scientific reports

Check for updates

OPEN

A multi-dilated convolution network for speech emotion recognition

Samaneh Madanian[1]✉, Olayinka Adeleye[1], John Michael Templeton[2], Talen Chen[1], Christian Poellabauer[3], Enshi Zhang[3] & Sandra L. Schneider[4]

Speech emotion recognition (SER) is an important application in Affective Computing and Artificial Intelligence. Recently, there has been a significant interest in Deep Neural Networks using speech spectrograms. As the two-dimensional representation of the spectrogram includes more speech characteristics, research interest in convolution neural networks (CNNs) or advanced image recognition models is leveraged to learn deep patterns in a spectrogram to effectively perform SER. Accordingly, in this study, we propose a novel SER model based on the learning of the utterance-level spectrogram. First, we use the Spatial Pyramid Pooling (SPP) strategy to remove the size constraint associated with the CNN-based image recognition task. Then, the SPP layer is deployed to extract both the global-level prominent feature vector and multi-local-level feature vector, followed by an attention model to weigh the feature vectors. Finally, we apply the ArcFace layer, typically used for face recognition, to the SER task, thereby obtaining improved SER performance. Our model achieved an unweighted accuracy of 67.9% on IEMOCAP and 77.6% on EMODB datasets.

Keywords Speech emotion recognition, Deep learning, Convolution neural network, Loss layer, Spectrogram, Emotion recognition

Speech signals are an important carrier of emotions in human interactions, i.e., they convey affective information about human emotional states via explicit linguistic and implicit paralinguistic cues. In human speech interactions, paralinguistic characteristics such as intonations, emotions, and context are crucial in transmitting the underlying intent of messages[1]. Speech emotional intelligence is considered a key requirement for effective human-computer interactions[2]. Consequently, Speech Emotion Recognition (SER), has garnered increasing attention from researchers. SER is the extraction of emotional information from speech signals to recognize the emotional states of users. This has been used in many Human-Computer Interaction (HCI) applications such as storytelling, e-tutoring[3], customer service and call attendants appraisal[4,5]. More recently, SER has been suggested as a decision support tool in mental health[6,7].

SER applications and their potential benefits motivate many efforts in developing Artificial Intelligence (AI) models for automatic emotion recognition. SER model development differs in the use of speech features including *low-level descriptors* (LLD)[8–10] and *spectrograms*[2,11,12]. LLDs are closely related to the signal itself and require hand-crafted procedures in extracting the audio features like pitch and Zero-Crossing Rate (ZCR) for training a classifier[13]. The performance of *deep representation learning* techniques in automatically learning high-level features from low-level data in this domain has accelerated the progress of detecting emotions from speech signals[14,15]. As a result, more attention has been given to *spectrograms*, which are often associated with the automatic extraction of deep emotional patterns using Deep Neural Networks (DNNs). A *spectrogram* is a 2D visual representation of the speech energy distribution across frequency bands over time. Speech signal can be transformed to typical 2D *spectrograms* by representing using Short-Time Fourier Transform (STFT). A signal's 2D representative characteristic is often used to leverage the advantage of the Convolution Neural Network (CNN) and its derivatives in computer vision and image recognition.

While CNNs with spectrograms have been used in SER, existing approaches face several key challenges: (1) the gridding effect in dilated CNNs that leads to information loss, (2) inconsistent performance across different emotional states due to class imbalance, and (3) limited generalizability across different datasets and speaker populations. Therefore, our work aimed at enhancing the performance of an SER system using a multi-dilated

[1]Department of Data Science and Artificial Intelligence, Auckland University of Technology, Auckland, New Zealand. [2]College of Engineering, University of the South Florida, Tampa, FL, USA. [3]Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL, USA. [4]Department of Communication Sciences and Disorders, Saint Mary's College, Notre Dame, IN, USA. ✉email: sam.madanian@aut.ac.nz

CNN structure. We explored state-of-the-art loss layer, commonly used for face recognition, to determine whether it can also be effective for the SER task. This study has the following contributions:

1. A novel multi-dilated CNN model SER feature extraction that systematically prevents information loss from the gridding effect.
2. An adaptive feature extraction framework combining a *spatial pyramid pooling* (SPP) and *ArcFace* loss that significantly improves class-balanced performance. This also address the spectrogram learning issue associated with varied lengths in the time domain.
3. Adapting a state-of-the-art loss layer traditionally used for face recognition to the SER model to further improve its performance.

Building upon existing techniques, our work represents a novel combination and adaptation of these methods specifically for SER. Based on the improvements in Unweighted Accuracy (UA) and the detailed analysis of our approach, it can be concluded that the proposed multi-dilated CNN architecture is a more effective feature extraction method.

The structure of this paper is as follows. Section "Related work" provides a review of the relevant literature. Section "Proposed framework" explains the adapted research methodology, followed by the description of the experiments in section "Experiments". The research results are discussed in section "Discussion", and the paper concludes with Section "Conclusions and future work".

Related work

Spectrogram-based SER has increasingly relied on deep learning, particularly leveraging CNN architectures. This is attributed to the inherent 2D presentation of spectrograms and the robust 2D learning capabilities of CNNs. Several studies have tried to develop CNN-based models to enhance an SER's performance.

A CNN-based model, combined with an Extreme Learning Machine (ELM) structure, was developed in[16] to extract emotional patterns from spectrograms for classification. The spectrogram was converted to a 3D presentation in[15]. The 3D presentation consisted of the raw spectrogram, its delta, and the delta of the delta. These researchers incorporated the first and the second derivative of energy distribution over time and developed a CNN-Bidirectional (Bi) Long short-term memory (LSTM) structure. Their model achieved 64% Unweighted Accuracy (UA) on the Interactive Emotional Dyadic Motion Capture Corpus (IEMOCAP) dataset. The work in[17] further improved the CNN-BiLSTM structure to a Dilated CNN-BiLSTM structure by applying a Dilated CNN structure, with residual block[18], and then centre loss[19] to improve SER performance. This increased the UA model on IEMOCAP to 69%.

In another attempt[2], FaceNet structure[20] and the area attention mechanism were used, obtaining 63% UA on the IEMOCAP dataset. FaceNet is well-known for its multi-branch convolution structure and great performance in face recognition for emotion recognition based on the speech spectrogram. Despite all these efforts, some open issues remain to be addressed in developing SER systems. The first issue is feature learning based on different sizes of speech spectrograms. Since the spectrogram is constructed with a range of frequency bands over a time period, the longer the raw audio is, the prolonged the spectrogram's time dimension is. This results in some issues when developing an audio dataset for SER. As the lengths of the audio files are usually different, their spectrograms have different lengths over the time dimension. This makes training the fully connected (FC) layer extremely challenging. FC layers only accept feature maps of the same size and normally form the last classification section of deep CNN models. A standard method to address this issue is using a global pooling strategy that only extracts the most prominent feature across a batch of feature maps of different sizes, such as maximal and average, to form a uniform size of a feature vector to train the FC layer. However, this could cause a loss of information.

Previous studies handle this issue using various strategies. For example, some studies split the spectrograms into multiple segments with a fixed length and zero-pad the short ones[17,21]. The authors of[22] segmented each audio file into different numbers of uniformed sub-segments with a fixed overlapping and leveraged the advantage of the Recurrent Neural Network (RNN) with LSTM structure to handle different lengths of timestep learning. Fully Convolution Networks (FCNs) were used in[20,23] to learn from raw spectrograms of various lengths in a time dimension, accredited to the characteristics of the convolutional kernel. However, the splitting or segmenting of the audio files could break the natural consistency of the semantic information inside the raw speech. Although FCNs can better preserve the raw speech data, they do not seem to lead to better performance compared to the splitting/segmenting strategy. Instead, in our work, we use Spatial Pyramid Pooling (SPP) introduced in[24] and also used in[25] to handle the issue of learning from the images of different sizes to address a similar problem in spectrogram learning for the SER.

The second issue concerns the identification of a useful CNN-based structure to learn emotional information from the speech spectrogram.[20] suggested transferring the multibranch structure of FaceNet, from face recognition, into the spectrogram learning for SER. However, this complicated structure fails to challenge a simple model developed by[17] that achieved a 6% higher UA than[20]. Nevertheless, the dilated CNN structure in[17] had some issues with SER performance. Using multiple layers of convolution layers with the same dilation rates or the dilation rates having the greatest common divisors rather than 1 caused the gridding effect (Fig. 1) that led to the loss of information consistency[26]. We applied an iterative schema to use different dilation rates to address this issue. However, in some SER experiments, we find that the convolution layers with the same dilation rates can lead to better performances when testing on some of the IEMOCAP audio sessions rather than the schema of a series of iterative dilation rates (in Supplementary Material). Thus, to achieve a synergy of the two structures, we apply and merge them across the whole CNN and name it the multi-dilated CNN structure. The details are explained in the Multi-Dilated CNN and SPP sub-section.

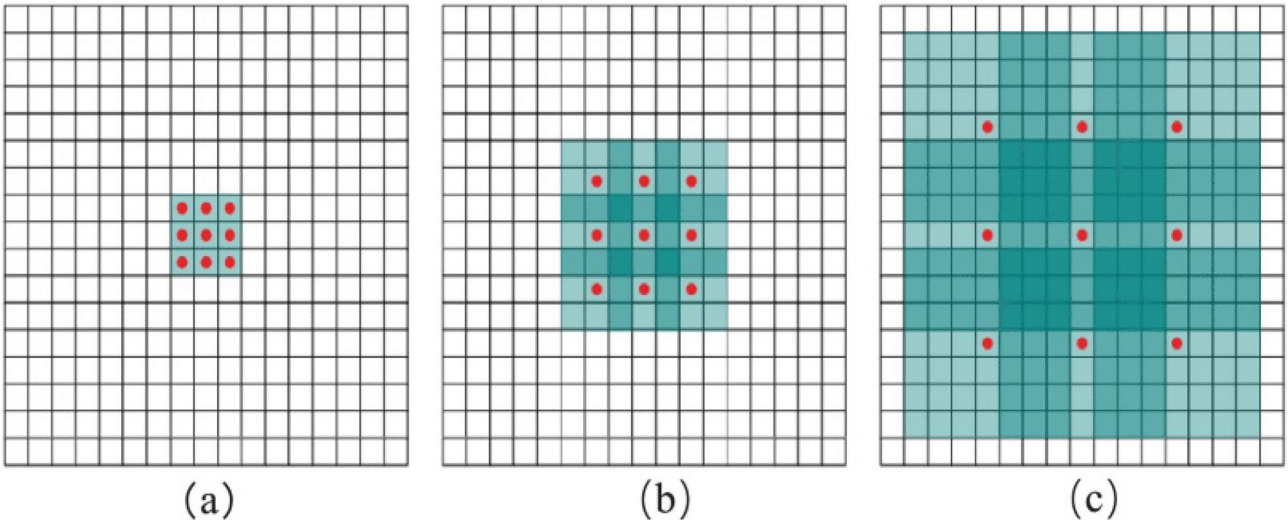

Figure 1. The gridding effect of the dilated CNN structure, adapted from[29].

The last issue is associated with using the advanced loss function to further improve the SER model's final performance. The softmax is the traditional loss function for training the DNNs for classification tasks. Still, it lacks the optimization of the feature embedding to explicitly encourage the intra-class similarity while diversifying the inter-class samples[27]. To address this issue,[17] applied centre loss, together with softmax loss, can effectively enlarge the inter-class distance while penalizing the intra-class error to improve the SER model's performance.[20] also experimented with the centre and triple loss, which were proven to benefit the face recognition model's performance and the general SER model's accuracy.[28] used affinity loss to reinforce the discriminability of the features to reduce emotion ambiguity. Motivated by the success of using the advanced loss function to enhance the discriminability of the SER model, we apply the ArcFace loss[27] to maximize the separability of classes by an angular margin, borrowed from the face recognition field to check if it can help improve the SER model's performance. The main idea is to add an additive angular margin penalty between the embedding features and the corresponding weight in the last classification layer, thus compacting the intra-class distance while increasing the inter-class discrepancy. This has led to better performance and less computationally expensive than other loss functions. We present the details of ArcFace loss in section "ArcFace loss".

To address these issues, we propose a multi-dilated CNN framework with the SPP and the ArcFace layer. Inspired by[17,21], we used a 3D spectrogram to train our model to enhance performance compared to the raw (1D) spectrogram.

Proposed framework

This section discusses our proposed SER model, including the speech features for training, the multi-dilated CNN structure, SPP, and the ArcFace layer. The main workflow of the proposed framework can be described as follows. First, similar to[17,21], we generate the spectrogram from raw speech files and compute its delta and the delta of the delta to concatenate into a 3D spectrogram. Second, the 3D spectrogram, after standardization, is sent to our novel multi-dilated CNN structure to extract emotional features. Third, the SPP is deployed to obtain the most prominent features in multiple scales, followed by the emergence of these features and passing through the attention mechanism. Finally, after an initial FC layer, the dimensionally reduced feature vector is sent to the ArcFace layer to perform emotion classification. This framework is illustrated in Fig. 2.

3D Log-Mel spectrogram

Spectrogram learning has been quite popular in previous SER studies, as mentioned in section "Relatedwork". The spectrogram is a 2D representation of the speech energy across different frequency bands over the whole period[30]. It is extracted based on two steps:

Step 1: Transform the audio signal from the time domain into the frequency domain using the Short-Time Fourier Transform (STFT). The main objective is to incorporate the frequency information, as the raw audio signal contains only the amplitude and the time features. The Fourier Transform (FT) is designed to replace any type of signal with a series of sinusoids of different frequencies and phases, to convert the amplitude information in the signal into the frequency representation, which can be described by:

$$f(t) = \frac{a_0}{2} + \sum a_n * \sin\left(n\omega t + \varphi_n\right) \quad or \quad \frac{a_0}{2} + \sum a_n * \sin\left(nwt\right) + \sum b_n * \cos\left(n\omega t\right) \tag{1}$$

$$F_t = \int_{-\infty}^{+\infty} f(t) * e^{-j\omega t} * dt \tag{2}$$

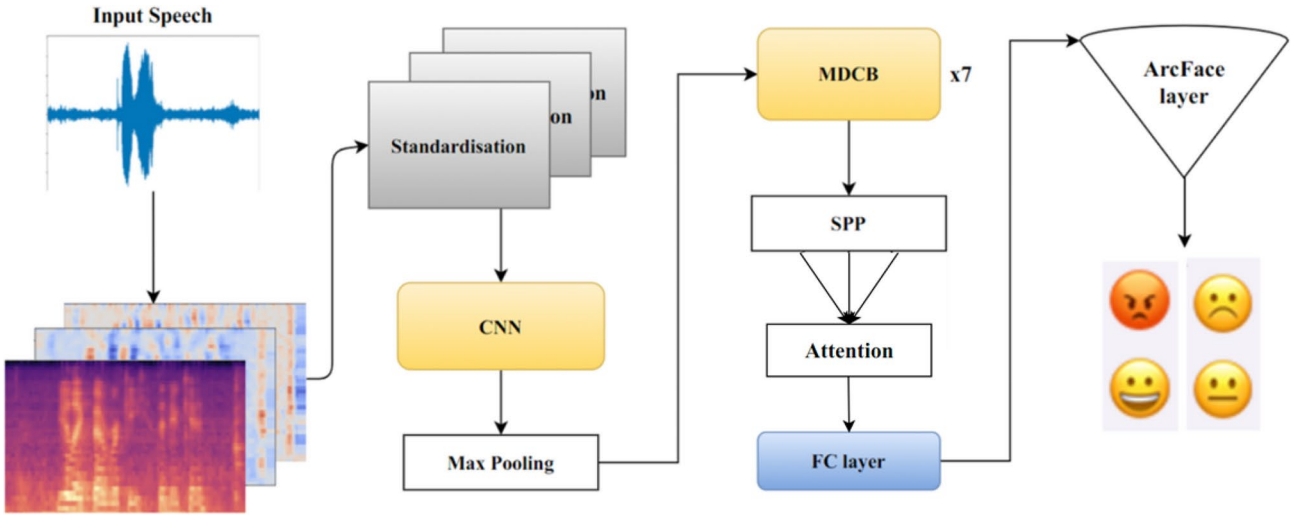

Figure 2. The multi-dilated CNN, SPP, and ArcFace SER framework.

where *f*(*t*) is the amplitude. ω represents the frequency, and φ stands for the phase. The outcome of Eq. 1 will be integrated into Eq. 2 by the orthonormal basis $e^{\left(-j\omega t\right)}$, which represents the signal points denoted by $cos\theta$ and $isin\theta$ in the complex number coordinate system. Furthermore, the STFT additionally considers the time information by recurrent multiplication of the signal with a sliding window while applying the FT to minimize stationarity[31].

Step 2: Scale the raw spectrogram of linear frequencies into their $log-mel$ ranges. This is a simulation of the human's hearing pattern, which does not follow a linear pattern, while a *mel-scale* can better reflect the hearing process by augmenting the small changes at low frequencies and alleviating them at high frequencies[32]. The core component of this conversion is the $mel-filter$ bank. It is created by transforming polar (the lowest and the highest) frequencies into the *mel-scale*, formulating the frequency band with evenly distributed points within the range of the *mel-scale* polar frequencies, inverse conversion of the *mel-scale* points into frequencies, the rounding of the frequency points to their nearest frequency bins, and the drawing of the triangular filters. The conversion between $mel-spectrum$ and the original frequency range can be described by Eq. 3:

$$f_{Mel} = 2595 * \log\left(1 + \frac{f}{700}\right) \tag{3}$$

Two important parameters in Step 1 and Step 2 are the length of the STFT and the number of filter banks. We set the length of the STFT as 512 (common length in STFT[33,34]), and the number of filter banks as 40 based on the outcome of[35] experiment. Before this, the hamming window with a width of 25ms and an overlap of 15ms is applied to frame the signal. Then, the signal is fed into the filter banks to realize the log-mel spectrum via Eq. 4:

$$LMSpec = \log\sum\left(FB(i) * X(i)^2\right) \tag{4}$$

where *FB*(*i*) stands for the output of the filter bank at the point i of the Fast FT (FFT). $X(i)^2$ is the power of the magnitude, in other words, the signal's energy.

In addition to the 40 log-mel spectrogram features, we also extract the delta and delta-delta features from the log-mel spectrum, following the approach described in[17,21]. However, we set $N = 9$ to capture dynamic information over a longer range, as our goal is to develop an utterance-level speech emotion recognition (SER) model rather than one based on shorter audio segments.

$$\Delta(LMSpec) = \frac{\sum_{k=1}^{N} k\left(LMSPec_{t+k} - LMSPec_{t-k}\right)}{2 * \sum_{k=1}^{N} k^2} \tag{5}$$

$$\Delta\Delta(LMSpec) = \frac{\sum_{k=1}^{N} k\left(\Delta(LMSPec)_{t+k} - \Delta(LMSPec)_{t-k}\right)}{2 * \sum_{k=1}^{N} k^2} \tag{6}$$

where t is the number of frames. Finally, the raw log-mel spectrum, its delta, and the delta of its delta are concatenated to form a 3D representation $X \in R^{(f,t,c)}$, where *f* represents the 40 $mel-filter$ banks, *t* is the time, and *c* is like the image *RGB* channel which consists of three feature maps: the spectrum, its delta, and the delta of the delta respectively.

Multi-dilated CNN and SPP

In this study, we propose a novel multi-dilated CNN structure for the SER task. Our choice of a CNN-based architecture, particularly the multi-dilated CNN structure, was driven by several key considerations related to SER's unique challenges. While RNNs and traditional DNNs are viable alternatives for processing sequential data, CNNs offer distinct advantages for processing speech spectrograms[36]. The 2D nature of spectrograms, representing frequency distributions over time, aligns naturally with CNN's ability to capture spatial relationships and hierarchical patterns. This is particularly relevant for emotion recognition, where emotional cues often manifest as specific patterns across both frequency bands and temporal sequences.

The dilated CNN (Fig. 3) is introduced by[37] to replace the traditional pooling layers for semantic segmentation. It has the advantage of extracting features from a larger receptive field while training and can well capture the fine details of the feature map. The important hyper-parameter of the dilated CNN is the dilation rate, which directly impacts the size of the receptive field. However, if stacking more dilated CNN layers, the dilation rates thereof should be well determined to avoid the gridding effect (Fig. 1). This issue is caused by passing the feature map through multiple dilated CNN layers, with the dilation rates having the highest common divisor except for 1. In this case, particular pixels in a feature map will always be ignored across multiple layers. To address this issue, we define the dilation rates of a stack of 3 CNN layers with various dilation rates.

Theoretical analysis of multi-dilated architecture

The effectiveness of our multi-dilated CNN architecture can theoretically be justified by analyzing the coverage of the reception field (RF). The dilation rates are determined following two criteria: one is avoiding having the greatest common divisor (excluding 1); two is having the same RF to the same number (3) layers of dilation rates of 2. The RF is calculated as follows:

$$RF = 2 * (K - 1) * \sum_{n}^{l} dr_l + 1 \tag{7}$$

where K stands for kernel size. dr_l is the dilation rate of the l_th convolution layer. n is the total number of convolution layers that RF wants to calculate based on.

When the kernel size and the number of convolution layers are fixed, the RF is only influenced by the sum of the dilation rates across all the convolution layers. RF grows exponentially with layer depth L and this growth comes at the cost of the gridding effect, where some pixels in the receptive field are systematically ignored. Our multi-dilated approach addresses this through controlled dilation rate variation, suggested by[26,37] where d_i and d_j represent the dilation rates of different layers in our multi-dilated CNN architecture, and i, j are layer indices.

$$\gcd(d_i, d_j) = 1 \text{ for } i \neq j \tag{8}$$

This guarantees complete coverage of the receptive field while maintaining the benefits of enlarged RF. We demonstrate this through visualisation of attention maps in section "Discussion". Furthermore, it is easy to identify that a 3-convolution-layer block with dilation rates of 1, 2, and 3 will have the same RF as the one with the uniformed dilation rates of 2, as both of the two dilation-scheme have a summation rate 6. Thus, we establish a block of 3 dilated convolution layers with dilation rates of 1, 2, and 3 to replace the fixed learning rates of 2. However, some of our practical experiments show that the dilated convolution layers with a fixed learning rate

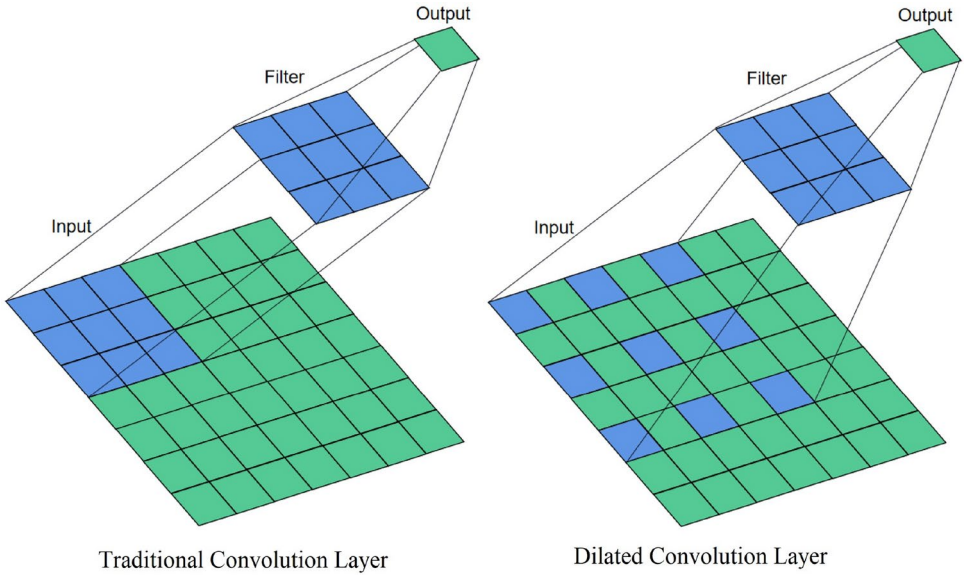

Figure 3. The traditional convolution layer versus the dilated convolution layer adapted from[38].

of 2 lead to better testing performance on some speakers of some sessions of the IEMOCAP dataset. This could be duedifferencesence between the speech spectrogram and the standard RGB image, which will be further discussed in section "Discussion". Therefore, we innovate a multi-dilation schema to incorporate the advantages of both dilation schemes.

The core component of our multi-dilation schema is the multi-dilation convolution block (MDCB), which comprises two dilated convolution layers in a single or multi-branch structure (Fig. 4). Each group of two dilated convolution layers in one block is a result of the Filter Factorization (FF) of a $5 * 3$ convolutional kernel. This

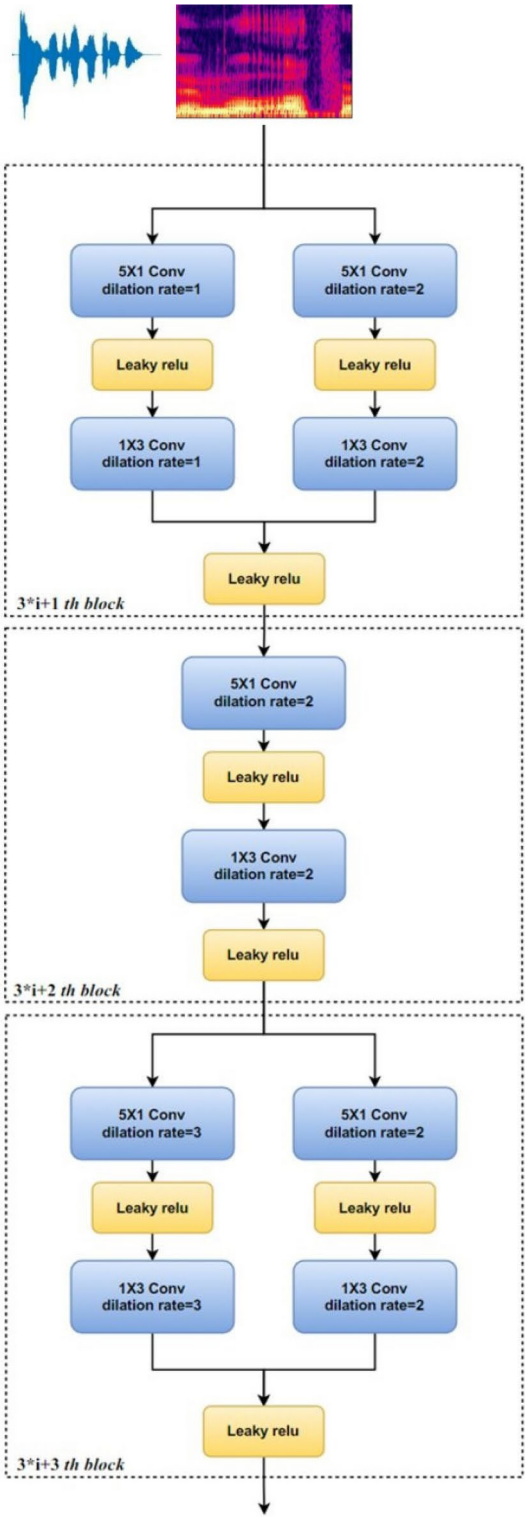

Figure 4. Multi-dilated CNN blocks.

treatment in our study is motivated by[39], which reduces computational complexity and leads to slightly better performance in multiple image classification tasks. This schaggregatesting the extracted feature maps foth the convolution layer with dilation rate 2 and the one with dilation rate 1 or 3. This is realized by defining the dilation rate iteratively according to the block's position. For the $3 * i + 1$ block, there are two branches of factorized convolution operations, with a dilation rate of 1 and 2, respectively. For the $3 * i + 2$ block, only one branch of the convolution operation with a dilation rate of 2 is deployed. The structure of the $3 * i + 3$ block is similar to the $3 * i + 1$ block, and the outputs of the two branches of these two blocks are merged to realize the combined learning effect, the only difference is the dilation rate 1 being replaced with 3. Molayers, the leaky rectified linear unit activation function is deployed after each convolution ope to incorporate the non-linear operation across different convolution layersration. There are 7 MDCBs in total for the whole convolution section.

Finally, the output from this section is passed through the SPP layer, which incorporates a multi-scale pooling strategy to extract different numbers of prominent features from each feature map. The SPP is introduced by[25] to optimize deep learning for images of varying sizes. Traditional deep CNNs rely on fixed image sizes, which affects recognition accuracy when faced with scale variations. Moreover, the FC layers, which often form the final classification section of the deep CNNs, can be only trained by the feature vectors of the fixed lengths. However, images of varied sizes passed through the CNN section will generate feature vectors of different lengths. To address these two problems, a pooling strategy is used between the CNN section and the FC layers to ensure the extracted feature vector has the unformed size according to the particular schema, e.g., max, and average. The traditional pooling strategy is based on a global level over the whole feature map, like global maximal pooling. This can address the issue regarding the FC layer learning but is still useless in alleviating the issue associated with deep CNN learning. Regarding this, the SPP strategy is designed to address these two problems. The SPP will extract the prominent feature vectors based on different pooling scales on the entire feature map, which is different from the global pooling schema. A comparison between the global pooling and the SPP with (2, 2) schema is shown in Fig. 5. The SPP (2, 2) means partitioning the feature map into $2 * 2$ grids and then applying the pooling schema for each grid. Generally, if the partition is (k, j), the number of feature vectors extracted from the SPP is $k * j$. If $k = j = 1$, the SPP is the same as the global pooling, which is an extreme case. The SPP strategy uses a group of spatial bins of multiple levels to split the feature map into divisions from finer to coarser levels, which can extract a variety of local information rather than global pooling.

For the SER task, a similar issue arises, as described above in the context of image recognition. The utterance-level spectrogram has various sizes for different speeches, as their periods are different. The SPP is expected to address the issue by extracting a series of feature vectors from different scales of frequency-time bins. This enables the classification section (FC layers) to incorporate a broader range of localized information derived from the extracted feature map.

Attention

To guide the feature extraction in the SPP layer, we also add a simple attention layer between the SPP and FC layers. The attention model has been proposed by[40] and is to follow the rationale of:

$$v = \sigma(w^T \cdot x + b) \tag{9}$$

$$alpha = Softmax(v \cdot u) \tag{10}$$

$$y = x \cdot alpha \tag{11}$$

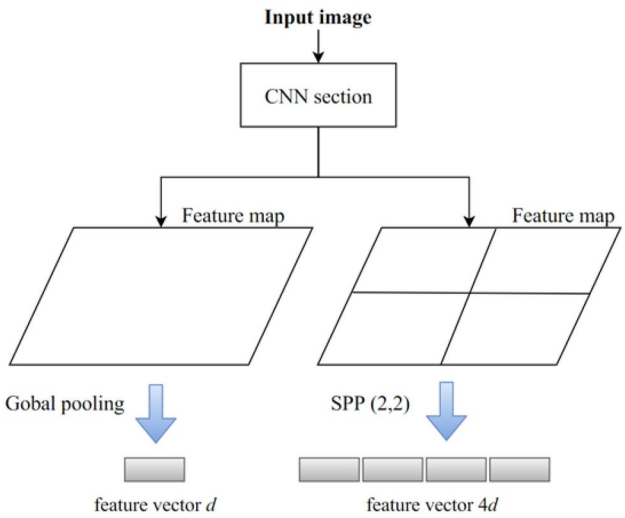

Figure 5. Global pooling versus SPP.

where *w* and *b* are the weight and bias vectors to measure the input feature importance. *u* is the context vector for computing the normalized importance weight *alpha*. The output *y* results from applying the importance weight alpha onto the input $x \cdot \sigma$ and *alpha* are the *sigmoid* and *Softmax* activation functions, respectively.

ArcFace loss

As mentioned in section two, traditional softmax loss functions have been widely used in deep learning while they face fundamental limitations in tasks requiring fine-grained feature discrimination such as face recognition and SER. The conventional softmax loss cannot explicitly optimize feature embedding to promote intra-class compactness while maximizing inter-class separation simultaneously[27]. This limitation is particularly critical in SER, where emotional expressions can have subtle variations within the same class (e.g., different intensities of "happy") or sharing some acoustic characteristics across classes (e.g., high arousal in both "angry" and "excited" states).

To address this issue, we treated speech signals analysis as an image analysis. For this, as discussed in section "3D Log-Mel spectrogram", we converted the speech signal into the 2D presentation of a Spectogram (image of the speech energy across different frequency bands over the whole time period—Fig. 2). Then we adopted ArcFace loss for SER, motivated by several theoretical considerations. First, like face recognition, SER requires the model to capture fine-grained discriminative features that can distinguish between closely related classes while being robust to intra-class variations[41,42]. Second, both domains benefit from learning angular relationships in their feature spaces. For example, in face recognition, this relates to facial geometry, while in SER, it corresponds to the distribution of spectro-temporal patterns characteristic of different emotions.

The ArcFace loss achieves this through its innovative approach to feature space optimization. Its ingenuity transforms the dot product between the feature vector extracted by the deep CNN and the weight of the output FC layer into a cosine distance. This transformation allows for identifying the angle between the feature vector and the target weight of the ground truth label by an arc-cosine function. Then, an angular margin can be added to the angle to realize the separability of the classification boundary.[27] provided an intuitive comparison between the softmax loss and ArcFace regarding their decision boundaries in a binary classification case. This type of decision boundary separation can effectively increase the inter-class distance while compacting the $intra - class$ error.

The implementation of the ArcFace loss was mainly based on modifying the FC output layer of a deep CNN model. The aim of the modification was to derive the angle between the input feature vector and the target weight in the last FC layer to realize the additional angular margin. The modified FC layer was called the ArcFace layer. It was based on the calculation of the logit of the output layer:

$$logit = W_j^T * x_i = \left|\left|W_j\right|\right| * \left|\left|x_i\right|\right| * \cos\theta_j \tag{12}$$

where x_i is the input feature. W_j is the weight of the *FC* layer. θ_j is the angle between them. To obtain θ_j and apply the additive penalt, several transformations are required to convert the last *FC* layer into the *ArcFace* layer. First, the feature vector x_i and the last *FC* layer's weight W_j are scaled with l_2 normalization to ensure $\left|\left|x_i\right|\right| = 1, \left|\left|W_j\right|\right| = 1.$

Next, the bias of the last FC layer is set to 0 to make the logit equal to θ_j. To calculate θ_j, the arccosine of the logit is taken. A margin penalty m is then added to the target angle θ_{y_i}, which represents the angle between x_i and its ground truth weight W_{y_i}.

The cosine of the adjusted angle θ_j, including θ_{y_i}, is taken to transform it back to logits. Finally, the logits are multiplied by a feature scale *s* to compensate for the scale loss due to l_2 normalisation:

$$logit = \begin{cases} S * \cos\theta_j \\ S * \cos(\theta_{yi} + m) \end{cases} \tag{13}$$

The logit from the ArcFace layer will be used to calculate the softmax loss:

$$L = -\frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{e^{s\cdot\cos\left(\theta_{yi}+m\right)}}{e^{s\cdot\cos\left(\theta_{yi}+m\right)}} + \sum_{j=1,j\neq y_i}^{c} e^{s\cdot\cos\theta_j}\right) \tag{14}$$

where *B* is the batch size, *c* is the number of classes for prediction. The larger *m* is, the more angular penalty will be added to the training set. It will significantly influence the final performance. *s* is the loss scale factor; the higher it is, the larger the loss it generates and thus the larger gradients of the model weights will result. Due to time limitations, exhaustively experimenting on these two hyper-parameters is impossible. So, after several experiments on different combinations of the two hyper-parameters and observing the respective training curve, we determined to use $m = 0.5$, which is the default value used in[27]. However, we define the feature scale $s = 2$, considering slowing and thus stabilizing the training process for the class imbalance training.

Experiments

The experimental setup, dataset, procedure, and results are discussed in this section.

Datasets

We evaluated our proposed SER framework using two datasets: IEMOCAP[43] and Berlin Database of Emotional Speech (EMODB)[44]. The IEMOCAP dataset comprises 10,039 utterances in dyadic sessions (five sessions with two actors each). Each session has a pair of male and female actors generating the speech data based on two methods. One method is to act according to the theatrical scripts; the other is acting in improvised affective scenarios. Each utterance in the dataset is labeled in seven emotions—*angry, excited, frustrated, happy, neutral, sad, and surprise*—by at least 3 human annotators. The average duration of the speeches is 4.5 seconds with 16kHz sampling rate. The EMODB contains 535 audio recordings from 5 male and 5 female actors. These recordings are all from daily life and are predefined by seven emotion labels, including *joy, anger, fear, sadness, disgust, boredom, and neutral*. The sampling rate of the EMODB speech data is 44.1kHz, which is down-sampled to 16kHz in all the experiments. In our study, only the improvised utterances were selected for all the experiments. That is, all the theatrical scripts were withdrawn and we only considered the speech data for four emotions: *angry, happy, sad, and neutral* in the IEMOCAP dataset similar to some previous SER studies[45–47]. On the other hand, all the utterances of the seven emotion classes in the EMODB were adopted in our experiments.

Experiment setup

In all our experiments, the $3D$ spectrogram extracted from the utterance level was used, as we wanted to maximally preserve the natural audio information. Hence, we neither employ uniform length segmentation (with zero-padding for shorter samples) as done by[17,21] nor did we divide the entire signal into overlapping segments as suggested in[22]. The raw spectrogram, its delta, and the delta of the delta are standardized by their respective global mean and standard deviation. The utterance-level $3D$ spectrogram was sent to our proposed multi-dilated CNN structure, passed through the SPP, and finally classified by the *FC* layer and the *ArcFace* layer. All experiments were carried out with the *NVIDIA* Graphics card *RTX* 3080. The model was implemented with the *Tensorflow* framework[48].

Model hyper-parameter determination

Given the complexity of the model and the high number of involved hyper-parameters, exhaustive fine-tuning to determine their values was almost impractical. Therefore, we conducted experiments to identify the main structural components:

1. The number of convolution layers (not to be filter factorized), including the initial convolution layers. We experiment on $4, 6, 8, 10, 12, 16, 20 - layer$ the structure of the plain CNN model for the SER. The 8-layer model resulted in the best performance. Then, we upgraded the plain model gradually to form our proposed model.
2. The number of partitions in the SPP layer. We experimented with global pooling, together with the group of two or three values in 2, 5, 6, 8, and 10 for height and width divisions, for a few epochs to observe the training curves. Then, we selected the most promising combinations and "ran" the whole training. The combination of the global pooling with the 5, 10 partitions resulted in the best validation performance and was selected by us for all the SPP experiments.
3. For m and s of the *ArcFace* layer, we experimented with $m = 0.3, 0.5,$ and 0.7; $s = 60$ (the default value in[27]), 2, and 1.5. By observing the training curves, we found when $m = 0.5$ and $s = 2$, the model training is most stable and promising.

For the number of filters in the convolution layers, we set it as 128, which followed the configuration in many previous SER studies[49,50]. The FC layer after the SPP layer was set to have 32 neurons.

Speaker-independent experiment

We exclusively conducted all the experiments in speaker-independent (SI) mode to address the generalizability challenge indicated in[52,52]. In contrast, the speaker-dependent (SD) style allows the model to learn specific characteristics of individual speakers during training, potentially making it easier to achieve high accuracy on the test data. Thus, evaluating a model's performance in the SI style provides a more rigorous assessment of its ability to generalize to unseen speakers. The SI experiment can also reflect the effectiveness of our model in practical scenarios, as it is impossible to gain the speech of all people in the world to train an SER model. For the IEMOCAP dataset, we trained the model with the speech data from four sessions and used the data of one actor in the rest session as the validation set and the other as the testing set. For EMODB, we also used the speech of eight out of the ten actors as the training set, one as the validation, and one as the testing set. All experiments were performed in a 10-fold cross-validation style and the average UA was reported.

We used UA as the metric to evaluate the model's performance. This is because the speech data in the IEMOCAP dataset is highly imbalanced[53]. For example, the quantity of ambiguous samples between anger and frustration is abundant, while that between anger and sadness is very few. It is more reliable to evaluate the model's performance based on the average accuracy across different emotion classes. The model is trained until the UA reaches its best on the validation set. The testing UA is then reported based on the trained model.

Furthermore, to reflect the effectiveness of the techniques we proposed in section "Proposed framework", an ablation study was conducted to compare the models with and without the techniques. We first developed an 8-layer plain CNN structure with the Global Maximal Pooling (GMP) and 2 FC layers, trained with the Softmax loss function as the baseline. Then we gradually applied the Filter Factorization (FF) to the convolution kernel, substituted the GMP with the SPP, upgraded the normal convolution kernel with a dilation strategy of a fixed rate 2, used the multi-dilated convolution structure, and replaced the last FC with the ArcFace layer. The UA of

CNN section	Pooling	Loss function	Unweighted accuracy	
			IEMOCAP	EMODB
Traditional CNN	Global Maximal Pooling (GMP)	Softmax	59.6	67.7
CNN with Filter Factorization (FF)	GMP	Softmax	60.3	62.7
CNN with FF	Spatial Pyramid Pooling (SPP)	Softmax	66.3	77.8
Dilated CNN (dilation rate = 2) with FF	SPP	Softmax	66.7	75.8
Multi-dilated CNN with FF	SPP	Softmax	67.5	75.8
Multi-dilated CNN with FF	SPP + Attention	Softmax	67.7	73.2
Multi-dilated CNN with FF	SPP + Attention	ArcFace	67.9	72.1
Multi-dilated CNN with FF	SPP	ArcFace	-	77.6

Table 1. The unweighted accuracy for our proposed SER framework in the ablation study of the IEMOCAP and EMODB datasets.

Research works	Unweighted accuracy	
	IEMOCAP	EMODB
[22]	63.98	-
[23]	63.9	-
[21]	64.74	82.82
[17]	69.32	85.39
[54]	66.5	-
[55]	66.1	-
[2]	68.96 (spectrogram + waveform)	-
	63.13 (only spectrogram)	-
Our model	67.9	77.6

Table 2. The comparison between our model and other SER models across both the IEMOCAP and EMODB datasets.

the above models for the two datasets is shown in Table 1. We also compare our model and the models in other SER studies based on the spectrogram learning in Table 2.

Model training

For training the model, we considered two important criteria. One was the imbalanced training, as the IEMOCAP dataset is highly imbalanced. The model was expected to pay more attention to the minority class and pay less attention to the majority class. We followed the screw-robust training strategy mentioned in[56] to assign class weights to the training losses. The class weight r_c is determined based on:

$$r_c = \alpha * \frac{N}{N_c} \tag{15}$$

where N is the total number of samples. N_c stands for the number of samples under class c. α is a scale factor. To optimize the training process, we use the Cyclical Learning Rate (CLR) strategy[57] to determine the learning rate. This was introduced to avoid the issue of a fixed learning rate training process. Too large a learning rate can lead to the fluctuation of the optimizer around the optimal loss point and thus, the model cannot converge or even step towards the undesired direction.

The minimal learning rate slows the training process, and the model can easily be trapped in the local minimal point. The CLR method runs an initial number of epochs with a series of respective learning rates in each epoch to observe the training curves. The learning rate leading to the obvious increase of the UA and the one resulting in the decrease of the UA are selected as the minimal and the maximal learning rates. Then, train the model with an iterative series of the learning rates between the minimal lr_{min} and maximal lr_{max} learning rates according to:

$$lr_t = lr_{\min} + (lr_{\max} - lr_{\min}) \cdot \max(0, (1 - x)) \tag{16}$$

where lr_t is the learning rate in t_{th} batch x can be defined by:

$$x = \left| \left| \frac{t}{stepsize} - 2 * cycle + 1 \right| \right| \tag{17}$$

where the *cycle* is the number of batches in a cycle of learning rates change.

$$cycle = \left\lfloor 1 + \frac{t}{2 * stepsize} \right\rfloor \tag{18}$$

where *stepsize* is defined as:

$$stepsize = \left[\frac{N}{Batchsize} \right] \tag{19}$$

where N is the number of training samples, and batch size is the number of samples in one training batch.

Discussion

Unlike existing dilated CNN approaches that often suffer from information loss due to "gridding" effects, in this research, we introduce a multi-dilated convolution block. This block dynamically adapts dilation rates to preserve semantic information and prevent loss of pixel information. Unlike existing approaches using dilated CNNs that often experience information loss due to "gridding" effects, our research introduces a multi-dilated convolution block. This block dynamically adjusts the dilation rates to preserve semantic information better and prevent pixel data loss. This method transforms feature extraction in speech spectrograms while directly addressing the challenge of maintaining information integrity in dilated convolution networks.

Furthermore, we incorporate Adaptive Spatial Feature Extraction through Intelligent Pooling, also known as SPP, and utilize ArcFace loss-a technique originally developed for face recognition. This combination represents a paradigm shift in how we conceptualize emotion as a computational problem, applying insights from biometric recognition to enhance our understanding of speech emotion.

From the experimental results, we can identify that the models gain significant improvement after replacing the GMP with the SPP. This could be associated with the functionality of the pooling layer. The GMP only extracts one prominent feature vector across the whole feature map generated from the CNN section during the feed-forward operation. This forces the CNN section to update the weights only relevant to the prominent feature vector in the back-propagation process. On the other hand, the SPP strategy, in our case, not only considers the feature at a global level but also considers multiple local features from finer to coarser levels. These local features on the feature map mean a series of time-frequency regions in various scales. This allows the model to select the emotion-related features from multiple regions of different ranges, as we do not know which time-frequency region of what size, or even the whole time-frequency map, in the speech spectrogram has the most emotional characteristics beforehand. Various feature vectors are extracted from the SPP based on the global pooling schema and the regions of multiple scales and are sent to the FC layer to automatically weigh the importance of these features for emotional classification purposes. Furthermore, when we investigate different SPP schemes, the selection of the pooling partitions substantially affects the final performance (Table 3).

We conducted the experiments using the plain 8-layer CNN structure to train and validate the IEMOCAP dataset. Two informative observation results could be derived. First, the number of partitions in the SPP layer is preferably the divisor of the feature map dimensions. To demonstrate the point, a representative plot for four SPP schemes (all based on our final scheme) is shown in Fig. 6. In this figure, the training set is sessions 2 to 5, and the validation is the female actor in session 1.

In this research, the input feature map from the CNN section is 20 * t (the time length varies according to different speech lengths). When the 6- or 8-partition schema (which are not the divisors of 20) schema is added to the SPP, the training curve seems more unstable, especially for 6 as it introduces factor 3, than not using them. This means the model design could not capture the features well for classification. Then, we select the relative promising scheme SPP(1,2,5,10), SPP(1,2,5,8,10), and SPP (1,5,10) to do full training and present the results in Table 3. Compared to the other two schemes, the SPP(1,5,10) leads to the best results, as its training/validation curve suggests.

Second, a too large frequency-time region in the SPP may generate a loss for the model. When comparing the SPP(1,2,5,10) and SPP(1,5,10), the latter has a better result than the former. This could be due to separating the feature map of varied length (in time dimension) into two partitions, which is thetic.

For the dilation strategy in the CNN section, the multi-dilated schema leads to a better result than the fixed dilation rate one in general. The multi-dilated structure can both leverage the advantage of the dilated CNN (of fixed dilation rate) and avoid the grinding effect. Especially for the IEMOCAP dataset, which is a difficult dataset to achieve desirable results in SER studies, the multi-dilated model improves the UA by around 2% rather than the CNN with a fixed dilation rate. For the EMODB dataset, the two dilation schemes lead to the same average UA. The reason is inferred to be associated with the ineffectiveness of the dilated CNN for the EMODB dataset.

SPP configurations with different Pooling partitions	Unweighted Accuracy	Weighted Accuracy
1 (Baseline)	52.02	63.72
1, 5, 10	68.48	73.02
1, 2, 5, 10	66.51	70.23
1, 2, 5, 8, 10	65.89	67.91

Table 3. CNN with FF comparison and experiments on SPP strategies.

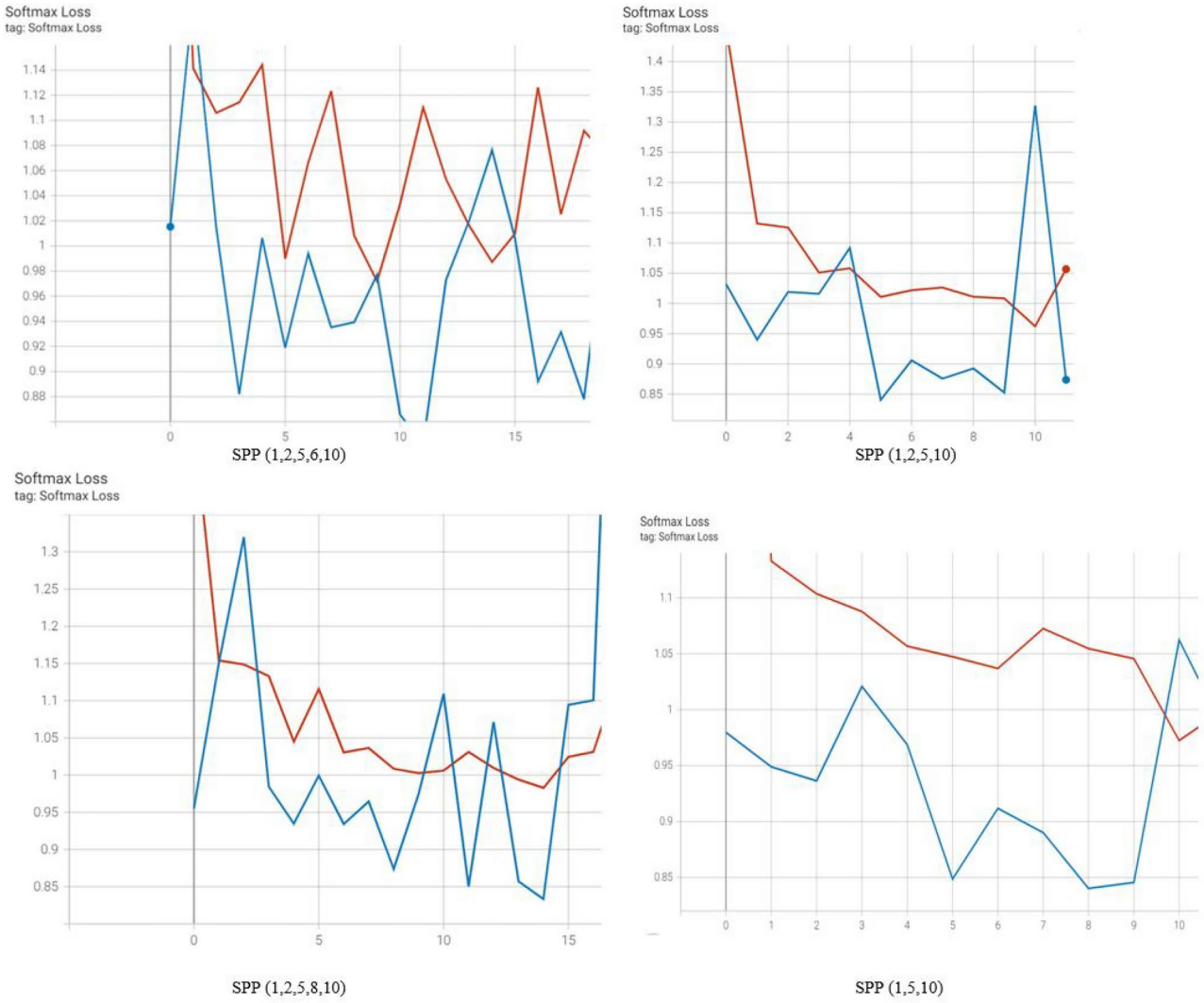

Figure 6. Performance comparison of 8-layer plain CNN with different SPP schema—training (red line)/ validation (blue line).

As Table 1 shows, the raw CNN kernel (with FF) results in a slightly higher UA than both of the two dilated CNN models, including the fixed dilation rate and the multi-dilation rates. Hence, using the latter one does not necessarily lead to better performance than the former one. On the other hand, the lower effectiveness of the dilated CNN for the EMODB dataset contradicts the similar results in[17], which demonstrates the dilation strategy can enhance the UA of the plain CNN up to 1.76% for the EMODB dataset. The reason could be due to data insufficiency and different data processing schemes.[17] segments the audio files into 3s, the shorter ones are zero-padded and then normalized by the mean and standard deviation. This could act as a data augmentation strategy to increase the sample size drastically. In our experiment, we use the original EMODB dataset and develop the SER model based on the utterance-level speech features. The total number of samples is 535, and the training samples we use to develop the model are 428 for 7 emotion classes. It is obvious that our model has more complexity (or deeper and wider structure) compared to that of[17], hence requiring more data to feed. Especially, the dilated CNN, which has a larger receptive field, probably requires more data to capture the desired pattern. Thus, when using the dilation strategy, both the fixed dilation rate and the multi-dilation rate scheme lead to worse performance than the plain CNN structure. Similarly, the attention mechanism leads to worse performance on the EMODB dataset, which is also estimated to be associated with the insufficient number of training samples. Thus, increasing the model complexity will be detrimental in this case. On the contrary, the IEMOCAP dataset has enough training size and thus results in better performance when adding the attention layer.

For the ArcFace layer, it successfully improves the model's performance for both of the two datasets. The ArcFace loss is applicable for the SER model based on spectrogram learning. It reduces the intra-class distance and increases the decision boundary between different classes to enhance the model's discriminability. It seems performing even better for the EMODB dataset, which has seven emotional labels. Although the dilation strategy reduces the UA of the model for the EMODB dataset, as discussed in the last paragraph, the ArcFace

layer drags the model's UA back by nearly 2%. As the paper of the ArcFace suggests, it was originally designed to address the ambiguity of recognizing the infinite number of human faces. This indicates it is naturally suitable for the recognition problem of more classes. In a nutshell, our framework proves the suitability of transplanting the ArcFace layer from the face recognition sphere to the SER area, and even better for more emotion-class recognition.

This model could be more suitable for practical use-cases as it is trained by utterance-level speech spectrograms of different sizes. Particularly, for the IEMOCAP dataset, the spectrograms originated from the audio clips from 0.58s to around 29s. This means there is also a huge discrepancy in the speech spectrogram sizes regarding the time axis. Our model can well learn from these spectrograms and make emotion classification to achieve the UA (0.679), which is close to that of[17], which achieves a great UA (0.69) among previous SER studies. The study also, to some extent, uses some data augmentation (as discussed in the paragraph before the last) to increase the sample size, which can improve the deep learning model's performance in common sense. Besides, our model also has higher performance than other previous models, as Table 2 shows. Especially compared with[2], which is a recent study that uses the well-known face recognition model - FaceNet and also trains the model only using the utterance-level spectrogram, our model leads to nearly 4.8% higher UA on the IEMOCAP dataset.

Conclusions and future work

In this study, we propose a novel SER model based on utterance-level spectrogram learning. Our model leveraged the advantage of the dilated CNN and well addresses the gridding effect. After using it to extract the feature maps from the spectrogram, the SPP layer is deployed to extract both the global-level prominent feature vector and multi-local-level feature vectors, followed by an attention model to weigh the feature vectors. Then, the FC layer selects the useful feature vectors and thus reduces the feature dimensionality. Finally, the ArcFace layer makes emotion classification based on the feature vectors, with the style of compacting the intra-class error and enlarging the inter-class distance. The final performance of our model is better than that of many previous SER studies and is comparable to that of the best SER model.

While our results are promising, we acknowledge several limitations in our current evaluation framework. Although IEMOCAP and EMODB are widely used benchmarks in the field, validation on additional datasets such as RAVDESS, CREMA-D, or MSP-IMPROV would provide more comprehensive insights into model generalization. The imbalanced nature of IEMOCAP's emotion classes also presents particular challenges for model evaluation, and while we addressed this through weighted training strategies, future work would benefit from more detailed per-class performance metrics. For the same reason, UA was chosen as the primary evaluation metric to have a fair evaluation across emotion classes. Therefore, future work would benefit from more detailed performance metrics, including confusion matrices and per-class performance measures, to better understand the model's strengths and weaknesses across different emotion categories. Additionally, cross-dataset evaluation, where models trained on one dataset are tested against others, would provide stronger evidence of generalization capabilities.

The model might also be further improved by incorporating the temporal learning model. As our study mainly focuses on extracting the emotional features based on the CNN-based model, the Recurrent Neural Net-based model, like Long-Short Term Memory, was not explored. Future research can also explore data augmentation strategies. This is associated with one limitation of our model on the EMODB dataset. Our MDCB even leads to slightly lower UA, as the EMODB has a lower training sample size. Therefore, augmenting the data can fully leverage our MDCB's capacity to capture the emotional feature. Finally, incorporating our proposed framework with linguistic information could possibly lead to better performance.

Data availability

The data that support the findings of this study are openly available in the EMO-DB repository at https://www.kaggle.com/datasets/revolutionarybukhari/iemocap reference number[44] and in the IEMOCAP repository at https://sail.usc.edu/iemocap/ reference number[43].

Received: 15 August 2024; Accepted: 3 March 2025
Published online: 10 March 2025

References

1. De Silva, U. et al. Clinical decision support using speech signal analysis: Systematic scoping review of neurological disorders. *J. Med. Internet Res.* **27**, 63004. https://doi.org/10.2196/63004 (2025).
2. Liu, S. et al. Speech emotion recognition based on transfer learning from the facenet framework. *J. Acoust. Soc. Am.* **149**(2), 1338–1345 (2021).
3. Zhu, L., Chen, L., Zhao, D., Zhou, J. & Zhang, W. Emotion recognition from Chinese speech for smart affective services using a combination of svm and dbn. *Sensors* **17**(7), 1694 (2017).
4. Mekruksavanich, S., Jitpattanakul, A., & Hnoohom, N. Negative emotion recognition using deep learning for thai language. In *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)* 71–74 (IEEE, 2020).
5. Fahad, M. S., Ranjan, A., Yadav, J. & Deepak, A. A survey of speech emotion recognition in natural environment. *Digit. Signal Process.* **110**, 102951 (2021).
6. Madanian, S., Parry, D., Adeleye, O., Poellabauer, C., Mirza, F., Mathew, S., & Schneider, S. Automatic speech emotion recognition using machine learning: Digital transformation of mental health (AIS, 2022).
7. Adeleye, A., Madanian, S., & Adeleye, O. Emotion variation detection in discrete english speech: A wavelet transform use case in mental health monitoring. In *Proceedings of the 2024 Australasian Computer Science Week. ACSW '24* 115–119 (Association for Computing Machinery, 2024). https://doi.org/10.1145/3641142.3641167.
8. Han, J., Zhang, Z., Ringeval, F., & Schuller, B. Reconstruction-error-based learning for continuous emotion recognition in speech. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2367–2371 (IEEE, 2017).

9. Huang, Z., Xue, W., Mao, Q. & Zhan, Y. Unsupervised domain adaptation for speech emotion recognition using pcanet. *Multim. Tools Appl.* **76**(5), 6785–6799 (2017).
10. Mirsamadi, S., Barsoum, E., & Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2227–2231 (IEEE, 2017).
11. Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service (PlatCon)* 1–5 (IEEE, 2017).
12. Xu, M., Zhang, F., Cui, X., & Zhang, W. Speech emotion recognition with multiscale area attention and data augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 6319–6323 (IEEE, 2021).
13. Madanian, S. et al. Speech emotion recognition using machine learning - a systematic review. *Intell. Syst. Appl.* **20**, 200266. https://doi.org/10.1016/j.iswa.2023.200266 (2023).
14. Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B.W. Survey of deep representation learning for speech emotion recognition. *IEEE Trans. Affect. Comput.* (2021)
15. Zhang, S., Zhang, S., Huang, T. & Gao, W. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans. Multim.* **20**(6), 1576–1590 (2017).
16. Guo, L., Wang, L., Dang, J., Zhang, L., & Guan, H. A feature fusion method based on extreme learning machine for speech emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2666–2670 (IEEE, 2018).
17. Meng, H., Yan, T., Yuan, F. & Wei, H. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access* **7**, 125868–125881 (2019).
18. He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016)
19. Wen, Y., Zhang, K., Li, Z., & Qiao, Y. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision* 499–515 (Springer, 2016).
20. Schroff, F., Kalenichenko, D., & Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 815–823 (2015)
21. Chen, M., He, X., Yang, J. & Zhang, H. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Process. Lett.* **25**(10), 1440–1444 (2018).
22. Luo, D., Zou, Y., & Huang, D.: Investigation on joint representation learning for robust feature extraction in speech emotion recognition. In *Interspeech* 152–156 (2018)
23. Zhang, Y., Du, J., Wang, Z., Zhang, J., & Tu, Y.: Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* 1771–1775 (IEEE, 2018).
24. Lazebnik, S., Schmid, C., & Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2169–2178 (IEEE, 2006).
25. He, K., Zhang, X., Ren, S. & Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916 (2015).
26. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., & Cottrell, G. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* 1451–1460 (2018). https://doi.org/10.1109/WACV.2018.00163.
27. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 4685–4694 (2019). https://doi.org/10.1109/CVPR.2019.00482.
28. Zhang, Y., Zou, Y., Peng, J., Luo, D., & Huang, D. Discriminative feature learning for speech emotion recognition. In *Artificial Neural Networks and Machine Learning-ICANN 2019: Text and Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV 28* 198–210 (Springer, 2019).
29. Lei, X., Pan, H. & Huang, X. A dilated cnn model for image classification. *IEEE Access* **7**, 124087–124095. https://doi.org/10.1109/ACCESS.2019.2927169 (2019).
30. Rojas, F., Madanian, S., Templeton, J. M., Poellabauer, C., & Schneider, S. L. Exploring deep learning and grad-cam for speech-based detection of mild traumatic brain injury. In *2024 IEEE International Conference on Big Data (BigData)* 6108–6116 (2024). https://doi.org/10.1109/BigData62323.2024.10825360.
31. Kerkeni, L. et al. Automatic speech emotion recognition using an optimal combination of features based on emd-tkeo. *Speech Commun.* **114**, 22–35 (2019).
32. Singh, C., Kumar, A., Nagar, A., Tripathi, S., & Yenigalla, P. Emoception: An inception inspired efficient speech emotion recognition network. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* 787–791 (IEEE, 2019).
33. Islam, M. R., Akhand, M. A. H., Kamal, M. A. S., & Yamada, K. Recognition of emotion with intensity from speech signal using 3d transformed feature and deep learning. *Electronics* **11**(15). https://doi.org/10.3390/electronics11152362 (2022).
34. Chen, X., Wu, L.-J., Mao, A., & Zhan, Z.-H. A new learning scheme of emotion recognition from speech by using mean Fourier parameters. In *2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI)* 96–101 (2019). https://doi.org/10.1109/ICACI.2019.8778548
35. Huang, Y., Wu, A., Zhang, G. & Li, Y. Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition. *IET Signal Processing***9**(4), 341–348. https://doi.org/10.1049/iet-spr.2013.0446. https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-spr.2013.0446 (2015).
36. Arias-Vergara, T. et al. Multi-channel spectrograms for speech processing applications using deep learning methods. *Pattern Anal. Appl.* **24**(2), 423–431. https://doi.org/10.1007/s10044-020-00921-5 (2021).
37. Yu, F., & Koltun, V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015).
38. He, J. et al. Bearing fault diagnosis via improved one-dimensional multi-scale dilated cnn. *Sensors* https://doi.org/10.3390/s21217319 (2021).
39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).
40. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1480–1489 (2016).
41. Harby, F., Alohali, M., Thaljaoui, A. & Talaat, A. S. Exploring sequential feature selection in deep bi-lstm models for speech emotion recognition. *Comput. Mater. Continua* **78**(2), 2689–2719. https://doi.org/10.32604/cmc.2024.046623 (2024).
42. Al-Tameemi, I. K. S., Feizi-Derakhshi, M.-R., Pashazadeh, S. & Asadpour, M. Multi-model fusion framework using deep learning for visual-textual sentiment classification. *Comput. Mater. Continua* **76**(2), 2145–2177. https://doi.org/10.32604/cmc.2023.040997 (2023).
43. Busso, C. et al. Iemocap: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **42**, 335–359 (2008).
44. Burkhardt, F. et al. A database of German emotional speech. *Interspeech* **5**, 1517–1520. https://doi.org/10.1007/s10579-008-9076-6 (2005).
45. Fahad, M. S., Deepak, A., Pradhan, G. & Yadav, J. Dnn-hmm-based speaker-adaptive emotion recognition using mfcc and epoch-based features. *Circuits Syst. Signal Process.* **40**, 466–489. https://doi.org/10.1007/s00034-020-01486-8 (2021).
46. Oflazoglu, C. & Yildirim, S. Recognizing emotion from Turkish speech using acoustic features. *EURASIP J. Audio Speech Music Process.* **2013**, 1–11. https://doi.org/10.1186/1687-4722-2013-26 (2013).

47. Lee, J., & Tashev, I. High-level feature representation using recurrent neural network for speech emotion recognition. In *Interspeech* (2015)
48. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016).
49. Middya, A. I., Nag, B. & Roy, S. Deep learning based multimodal emotion recognition using model-level fusion of audio-visual modalities. *Knowl.-Based Syst.* **244**, 108580. https://doi.org/10.1016/j.knosys.2022.108580 (2022).
50. Mountzouris, K., Perikos, I. & Hatzilygeroudis, I. Speech emotion recognition using convolutional neural networks with attention mechanism. *Electronics* https://doi.org/10.3390/electronics12204376 (2023).
51. Rybka, J. & Janicki, A. Comparison of speaker dependent and speaker independent emotion recognition. *Int. J. Appl. Math. Comput. Sci.* **23**(4), 797–808. https://doi.org/10.2478/amcs-2013-0060 (2013).
52. Shah Fahad, M., Ranjan, A., Yadav, J. & Deepak, A. A survey of speech emotion recognition in natural environment. *Digit. Signal Process.* **110**, 102951. https://doi.org/10.1016/j.dsp.2020.102951 (2021).
53. Liang, X., Jiang, H., Xu, W., & Zhou, Y. Gaussian-smoothed imbalance data improves speech emotion recognition. ArXiv:2302.08650 (2023).
54. Zhao, Z. et al. Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. *IEEE Access* **7**, 97515–97525 (2019).
55. Zhao, Z. et al. Self-attention transfer networks for speech emotion recognition. *Virtual Reality Intell. Hardware* **3**(1), 43–54 (2021).
56. Shih, P.-Y., Chen, C.-P., & Wang, H.-M. Speech emotion recognition with skew-robust neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2751–2755 (2017). https://doi.org/10.1109/ICASSP.2017.7952657.
57. Smith, L. N. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 464–472 (2017). https://doi.org/10.1109/WACV.2017.58.

Acknowledgements

This research is supported by the AUT Faculty of Design and Creative Technology contestable grant 2023 and the AUT School of Engineering, Computer and Mathematical Sciences Emerging Researcher Fund 2022.

Author contributions

SM conceptualised the research and advised on the applied research methodology. CT worked as the research assistant under SM supervision and guidance in this project. This research is the outcome of research collaboration among SM, OA, JT, CP and SS. CT prepared the first draft of the project report which was commented on and enhanced by SM. OA prepared the first draft of the manuscript which was commented on and enhanced by SM, JT, CP, EZ, and SS. SM addressed the feedback and prepared the final version of the manuscript for the submission.

Declarations

Competing interest

The authors declare no competing interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-92640-2.

Correspondence and requests for materials should be addressed to S.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.