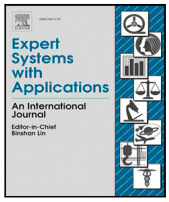
Expert Systems With Applications 230 (2023) 120608

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition

Nhat Truong Pham [a], Duc Ngoc Minh Dang [b], Ngoc Duy Nguyen [c], Thanh Thi Nguyen [d], Hai Nguyen [e], Balachandran Manavalan [a], Chee Peng Lim [f], Sy Dzung Nguyen [g,h,*]

[a] Computational Biology and Bioinformatics Laboratory, Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon, 16419, Gyeonggi-do, Republic of Korea
[b] Computing Fundamental Department, FPT University, Ho Chi Minh, Viet Nam
[c] Inveto Research, Brisbane, Queensland, Australia
[d] School of Information Technology, Deakin University, Victoria, Australia
[e] Khoury College of Computer Sciences, Northeastern University, Boston, USA
[f] Institute for Intelligent Systems Research and Innovation, Deakin University, Victoria, Australia
[g] Laboratory for Computational Mechatronics, Institute for Computational Science and Artificial Intelligence, Van Lang University, Ho Chi Minh City, Viet Nam
[h] Faculty of Mechanical–Electrical and Computer Engineering, School of Technology, Van Lang University, Ho Chi Minh City, Viet Nam

ARTICLE INFO

Keywords:
Speech emotion recognition
Mel spectrogram features
Generative adversarial networks
Attention mechanism
Dilated convolutional neural networks
Dilated recurrent neural networks
Long short-term memory
Hybrid data augmentation
Short-time Fourier transform

ABSTRACT

Recently, speech emotion recognition (SER) has become an active research area in speech processing, particularly with the advent of deep learning (DL). Numerous DL-based methods have been proposed for SER. However, most of the existing DL-based models are complex and require a large amounts of data to achieve a good performance. In this study, a new framework of deep attention-based dilated convolutional-recurrent neural networks coupled with a hybrid data augmentation method was proposed for addressing SER tasks. The hybrid data augmentation method constitutes an upsampling technique for generating more speech data samples based on the traditional and generative adversarial network approaches. By leveraging both convolutional and recurrent neural networks in a dilated form along with an attention mechanism, the proposed DL framework can extract high-level representations from three-dimensional log Mel spectrogram features. Dilated convolutional neural networks acquire larger receptive fields, whereas dilated recurrent neural networks overcome complex dependencies as well as the vanishing and exploding gradient issues. Furthermore, the loss functions are reconfigured by combining the SoftMax loss and the center-based losses to classify various emotional states. The proposed framework was implemented using the Python programming language and the TensorFlow deep learning library. To validate the proposed framework, the EmoDB and ERC benchmark datasets, which are imbalanced and/or small datasets, were employed. The experimental results indicate that the proposed framework outperforms other related state-of-the-art methods, yielding the highest unweighted recall rates of 88.03 ± 1.39 (%) and 66.56 ± 0.67 (%) for the EmoDB and ERC datasets, respectively.

1. Introduction

Speech emotion recognition (SER) is important in a variety of applications, such as e-learning, human–computer interaction, robotics, and mobile services (Cen, Wu, Yu, & Hu, 2016; Huahu, Jue, & Jian, 2010; Yoon, Cho, & Park, 2007). Over the last few decades, SER has become one of the most active research areas in speech processing. Feature extraction and classification are the two primary components of an SER systems. Identifying the most representative features and addressing the imbalanced labeling issue are the key challenges in SER.

Feature extraction and selection problems in SER have been investigated over the past decade. While researchers have attempted to discover the important features, it remains challenging to identify the most crucial ones. As such, the majority of existing research studies have used a variety of techniques and parameters for SER, including pitch, energy, zero-crossing rate (ZCR), formants, root mean square error (RMSE), and prosodic (Arias, Busso, & Yoma, 2014; Cao, Verma, & Nenkova, 2015; Chen, Mao, Xue, & Cheng, 2012; Dai, Han, Dai, &

* Corresponding author at: Laboratory for Computational Mechatronics, Institute for Computational Science and Artificial Intelligence, Van Lang University, Ho Chi Minh City, Viet Nam.
E-mail addresses: truongpham96@skku.edu (N.T. Pham), ducdnm2@fe.edu.vn (D.N.M. Dang), duy.nguyen@inveto.ai (N.D. Nguyen), thanh.nguyen@deakin.edu.au (T.T. Nguyen), nguyen.hai1@northeastern.edu (H. Nguyen), bala2022@skku.edu (B. Manavalan), chee.lim@deakin.edu.au (C.P. Lim), dung.nguyensy@vlu.edu.vn (S.D. Nguyen).

https://doi.org/10.1016/j.eswa.2023.120608
Received 19 January 2023; Received in revised form 20 April 2023; Accepted 27 May 2023
Available online 1 June 2023

Xu, 2015; Lee, Mower, Busso, Lee, & Narayanan, 2011; Wu & Liang, 2010), Mel-frequency cepstral coefficient (MFCC) (Albornoz, Milone, & Rufiner, 2011), linear predictive coding and log frequency power coefficients (Yeh, Pao, Lin, Tsai, & Chen, 2011). Furthermore, researchers have designed a variety of classifiers to discriminate emotions, such as the hidden Markov model (Albornoz et al., 2011), Gaussian mixture model (Albornoz et al., 2011; Wu & Liang, 2010), support vector machine (SVM) (Cao et al., 2015; Chen et al., 2012; Wu & Liang, 2010), k-nearest neighbors (Yeh et al., 2011), and Bayesian logistic regression (Lee et al., 2011).

Deep neural networks (DNNs), in conjunction with the advancement in deep learning (DL), have become an effective method for automatically extracting SER features as opposed to manual feature selection (Huang, Tian, Wu, & Zhang, 2019; Tzirakis, Zhang, & Schuller, 2018). In this regard, the time-frequency domain characteristics from spectrograms can be effectively extracted and utilized with deep convolutional neural networks (CNNs) and long short-term memory (LSTM) models (Issa, Demirci, & Yazici, 2020; Zhang, Zhang, Huang, & Gao, 2017; Zhao, Mao, & Chen, 2019). Several studies (Chen, He, Yang, & Zhang, 2018; Peng et al., 2020) have applied an attention mechanism to obtain the most representative utterance features that corresponding to certain emotions. However, the existing DL methods are typically complex, and require large labeled datasets to achieve a good performance.

Because each speech can stem from several confusing emotions, collecting and annotating data samples is costly and laborious. Consequently, both conventional and cutting-edge data augmentation techniques (Lalitha, Gupta, Zakariah, & Alotaibi, 2020; Qian, Hu, & Tan, 2019; Tiwari, Soni, Chakraborty, Panda, & Kopparapu, 2020) have been used to create and synthesize training data. Existing research studies employ either noise or over-sampling methods to produce and balance datasets or using generative adversarial networks (GANs) to process data from the feature space. However, most of these methods are susceptible to data sampling and sequential modeling issues.

In this paper, we present a hybrid data augmentation (HDA) method that combines conventional and GAN-based models to produce additional data samples based on a labeled dataset. A deep dilated convolution-recurrent neural network (DCRNN) is constructed using the three-dimensional (3D) log Mel spectrogram (MelSpec) low-level features as its inputs. To exploit the utterance-level characteristics, the deep DCRNN model extracts high-level features as the inputs to the attention layer. To facilitate classifying emotions in audio signals, both the SoftMax and center-based loss functions are merged. To demonstrate the HDA and emotion recognition capabilities, we employed the EmoDB (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005) and the ERC datasets for performance evaluation, respectively. Finally, for the purpose of reproducibility, we have made the implementation of our proposed DL-based framework would be publicly available at https://github.com/nhattruongpham/hda-adcrnn-ser. This would enable researchers and practitioners to replicate our experimental results and use our framework for their own research and applications.

The research contributions of this study are as follows:

(1) We propose a hybrid data augmentation (HDA) method based on time shifting, pitch shifting (Haghparast, Penttinen, & Välimäki, 2007; Lent, 1989), and WaveGAN (Donahue, McAuley, & Puckette, 2019) to generate new data samples from a given dataset;
(2) We devised a modified attention-based dilated convolutional and recurrent neural network (mADCRNN) model that combines dilated CNNs and dilated LSTM models with an attention mechanism to learn and extract utterance-level features from 3D log MelSpec low-level features. To the best of our knowledge, this combination is novel in the SER literature;
(3) We reconfigured the loss function by combining both the SoftMax and center-based losses to classify emotional speech signals from the original and augmented data samples. A systematic evaluation on the loss functions and data augmentation method is conducted;
(4) The empirical results confirm that the proposed DL framework comprising the mADCRNN and HAD outperforms state-of-the-art methods in SER tasks, with unweighted recall scores of 88.03% and 66.56% for the EmoDB and ERC datasets, respectively.

The methodology of our proposed DL-based framework is as follows:

- The HDA method, which consists of time shifting, pitch shifting (Haghparast et al., 2007; Lent, 1989), and WaveGAN (Donahue et al., 2019) is utilized to generate conditional data samples based on a given dataset;
- The mADCRNN model was devised by combining dilated CNNs and dilated LSTM models with an attention mechanism to learn and extract utterance level features from 3D log MelSpec low-level features;
- The loss function is reconfigured by combining both the SoftMax and center-based losses to classify emotional speech signals based on the original and augmented data samples;
- The experimental results on two benchmark datasets confirm that our proposed DL framework is useful and outperforms state-of-the-art methods in tackling SER tasks.

The remaining part of this paper is organized as follows: The related studies are presented in Section 2 of the literature presents the relevant research. The proposed method is described in Section 3. The experimental findings and performance comparisons are analyzed and discussed in Section 4. Finally, the conclusions and recommendations for further research are presented in Section 5.

2. Related studies

DL models, such as deep CNNs and long short-term memory (LSTM) networks, have been studied to extract features from spectrograms of raw audio data and identify emotions. Zhang et al. (2017) used CNNs to extract the 3D log MelSpec characteristics. Different emotions were identified using a support vector machine (SVM) classifier, and a discriminant temporal pyramid matching technique was developed to concatenate the learned segment-level information. Zhao et al. (2019) created 1D and 2D CNNs along with LSTM networks to perform SER tasks. Crucial sequence segments were grouped together by Sajjad, Kwon, et al. (2020) using an RBFN (radial basis function network). Using a bidirectional LSTM network, the CNN could extract representative features and related temporal data for SER from the selected sequences after being transformed into spectrograms. Meng, Yan, Yuan, and Wei (2019) presented a new architecture based on the attention mechanism, employing a dilated CNN with a residual block and bidirectional LSTM (i.e., ADRNN). A 3D log MelSpec was used to extract the features and train the ADRNN on the feature representation. Using a loss function comprising the center loss and the SoftMax losses, several emotion categories were established for the SER.

As not all features contribute equally toward recognizing emotions from speech signals, several recent studies have focused on the use of an attention mechanisms in SER. In Chen et al. (2018) and Meng et al. (2019), an attention-based LSTM model was developed to learn relevant high-level features for representing emotional states. Peng et al. (2020) devised a sliding recurrent neural network (RNN) with an attention mechanism was devised to extract segment-level features, relying only on the important emotional parts associated with speech features. Ho, Yang, Kim, and Lee (2020) adopted a self-attention scheme for an RNN to exploit the context information at each time step, while a multi-headed attention method was used to predict emotions.

To address imbalanced data and avoid overfitting, researchers have leveraged data augmentation methods to generate or synthesize additional data samples. Park et al. (2019) proposed a SpecAugment method that included various features, such as warping, frequency masking, and time masking, as inputs to a neural network. To increase the SER performance, Rebai, BenAyed, Mahdi, and Lorré (2017) suggested

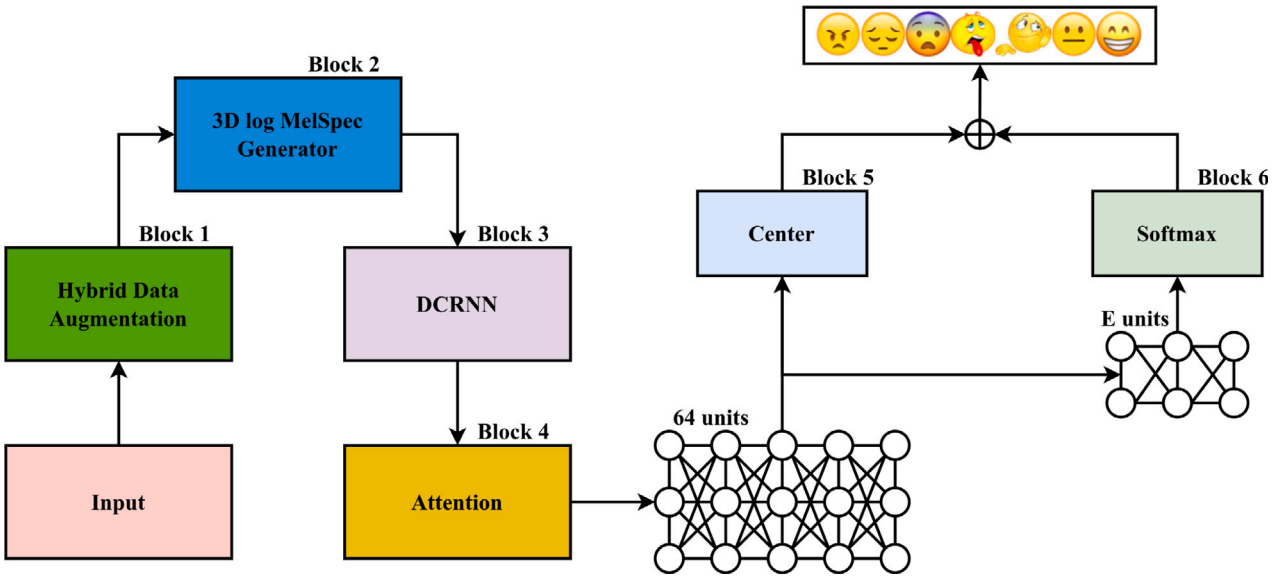


Fig. 1. Architecture of the proposed methods.

a new DNN architecture consisting of both data augmentation and ensemble methods. Recently, GAN-based methods have been used as a data augmentation approaches for SER. A GAN model, an autoencoder for feature extraction, and an auxiliary classifier for SER made up the adversarial data augmentation network (ADAN) by Yi and Mak (2020). By leveraging the cross-entropy loss function, the ADAN model was used for GAN training, producing feature vectors in both the original feature space and the latent space using the Wasserstein divergence approach. To enhance the SER performance, Bao, Neumann, and Vu (2019) explored a cycle-consistent adversarial network (CycleGAN) to transfer feature vectors from a sizable speech corpus without labels to the synthetic features of emotion styles. Whereas Alzubi et al. (2020a) proposed a modified version of GAN for processing text and identifying similar and dissimilar sentences using collaborative and adversarial learning. Denoted as collaborative adversarial network (CAN), the proposed model utilized a feature extractor to boost the relationships between sentences by obtaining similar characteristics within a word pair. CAN outperformed MaLSTM (LSTM with the Manhattan distance) and other state-of-the-art models in a series of evaluations using the Quora Question Pairs dataset.

3. Proposed approach

This study is motivated by the success of several existing methods, namely WaveGAN (Donahue et al., 2019), pitch-shifting (Haghparast et al., 2007; Lent, 1989), DL for handling 3D log MelSpec (Meng et al., 2019), and effective loss functions for SER, which include along with the contrastive-center (CT-C) loss and SoftMax losses proposed by Pham, Dang, and Nguyen (2020). By exploiting the advantages of these methods and alleviating their shortcomings, we formulated a new DL-based framework for tackling SER tasks, as follows:

- We leveraged WaveGAN, pitch shifting, and time shifting to devise an HDA method for generating and synthesizing data samples.
- We enhanced the ADRNN by removing batch normalization (BN) and utilizing a fully connected (FCN) layer to obtain the reconfigured loss functions.
- We validated the proposed DL framework comprehensively with a variety of loss functions, including the SoftMax loss, reconfigured center-based loss with SoftMax loss, SoftMax loss with center loss (Meng et al., 2019), and the CT-C loss with a SoftMax loss (Pham et al., 2020).

As depicted in Fig. 1, the baseline architecture consists of six elements: Hybrid Data Augmentation (Block 1), 3D log MelSpec Generator (Block 2), DCRNN (Block 3), Attention (Block 4), Center (Block 5), and SoftMax (Block 6). The information flow is as follows. Block 1 receives the speech signals and generates additional data samples in the latent space. The generated data samples are passed to Block 2 to extract the 3D log MelSpec features. The extracted features are fed into Blocks 3 and 4 for learning the representation of the 3D log MelSpec information and extracting the associated high-level features, which are then passed to Blocks 5 and 6 for the final classification.

To learn and extract high-level representations of the 3D log MelSpec low-level features for SER, we devised a deep mADCRNN model, as shown in Fig. 1. The ADRNN serves as the foundation for the deep mADCRNN model, but with all BN levels following the removal of dilated CNN layers. A dilated LSTM model was used instead of a bidirectional LSTM (BiLSTM) to address the recognized difficulties in training RNNs, including complicated dependencies, vanishing, and exploding gradient information, as stated in Chang et al. (2017). In addition, to compute the SoftMax loss function, we first reshaped the FCN layer to one with E units corresponding to E classes after calculating the center-based loss function using an FCN layer of 64 units. The center loss and SoftMax loss functions in the ADRNN were computed after forming an FCN layer of E units corresponding to E classes, which is the difference between our devised network and a standard ADRNN model. The baseline architecture shown in Fig. 1 was designed to operate as follows: Firstly, based on the HDA method, the audio signals are used to generate additional data samples. Secondly, an CNN layer was employed to process the low-level 3D log MelSpec features extracted by using a 3D log MelSpec generator. Thirdly, to extract temporal information, a residual block was utilized along with three dilated CNN layers. All feature maps were then used as inputs to the dilated LSTM networks to learn the consecutive features. To exploit the utterance-level characteristics from the sequential features, an attention layer was included. Finally, by combining the SoftMax loss and center-based loss functions, several loss functions are formulated for undertaking SER tasks.

3.1. The proposed hybrid data augmentation (HDA) method

HDA, which leverages both traditional and GAN-based methods, was employed in this study to generate additional data samples based on the original data or from the latent-space vector. Specifically, the

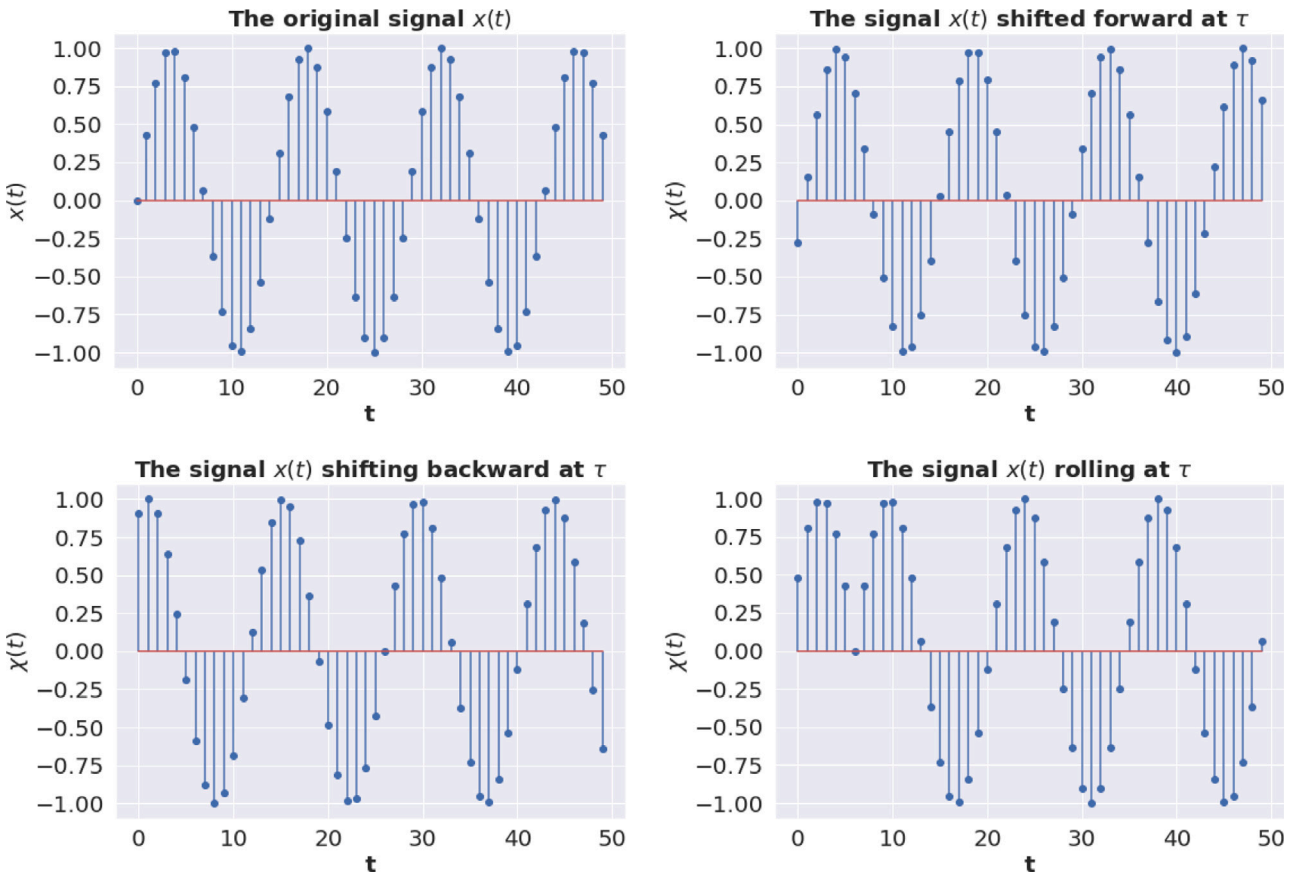


Fig. 2. Examples of signal time shifting and rolling.

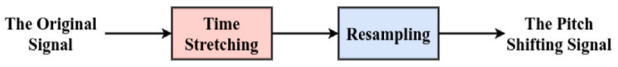


Fig. 3. The process of the pitch shifting.

traditional time shifting and pitch shifting methods as well as the GAN-based model for audio and speech signals, namely WaveGAN, were utilized in HAD, as detailed in the following sub-sections.

3.1.1. Traditional methods

Time shifting. Given a signal $x(t)$, its waveform can be shifted forward or backward by adding or subtracting a finite time, τ, respectively. Output $\chi(t)$,after shifting is:

$$\chi(t) = x(t \pm \tau), \tag{1}$$

where $\tau = sr/100$ and sr is the sampling rate of the signal.

With time shifting, we can only change the signal's location, but not its amplitude, forward or backward. In this study, we aim to roll and shift the signal along its time base. Examples of time shifting and rolling are show in Fig. 2.

Pitch shifting. Pitch shifting is an efficient technique developed by Lent (1989) based on time stretching and re-sampling (Haghparast et al., 2007), as shown in Fig. 3.

Time stretching was calculated by computing the time stretching ratio, S_{ratio}, as follows. Given the number of half-steps num_{hstep} and the number of bins num_{bins} in each octave section, time stretching is achieved by

$$S_{ratio} = 2^{-\left(\frac{num_{hstep}}{num_{bins}}\right)}. \tag{2}$$

while the resampling is obtained by computing the re-sampling ratio R_{ratio} as follows:

$$R_{ratio} = \frac{T_{sr}}{S_{sr}}, \tag{3}$$

where the sampling rates of the source and target signals, respectively, are S_{sr} and T_{sr}. The pitch-shifted signals are in either an accelerated or decelerated form, depending on whether R_{ratio} is greater than one or otherwise. For example, Fig. 4 depicts the pitch shifting $\Gamma(t)$ of the source signal $x(t)$ obtained using the Librosa library (McFee et al., 2015).

3.1.2. GAN-based method with waveform

Donahue et al. (2019) introduced the first GAN architecture for unsupervised audio synthesis. Denoted as WaveGAN, this method was built on a deep convolutional GAN architecture (DCGAN) (Radford, Metz, & Chintala, 2016) to produce images. To create WaveGAN, we modified DCGAN, as follows:

- Changing the DCGAN properties to support audio waves instead of just images;
- Using the 1D filters with a length of 25 instead of 5×5 2D filters;
- Increasing and using a stride of 4 instead of 2×2;
- Eliminating the BN mechanism in both generator and discriminator;

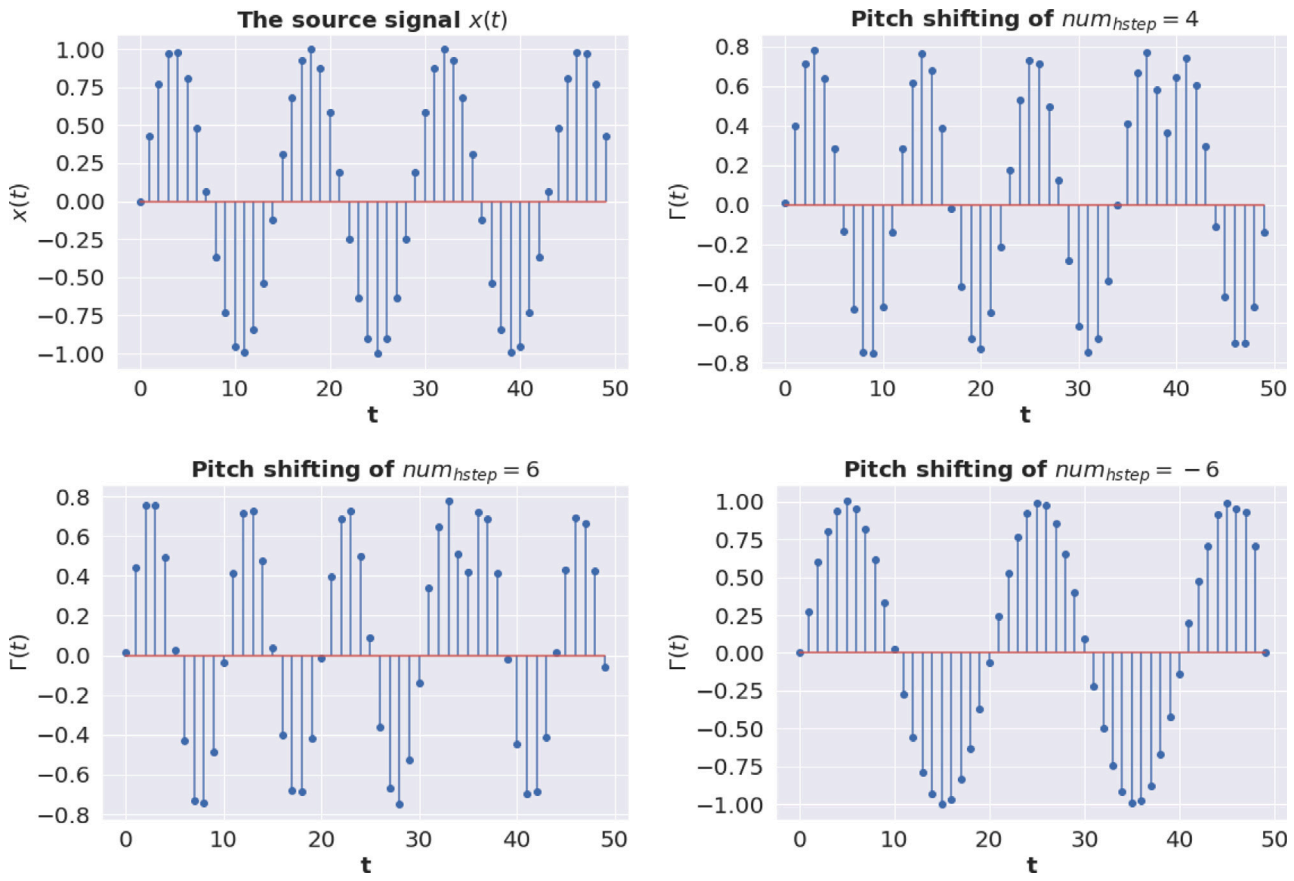


Fig. 4. Pitch shifting samples with different num_{hstep}.

- Leveraging the Wasserstein GAN and its variant with a gradient penalty (WGAN-GP) as proposed by Guyon et al. (2017) to achieve Lipschitz continuity during training;
- Employing only the phase shuffle operation in the discriminator to randomly perturb the phase of each layer using $[-n, n]$ samples, where n is a hyperparameter.

While the transposed convolution operation in the discriminator uses down-sampling instead of up-sampling, that in the generator is based on up-sampling. The output dimensions were the same as those of DCGAN. WaveGAN still has the same number of parameters, despite our modification. Table 1 provides a detailed explanation of the WaveGAN parameters presented by Radford et al. (2016).

3.2. 3D log MelSpec extraction

We used a deep mADCRNN model with the 3D log MelSpec, including low-level features as inputs. The static, delta, and delta-delta coefficients that comprise these 3D log MelSpec low-level features were derived as follows:

- Firstly, the audio samples were converted into MelSpec using a short-time Fourier transform (STFT) with a window length of 25 ms, an overlap between the successive windows of 10 ms, 40 filter banks, and a frame rate of 16 kHz. This produced 512 frequency bins, corresponding to a fast Fourier transform (FFT) size of 512, with linear spaces from 300 Hz to 8 kHz.
- Next, the static coefficients were obtained by scaling MelSpec logarithmically.
- Then, the delta coefficients were obtained by computing the derivatives of static coefficients.
- Finally, the delta-deltas coefficients were obtained by computing their derivatives of delta coefficients.

Figs. 5 and 6 depict the waveform signal, log MelSpec, delta, and delta-deltas coefficients of a sample of the sad emotion from the original and WaveGAN-processed EmoDB datasets, respectively. Figs. 7 and 8 show the waveform, log MelSpec, delta, and delta-deltas coefficients of another sample of the angry emotion from EmoDB, with time shifting and pitch shifting operations, respectively.

3.3. Deep learning framework

The proposed DL framework comprises three main components: a deep mADCRNN architecture and attention-based layer, and loss function. The deep mADCRNN model was designed to extract high-level representative features from the 3D log MelSpec information. The attention-based layer was applied to the deep mADCRNN model to emphasize important features for capturing emotions, whereas the loss function was used to effectively discriminate different emotions. The details of each component are explained as follows:

3.3.1. Deep mADCRNN architecture

The deep mADCRNN model was used to learn and extract the high-level representations from the 3D log MelSpec low-level features. It consisted of three dilated CNN layers with a skip-dilated CNN connection, one linear layer, two dilated LSTM layers, and one regular CNN layer. The first CNN layer has a stride of 1, a 3×3 kernel size, 128 feature mappings, and a *valid* padding. Each dilated CNN layer comprised 256 feature maps with a 3×3 kernel size and the *same* padding. In this study, the dilation rate was specified as a list of (1, 2) for the dilated LSTM network, and as a value of 2 for the dilated CNN model. To down-sample the feature maps, only the max-pooling layer is added after the first CNN layer. The max-pooling layer was formed with a 2×4 kernel size, 128 feature mappings, a 2×4 stride, and *valid* padding. We added a 512-output linear layer before fitting all the

Table 1
Description of the WaveGAN architecture (Donahue et al., 2019).

WaveGAN Architecture					
Generator			Discriminator		
Operation	Kernel size	Output shape	Operation	Kernel size	Output shape
Input z ~Uniform(−1,1)	—	(Bs, 100)	Input x or G(z)	—	(Bs, 16384, C)
FCN	(100, 256D)	(Bs, 256D)	Conv1D 1 (S = 4)	(25, C, D)	(Bs, 4096, D)
Reshape	—	(Bs, 16, 16D)	LeakyReLU 1 (β = 0.2)	—	(Bs, 4096, D)
ReLU 1	—	(Bs, 16, 16D)	Phase Shuffle 1 (BS = 2)	—	(Bs, 4096, D)
Transpose Conv1D 1 (S = 4)	(25, 16D, 8D)	(Bs, 64, 8D)	Conv1D 2 (S = 4)	(25, D, 2D)	(Bs, 1024, 2D)
ReLU 2	—	(Bs, 64, 8D)	LeakyReLU 2 (β = 0.2)	—	(Bs, 1024, 2D)
Transpose Conv1D 2 (S = 4)	(25, 8D, 4D)	(Bs, 256, 4D)	Phase Shuffle 2 (BS = 2)	—	(Bs, 1024, 2D)
ReLU 3	—	(Bs, 256, 4D)	Conv1D 3 (S = 4)	(25, 2D, 4D)	(Bs, 256, 4D)
Transpose Conv1D 3 (S = 4)	(25, 4D, 2D)	(Bs, 1024, 2D)	LeakyReLU 3 (β = 0.2)	—	(Bs, 256, 4D)
ReLU 4	—	(Bs, 1024, 2D)	Phase Shuffle 3 (BS = 2)	—	(Bs, 256, 4D)
Transpose Conv1D 4 (S = 4)	(25, 2D, D)	(Bs, 4096, D)	Conv1D 4 (S = 4)	(25, 4D, 8D)	(Bs, 64, 8D)
ReLU 5	—	(Bs, 4096, D)	LeakyReLU 4 (β = 0.2)	—	(Bs, 64, 8D)
Transpose Conv1D 5 (S = 4)	(25, D, C)	(Bs, 16384, C)	Phase Shuffle 4 (BS = 2)	—	(Bs, 64, 8D)
Tanh	—	(Bs, 16384, C)	Conv1D 5 (S = 4)	(25, 8D, 16D)	(Bs, 16, 16D)
			LeakyReLU 5 (β = 0.2)	—	(Bs, 16, 16D)
			Reshape	—	(Bs, 256D)
			FCN	(256D, 1)	(Bs, 1)

feature maps into the dilated LSTM network to efficiently minimize the parameters. As each LSTM cell contained 512 units, we produced high-level representations with 512 consecutive dimensions. To enhance training efficiency, we used an additional BN layer following the linear layer. Fig. 9 illustrates the deep DCRNN architecture.

3.3.2. Attention-based layer

Not all sequential high-level representations contribute equally toward capturing emotions in speech signals. Therefore, after extracting the high-level representations, an attention-based layer is added to leverage the utterance-level characteristics for SER. A BiLSTM with an attention layer is defined as follows:

$$Att = \sum_{t=1}^{T} \alpha \times h_t, \tag{4}$$

where Att is the attention output, $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$ denotes the hidden state of the BiLSTM output at time step t, T is the total number of time steps, and α is the normalized attention weight computed as follows:

$$\alpha = \frac{\exp(W \cdot h_t)}{\sum_{j=1}^{T} \exp(W \cdot h_t)}, \tag{5}$$

where $(\cdot)$ denotes the element-wise product and W is the vector of trainable weights.

The SoftMax loss function was then used to efficiently transfer the utterance-level features into various emotion states from E classes. To this end, we added an FCN layer with 64 output units to compute the center loss function. After the FCN layer, only one dropout layer was applied. Our approach differs from that of Meng et al. (2019), because the center loss function is generated using 64 units in our method, rather than from the FCN layer with E units corresponding to E classes. When calculating the loss function shown in Section 3.3.3, the center loss and SoftMax loss functions were used as inputs.

3.3.3. Loss function

To classify different emotions, we blended the SoftMax loss and center-based loss functions to update the network weights along the way. We employed the center loss function to reduce the within-class distance and the SoftMax loss function to maximize the between-class distance in order to separate the features and discriminate them for SER.

The SoftMax/cross-entropy loss function is defined as follows:

$$\mathcal{L}_{SM} = -\sum_{n=1}^{bs} \log\left(\frac{e^{W_{y_n}^T \times x_n + b_{y_n}}}{\sum_{m=1}^{E} e^{W_m^T \times x_n + b_m}}\right), \tag{6}$$

where $\mathcal{L}_{SM}$ is the SoftMax loss function and bs is the batch size or the number of samples in a mini-batch.

The following definition describes the calculation of the center loss function, in order to determine how far apart the feature centroids are from their respective class centroids:

$$\mathcal{L}_{CT} = \frac{1}{2}\sum_{n=1}^{bs} \left|\left|x_n - C_{y_n}\right|\right|_2^2, \tag{7}$$

where $\mathcal{L}_{CT}$ is the center loss function and C_{y_n} is the centroid of the class that the nth sample belongs to.

The loss function combines the SoftMax loss and center loss functions, as shown in Eq. (8).

$$\mathcal{L}_T = \epsilon \times \mathcal{L}_{CT} + \mathcal{L}_{SM}, \tag{8}$$

where $\mathcal{L}_T$ is the total loss function and $\epsilon \in (0, 1)$ is a weighting factor to balance between the center and SoftMax loss functions. If $\epsilon = 0$, the loss function becomes the SoftMax loss function.

The SoftMax loss function in our proposed model is defined in the same manner as that in Meng et al. (2019). However, in this study, we calculated the center-based loss function after an FCN of 64 units, instead of from an FCN with E units (output classes) as in Meng et al. (2019).

4. Experimental results and comparison

4.1. Datasets

4.1.1. EmoDB dataset

In this study, the Berlin Database of Emotional Speech (EmoDB) (Burkhardt et al., 2005) was used to train the developed HDA method and recognize emotions from speech signals. EmoDB consists of recorded 535 audio data samples pertaining to sentences uttered by five males and five females aged 25–32 years under seven emotions, *viz.*, happiness, sadness, anger, neutral, fear, disgust, and boredom. Original data samples were recorded at 44.1 kHz and then re-sampled to 16 kHz. The distribution of EmoBD data is shown in Fig. 10.

4.1.2. ERC dataset

The emotion recognition (ERC) dataset is a subset of the Crema-D dataset created by Cao et al. (2014). The Crema-D dataset was recorded for six distinct emotional states, i.e., anger, sadness, happiness, neutral, disgust, and fear, using a selection of 12 predefined sentences. Samples were collected from 48 men and 43 women, aged between 20 and 74 years, and from different racial and ethnic backgrounds, including

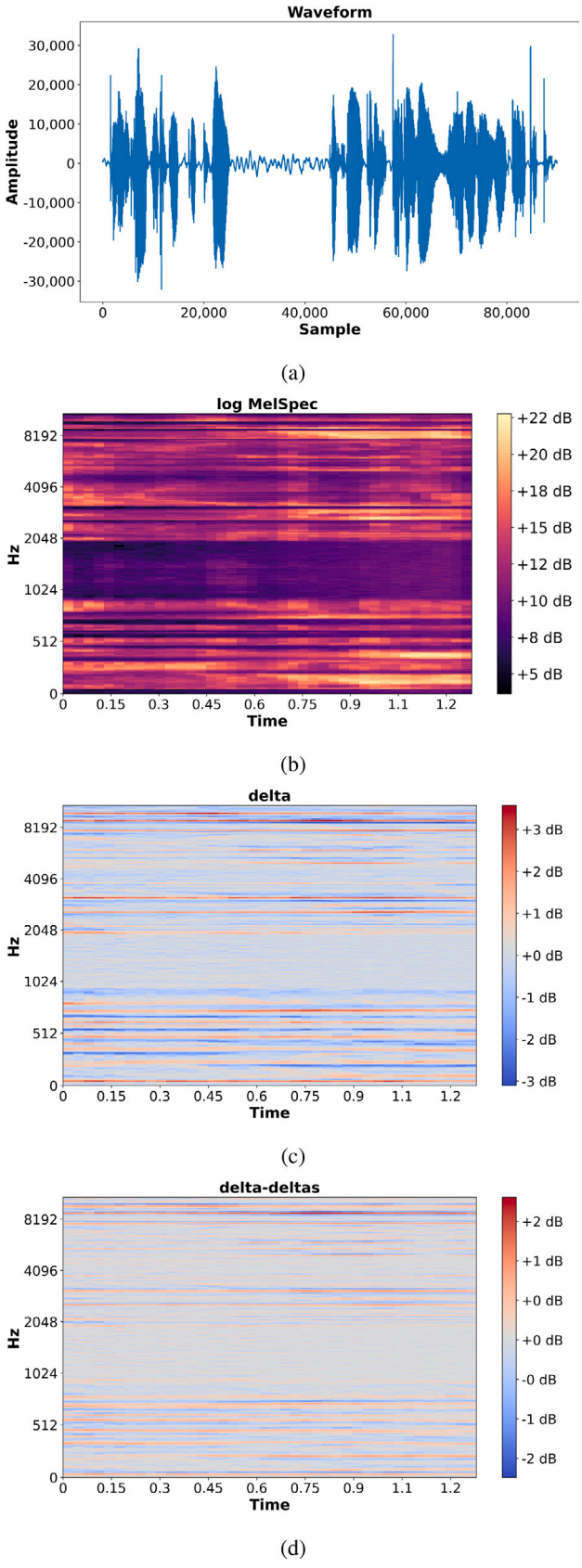


Fig. 5. Visualization of the 3D log MelSpec features of the sad emotion sample in the original EmoDB dataset: (a) Waveform, (b) log MelSpec, (c) delta, and (d) delta-deltas features.

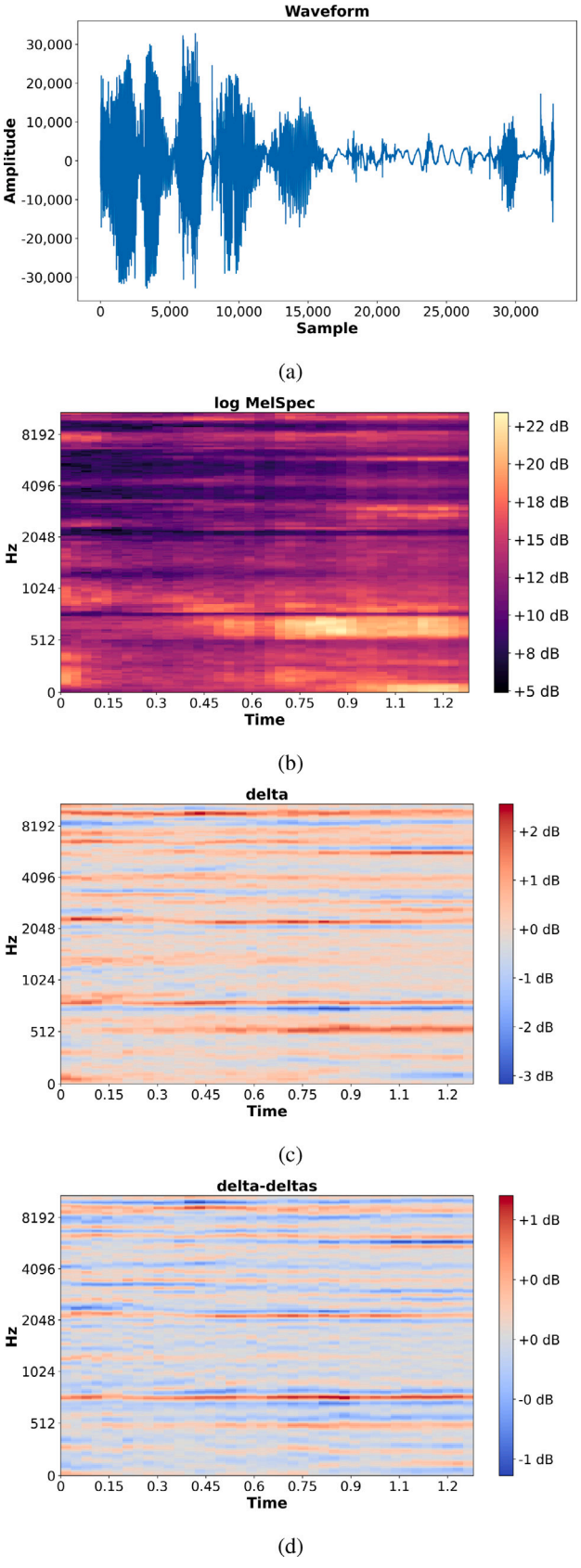


Fig. 6. Visualization of the 3D log MelSpec features of sad emotion sample in the applied WaveGAN on the EmoDB dataset: (a) Waveform, (b) log MelSpec, (c) delta, and (d) delta-deltas features.

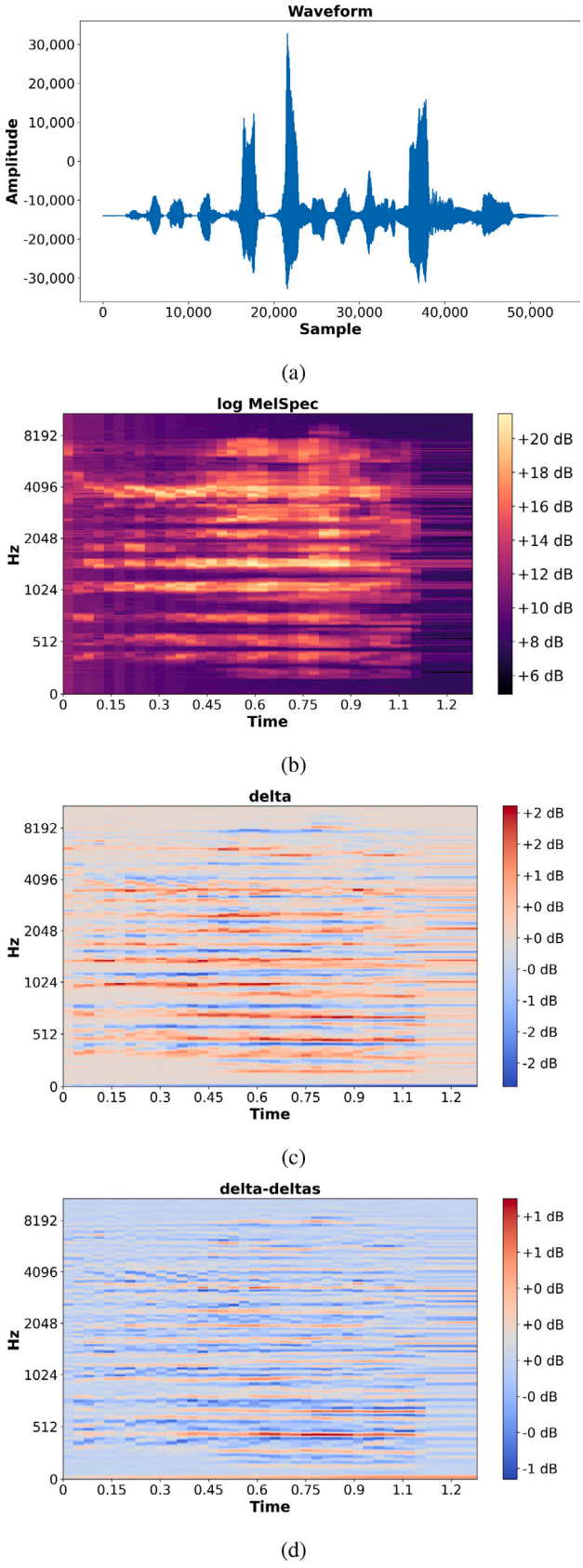


Fig. 7. Visualization of the 3D log MelSpec features of the angry emotion sample with time shifting from the EmoDB dataset: (a) Waveform, (b) log MelSpec, (c) delta, and (d) delta-deltas features.

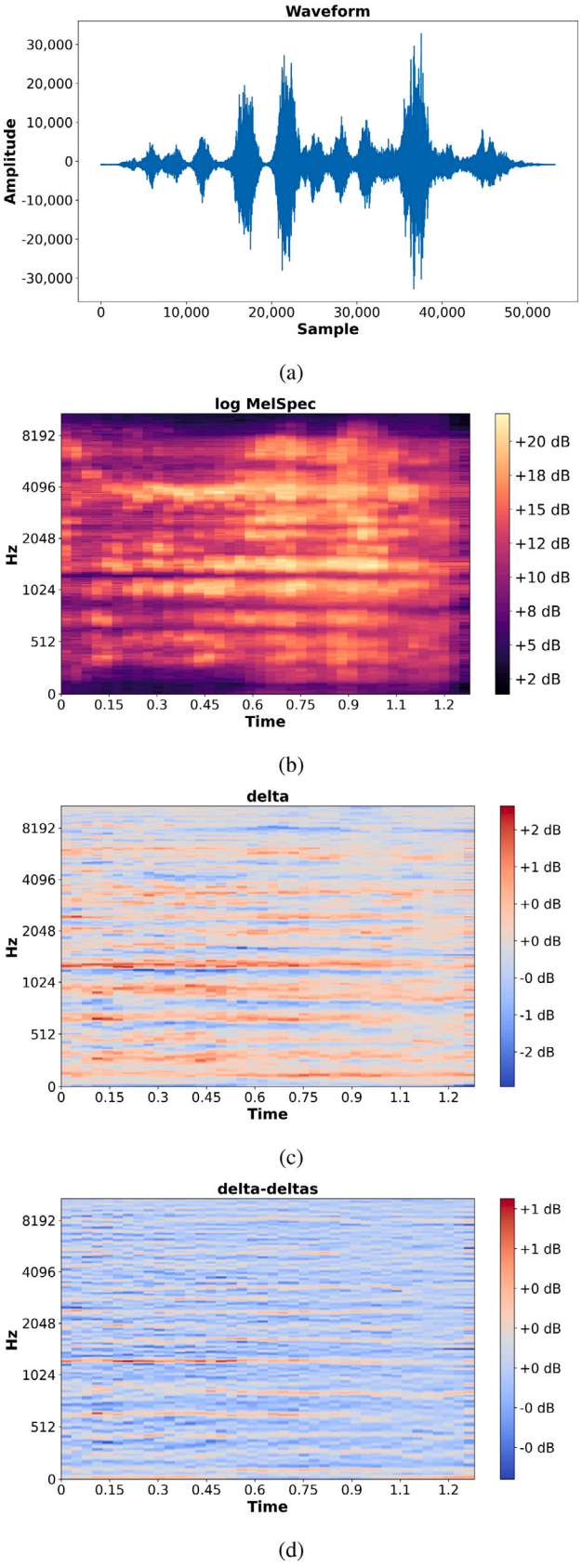


Fig. 8. Visualization of the 3D log MelSpec features of the angry emotion sample with pitch shifting from the EmoDB dataset: (a) Waveform, (b) log MelSpec, (c) delta, and (d) delta-deltas features.

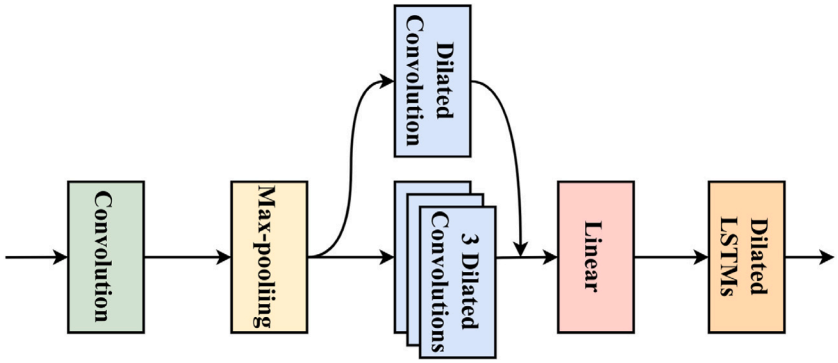


Fig. 9. Deep DCRNN architecture.

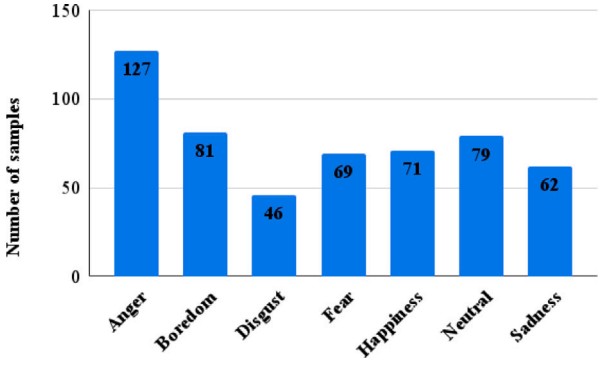


Fig. 10. Detailed sample distribution of the EmoDB dataset.

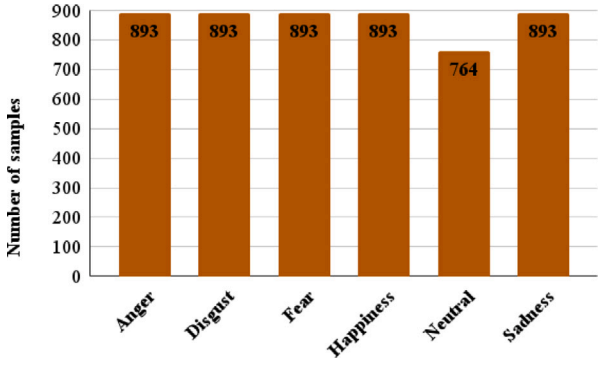


Fig. 11. Detailed distribution of the ERC dataset.

Asian, African, American, Hispanic, Caucasian, and unspecified. The ERC dataset contained 5,229 samples, with a total of 893 samples for each emotion, except for the neutral emotion, which had 764 samples. It was used only to train and validate the proposed deep mADCRNN model in this study. Fig. 11 depicts the distribution of the emotional states in the ERC dataset.

4.2. Experimental setup

The HDA method was employed as an upsampling technique by generating additional data samples per class for each dataset. Using the HDA method, the number of generated data samples is the same as that in the original dataset. Based on the original data samples, three augmented datasets were formed: (i) applying time shifting (denoted as time-shifting-augmented dataset), (ii) applying pitch shifting (denoted as pitch-shifting-augmented dataset), and (iii) applying WaveGAN (denoted as WaveGAN-augmented dataset). The proposed DL framework

Table 2
The experimental results on the EmoDB using the proposed deep learning framework.

Case	Loss function	Unweighted recall (%)
Origin	L_{f1}	83.58 ± 2.17
	L_{f2}	**86.90 ± 1.16**
	L_{f3}	86.84 ± 3.86
	L_{f4}	85.14 ± 1.60
AppliedWaveGAN	L_{f1}	85.48 ± 1.79
	L_{f2}	**88.03 ± 1.39**
	L_{f3}	87.82 ± 1.90
	L_{f4}	86.08 ± 2.79
AppliedTime Shifting	L_{f1}	76.96 ± 3.37
	L_{f2}	**82.01 ± 2.46**
	L_{f3}	77.38 ± 4.40
	L_{f4}	75.94 ± 3.32
AppliedPitch Shifting	L_{f1}	79.62 ± 5.71
	L_{f2}	**85.31 ± 1.77**
	L_{f3}	80.00 ± 2.65
	L_{f4}	78.79 ± 4.40

Table 3
Ablation studies on the ERC dataset.

Loss function	Unweighted recall (%)
L_{f1}	65.83 ± 0.82
L_{f2}	**66.56 ± 0.67**
L_{f3}	65.58 ± 0.60
L_{f4}	64.86 ± 0.82

was implemented using TensorFlow (Abadi et al., 2016) and trained on a single NVIDIA GeForce GTX 1050Ti with 4 GB of VRAM and 16 GB of RAM. Our deep mADCRNN model was trained using an Adam (Kingma & Ba, 2015) optimizer with a learning rate of 0.0001, batch size of 16, 3,000 epochs, and a dropout rate of 0.5. The results from 5-fold cross-validation were computed.

We investigate the effect of the following loss functions:

- L_{f1}: only the SoftMax loss function;
- L_{f2}: the reconfigured center-based loss and SoftMax loss functions;
- L_{f3}: The SoftMax loss and center loss functions, as in Meng et al. (2019);
- L_{f4}: The SoftMax loss and CT-C loss functions, as in Pham et al. (2020).

We perform the experiments in two scenarios, as follows:

- Without using the HDA method on the ERC dataset;
- Using the HDA method on the EmoDB dataset, and time shifting, pitch shifting, and WaveGAN were applied to the original EmoDB dataset. Subsequently, both the original and augmented data samples were used to train and evaluate the DL framework.

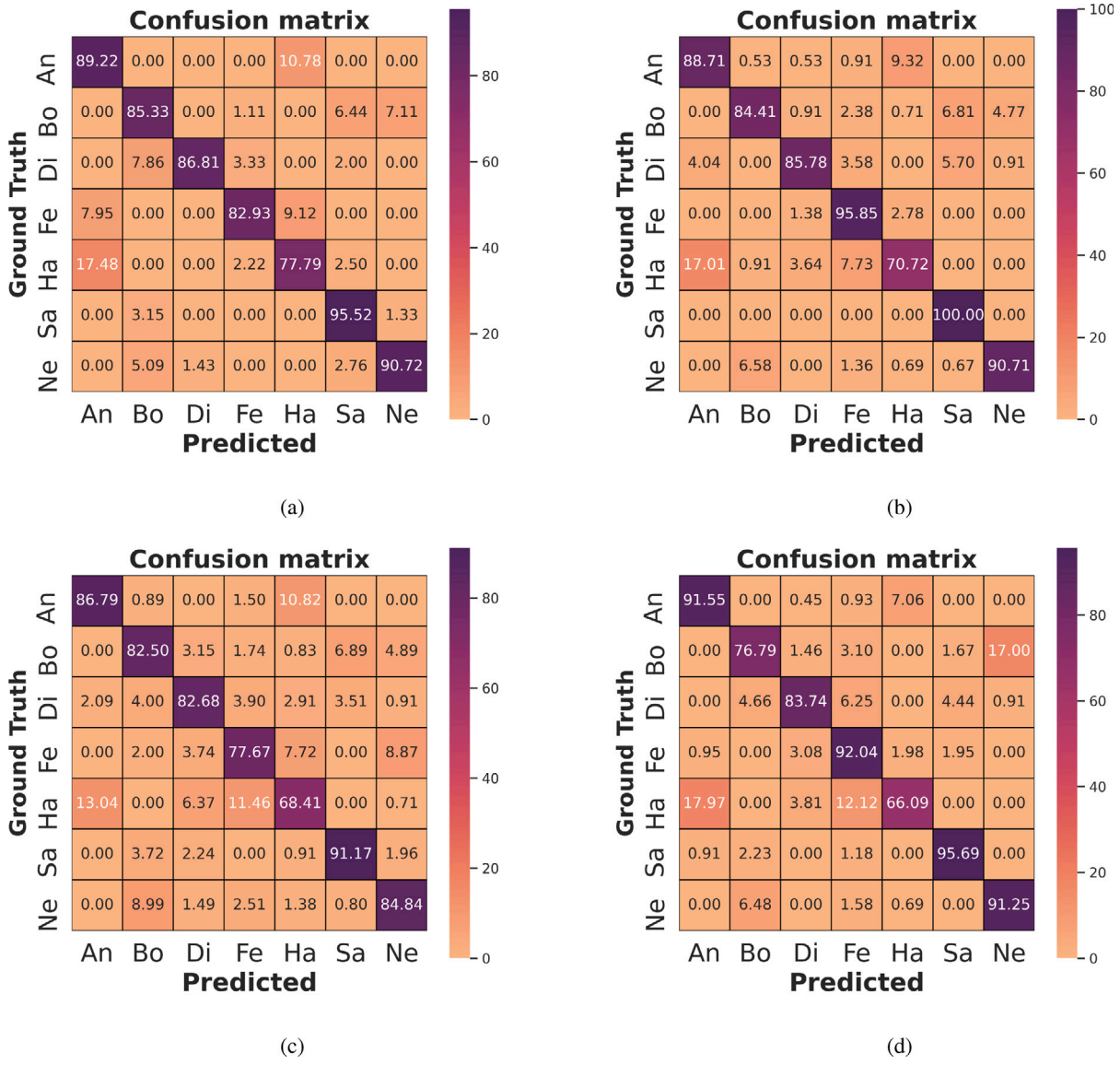


Fig. 12. Confusion matrices using the deep mADCRNN model with the reconfigured SoftMax loss and center loss L_{f2} on the EmoDB dataset: (a) The original dataset; (b) the WaveGAN-augmented dataset; (c) the time shifting-augmented dataset; and (d) the pitch shifting-augmented dataset.

In this study, all datasets were split into training and test sets at a ratio of 80:20.

For 3D log MelSpec extraction, we used the method described by Lyons et al. (2020) to extract the static, delta, and delta-deltas coefficients of log MelSpec. In this study, WaveGAN was trained for up to 90×10^3 iterations, which was then used to generate the augmented datasets.

For performance evaluation, the unweighted recall (UR) metric was used, as in Eq. (9):

$$UR = \frac{1}{nE} \sum_{e=1}^{nE} \frac{TP_e}{AP_e}, \tag{9}$$

where nE denotes the number of emotions, TP_e is the number of true-positive classified samples of emotion e, and AP_e is the total number of actual positive samples of emotion e.

In most classification problems, particularly SER tasks, researchers often use a confusion matrix to evaluate the performance (Chen et al., 2018; Jeon, Hasan, Park, Lee, & Manavalan, 2022; Meng et al., 2019; Pham et al., 2020; Pham, Nguyen, Nguyen, Pham, & Dang, 2023; Zhang et al., 2023). We also used a confusion matrix to show the results of the proposed DL framework with different configurations and loss functions.

4.3. Results

4.3.1. Experiments on the EmoDB dataset using HDA and mADCRNN

The experimental results on the EmoDB dataset with different loss functions are reported in Table 2, where "Origin" denotes the original dataset, while the rest are augmented datasets using the HDA method. The proposed method yields the highest UR scores of 86.90 ± 1.16, 88.03 ± 1.39, 82.01 ± 2.46, and 85.31 ± 1.77 (%) using the L_{f2} loss function on the original, WaveGAN-augmented, time shifting-augmented, and pitch shifting-augmented EmoDB datasets, respectively.

Fig. 12 presents the confusion matrices of the proposed methods using the HDA, deep mADCRNN, and reconfigured loss function L_{f2}, where *An*, *Bo*, *Di*, *Fe*, *Ha*, *Sa*, and *Ne* denote Anger, Boredom, Disgust, Fear, Happiness, Sadness, and Neutral emotions, respectively. The proposed method performed best in recognizing Sadness, achieving UR scores of 95.52, 100.00, 91.17, and 95.69 (%) in recognizing Sadness with the original, augmented WaveGAN, augmented time shifting, and

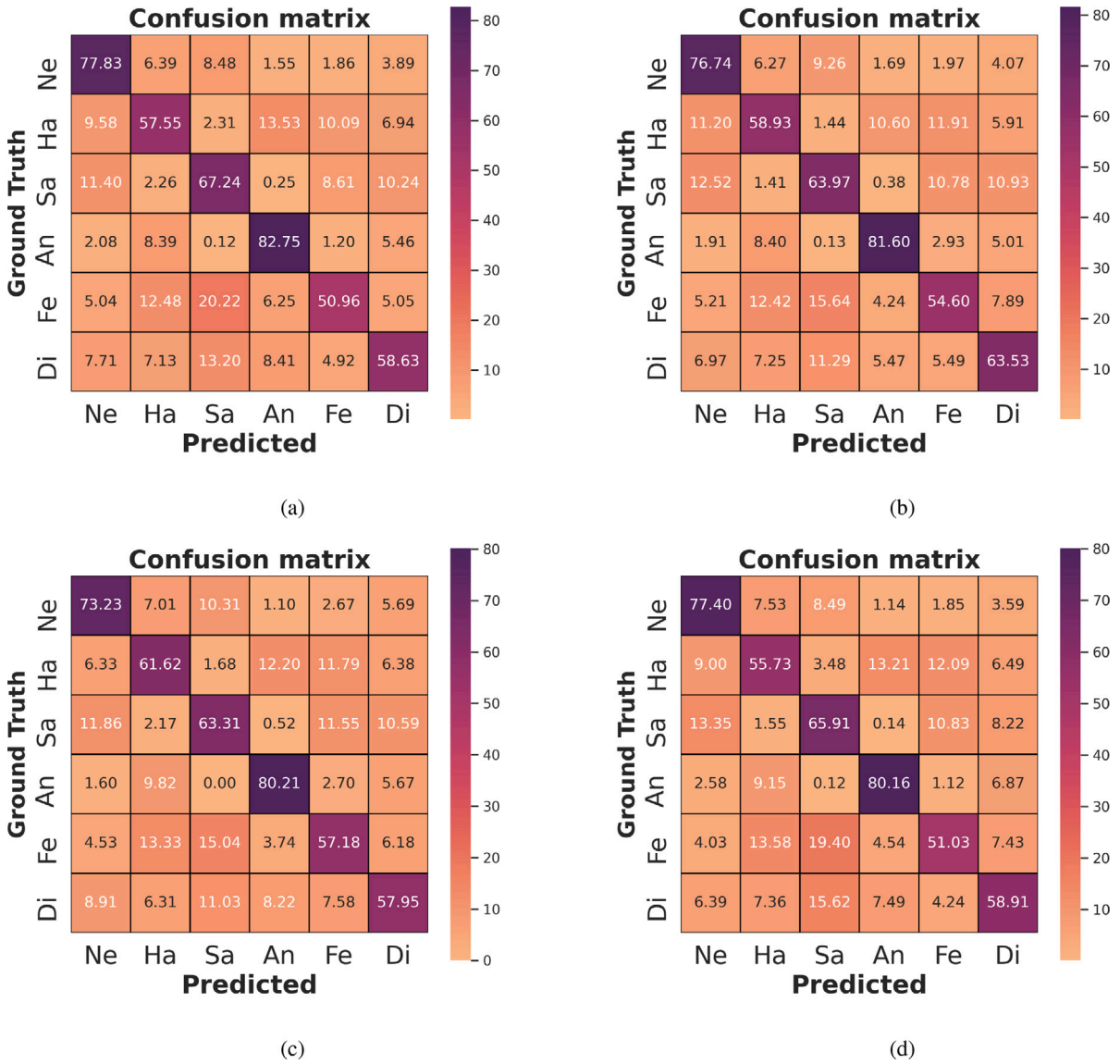


Fig. 13. Confusion matrices using the deep mADCRNN with four loss functions on the ERC dataset: (a) L_{f1}; (b) L_{f2}; (c) L_{f3}; and (d) L_{f4}.

augmented pitch shifting EmoDB datasets, respectively. In contrast, the proposed method yielded the lowest performance in recognizing Happiness in all experiments. Based on Table 2 and Fig. 12, using the WaveGAN-augmented dataset can improve the SER performance, whereas the others exhibit a decreased performance in recognizing emotional states from speech signals.

4.3.2. Ablation studies on the ERC dataset

The experimental results for the ERC dataset with different loss functions are presented in Table 3, indicating the effectiveness of the deep mADCRNN architecture. Note that HDA was not used in these ablation studies. As shown in Table 3, using the deep mADCRNN model with the L_{f2} loss function achieves the highest UR scores of 66.56 ± 0.67 (%) on the ERC dataset. The associated confusion matrices are shown in Fig. 13. The proposed mADCRNN model with the L_{f2} loss function produces the best performance in recognizing Anger, while it yields the poorest performance in recognizing Fear.

4.4. Comparison and discussion

Table 4 presents a comparison between the proposed and existing state-of-the-art methods for the EmoDB dataset. Note that 3D ACRNN stands for 3D attention-based CRNN, and WADAN is the Wasserstein adversarial data augmentation network. As shown in Table 4, our HDA and mADCRNN framework achieves the highest UR score, i.e., 88.03 ± 1.39 (%) with WaveGAN-augmented EmoDB dataset. The number of trainable parameters for each method, which indicates the complexity of each algorithm, is listed in Table 4. Based on Table 4, our proposed DL framework has 7,450,121 trainable parameters, which are fewer than those of 3-D ACRNN and ADCRNN. The results of our proposed DL framework exhibit a significant improvement. Notably, the number of trainable parameters was based solely on the classification models. Moreover, owing to computational limitations and conflicting model variants, the number of trainable parameters reported by Yi and Mak (2020) and the training time for all the methods are not included in Table 4.

From Tables 2 and 3 as well as Figs. 12 and 13, it can be observed that the reconfigured loss function L_{f2} is the most effective for the deep mADCRNN model. All experiments on the EmoDB and ERC datasets yielded the best performance on the L_{f2} loss function using the deep mADCRNN with and without HDA, respectively.

Based on several experiments on the EmoDB dataset and the ablation studies on the ERC dataset, the results were consistent, indicating

Table 4
Comparison of the proposed and existing state-of-the-art methods on the EmoDB dataset.

Reference	Method	Result (%)	Trainable parameters
Chen et al. (2018)	3-D ACRNN	82.82 ± 4.99 (Unweighted Recall)	8,188,486
Meng et al. (2019)	ADRNN	85.39 ± 1.86 (Unweighted Accuracy)	4,563,977
Pham et al. (2020)	ADCRNN	87.95 ± 3.45 (Unweighted Recall)	9,127,954
Yi and Mak (2020)	WADAN + DNN	83.31 ± 0.20 (Unweighted Recall)	–
Ours	HDA + mADCRNN	88.03 ± 1.39 (Unweighted Recall)	7,450,121

that the proposed framework can generalize well with and without HDA.

Although the proposed method shows the lowest performance in recognizing Happiness in all experiments conducted on both the EmoDB and ERC datasets, this finding is consistent with the results reported in other studies, including those by Meng et al. (2019), Chen et al. (2018), and Pham et al. (2020).

In addition, WaveGAN is effective at generating waveforms (Donahue et al., 2019). Therefore, the experiments using deep mADCRNN on the WaveGAN-augmented EmoDB dataset yield the highest performance in all loss functions. Although using time shifting and pitch shifting leads to a decreased performance, these techniques always exist in practice, either when speaker speaks faster or slower, or when playing recordings in a forward or backward manner.

Furthermore, dilated CNNs and dilated LSTMs have been shown to be more efficient and effective than their conventional counterparts (Chang et al., 2017; Li, Liu, Drossos, & Virtanen, 2020; Meng et al., 2019; Pham et al., 2020; Zhu, Li, Chen, Herrero, & Georgiou, 2020). The proposed DL framework achieved a superior performance as compared with those from other state-of-the-art methods on the EmoDB dataset.

Finally, dilated LSTM, instead of the BiLSTM, mADCRNN, was more stable and less complex than ADCRNN. This is because the mADCRNN has a lower deviation in the results and fewer trainable parameters than the ADCRNN.

5. Conclusion

In this study, we have designed an HDA method that combines both traditional and GAN-based models to generate additional data samples for SER. In addition, a deep mADCRNN model was implemented to learn and extract the utterance-level features from the 3D log MelSpec low-level features. Several combinations of the SoftMax and center-based loss functions have been investigated to improve the SER performance. The experimental results indicate that the devised HDA method can achieve better results than the state-of-the-art methods when dealing with limited and imbalanced datasets.

Although our proposed DL framework comprising HDA and mADCRNN for SER has proven to be useful, several aspects can be further improved. In future work, we will investigate multi-feature and multi-modality fusion (Pham et al., 2022) to identify robust and optimal features for SER. In addition, keyword spotting will be employed for integration into a real-time SER system. Because feature extraction and selection are key elements in SER, clustering methods, e.g., Nguyen, Nguyen, and Pham (2022) can be considered to pre-define the feature set. Neural-fuzzy algorithms, such as that of Nguyen, Choi, and Seo (2017), can be integrated to formulate a new SER classifier. In addition, ensemble learning methods will be investigated to improve the classification performance. However, the proposed mADCRNN model is complex in terms of the number of layers and parameters. Consequently, it would be useful to leverage optimization methods, such as invasive weed optimization integrated with differential evolutionary (Movassagh et al., 2021) and dynamic programming-based ensemble design (Alzubi et al., 2020b), to reduce the ensemble size while maximizing ensemble diversity to improve classification accuracy and identify the optimum model parameters for undertaking SER tasks.

CRediT authorship contribution statement

Nhat Truong Pham: Conceptualization, Methodology, Software, Validation, Data curation, Writing – original draft, Writing – review & editing. **Duc Ngoc Minh Dang:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Ngoc Duy Nguyen:** Methodology, Software, Writing – original draft. **Thanh Thi Nguyen:** Methodology, Software, Writing – original draft. **Hai Nguyen:** Methodology, Software, Writing – original draft. **Balachandran Manavalan:** Writing – review & editing, Supervision. **Chee Peng Lim:** Writing – original draft, Writing – review & editing, Supervision. **Sy Dzung Nguyen:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared the link to our code in the manuscript.

Acknowledgments

Nhat Truong Pham and Sy Dzung Nguyen would like to thank the Vietnam National Foundation for Science and Technology Development (NAFOSTED) for their support under grant number 107.01-2019.328.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). TensorFlow: A system for large-scale machine learning. In K. Keeton, T. Roscoe (Eds.), *12th USENIX symposium on operating systems design and implementation, OSDI 2016, Savannah, GA, USA, November (2016) 2-4* (pp. 265–283). USENIX Association, https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

Albornoz, E. M., Milone, D. H., & Rufiner, H. L. (2011). Spoken emotion recognition using hierarchical classifiers. *Computer Speech and Language*, *25*, 556–570.

Alzubi, O. A., Alzubi, J. A. A., Alweshah, M., Qiqieh, I., Al-Shami, S., & Ramachandran, M. (2020b). An optimal pruning algorithm of classifier ensembles: dynamic programming approach. *Neural Computing and Applications*, *32*, 16091–16107. http://dx.doi.org/10.1007/s00521-020-04761-6.

Alzubi, J. A. A., Jain, R., Kathuria, A., Khandelwal, A., Saxena, A., & Singh, A. (2020a). Paraphrase identification using collaborative adversarial networks. *Journal of Intelligent & Fuzzy Systems*, *39*, 1021–1032. http://dx.doi.org/10.3233/JIFS-191933.

Arias, J. P., Busso, C., & Yoma, N. B. (2014). Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech and Language*, *28*, 278–294.

Bao, F., Neumann, M., & Vu, N. T. (2019). Cycle GAN-based emotion style transfer as data augmentation for speech emotion recognition. In G. Kubin, & Z. Kacic (Eds.), *Interspeech 2019, 20th annual conference of the international speech communication association, Graz, Austria, 15-19 2019* (pp. 2828–2832). ISCA, http://dx.doi.org/10.21437/Interspeech.2019-2293.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of german emotional speech. In *INTERSPEECH 2005 - Eurospeech, 9th European conference on speech communication and technology, Lisbon, Portugal, September (2005) 4-8* (pp. 1517–1520). ISCA, http://www.isca-speech.org/archive/interspeech_2005/i05_1517.html.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, *5*, 377–390. http://dx.doi.org/10.1109/TAFFC.2014.2336244.

Cao, H., Verma, R., & Nenkova, A. (2015). Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. *Computer Speech and Language*, *29*, 186–202.

Cen, L., Wu, F., Yu, Z. L., & Hu, F. (2016). A real-time speech emotion recognition system and its application in online learning. In *Emotions, technology, design, and learning* (pp. 27–46). Elsevier.

Chang, S., Zhang, Y., Han, W., Yu, M., Guo, X., Tan, W., et al. (2017). Dilated recurrent neural networks. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December (2017) 4-9, Long Beach, CA, USA* (pp. 77–87). https://proceedings.neurips.cc/paper/2017/hash/32bb90e8976aab5298d5da10fe66f21d-Abstract.html.

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, *25*, 1440–1444.

Chen, L., Mao, X., Xue, Y., & Cheng, L. L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Processing*, *22*, 1154–1160.

Dai, W., Han, D., Dai, Y., & Xu, D. (2015). Emotion recognition and affective computing on vocal social media. *Information & Management*, *52*, 777–788.

Donahue, C., McAuley, J. J., & Puckette, M. S. (2019). Adversarial audio synthesis. In *7th International conference on learning representations, ICLR 2019, New Orleans, la, USA, May (2019) 6-9*. OpenReview.net. https://openreview.net/forum?id=ByMVTsR5KQ.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein GANs. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December (2017) 4-9, Long Beach, CA, USA* (pp. 5767–5777). https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html.

Haghparast, A., Penttinen, H., & Välimäki, V. (2007). Real-time pitchshifting of musical signals by a time-varying factor using normalized filtered correlation time-scale modification (NFC-TSM). In *Proceedings of the international conference on digital audio effects (DAFx), Bordeaux, France* (pp. 10–15). Citeseer.

Ho, N. H., Yang, H. J., Kim, S. H., & Lee, G. (2020). Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access*, *8*, 61672–61686.

Huahu, X., Jue, G., & Jian, Y. (2010). Application of speech emotion recognition in intelligent household robot. In *2010 International conference on artificial intelligence and computational intelligence* (pp. 537–541). IEEE.

Huang, Y., Tian, K., Wu, A., & Zhang, G. (2019). Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. *Journal of Ambient Intelligence and Humanized Computing*, *10*, 1787–1798.

Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, *59*, Article 101894.

Jeon, Y., Hasan, M. M., Park, H. W., Lee, K. W., & Manavalan, B. (2022). TACOS: A novel approach for accurate prediction of cell-specific long noncoding rnas subcellular localization. *Briefings in Bioinformatics*, *23*, http://dx.doi.org/10.1093/bib/bbac243.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio, & Y. LeCun (Eds.), *3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May (2015) 7-9, Conference Track Proceedings*. http://arxiv.org/abs/1412.6980.

Lalitha, S., Gupta, D., Zakariah, M., & Alotaibi, Y. A. (2020). Investigation of multilingual and mixed-lingual emotion recognition using enhanced cues with data augmentation. *Applied Acoustics*, *170*, Article 107519.

Lee, C. C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, *53*, 1162–1171.

Lent, K. (1989). An efficient method for pitch shifting digitally sampled sounds. *Computer Music Journal*, *13*, 65–71.

Li, Y., Liu, M., Drossos, K., & Virtanen, T. (2020). Sound event detection via dilated convolutional recurrent neural networks. In *2020 IEEE international conference on acoustics, speech and signal processing, ICASSP 2020, Barcelona, Spain, May (2020) 4-8* (pp. 286–290). IEEE, http://dx.doi.org/10.1109/ICASSP40776.2020.9054433.

Lyons, J., Wang, D. Y. B., Gianluca, Shteingart, H., Mavrinac, E., Gaurkar, Y., et al. (2020). Jameslyons/python_speech_features: release v0.6.1. http://dx.doi.org/10.5281/zenodo.3607820.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., et al. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (pp. 18–25).

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D log-Mel spectrograms with deep learning network. *IEEE Access*, *7*, 125868–125881.

Movassagh, A. A., Alzubi, J. A., Gheisari, M., Rahimi, M., Mohan, S., Abbasi, A. A., et al. (2021). Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model. *Journal of Ambient Intelligence and Humanized Computing*, 1–9.

Nguyen, S. D., Choi, S. B., & Seo, T. I. (2017). Recurrent mechanism and impulse noise filter for establishing anfis. *IEEE Transactions on Fuzzy Systems*, *26*, 985–997.

Nguyen, S. D., Nguyen, V. S. T., & Pham, N. T. (2022). Determination of the optimal number of clusters: A fuzzy-set based method. *IEEE Transactions on Fuzzy Systems*, *30*, 3514–3526. http://dx.doi.org/10.1109/TFUZZ.2021.3118113.

Park, D. S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E. D., et al. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. In G. Kubin, Z. Kacic (Eds.), *Interspeech 2019, 20th Annual conference of the international speech communication association, Graz, Austria, 15-19 2019* (pp. 2613–2617). ISCA, http://dx.doi.org/10.21437/Interspeech.2019-2680.

Peng, Z., Li, X., Zhu, Z., Unoki, M., Dang, J., & Akagi, M. (2020). Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks with auditory front-ends. *IEEE Access*, *8*, 16560–16572.

Pham, N. T., Dang, D. N. M., & Nguyen, S. D. (2020). A method upon deep learning for speech emotion recognition. *Journal of Advanced Engineering and Computation*, *4*, 273–285.

Pham, N. T., Nguyen, S. D., Nguyen, V. S. T., Pham, B. N. H., & Dang, D. N. M. (2023). Speech emotion recognition using overlapping sliding window and Shapley additive explainable deep neural network. *Journal of Information and Telecommunication*, 1–19.

Pham, N. T., Tran, A. T., Pham, B. N. H., Dang-Ngoc, H., Nguyen, S. D., & Dang, D. N. M. (2022). Speech emotion recognition: A brief review of multi-modal multi-task learning approaches. In *International conference on advanced engineering theory and applications* (pp. 563–572). Springer.

Qian, Y., Hu, H., & Tan, T. (2019). Data augmentation using generative adversarial networks for robust speech recognition. *Speech Communication*, *114*, 1–9.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In Y. Bengio, & Y. LeCun (Eds.), *4th International conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May (2016) 2-4, Conference Track Proceedings*. http://arxiv.org/abs/1511.06434.

Rebai, I., BenAyed, Y., Mahdi, W., & Lorré, J. P. (2017). Improving speech recognition using data augmentation and acoustic model fusion. *Procedia Computer Science*, *112*, 316–322.

Sajjad, M., Kwon, S., et al. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access*, *8*, 79861–79875.

Tiwari, U., Soni, M. H., Chakraborty, R., Panda, A., & Kopparapu, S. K. (2020). Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions. In *2020 IEEE international conference on acoustics, speech and signal processing, ICASSP 2020, Barcelona, Spain, May (2020) 4-8* (pp. 7194–7198). IEEE, http://dx.doi.org/10.1109/ICASSP40776.2020.9053581.

Tzirakis, P., Zhang, J., & Schuller, B. W. (2018). End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing, ICASSP 2018, Calgary, AB, Canada, April (2018) 15-20* (pp. 5089–5093). IEEE, http://dx.doi.org/10.1109/ICASSP.2018.8462677.

Wu, C. H., & Liang, W. B. (2010). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, *2*, 10–21.

Yeh, J. H., Pao, T. L., Lin, C. Y., Tsai, Y. W., & Chen, Y. T. (2011). Segment-based emotion recognition from continuous Mandarin Chinese speech. *Computers in Human Behavior*, *27*, 1545–1552.

Yi, L., & Mak, M. W. (2020). Improving speech emotion recognition with adversarial data augmentation network. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Yoon, W., Cho, Y., & Park, K. (2007). A study of speech emotion recognition and its application to mobile services. In J. Indulska, J. Ma, L. T. Yang, T. Ungerer, & J. Cao (Eds.), *Ubiquitous intelligence and computing, 4th international conference, UIC 2007, Hong Kong, China, July (2007) 11-13, Proceedings* (pp. 758–766). Springer, http://dx.doi.org/10.1007/978-3-540-73549-6_74.

Zhang, X., Wei, L., Ye, X., Zhang, K., Teng, S., Li, Z., et al. (2023). Siamese CPP: A sequence-based siamese network to predict cell-penetrating peptides by contrastive learning. *Briefings in Bioinformatics*, *24*, http://dx.doi.org/10.1093/bib/bbac545.

Zhang, S., Zhang, S., Huang, T., & Gao, W. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, *20*, 1576–1590.

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323.

Zhu, T., Li, K., Chen, J., Herrero, P., & Georgiou, P. (2020). Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. *Journal of Healthcare Informatics Research*, *4*, 308–324. http://dx.doi.org/10.1007/s41666-020-00068-2.