Research Paper

Fine-Grained Recognition of Fidgety-Type Emotions Using Multi-Scale One-Dimensional Residual Siamese Network

Jiu SUN*, Jinxin ZHU, Jun SHAO

School of Information Technology, Yancheng Institute of Technology
Yancheng, Jiangsu, China

*Corresponding Author e-mail: sunjiu@ycit.edu.cn

(*received August 12, 2023; accepted April 4, 2024; published online July 15, 2024*)

Fidgety speech emotion has important research value, and many deep learning models have played a good role in feature modeling in recent years. In this paper, the problem of practical speech emotion is studied, and the improvement is made on fidgety-type emotion using a novel neural network model. First, we construct a large number of phonological features for modeling emotions. Second, the differences in fidgety speech between various groups of people were studied. Through the distribution of features, the individual features of fidgety emotion were studied. Third, we propose a fine-grained emotion classification method, which analyzes the subtle differences between emotional categories through Siamese neural networks. We propose to use multi-scale residual blocks within the network architecture, and alleviate the vanishing gradient problem. This allows the network to learn more meaningful representations of fidgety speech signal. Finally, the experimental results show that the proposed method can provide the versatility of modeling, and that fidgety emotion is well identified. It has great research value in practical applications.

Keywords: residual convolutional neural network; multi-scale neural network; fidgety speech emotion; fine-grained emotion classification; Siamese neural networks.

Acronyms

1-D	– one-dimensional,
AI	– artificial intelligence,
CNN	– convolutional neural network,
GMM	– Gaussian mixture model,
LSTM	– long short-term memory,
PCM	– pulse code modulation,
RNN	– recurrent neural network,
SEU	– Southeast University,
SVM	– support vector machine,
USB	– Universal Serial Bus,
WAV	– Waveform Audio File Format.

1. Introduction

Emotion recognition is a fundamental aspect of human communication and understanding. It plays a crucial role in various domains, including psychology, human-computer interaction, and social robotics. Traditional approaches to emotion recognition have primarily focused on categorical classification (Latif *et al.*, 2023; Yan *et al.*, 2013), but there is a growing need for more detailed analysis, especially in capturing subtle variations and specific types of emotions.

Various feature analysis and modeling algorithms have been applied to speech emotion recognition, including feature normalization, stochastic parameter optimization, neural networks and Gaussian mixture models (Jin *et al.*, 2009; 2014; Huang *et al.*, 2009a; Wang, Tashev, 2017; Lieskovská *et al.*, 2021). Chen and Huang (2021), proposed to study hybrid features in speech emotion recognition. Dupuis and Pichora-Fuller (2014) recommended to study behavioural features in emotional speech. Atila and Şengür (2021) proposed to use the novel convolutional neural network and long-short term memory network for emotion recognition. In their study, deep neural network structures were reviewed and studied. Large amount of data is required for deep learning.

Praseetha and Vadivel (2018) also studied deep learning models. In their studies only basic emotions were investigated.

Other researchers studied practical problems in emotion recognition, including text and speaker independent emotion recognition, practical types of emotions, cognitive related states, and language specific emotion models (Huang *et al.*, 2013a; 2013b; 2016; Wu *et al.*, 2018; Jin *et al.*, 2011; Xu *et al.*, 2014; Zou *et al.*, 2011).

Zhou *et al.* (2021) suggested to study a cough sound event using acoustic features. In their study, a Mel-spectrogram was used for feature analysis and a convolutional neural network was used for modeling. COVID-19 influenced cough sound recognition has wide potential applications. Atsavasirilert *et al.* (2019) proposed to study the computational efficiency in speech emotion recognition. In their study, the light weight convolutional neural network was proposed, and the real world challenges in computing resources were given their work has important practical value. They further studied Mel-spectrograms and treated the speech signal processing problem as 2-D information processing. However, in their work, emotion types were limited.

Emotion recognition is an important field in understanding human behavior, with traditional machine learning models and deep neural networks being widely used for classification. However, limited research has been conducted on emotions with specific practical values, such as fidgety emotions, which have unique significance.

This research paper addresses the gap in fine-grained practical speech emotion recognition by providing a more detailed categorization of emotions. While the traditional approach considers six main emotional categories (sadness, joy, anger, disgust, surprise, and fear), this paper aims to explore emotions with special practical value, including fidgety emotions. By considering specific application scenarios, fine-grained subtypes, and composite types of emotions, this paper offers a comprehensive framework for emotion detection in practical applications.

Fidgety emotion represents a significant emotional category distinct from traditional emotion research, which primarily focuses on basic emotional categories. Fidgety is a complex emotion with practical value, playing a crucial role in the realms of learning and cognition. It holds particular significance in influencing cognitive abilities, behavioral control, and psychological stability. While conventional emotion recognition research extensively explores the six basic emotions, happiness, anger, surprise, sadness, fear, and disgust, there has been limited investigation into complex emotions.

Fidgety emotion, characterized by its complexity, is particularly triggered in repetitive and tedious cognitive tasks, especially during prolonged periods of repetitive work. It remains a complex emotion with practical implications, significantly impacting cognitive abilities, behavioral control, and psychological stability within the processes of learning and cognition.

The paper explores the use of a Siamese neural network architecture, which excels in metric distance learning, for comparing and classifying fidgety-type emotions. We further propose to use a 1-D convolutional residual neural network, to improve the Siamese network structure. By constructing a large number of phonological features and analyzing group differences, the model captures individual characteristics and enables precise identification of emotional subcategories.

The empirical prowess of 1-D convolutional networks has been well-documented, asserting their supremacy in diverse time-serial feature extraction and modeling tasks. Numerous instances have showcased their state-of-the-art performance in extracting intricate patterns from temporal data streams, such as vibration signal processing, fault detection, and ECG signal processing (Abdeljaber *et al.*, 2017; Avci *et al.*, 2018; 2019; Kiranyaz *et al.*, 2019; Xiong *et al.*, 2017). However, the use of residual shortcut and multiscale receptive fields in specific emotion recognition has not been studied yet.

The proposed approach takes into account the nuances and complexities of fidgety emotions, which have important practical implications. By providing a more detailed understanding of these emotions, the research contributes to the development of effective emotion recognition systems. Additionally, by considering the specific contexts and characteristics of fidgety emotions, the proposed framework is tailored to address their unique practical challenges. This research serves as a valuable contribution to the field of fine-grained practical speech emotion recognition, providing insights and techniques for improved detection and understanding of fidgety emotions.

The key contribution of this research lies in its practical application of fine-grained fidgety-type emotion recognition using the improved Siamese network structure. The proposed method demonstrates versatility in modeling emotions across different ages and genders, showcasing its potential for real-world applications. The experimental results validate the effectiveness of the approach, giving promising practical implications in emotion recognition.

The paper is structured as follows: Sec. 2 provides an overview of the database used for training and evaluation. Section 3 presents the methodology employed for fine-grained fidgety-type emotion recognition as a few-shot learner. Section 4 presents the experimental results obtained from applying the proposed methodology. Finally, Sec. 5 concludes the paper by summarizing the key findings and discussing the implications and future directions of the research.

2. Database

We have employed a local database from Southeast University (SEU) to validate our method of emotion recognition (Huang *et al.*, 2009b; 2011; 2014; 2020), for fidgety-type emotions.

The recording software uses Adobe Audition. During recording, a monaural channel is used. The recorded speech signals are saved in the WAV (Waveform Audio Format) format encoded with PCM (pulse-code modulation). The recording hardwares include: one high-performance computer, one M-audio MobilePre USB sound card, one large-diaphragm condenser microphone, and one pair of monitoring headphones.

The recording process takes place in a quiet laboratory. After each recording, data verification and supplementation should be carried out. The recorded speech files should be manually checked promptly to eliminate any possible errors that may occur during the recording process. For example, inspecting and removing segments with signal overload, irregular noises (such as coughing), and long periods of silence caused by abnormal pauses. If the recording files have significant errors, supplementary recording may be necessary.

The collected data within this database encompasses speech-based emotions of a cognitive nature, encompassing emotions such as annoyance, fatigue, confidence, and joy.

For the purpose of this paper, a specific subset of utterances are chosen from the SEU database. To capture elicited emotional speech, negative emotions are induced through mathematical calculation tasks, involving the verbal reporting of calculated results and recording emotional speech, all conducted in Chinese. In the experimental dataset, 8 male and 8 female native Chinese-speaking participants volunteered, with careful selection to ensure gender balance, resulting in 3000 utterances for each gender category. The induction experiment avoided a standardized text, opting for the emotional speech collection in a natural state, in contrast to the scripted nature of a standardized text often used in acted speech recording. The recorded dataset comprises 6000 sentences, totaling 18 662 seconds, distributed across 2000 samples for fidgetiness, 2000 for happiness, and an additional 2000 for a neutral emotional state, forming a comprehensive subset of 6000 samples.

In addition to utilizing the SEU database for our research on speech emotion recognition, we have undertaken the task of manual annotation to achieve a fine-grained level of more detailed emotion types, as shown in Fig. 1. This meticulous process adds significant value to our research problem. By annotating the data ourselves, we ensure a comprehensive and nuanced understanding of the emotions expressed in the speech samples. This granular approach enables us to capture subtle variations and nuances within emotions, contributing to a more accurate and comprehensive analysis. The annotators are carefully selected with a background in psychology study and proper training of emotion utterance annotation. The annotation results are cross confirmed. We adopted a multiple annotation approach with a voting strategy.

The five fidgety levels are divided into five categories based on the general discriminative ability of human annotators, using ratings of 1, 3, 5, 7, and 9. Different intensity levels are assigned based on the strength of emotions. This annotation is employed to distinguish fine-grained emotional intensities, facilitating supervised learning to differentiate between specific emotional nuances.

3. Methodology

3.1. Few-shot fine-grained fidgety-type emotion recognition

Fine-grained fidgety-type emotion recognition refers to the accurate detection and classification of subtle variations in emotions, particularly those expressed through fidgety behavior. Few-shot learning is a machine learning approach that can generalize from a small number of training examples, which is crucial for emotion recognition tasks where obtaining large labeled datasets is challenging.

In our paper, a few-shot learning framework is applied to fine-grained fidgety-type emotion recognition. The paper introduces the concept of a Siamese neural

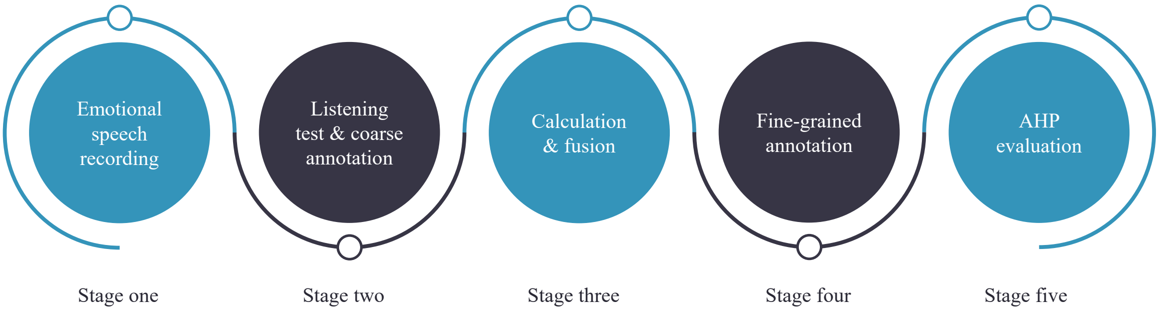

Fig. 1. Flow chart of the fine-grained annotation for emotional speech.

network, which is well-suited for metric distance learning. The Siamese network compares the input samples with templates and learns to measure the similarity or dissimilarity between them, as shown in Fig. 2.

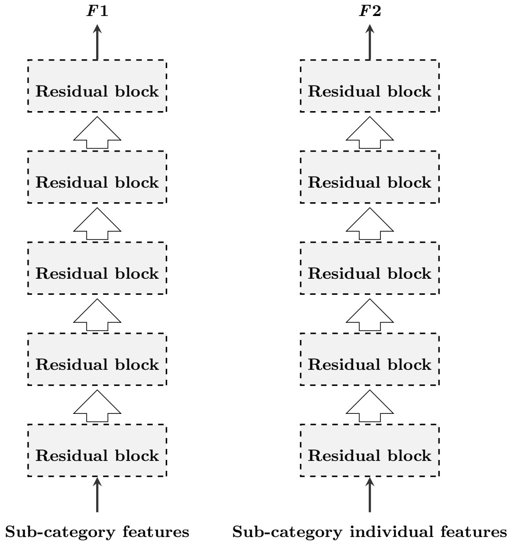

Fig. 2. Siamese network for fine-grained fidgety-type emotion recognition.

The novelty of the proposed solution lies in the utilization of the Siamese network as a few-shot learner. By leveraging this architecture, the model can effectively learn to recognize and classify fidgety-type emotions, even with limited training data. The Siamese network ability to learn meaningful representations of emotional features, combined with the few-shot learning approach, enhances the accuracy of fine-grained identification.

By constructing a large number of phonological features, analyzing group differences, and utilizing residual connections to address the vanishing gradient problem, the proposed method in the paper achieves a fine-grained emotion classification. This approach allows for the precise analysis of subtle differences between emotional categories. The experimental results demonstrate the versatility of the proposed method, highlighting its potential for practical applications in emotion recognition tasks involving fidgety speech.

First, we generate pairs of emotional samples, divided into positive and negative matches, and when we collect a small number of fidgety emotional types of specific speakers, we randomly select samples that are not sub-category and pair them to produce a negative training dataset. It is necessary to focus on the generation of negative sample pairs of similar sub-categories to improve fine-grained identification.

Within this system, we have incorporated a speaker recognition module that utilizes MFCC features with an i-vector approach. The i-vector approach is a commonly used technique in speaker recognition. It is a statistical modeling method that represents speaker characteristics using a low-dimensional fixed-length vector called the i-vector. This module enables fine-grained matching of fidgety subcategories, specifically within the sample range of individual speakers. The goal is to enhance the accuracy of fine-grained identification.

Not all components of the output contribute equally to the comparison process of the Siamese network outputs. As a result, we have implemented a fully connected layer that takes the outputs of the two subnetworks and generates the final classification output.

3.2. Improved Siamese network based on multi-scale residual network

In the realm of fine-grained modeling and recognition, our proposed incorporation of "multi-scale" architectures with various receptive fields is a promising avenue. This approach allows for a more intricate understanding of intricate details within data. Alongside this, the fusion of few-shot learning principles with distance learning methodologies has proven to be a potent combination in the pursuit of enhancing recognition capabilities.

In a typical Siamese network, we have two identical subnetworks that process input examples independently and produce fixed-length embeddings. These embeddings are then compared to determine their similarity or dissimilarity. In an improved Siamese network, to enhance the network's performance, we can incorporate residual connections within each subnetwork.

3.2.1. 1-D Convolution

1-D convolution is a fundamental operation in signal processing and data analysis, particularly for analyzing time-series signals. It involves combining two input signals to produce an output signal by sliding one signal (known as the kernel or filter) over the other, element by element, and computing the sum of element-wise products at each step. This operation is often used for various tasks such as feature extraction, filtering, and pattern recognition within time-series data.

The input speech signal is denoted as $x[n]$. The convolution kernel is denoted as $h[k]$.

Sliding operation: the convolution operation involves sliding the kernel over the input signal. At each step, the kernel is aligned with a portion of the input signal, and an element-wise multiplication is performed between the kernel and the overlapping portion of the input signal.

The convolution operation at a given time index n is calculated by sliding the kernel $h[k]$ over the speech signal $x[n]$ and performing the element-wise multiplication followed by summation, as shown in Eq. (1):

$$y[n] = \sum_{k=-\infty}^{\infty} x[n-k] \cdot h[k]. \tag{1}$$

In practice, the summation is limited to the valid range of k where both $x[n-k]$ and $h[k]$ are defined.

The resulting convolved signal $y[n]$ is obtained by performing the above convolution operation for each time index n, as shown in Eq. (2):

$$y[n] = \sum_{k=-\infty}^{\infty} x[n-k] \cdot h[k] \quad \text{for all } n. \tag{2}$$

One of the primary applications of 1-D convolution in time-series analysis is feature extraction and filtering. Using compact 1-D convolution we can highlight specific patterns and features within the fidgety speech signal.

3.2.2. Multi-scale residual convolution

Let us consider a specific layer, denoted as the layer L. The output of the layer L can be represented as $\mathbf{H_L}(\mathbf{x})$, where $\mathbf{x}$ is the input to that layer. To introduce a residual connection, we define the residual function $\mathbf{R_L}(\mathbf{x})$, which captures the difference between the input and output of the layer L. The output of the layer L with the residual connection, denoted as $\mathbf{F_L}(\mathbf{x})$, is given by:

$$\mathbf{F_L}(\mathbf{x}) = \mathbf{H_L}(\mathbf{x}) + \mathbf{R_L}(\mathbf{x}), \tag{3}$$

where $\mathbf{F_L}(\mathbf{x})$ represents the desired output of the layer L. By adding the residual function $\mathbf{R_L}(\mathbf{x})$ to the input $\mathbf{x}$, we allow the network to learn the residual mapping.

The residual function $\mathbf{R_L}(\mathbf{x})$ can be defined as:

$$\mathbf{R_L}(\mathbf{x}) = \mathbf{W_L} \cdot \mathbf{x}, \tag{4}$$

where $\mathbf{W_L}$ represents the weights of the residual connection, which are learned during the training process. Multiplying the input $\mathbf{x}$ by $\mathbf{W_L}$ allows the network to capture the residual information that needs to be added to the output.

With the addition of residual connections, the output of layer $L+1$ can be expressed as:

$$\mathbf{H_{L+1}}(\mathbf{F_L}(\mathbf{x})) = \mathbf{H_{L+1}}(\mathbf{H_L}(\mathbf{x}) + \mathbf{R_L}(\mathbf{x})). \tag{5}$$

Convolutional kernels of different scales can extract features of varying precision, with smaller kernels capturing finer details. If a single layer uses only kernels of the same scale, it may overlook features of other precisions, resulting in incomplete information being represented by the extracted features. Consequently, we have designed three distinct resolutions for feature extraction, as illustrated in Fig. 3.

By incorporating residual connections in this manner, the gradient can flow directly from the output of a layer to its input, facilitating the flow of gradients during training. This alleviates the vanishing gradient problem and enables the network to learn more meaningful representations.

In the improved Siamese network, multiple residual connections can be added at different layers. By utilizing residual connections, the improved Siamese network can effectively learn complex patterns and relationships in the input data, leading to better similarity or distance measurements and improved performance in fine-grained emotion recognition. The overall framework is shown in Fig. 4.

Our innovative approach to the fine-grained fidgety emotion recognition challenge involves the utilization

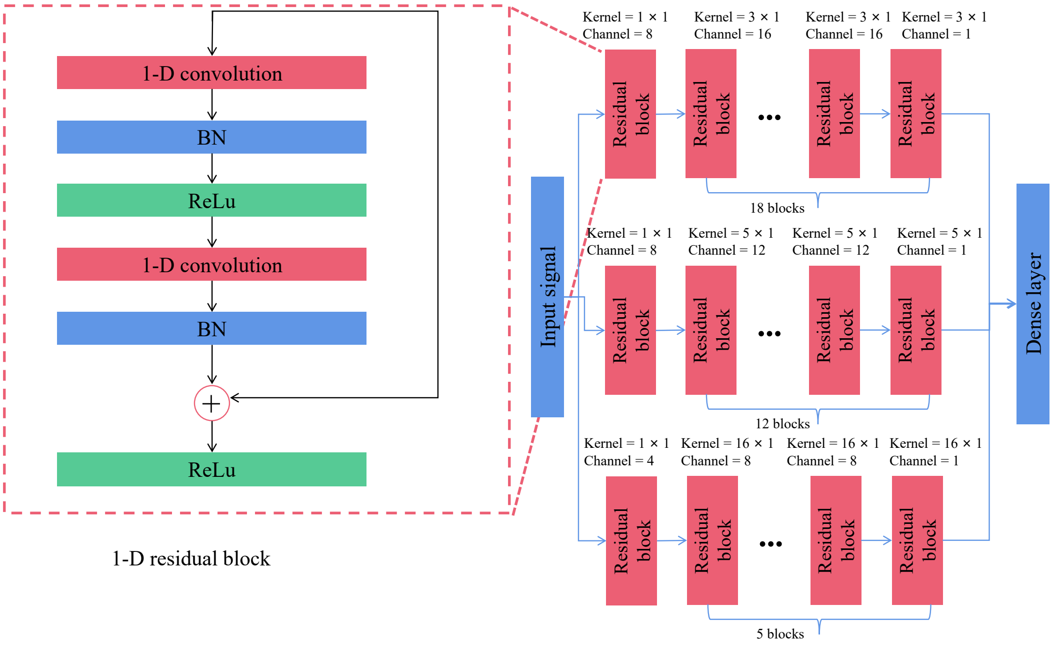

Fig. 3. Proposed multi-scale residual Siamese network structure.

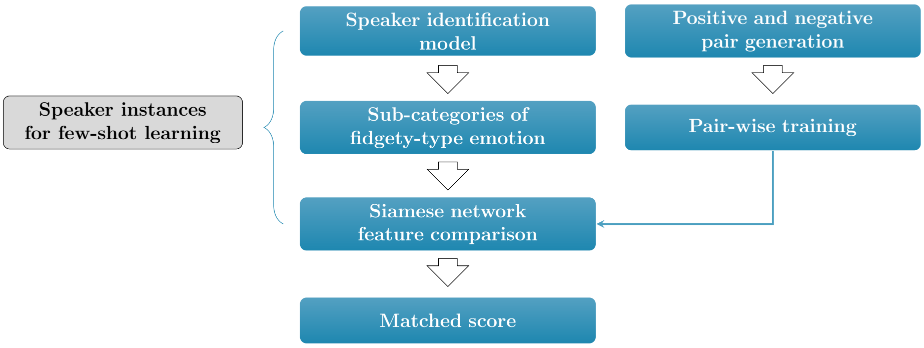
Speaker instances for few-shot learning
Speaker identification model
Sub-categories of fidgety-type emotion
Siamese network feature comparison
Matched score
Positive and negative pair generation
Pair-wise training

Fig. 4. Metric learning using speaker instances as few-shot learning.

of a 1-D convolutional residual neural network, strategically designed to augment the traditional Siamese network. The integration of residual blocks within this framework plays a pivotal role in enhancing convergence during the training process. By capitalizing on the inherent advantages of 1-D convolutions, particularly their proficiency in processing time-serial signals, our architecture demonstrates remarkable potential.

3.3. Training samples generation

Let S be a speech sample, C_{fid} be the main category of the fidgety emotion, and c_{fid}^{j} be the subcategory of the fidgety emotion within the main category. The method for generating positive and negative sample pairs is as follows:

$$S_1 \in C_{\text{fid}}, \tag{6}$$

$$S_2 \notin C_{\text{fid}}. \tag{7}$$

Neg_coarse = $\{S_1, S_2\}$ forms a negative sample pair:

$$S_1 \in c_{\text{fid}}^{j}, \tag{8}$$

$$S_2 \in c_{\text{fid}}^{k}, \tag{9}$$

$$j \neq k. \tag{10}$$

Neg_fine = $\{S_1, S_2\}$ forms a negative sample pair, representing samples that require fine-grained distinction. Neg_fine : Neg_coarse > 3 : 1 This ensures that the model has a higher resolution for fine-grained samples:

$$S_1 \in c_{\text{fid}}^{j}, \tag{11}$$

$$S_2 \in c_{\text{fid}}^{j}. \tag{12}$$

Pos = $\{S_1, S_2\}$ forms a positive sample pair, used to supervise the output results of the Siamese network. The distance between samples in the same fine-grained subclass should be relatively close.

4. Experimental results

4.1. Experimental data

In our experiments on speech emotion recognition, we have recognized the critical role of the emotion corpus. While basic emotion types have received considerable attention, the study of emotions with practical value remains insufficient. Particularly, the scarcity of negative practical emotions in existing databases poses a challenge. Therefore, we have made a deliberate decision to exclusively employ the SEU database for our research. Unlike other databases that predominantly focus on basic or positive emotions within ordinary settings, the SEU database offers a unique advantage by providing a comprehensive collection of practical emotions, including the elusive fidgety-type emotion. This strategic selection enables us to delve deeper into understanding and accurately recognizing the nuanced emotions encountered in real-world scenarios.

4.2. Models comparison

In this research study, we aim to investigate the effectiveness of the proposed multi-scale residual Siamese network for fine-grained fidgety-type emotion recognition. We compare it against four other classifiers: baseline Siamese network, LSTM, support vector machine (SVM), and Gaussian mixture model (GMM).

The baseline Siamese network is a deep neural network architecture that learns to measure similarity between input samples. It consists of two identical subnetworks that share weights, enabling it to compute a similarity metric between two inputs. The residual Siamese network builds upon this architecture by incorporating residual connections, which help alleviate the vanishing gradient problem and enable easier optimization.

Long short-term memory (LSTM) is a widely used recurrent neural network (RNN) architecture that has shown remarkable success in various sequence-based tasks, including natural language processing and speech recognition. Unlike traditional RNNs, LSTM incorporates specialized memory cells that can capture and retain information over long periods. This unique characteristic enables LSTM to effectively learn and model complex temporal dependencies in sequences.

SVM is a supervised machine learning algorithm used for classification tasks. It aims to find an optimal hyperplane that maximally separates different classes in the feature space. SVMs are known for their ability to handle high-dimensional data and work well when there is a clear margin of separation between classes.

GMM is a probabilistic model that represents the distribution of data points as a mixture of Gaussian distributions. It can capture complex data patterns by estimating the parameters of Gaussian components. GMMs are versatile and can handle a wide range of data distributions, making them suitable for modeling fine-grained emotions.

4.3. Parameter settings

For the Siamese networks, we use a learning rate of 0.001, batch size of 32, and training for a fixed number of epochs (100). For training the LSTM model, the chosen parameter setting was a learning rate of 0.001, a batch size of 64, training for approximately 50 epochs. SVM parameter settings: $C = 1$, kernel = radial basis function (RBF), gamma = 0.1. GMM parameter settings: number of Gaussian components = 12, mean and covariance initialization based on data, maximum number of iterations = 100.

"Epoch" refers to a single pass through the entire training dataset, and it is used to optimize the model's parameters by adjusting them based on the accumulated error to improve overall performance during training. "Learning rate" in the context of machine learning is a hyperparameter that determines the size of the steps taken during the optimization process, influencing how quickly or slowly a model converges to the optimal set of parameters. "Batch size" refers to the number of training examples utilized in one iteration, influencing the efficiency of model training and the amount of computational resources required.

The training-to-validation-to-testing ratio is 6:1:3, totaling 6000 samples. Training dataset consists of 3600 utterances; validation dataset consists of 600 utterances; testing dataset consists of 1800 utterances.

In the experimental process of comparing models, we utilized different parameters to obtain the empirically optimal performance for each model. For example, we conducted a search for SVM parameters to set the optimal values. We compared different kernel functions, including RBF, linear, and polynomial, and the results indicated that RBF performed the best. We optimized the values of C and gamma through the grid search. For GMM, we experimented with different values for the number of mixture components (4, 12, 16, 24) and employed a diagonal matrix initialization method to optimize the empirically best results for the model. In the case of LSTM, we compared different optimizers, with Adam yielding the best results. We conducted a search for different learning rates, selecting the empirically optimal learning rate based on $F1$ scores.

The purpose of comparing these models is to evaluate the efficacy of the proposed multi-scale residual Siamese network for fine-grained fidgety-type emotion recognition. By contrasting its performance with other established classifiers, such as the baseline Siamese network, LSTM, SVM, and GMM, we can determine whether the additional architectural enhancements of the residual Siamese network yield improved accuracy and robustness in recognizing fine-grained emotions characterized by fidgety behaviors.

4.4. Results

In our experiment, we adopt the confusion matrix as a crucial tool for evaluating and comparing different emotion recognition models. As show in Tables 1–5, the confusion matrix provides a comprehensive summary of the models' predictions, enabling us to analyze the true positives, true negatives, false positives, and false negatives in classifying emotions. By utilizing the confusion matrix, we can gain insights into the performance of each model in accurately recognizing and classifying different emotions. This evaluation allows us to compare the effectiveness of various models and make informed decisions regarding their suitability for emotion recognition tasks.

As shown in Fig. 5, we compared various popular machine learning models to gain insights into their performance and effectiveness in our study. By examining

Table 1. Confusion matrix for fine-grained fidgety-type emotion recognition using multi-scale residual Siamese network.

Actual emotion	Predicted emotion					
	Fidgety level 1	Fidgety level 2	Fidgety level 3	Fidgety level 4	Fidgety level 5	Neutral
Fidgety level 1	80.1	8.4	4.5	3.5	1.0	2.5
Fidgety level 2	7.5	81.9	3.5	0.8	1.2	5.1
Fidgety level 3	3.9	6.1	75.2	5.0	2.5	7.3
Fidgety level 4	2.5	5.5	6.0	77.8	4.0	4.2
Fidgety level 5	1.8	2.2	8.0	7.5	75.1	5.4
Neutral	1.5	2.1	2.1	3.9	3.1	87.3

Table 2. Confusion matrix for fine-grained fidgety-type emotion recognition using baseline Siamese network.

Actual emotion	Predicted emotion					
	Fidgety level 1	Fidgety level 2	Fidgety level 3	Fidgety level 4	Fidgety level 5	Neutral
Fidgety level 1	72.7	10.4	6.3	3.1	4.5	2.0
Fidgety level 2	4.7	77.1	6.0	5.5	5.4	1.3
Fidgety level 3	5.8	8.0	70.2	6.4	4.5	5.1
Fidgety level 4	3.4	3.7	0.3	74.8	11.3	6.5
Fidgety level 5	4.5	3.3	7.9	8.5	70.3	5.5
Neutral	1.1	3.4	2.3	4.4	8.4	80.4

Table 3. Confusion matrix for fine-grained fidgety-type emotion recognition using LSTM.

Actual emotion	Predicted emotion					
	Fidgety level 1	Fidgety level 2	Fidgety level 3	Fidgety level 4	Fidgety level 5	Neutral
Fidgety level 1	70.1	8.6	8.2	7.1	1.2	4.8
Fidgety level 2	8.6	75.2	2.4	4.4	7.4	2.0
Fidgety level 3	6.5	7.4	64.9	7.4	6.3	3.5
Fidgety level 4	5.7	7.4	5.3	70.3	6.5	4.8
Fidgety level 5	6.3	6.3	5.2	7.8	64.8	9.6
Neutral	7.8	6.4	3.7	3.3	3.7	75.1

Table 4. Confusion matrix for fine-grained fidgety-type emotion recognition using SVM.

Actual emotion	Predicted emotion					
	Fidgety level 1	Fidgety level 2	Fidgety level 3	Fidgety level 4	Fidgety level 5	Neutral
Fidgety level 1	77.2	4.7	6.3	4.5	3.7	3.6
Fidgety level 2	8.3	75.4	7.4	6.3	2.1	0.5
Fidgety level 3	3.2	10.9	66.8	8.4	8.4	2.3
Fidgety level 4	4.6	8.6	7.3	70.1	5.5	3.9
Fidgety level 5	3.3	1.4	5.6	7.5	71.4	10.8
Neutral	6.4	4.5	2.4	3.3	3.1	80.3

Table 5. Confusion matrix for fine-grained fidgety-type emotion recognition using GMM.

Actual emotion	Predicted emotion					
	Fidgety level 1	Fidgety level 2	Fidgety level 3	Fidgety level 4	Fidgety level 5	Neutral
Fidgety level 1	70.2	7.3	6.3	5.5	3.1	7.6
Fidgety level 2	7.4	77.9	6.2	3.9	2.4	2.2
Fidgety level 3	7.8	8.4	70.8	7.9	3.4	1.7
Fidgety level 4	1.7	8.9	8.4	69.1	6.9	5.0
Fidgety level 5	4.1	1.9	7.3	7.4	70.9	8.4
Neutral	9.8	2.3	5.8	4.5	2.4	75.2

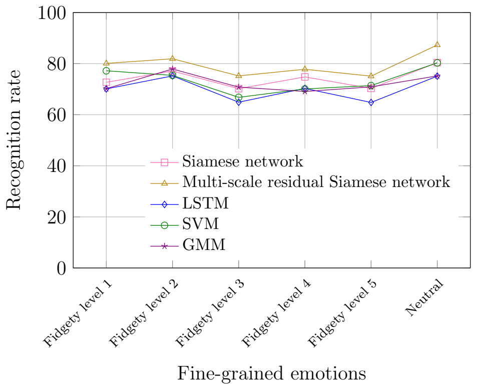

Fig. 5. Comparison among various modeling algorithms for averaged recognition rates.

and comparing the different curves generated by these models, we were able to assess their recognition rates and classification accuracy for the task at hand. This comparative analysis allowed us to evaluate the strengths and weaknesses of each model, identify areas of specialization, and uncover potential limitations.

From the experimental results, we can see that various modeling algorithms exhibit distinctive recognition rates for different fine-grained emotions.

Siamese network: the Siamese network exhibits moderate recognition rates across all fine-grained emotions, ranging from 70.2 % to 80.4 %. It achieves relatively higher rates for fidgety level 1 and fidgety level 2 compared to the other emotions.

Multi-scale residual Siamese network: the residual Siamese network demonstrates consistent performance, with recognition rates ranging from 75.1 % to 87.3 %. It achieves higher rates for fidgety level 1, fidgety level 2, and neutral emotions, indicating its effectiveness in recognizing these categories.

LSTM: the LSTM model showcases relatively lower recognition rates, ranging from 64.8 % to 75.2 %. It may require a much larger training database to capture the subtle distinctions between fine-grained emotions, resulting in slightly lower overall performance.

SVM: the SVM model demonstrates varied recognition rates, ranging from 66.8 % to 80.3 %. It performs relatively well for fidgety level 1 and fidgety level 2 emotions, but its performance drops for fidgety level 3 and fidgety level 4.

GMM: the GMM model achieves recognition rates ranging from 69.1 % to 77.9 %. It displays relatively lower rates compared to other models, particularly for fidgety level 1, fidgety level 3, and fidgety level 4 emotions.

Overall, the multi-scale 1-D residual Siamese network stands out with the highest recognition rates across various fine-grained emotions. The Siamese network and SVM models perform reasonably well, but their rates are slightly lower compared to the residual Siamese network. The LSTM and GMM models exhibit comparatively lower recognition rates, indicating the need for further improvement in capturing fine-grained emotional nuances.

4.5. Discussions

The advantages of the multi-scale residual Siamese network lie in its ability to enhance the model depth and, consequently, improve representation capability by introducing residual results. The use of the Siamese network structure enables fine-grained differentiation of emotion categories. However, its drawback is its reliance on a substantial amount of data for training, making it highly data-dependent.

The baseline Siamese network excels in distinguishing subtle differences between different emotions but lacks the introduction of residual structures, leaving room for improvement in representation capability.

LSTM's strength lies in its structure, which is conducive to modeling time series data. However, its computational complexity and convergence in modeling may not always achieve ideal results, especially under conditions of limited objective data.

SVM exhibits strong discriminative power under small-sample conditions, but it lacks the ability for automatic representation learning, making it challenging to fully exploit the value of training data.

GMM's advantage lies in its strong fitting capability and ability to model arbitrary feature distributions. However, this is contingent upon having sufficient and diverse data coverage, resulting in a high dependence on data.

The performance of a model is influenced by the characteristics of different input data because the statistical machine learning approach is inherently dependent on data. To address this challenge, a strategy is to separate training, validation, and testing data. This allows for objective and reasonable testing on an unseen test set, effectively demonstrating the model's generalization ability.

The multi-scale residual Siamese network proposed by us exhibits high reliability and stability. This is ensured through the separation of our testing, validation, and training sets. Experimental results indicate that its recognition performance surpasses that of other traditional models.

To substantiate the efficacy of our proposed model, we conducted a comprehensive comparative analysis. Our proposed network was meticulously pitted against the traditional Siamese network, as well as other prominent machine learning algorithms. Through rigorous experimentation and meticulous evaluation, our results unveil the prowess of our approach, demonstrating its superior performance in the realm of fine-grained fidgety-type emotions modeling and recognition tasks. This novel fusion of multi-scale architectures, few-shot learning, and distance learning principles, bolstered by the advancements of 1-D convolutional residual neural networks, introduces a pioneering stride towards unraveling the complexities of intricate data domains.

5. Conclusions

This paper focuses on the practical application of fidgety speech emotion recognition. Our contributions are centered around the development of phonological features and the implementation of a meticulous emotion classification method that utilizes Siamese neural networks with residual connections.

To enhance the precision of emotion classification, we have introduced a meticulous approach employing Siamese neural networks. By integrating residual connections, we have effectively addressed the challenge of the vanishing gradient, enabling the network to acquire more meaningful representations of fidgety speech emotions.

Experimental results have demonstrated the efficacy and adaptability of our approach, as we have successfully achieved accurate identification of fidgety emotions. Our proposed approach exhibits significant potential for practical applications and lays the foundation for further advancements in this field.

In future endeavors, it would be valuable to explore the integration of contextual information, such as situational cues and temporal dynamics, in order to enhance both the accuracy and contextual comprehension of fidgety speech emotion recognition.

Acknowledgments

The authors would like to thank Professor Li Zhao from the Southeast University for providing the raw speech emotional database used in this research.

References

1. Abdeljaber O., Avci O., Kiranyaz S., Gabbouj M., Inman D.J. (2017), Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks, *Journal of Sound and Vibration*, **388**: 154–170, doi: 10.1016/j.jsv.2016.10.043.

2. Atila O., Şengür A. (2021), Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition, *Applied Acoustics*, **182**: 108260, doi: 10.1016/j.apacoust.2021.108260.

3. Atsavasirilert K. *et al.* (2019), A light-weight deep convolutional neural network for speech emotion recognition using mel-spectrograms, [in:] *Proceedings of 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing* (*ISAI-NLP*), pp. 1–4, doi: 10.1109/isai-nlp48611.2019.9045511.

4. Avci O., Abdeljaber O., Kiranyaz S., Hussein M., Inman D.J. (2018), Wireless and real-time structural damage detection: A novel decentralized method for wireless sensor networks, *Journal of Sound and Vibration*, **424**: 158–172, doi: 10.1016/j.jsv.2018.03.008.

5. Avci O., Abdeljaber O., Kiranyaz S., Inman D.J. (2019), Convolutional neural networks for real-time and wireless damage detection, *Dynamics of Civil Structures*, **2**: 129–136, doi: 10.1007/978-3-030-12115-0_17.

6. Chen Q., Huang G. (2021), A novel dual attention-based BLSTM with hybrid features in speech emotion recognition, *Engineering Applications of Artificial Intelligence*, **102**: 104277, doi: 10.1016/j.engappai.2021.104277.

7. Dupuis K., Pichora-Fuller M.K. (2014), Recognition of emotional speech for younger and older talkers, *Ear & Hearing*, **35**(6): 695–707, doi: 10.1097/aud.0000000000000082.

8. Huang C., Chen G., Yu H., Bao Y., Zhao L. (2013a), Speech emotion recognition under white noise, *Archives of Acoustics*, **38**(4): 457–463, doi: 10.2478/aoa-2013-0054.

9. Huang C., Jin Y., Zhao Y., Yu Y., Zhao L. (2009a), Speech emotion recognition based on re-composition of two-class classifiers, [in:] *Proceedings of 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–3, doi: 10.1109/acii.2009.5349420.

10. Huang C., Jin Y., Zhao Y., Yu Y., Zhao L. (2009b), Recognition of practical emotion from elicited speech, [in:] *Proceedings of the First International Conference on Information Science and Engineering*, pp. 1–4, doi: 10.1109/icise.2009.875.

11. Huang C., Liang R., Wang Q., Xi J., Zha C., Zhao L. (2013b), Practical speech emotion recognition based on online learning: From acted data to elicited data, *Mathematical Problems in Engineering*, **2013**: 265819, doi: 10.1155/2013/265819.

12. Huang C., Song B., Zhao L. (2016), Emotional speech feature normalization and recognition based on speaker-sensitive feature clustering, *International Journal of Speech Technology*, **19**(4): 805–816, doi: 10.1007/s10772-016-9371-3.

13. Huang C., Zhao Y., Jin Y., Yu Y., Zhao L. (2011), A study on feature analysis and recognition of practical speech emotion, *Journal of Electronics & Information Technology*, **33**(1): 112–116, doi: 10.3724/sp.j.1146.2009.00886.

14. Jin Y., Huang C., Zhao L. (2011), A semi-supervised learning algorithm based on modified self-training SVM, *Journal of Computers*, **6**(7): 1438–1443, doi: 10.4304/jcp.6.7.1438-1443.

15. Jin Y., Song P., Zheng W., Zhao L. (2014), A feature selection and feature fusion combination method for speaker-independent speech emotion recognition, [in:] *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4808–4812, doi: 10.1109/icassp.2014.6854515.

16. Jin Y., Zhao Y., Huang C., Zhao L. (2009), Study on the emotion recognition of whispered speech, [in:] *Proceedings of 2009 WRI Global Congress on Intelligent Systems*, pp. 242–246, doi: 10.1109/gcis.2009.175.

17. Kiranyaz S., Gastli A., Ben-Brahim L., Al-Emadi N., Gabbouj M. (2019), Real-time fault detection and identification for MMC using 1-D convolutional neural networks, *IEEE Transactions on Industrial Electronics*, **66**(11): 8760–8771, doi: 10.1109/tie.2018.2833045.

18. Latif S., Shahid A., Qadir J. (2023), Generative emotional AI for speech emotion recognition: The case for synthetic emotional speech augmentation, *Applied Acoustics*, **210**: 109425, doi: 10.1016/j.apacoust.2023.109425.

19. Lieskovská E., Jakubec M., Jarina R., Chmulík M. (2021), A review on speech emotion recognition using deep learning and attention mechanism, *Electronics*, **10**(10): 1163, doi: 10.3390/electronics10101163.

20. Praseetha V.M., Vadivel S. (2018), Deep learning models for speech emotion recognition, *Journal of Computer Science*, **14**(11): 1577–1587, doi: 10.3844/jcssp.2018.1577.1587.

21. Wang Z.Q., Tashev I. (2017), Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks, [in:] *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 5150–5154, doi: 10.1109/icassp.2017.7953138.

22. Wu C., Huang C., Chen H. (2018), Text-independent speech emotion recognition using frequency adaptive features, *Multimedia Tools and Applications*, **77**(18): 24353–24363, doi: 10.1007/s11042-018-5742-x.

23. Xiong Z., Stiles M., Zhao J. (2017), Robust ECG signal classification for the detection of atrial fibrillation using novel neural networks, [in:] *Proceedings of 2017 Computing in Cardiology Conference* (*CinC*), **44**, doi: 10.22489/cinc.2017.066-138.

24. Xu X., Huang C., Wu C., Wang Q., Zhao L. (2014), Graph learning based speaker independent speech emotion recognition, *Advances in Electrical and Computer Engineering*, **14**(2): 17–22, doi: 10.4316/aece.2014.02003.

25. Yan J., Wang X., Gu W., Ma L. (2013), Speech emotion recognition based on sparse representation, *Archives of Acoustics*, **38**(4): 465–470, doi: 10.2478/aoa-2013-0055.

26. Zhou Q. *et al*. (2021), Cough recognition based on Mel-spectrogram and convolutional neural network, *Frontiers in Robotics and AI*, **8**: 1–7, doi: 10.3389/frobt.2021.580080.

27. Zou C., Huang C., Han D., Zhao L. (2011), Detecting practical speech emotion in a cognitive task, [in:] *Proceedings of 20th International Conference on Computer Communications and Networks* (*ICCCN*), pp. 1–5, doi: 10.1109/icccn.2011.6005883.