into the learned features. Most approaches to interpreting CNNs are based on the idea of highlighting relevant image aspects that contributed to a prediction. This is achieved by propagating the output signal back through the network to understand somehow what the output has encoded part of the input image. The fully connected layers in the CNN stand as black-boxes between the convolutional layers and the classifier, leading to loss of the spatial information of the image. The Class Activation Map (CAM) [43] algorithm averages each spatial unit across the last-convolutional layer feature maps, weighted by corresponding fully-connected layer weights for a particular class, gives its activation map. This map, when scaled to input image dimensions, highlights the discriminating image regions for that class. Fig. 6 shows the CAMs generated from the CASIA database, with raw speech signal input, Fig. 7 is the CAMs generated from the same database but with spectrogram input. The predicted class and its score are shown in the same figure. The experiment files are identical in Fig. 6 and 7. Therefore we can find hot zones located almost in the same period. Nevertheless, the hot zones appear narrower in raw signal and points to a more specific word from speech utterance. This phenomenon also explains why the accuracy is better in raw speech than the spectrogram. The algorithm can distinguish the emotion based on more useful and undisturbed information. From spectrogram figures, we can find that all of the hot zones are located in vowel regions of the speech utterance. Furthermore, according to the results of raw speech, we found that the hot zone is not only located in the vowel region

TABLE 14. Accuracy based on different noise sources and level. Train the network with noise sources, under SNR=20dB.

	CASIA	EMODB	IEMOCAP
Without Noise	**84.6%**	**90.3%**	**71.5%**
Pink Noise SNR 20dB	83.3%	89.8%	70.9%
Crowd Noise SNR 20dB	82.9%	89.1%	69.8%
Footsteps Noise SNR 20dB	83.1%	88.9%	70.3%
Car Noise SNR 20dB	83.5%	89.2%	69.5%
Pink Noise SNR 10dB	82.9%	89.5%	69.1%
Crowd Noise SNR 10dB	82.5%	88.9%	69.3%
Footsteps Noise SNR 10dB	82.8%	87.8%	67.9%
Car Noise SNR 10dB	82.4%	88.2%	68.2%
Pink Noise SNR 0dB	78.3%	80.1%	58.8%
Crowd Noise SNR 0dB	75.1%	77.6%	60.1%
Footsteps Noise SNR 0dB	73.6%	78.5%	59.5%
Car Noise SNR 0dB	72.1%	76.6%	58.7%

but a particular word in a sentence. Moreover, if we listen to the original speech files, the utterance hot zones indeed carry emotional contain, which indicates that the proposed deep learning network operates correctly. Previous studies [44], [45] also confirm that the vowel region will help the emotion recognition. The network is capable of considering the emotional content of different portions of speech.

V. CONCLUSION

In this paper, we proposed a novel emotion recognition algorithm that does not rely on any speech acoustic features and combines speaker gender information with the emerging R-CNN structure. While many previous works have focused on speech emotion recognition algorithms, most of them rely on the proper selection of speech acoustic features. The proposed work can benefit from the rich information of raw speech data without any artificial intervention, which prevents the omission of emotional information that cannot be direct mathematically modeled as a speech acoustic characteristic. We also add speaker gender information to further improve recognition accuracy. The proposed algorithm is evaluated on three public databases with different language systems. The experimental results show that the proposed algorithm provided significantly higher-accuracy predictions compared to existing speech emotion recognition algorithms in different language systems.

Although using a deep learning structure, we have excellent recognition performance, but this algorithm is still far from the real business application. Until now, the proposed algorithm has not been able to operate in real-time, which means we have to record complete speech utterance and then predict the emotion. Also, the computation power of the proposed algorithm is relatively large to be integrated on the mobile device, which also narrows down the application scenario.

In future work, we aim at incorporating more modalities, like image, in order to increase its accuracy. Also, we intend to training and experimenting with more conditions, including continuous emotion labels. It would also be interesting to explore the emotion recognition performance of different ages. We will also try to reduce the computation power of the current algorithm to fit the real application scenario.

APPENDIX

In order to visualize and better understand the impact of gender information, we did the following experiment:

1. Select anger and neutral categories from CASIA database, these two emotions are the easiest to distinguish.

2. Separately from the proposed algorithm, with and without gender information.

3. We pick the data before the softmax layer and use principal component analysis (PCA) to reduce the data dimension into two. The below Fig. 8(a) and Fig. 8(b) represent with and without gender information, respectively. We can clearly see that after adding gender information, the distribution of the data is clearly distinguished, which can significantly improve the accuracy.

Therefore, there are several facts that why the gender information can improve the accuracy of emotion recognition.

i) Previous researches indicate the acoustic features, and also physiologic are the difference between the male and female. [46, 47]

ii) Previous research points out that females are more emotionally expressive than males [48]

iii) Previous research demonstrated the improvement of emotion recognition accuracy among vocal cues in females compared with males. [49]

iv) Therefore, if gender information is added. We can make the emotional vocal features of the two genders as a basic grouping. Therefore, it reduces the mutual influence of each other. Just as our simple experiment shows. That improves overall accuracy.

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[2] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 97–108, Apr. 2015.

[3] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, pp. 577–580.

[4] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5688–5691.

[5] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2227–2231.

[6] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 26, no. 1, pp. 31–43, Jan. 2018.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[8] K. Wang, N. An, B. Nan Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 69–75, Jan. 2015.

[9] V. A. Petrushin, "Emotion recognition in speech signal: Experimental study, development, and application," in *Proc. 6th Int. Conf. Spoken Lang. Process.*, Beijing, China, 2000, pp. 222–225.

[10] R. Tato, R. Santos, R. Kompe, and J. M. Pardo, "Emotional space improves emotion recognition," in *Proc. Interspeech*, 2002, pp. 2029–2032.

[11] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falcão, "Spoken emotion recognition through optimum-path forest classification using glottal features," *Comput. Speech Lang.*, vol. 24, no. 3, pp. 445–460, Jul. 2010.

[12] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

[13] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. 4th Int. Conf. Spoken Lang. Process. (ICSLP)*, vol. 3, 1996, pp. 1970–1973.

[14] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm," in *Proc. IEEE Int. Conf. Multimedia Expo*, Amsterdam, The Netherlands, Jul. 2005, pp. 1500–1503.

[15] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[16] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Commun.*, vol. 49, no. 2, pp. 98–112, Feb. 2007.

[17] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 5, pp. 867–881, Oct. 2010.

[18] T. Glasmachers, "Limits of end-to-end learning," in *Proc. 9th Asian Conf. Mach. Learn.*, vol. 77, 2017, pp. 17–32.

[19] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5200–5204.

[20] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[21] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AL, Canada, Apr. 2018, pp. 5089–5093.

[22] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Jeju, South Korea, Dec. 2016, pp. 1–4.

[23] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proc. ACM Multimedia Conf. (MM)*, Mountain View, CA, USA, 2017, pp. 478–484.

[24] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.

[25] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

[26] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, and A. Sciarrone, "Gender-driven emotion recognition through speech signals for ambient intelligence applications," *IEEE Trans. Emerg. Topics Comput.*, vol. 1, no. 2, pp. 244–257, Dec. 2013.

[27] R. Xia, J. Deng, B. Schuller, and Y. Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 990–994.

[28] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: http://arxiv.org/abs/1512.03385

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[32] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, vol. 9, 2010, pp. 249–256.

[33] J. H. Tao, F. Z. Liu, M. Zhang, and H. B. Jia, "Design of speech corpus for mandarin text to speech," in *Proc. Blizzard Challenge Workshop*, 2008, p. 1.

[34] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.

[35] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[36] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2009, pp. 552–557.

[37] S. Deb and S. Dandapat, "Emotion classification using segmentation of vowel-like and non-vowel-like regions," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 360–373, Jul. 2017.

[38] L. Zao, D. Cavalcante, and R. Coelho, "Time-frequency feature and AMS-GMM mask for acoustic emotion classification," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 620–624, May 2014.

[39] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Commun.*, vol. 57, pp. 1–12, Feb. 2014.

[40] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Erlangen, Germany: University of Erlangen-Nuremberg Erlangen, 2009.

[41] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops*, Atlanta, GA, USA, 2006, p. 8.

[42] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[43] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," 2015, *arXiv:1512.04150*. [Online]. Available: http://arxiv.org/abs/1512.04150

[44] V. Sethu, E. Ambikairajah, and J. Epps, "Phonetic and speaker variations in automatic emotion classification," in *Proc. Interspeech*, 2008, pp. 617–620.

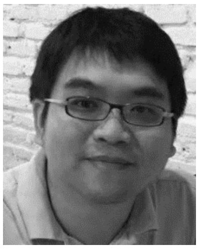
[45] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *Proc. Interspeech*, 2004, pp. 205–221.

[46] I. R. Titze, "Physiologic and acoustic differences between male and female voices," *J. Acoust. Soc. Amer.*, vol. 85, no. 4, pp. 1699–1707, 1989.

[47] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, vol. 87, no. 2, pp. 820–857, 1990.

[48] J. M. Vigil, "A socio-relational framework of sex differences in the expression of emotion," *Behav. Brain Sci.*, vol. 32, pp. 375–390, Oct. 2009.

[49] A. E. Thompson and D. Voyer, "Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis," *Cognition Emotion*, vol. 28, pp. 1164–1195, Oct. 2014.



TING-WEI SUN (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan, in 2006. He is currently pursuing the Ph.D. degree with the Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan.

His current research interests include machine learning, emotion detection, and signal processing.