BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 69(1), 2021, Article number: e136300
DOI: 10.24425/bpasts.2020.136300

Speech emotion recognition using wavelet packet reconstruction with attention-based deep recurrent neutral networks

Hao MENG[1], Tianhao YAN[1]*, Hongwei WEI[1], and Xun JI[2]

[1] Key laboratory of Intelligent Technology and Application of Marine Equipment (Harbin Engineering University),
Ministry of Education, Harbin, 150001, China

[2] College of Marine Electrical Engineering, Dalian Maritime University, Dalian, 116026, China

Abstract. Speech emotion recognition (SER) is a complicated and challenging task in the human-computer interaction because it is difficult to find the best feature set to discriminate the emotional state entirely. We always used the FFT to handle the raw signal in the process of extracting the low-level description features, such as short-time energy, fundamental frequency, formant, MFCC (mel frequency cepstral coefficient) and so on. However, these features are built on the domain of frequency and ignore the information from temporal domain. In this paper, we propose a novel framework that utilizes multi-layers wavelet sequence set from wavelet packet reconstruction (WPR) and conventional feature set to constitute mixed feature set for achieving the emotional recognition with recurrent neural networks (RNN) based on the attention mechanism. In addition, the silent frames have a disadvantageous effect on SER, so we adopt voice activity detection of autocorrelation function to eliminate the emotional irrelevant frames. We show that the application of proposed algorithm significantly outperforms traditional features set in the prediction of spontaneous emotional states on the IEMOCAP corpus and EMODB database respectively, and we achieve better classification for both speaker-independent and speaker-dependent experiment. It is noteworthy that we acquire 62.52% and 77.57% accuracy results with speaker-independent (SI) performance, 66.90% and 82.26% accuracy results with speaker-dependent (SD) experiment in final.

Key words: speech emotion recognition; voice activity detection; wavelet packet reconstruction; feature extraction; LSTM network; attention mechanism.

1. Introduction

Emotions are one of the unique characteristics of human beings distinguished from machines as machines are emotionless while human beings are not [1]. It is significant for humans to communicate with each other in emotional expression. Therefore, there is an indispensable branch that endows machine to recognize emotion and researchers increasingly throw themselves into the study of speech emotion recognition (SER) which plays an effective role in areas of human-machine interaction such as call center systems, health care systems, etc.

Speech signal is the most direct way to utilize in the SER problem. It is the most typical SER systems work by extracting features from raw speech, which include low-level descriptions and high-level statistics functions from five feature categories like pitch-related features, formant features, energy-related features, etc. [2]. In the last decade, human put their eyes on the aforementioned utterance features to accomplish the classification mask. However, these features often overlooked amounts of information either temporal domain or frequency domain that comprised the most common features like mel frequency cepstral coefficients (MFCCs) or log-mel spectrum, because this kind of features all went through fast Fourier transform (FFT) to obtain the majority information from frequency domain. It had drawn a great deal of attention to how to settle this dispute about the features which consisted of both of two domains at the same time. The algorithm of wavelet packet reconstruction was proposed to overcome the issue of losing temporal domain information. WPR is capable of dividing further different frequency band part and reconstruct wavelet sequence signal that has ability to acquire more valuable features from raw speech signal. In addition, the classification accuracy of SER made long-term progress benefited from the development of deep learning networks in recent years. Panagiotis et al. [3] used the deep architecture of convolutional neural networks (CNN) with the 2-layer long short-term memory (LSTM) to obtain great success in recognizing emotions for the RECOLA database. Liu et al. [4] proposed the extraordinary pooling algorithm named global k-max pooling with CNN structure to process the most salient frames in the IEMOCAP database. Jaebok et al. [5] adopted the deep 3-dimensional CNN framework accompanied with extreme learning machine (ELM) from raw speech signal without using sliding contextual window to extract spatio-temporal features in a seamless way. Chen [6] used the functional variables that included deltas and delta-deltas of the log mel-spectrogram (Log-Mels) as the CNN input based on the attention mechanism to help to link the time-frequency relationship.

Inspired by the positive result of wavelet packet decomposition coefficient, it appealed to a great deal of researchers in the issue of SER. Dengaonkar [7] proposed a method of using wavelet packet transform (WPT) to judge the number of coefficients better than threshold in diverse bands, and got an attractive performance in the three emotion classifications in the database of Marathi. Feng [8] utilized the low-level descriptions of features from the mel-scale wavelet packet decomposition to perform

*e-mail: 1825109095@hrbeu.edu.cn

Manuscript submitted 2020-07-28, revised 2020-10-29, initially accepted for publication 2020-12-21, published in February 2021

SER based on the LSTM network. Gupta et al. [1] employed the pitch value, wavelet packet feature vectors (energy and entropy) to consist of feature set, used principal components analysis (PCA) for reducing the dimension of feature and adopted the support vector machine (SVM), random forest (RF) algorithm for classification in the database of EMODB.

The vital contributions of our work are different from above research. In Section 2, we introduce the related works about development progress in the aspect of feature extraction and model structure in SER. In Section 3, the mixed feature set and networks structure of our contributions will be explained in detail. In Section 4, our works are validated by the performance in both of two databases, IEMOCAP and EMODB respectively. In Section 5, we present an ultimate conclusion from our experiment and look into the future work at last.

2. Related work

SER has appealed to a great deal of attention in human-centered signal processing research. In recent years, a large number of researchers put their eyesight on how to extract serviceable feature set and exploit these to acquire better performance in the layer of deep learning networks. In the aspect of feature extraction, features extraction of SER includes four categories: 1. Prosodic features (pitch, vibrancy, intensity, etc.); 2. Excitation features (short-time energy, zero-crossing rate, etc.); 3. Formant features (formant, bandwidth, etc.); 4. Spectrum features (mel-filterbank, MFCCs, etc.). It is also necessary for utilizing prosodic features to obtain the feature of tune and rhythm from speech signal. And the information about formant features and excitations can have enormous advantages to be generated huge merits to emotional recognition [1]. Yenigalla et al. [9] adopted a novel algorithm that extracts the feature set of phoneme sequence and spectrogram combined CNN model to recognize emotions on IEMOCAP corpus. Kim et al. [10] employed the spectro-temporal feature map based on the raw speech signal with a moderate number of parameters to import the CNN model. Jing et al. [11] proposed a novel type of feature that had ability to retain more emotional information at the word-level and affective expression as accurately as possible. Chen et al. [12] used two types of categories that were the electroglottography signal connected with speech signal to study feature extraction. Hook et al. [13] utilized paralinguistics features with the model of random forests and support vector machines to achieve emotional performance. Mencattini et al. raised novel feature maps accompanied by the amplitude modulation to predict the speech emotional state [14] and paralinguistic, and non-linguistic information [15, 16]. Besides, lexical and semantic play a crucial role that includes amounts of emotional information of human beings in SER [17]. Another aspect of model networks, Trigeorgis et al. [18] presented an innovative algorithm that combines the CNNs with LSTM networks to process the issue of ‘context-aware’ emotional feature and the model can learn the best affective representation of speech. Xie et al. [19] introduced the LSTM with attention-based dense connections which had ability to process time series favourably. Tao et al. [20] adopted a new variation of LSTM, advanced LSTM, for better temporal context modeling in weighted pooling RNN.

Huang et al. [21] presented a novel feature set that includes sub-band spectral centroid weighted wavelet packet cepstral coefficients (W-WPCC) to recognize emotion with the importance weights support vector machine (IW-SVM). The feature set achieved comparable performance with MFCC feature and enhances the robustness and generalization via adding the white Gaussian noise in the classification of SER. Firoz [22] used the wavelet packet decomposition features based on the ELM algorithm to acquire 76.66% accuracy in the database of AIR (ALL India Radio). Wang et al. [23] took advantage of the sequential floating forward search (SFFS) method to select the most efficient feature subset of wavelet packets decomposition based on the RBF kernel support vector machine for speaker-independent. Sekkate et al. [24] employed a relatively low-dimensional feature set that combines three features including mel frequency cepstral coefficients (MFCCs) derived from discrete wavelet transform (DWT) sub-band coefficients (DMFCC), and they utilized feature extraction algorithm from above feature set to acquire better results based on the clean conditions and several noised environments in specific emotional databases. Finally, we summarize above the aforementioned works shown in Table 1, where SI and SD are speaker-independent and speaker-dependent respectively.

In this paper, we propose a novel feature map from WPR to extract the wavelet packet sequence that includes the message of frequency domain and temporal domain at the same time. We utilize the information from time-frequency domain to be fed into the BiLSTM (bi-directional long short-term memory) network based on attention mechanism combined with conventional feature set to accomplish the emotion recognition objective. The major contributions of this paper are summarized as:

- Compared with exploiting a few acoustic frequency feature sets alone in [3], we utilize eight wavelet signal sequences from three specific layers of wavelet packet reconstruction to extract effective feature maps mixed with conventional feature set as the input of our proposed end-to-end network structure, and it aims at maintaining the robust frequency information connected with temporal information from speech signal in feature extraction.
- In order to enhance further serviceable feature zone, we avail our recommended framework that adds an upsampling layer after a BiLSTM (bi-directional-LSTM) layer and attention mechanism layer. Note that it outperforms 1.68%, 2.65% recognition accuracy results in two databases from speaker-independent experiment respectively.

3. Proposed methodology

In this section, we introduce our proposal of how to extract the original WPR feature map from the raw speech signal to be fed into the architecture of classification that is shown in Fig. 1.

3.1. Voice activity detection. In fact, we could encounter a serious problem of SER with silent frame and noise that have a neg-

Table 1
The Resume of realted works

Author	**Feture names**	**Feature size**	**Classification Technique**	**Dataset**	**Language**	**Class**	**Results [speaker dependency]**
P. Yengalla et al. [9]	Phoneme sequence & spectrogram	A set of 47 phonemes & 128*256 (spectrogram)	Multi-channel CNN and max-pooling	IEMOCAP	English	Neutral, happy, sad, angry	68.5% [SI]
J. Kim et al. [10]	Spectro-temporal	10*10*256 (Spectral)	3D-CNN-DNN-ELM and 3D Max pooling	LDC, eNTERFACE, EMODB, FAU-AIBO, IEMOCAP, SEMAINE, RECOLA	English, German, French	Neutral, happy, sad, angry	53.6% [SI]
S. Jing et al. [11]	Prominence features	290 Conventional feature set	Double layer feature dimension reduction (DLFDR) Model	Chinese dual-mode emotional speech database (CDESD)	Chinese	Sad, joy, fear, surprise, neutral, angry, disgust	69.65% [SI]/70.3% [SD]
J.Hock et al. [13]	Paralinguistic and MFCC coefficients	87 LLDs	Random forests & support vector machine	SAVEE, EMODB, PESD, GEES	English, German, Polish, Serbian	Angry, bored, disgust, fear, happy, sad, neutral	78.22% (SAVEE), 87.81% (EMODB), 75.41% (PESD), 93.41% (GEES) [SI]
F. Tao et al. [20]	MFCC, ZCR, energy, etc.	36 LLDs	Advanced LSTM based on weighted pooling	IEMOCAP	English	Angry, sad, happy, neutral	58.7% [SI]
Y.M. Huang et al. [21]	Weighted wavelet packet cepstral coefficients	39 (W-WPCC)	Importance-weighted support vector machine with Gaussian mixed model	EMODB	German	Angry, bored, fear, joy, neutral, sad	73.01% [SI]
F. Shah et al. [22]	Wavelet packet decoposition feature	20 coefficients	DNN-ELM	All India Radio (AIR) Trivandrum	India	Angry, sad, bored, fear, surprise, calm, neutral, anxiety	76.66% [SI]
K.X. Wang et al. [23]	Wavelet packet construction with sequential floating forward search	480 coefficients	Support vector machine based on radial basis function	EMODB	German	Angry, bored, anxiety, sad, happy, disgust, neutral	79.54% [SI]
Sekkate et al. [24]	MFCC, DMFCC, Pitch	303	NB, SVM	EMODB/IEMOCAP	German/ English	Angry, sad, bored, digust, fear, neutral, happy, surprise, frustration	81.32% (EMODB), 42.87% (IEMOCAP) [SI]

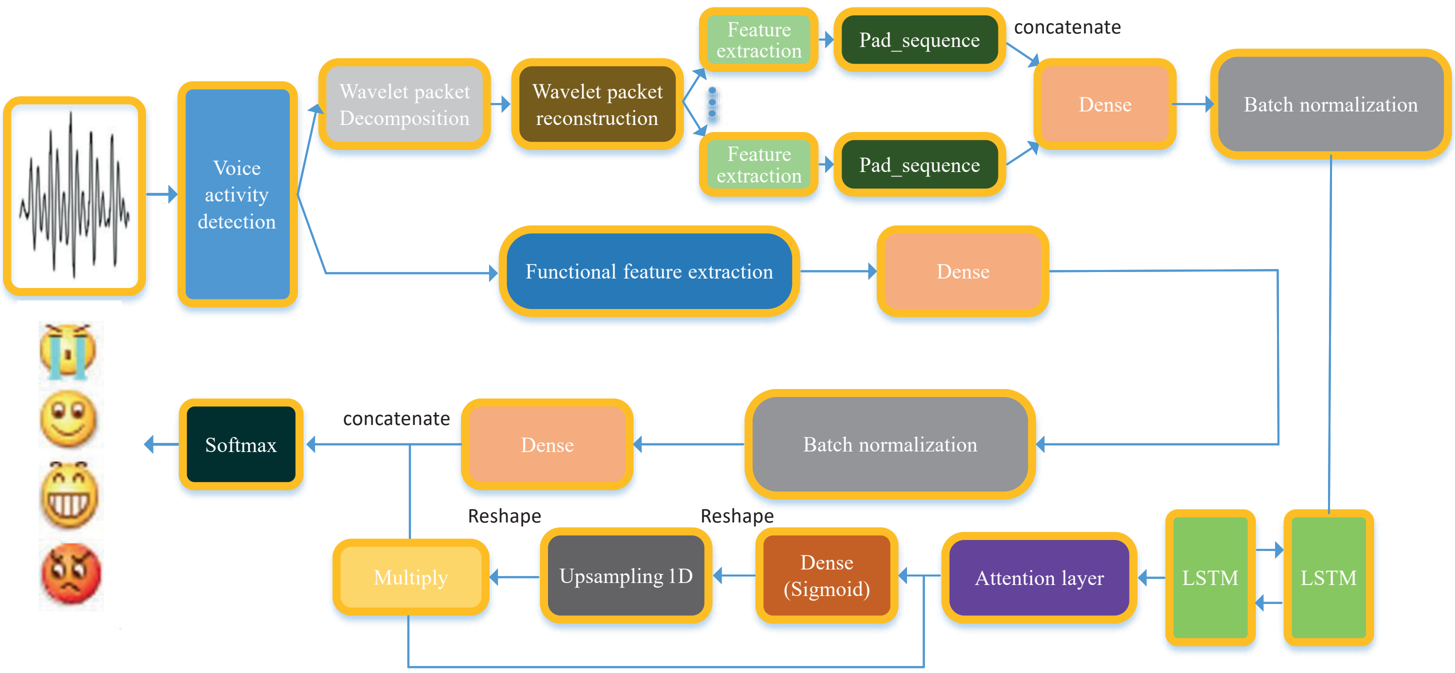

Fig. 1. Illustration of the whole structure to manage our proposed discriminating mixed feature maps from wavelet packet reconstruction for speech emotion recognition

ative effect in the processing of extracting feature. Therefore, it is a reason that the classification accuracy becomes lower and it is vital to eliminate the effect of above reason from utterance and retain effective speech sequence signal in the preprocessing operation of feature extraction.

Voice activity detector (VAD) is aimed at detecting the start point and end point of speech signal, so it is also described as start-stop recognition. The purpose of VAD is to separate speech signal from other signals such as silent frames in a segment of the whole of signal. It is crucial for SER to detect valuable speech signals in this process and it can be carried out correctly when the starting-ending points are accurately determined.

The speech signal owns one of the characteristics that is the periodicity of voiced sound while it is not usually described by silent frame and noise. If a signal is a periodic function, then its autocorrelation function also has periodicity and the period is same as the period of the signal. Therefore, it can be performed by the autocorrelation function of the speech signal in the process of VAD.

We adopt the method named autocorrelation function of double-threshold decision to achieve the purpose. Given a speech signal, we split the signal into short frames with hamming window of 25 ms and a 10 ms frame shift. Moreover, it is necessary to utilize the autocorrelation function that is shown in Eq. (1).

$$R_n(k) = \frac{\sum\limits_{m=1}^{N-k} x_n(m)x_n(m+k)}{\sum\limits_{m=1}^{N} x_n^2(m)} \qquad (0 \leq k \leq K), \tag{1}$$

where $x_n\ (m)$ denotes the raw speech signal, N and K represent the size of frame and the value of delay respectively. Moreover, we also make the process of normalization in this formula because it could avoid the effect from absolute energy in the speech voice activity detection.

Next, we set the value of double-threshold that represents the lower limit 'T_1' and superior 'T_2' in terms of the situation of silent frame and noise. T_1, T_2 both need to be set as thresholds, and it is determined as the part of speech when the maximum value of correlation function is greater than T_2 while it is decided as the voice activity of speech signal when the maximum value of correlation function is greater than or lower than T_1. The whole process of voice activity detection with utilizing autocorrelation function is accomplished by the Matlab 2014b software.

Therefore, it can judge the location of speech activity and remove the silent frame according to the threshold, whatever the beginning or ending part of speech signal, even this kind of frame presented in the form of noise could happen among the whole speech signal sequence. It is remarkable that we exploit a tiny value in the signal sequence instead of the aforementioned noise, because it is necessary to try method best to retain temporal domain information of speech signal. Although there are many robust VAD that can be used instead of this simple one, we also choose the autocorrelation function algorithm to perform VAD in terms of characteristics of briefness and effectiveness.

3.2. Wavelet packet reconstruction sequence generation. After adopting the voice activity detection of autocorrelation function to eliminate the emotional irrelevant frame due to the silent frame on both sides of speech signals, we set out wavelet packet reconstruction (WPR) to get feature map as the input of the deep networks after acquiring the pure speech signal.

In the paper above, many researchers could extract the signal feature to utilize FFT (fast Fourier transform). However, the weakness of FFT is obvious that it can be only analyzed in the frequency domain and be short of the time resolution, and FFT generates a bad performance with processing irregular speech signal like step signal or spike signal when this situation could happen in the speech emotion recognition. Therefore, WPR algorithm based on wavelet transform has better ability to deal with this kind of nonstationary speech signal [25]. It is our purpose to reconstruct new speech sequence signal under the process of WPR method, and it benefits to make some given signal under different frequency for the sake of retaining more temporal and frequency domain information in this section.

In the process of wavelet transform, we use the raw speech signal as input signal firstly, and then it is indispensable that we exploit the Daubechies (dbN) wavelet base to divide the signal into approximation and detail component respectively, where h represents low-frequency domain and g demonstrates high-frequency domain. Next, we divide further the signal processed above into low-high domain by that analogy. Besides, it is significant to select the suitable wavelet basis function that possesses enough capacity to match the raw signal in the time-frequency domain. We take multiple factors into account to choose the Daubechies (dbN) wavelet that is presented by Inrid Danbechies. The dbN wavelet has special characteristics that include simultaneously better orthogonality, compact support and bigger vanishing moment orders with the increase of the sequence ranks so that it is stronger for the ability of localization in the frequency domain.

The whole structure of wavelet packet decomposition is shown in Fig. 2. It is the purpose for us to acquire new wavelet packet reconstruction speech signal compared with raw signal, and we define the approximate component $\Phi\ (t)$ to be $d_1^0\ (t)$ and the detailed component to be $d_1^1\ (t)$ in the first layer, where subscript points at the number of wavelet decomposition layer and superscript points at the location of wavelet packet in the layer. Firstly, we compute the value of wavelet packet function base shown in Eq. (2).

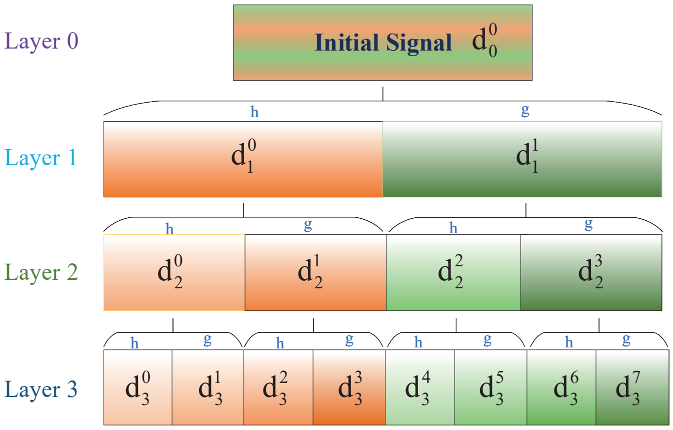

Fig. 2. Illustration of the whole structure in wavelet packet decomposition based on the dyadic wavelet transform

$$\begin{cases} d_{L-1}^{2n}(t) = \sum_k h_k d_L^n(t-k),\ d_L^n(t-k) = d_{L-1}^n(2t-k) \\ d_{L-1}^{2n+1}(t) = \sum_k g_k d_L^n(t-k),\ d_L^n(t-k) = d_{L-1}^n(2t-k) \end{cases} \tag{2}$$

The second formula is also described as Eq. (3):

$$\begin{cases} d^{2n}(t) = \sum\limits_k h_k d^n(2t-k) \\ d^{2n+1}(t) = \sum\limits_k g_k d^n(2t-k), \end{cases} \tag{3}$$

where h_k and g_k represent lowpass and highpass half-band filters respectively, and we adopt the dyadic wavelet transform that discretizes the scale by power series. The scale parameters are 2^i which i denote the number of layers. The parameters of d and k express wavelet packet coefficient and translation variable separately.

Then, we calculate the value of wavelet packet transformation that is used to match raw speech signal as the coefficient of wavelet packet function base shown in Eq. (4). The fourth formula demonstrates the meaning of acquiring the projection values in each wavelet packet function base from raw speech signal according to computing the inner product between above parameters. The more projection values are, the more percentages of feature information are carried by raw signal matched with wavelet signal.

$$\begin{cases} D_\Phi^{2n}(2^i,k) = < f(t), d^{2n}(t) > = \dfrac{1}{\sqrt{2^i}} \sum\limits_t \Phi^* \left(\dfrac{2t-k}{2^i} \right) \\ D_\Phi^{2n+1}(2^i,k) = < f(t), d^{2n+1}(t) > = \dfrac{1}{\sqrt{2^i}} \sum\limits_t \Psi^* \left(\dfrac{2t+1-k}{2^i} \right), \end{cases} \tag{4}$$

where $f(t)$ represents raw speech signal. Finally, we acquire eight new reconstructed speech signals with three layers after exploiting the value of wavelet packet transformation shown as the formula below Eq. (5):

$$\begin{cases} f_{new}^{*2n} = \sum \left(D_\Phi^{2n} \cdot d^{2n}(t) \right) \\ f_{new}^{*2n+1} = \sum \left(D_\Psi^{2n+1} \cdot d^{2n+1}(t) \right). \end{cases} \tag{5}$$

Therefore, the f_{news} are used to be new speech signals to extract further feature maps. In addition, it is a remarkable fact that we select eight new reconstructed signals to execute next feature extraction. Because if we choose less than eight leaves, it could happen that the distribution of frequency domain is not meticulous and it could lead to ignoring much useful frequency information. While if we choose more than eight leaves, the distribution is so redundant that it could generate much unnecessary frequency information and waste training time. It is more suitable for model networks to exploit three wavelet packet layers, so we append an extra experiment that verifies the effect of diverse layers on the performance results shown as Table 3.

3.3. Feature extraction. After acquiring eight new reconstructed speech signals under diverse frequency range, we aim at each reconstructed signal to extract further feature maps. Therefore, it is necessary to exploit frame-level descriptions from each reconstructed speech signal to extract more valuable information while WPR speech signals have a positive advantage on retaining the characteristic of time-frequency domain.

Firstly, we set some fundamental speech parameters that include Hamming windows of 25 ms and a 10 ms shift, and then we receive the confirmed size of windows in the whole speech signal sequences followed with 16 kHz sample rate. Next, in each reconstructed wavelet signal, we utilize these signal sequences to extract features of LLDs (low-level descriptions) that are 36-dimensional vectors include spectral centroid, 0–12th MFCC, chroma vector, etc. After that, the whole sequence should be accomplished the normalization in case of excessive numerical differences, and then we take advantage of the function of pad sequences to set up 1024 as equal-length speech sequence and zero-padding is applied for the sequences whose duration is less than 1024. Finally, we employ the algorithm of matrix dimension transition to constitute 8*36 feature map in a speech signal, so there are 8*36 feature sets generated in this process, and we gain 3-D tensor as a kind of input to put into networks that its shape is the size of (batch size, 1024, 36).

On the other hand, we utilize a set of dense layers to construct a skip-connection structure based on the conventional feature map in order to hold more frequency information directly as another layer of feature extraction. Therefore, we also extract 140 fused low-level descriptions (LLDs) and high-level statistical functions (HSFs) in total from OpenSmile Toolkit [26] as the other input shown in Table 2. The feature set includes different kinds of speech features such as: energy, pitch, voiced, formant and cepstrum. For example, pitch, also known as fundamental frequency F0, is a feature that corresponds to the frequency of vibration of the vocal folds [24]. Formant parameter is an important feature of acoustic parameters related to speech quality. The vocal tract can be regarded as a pipe with non-uniform cross-section. When sound excitation passes through the vocal tract, resonance occurs at a set of frequencies, which are the formant frequencies [2, 27]. In the end, we acquire 1-D tensor as another kind of input to import into model that its shape is the size of (batch size, 140).

3.4. Architecture of proposed neutral networks. After accomplishing the mission of extracting the phrase spectrum features, we will achieve the final classification in ter ms of networks and classifier. In this paper, we will utilize the BiLSTM with a strengthened attention mechanism model and residual block to distill further available emotional information. In the process of model design, we employ BiLSTM networks and attention mechanism structure to process 2-D WPR feature set below 1) and 2). And then, as a skip connection layer, we adopt a set of dense layers to extract 1-D conventional feature set below 3).

1) BiLSTM model: LSTM (long short-term memory) has ability to solve the current policy decision via holding useful historic information, for example, it has ability to exploit the foregone utterance frame with present frame and select expedient message frame in order to recognize emotional state clearly. The purpose of LSTM layers ai ms to retain the temporal and frequency domain information of WPR together as far as possible and take advantage of framework to build the information that does not emerge in the traditional networks. Compared to conventional RNN (recurrent neural network) structure of single tanh function, LSTM possesses particular structure that includes three "door", the gate of 'forget', 'input', 'output' individually. 'Forget' gate is able to assist the model to select information left which includes some weakness frame from the input of model together with previous time output Q_t^l. We take a hypothesis that Q_t^{l-1} belongs to a feature vector that contains more high-level affective frames from the model of DNN. The f_t, i_t, Q_t^l and

Table 2
Mixed conventional feature set

Family	Low level descriptions (LLDs)	High-level statistical functions (HSFs)
Excitation	Short-time Energy Vibrancy Linear regression coefficient 0–250 Hz band proportion	Max/Min/Mean/Std – Mean squared error (MSE) –
Voiced	Voiced frame in differential pitch The first-order vibrancy of pitch The second-order vibrancy of pitch	Max/Min/Mean/Std Max/Min/Mean/Std Max/Min/Mean/Std
Formant	First formant Second formant Third formant	Max/Min/Mean/Std/First-order vibrancy Max/Min/Mean/Std/First-order vibrancy Max/Min/Mean/Std/First-order vibrancy
Cepstrum	MFCC 0–12 First-order differential of MFCC 0–12	Max/Min/Mean/Std Max/Min/Mean/Std

C_t illuminate 'forget', 'input', 'output' gate and a cell with a self-recurrent correlation separately. Next, the formula of updating the LSTM layer is illustrated in Eqs. (6–10).

$$f_t = \sigma\left(W_f Q_t^{l-1} + U_f Q_{t-1}^{l} + b_f\right), \tag{6}$$

$$i_t = \sigma\left(W_i Q_t^{l-1} + U_i Q_{t-1}^{l} + b_i\right), \tag{7}$$

$$o_t = \sigma\left(W_o Q_t^{l-1} + U_o Q_{t-1}^{l} + b_o\right), \tag{8}$$

$$C_t = f_t \cdot C_{t-1} + \tanh\left(W_c Q_t^{l-1} + U_c Q_{t-1}^{l} + b_c\right), \tag{9}$$

$$Q_t^l = o_t \cdot \tanh(C_t), \tag{10}$$

where σ demonstrates the 'sigmoid' loss function, the W_f, U_f, W_i, U_i, W_o, U_o, W_c, U_c are the weight points of above formula respectively and b attributes the value of corresponding 'Bias'.

The bidirectional framework of LSTM makes the model attain the further sequence which makes a benefit in the present state. It indicates that the structure does not only retain the dominant feature information from the past signal sequence, but also inherits the feature pertinence from succedent speech sequence so that it turns the model into a more receptive field to obtain better performance [28].

In this paper, we perform the concatenate wavelet packet reconstruction from the layer of feature extraction into dense layer with 256 cells as a result of considering the feature of timing sequence and dimensionality reduction so that we can put processed data into the structure of RNN, different from [29, 30] etc. Next, we continue to abstract available features from DNN to BiLSTM for temporal summarization and BiLSTM layer with each direction contained 32 cells could elaborate its unique advantages in this stage.

2) Attention layer: The model based on the attention mechanism has generated widespread success in the area of machine translation, it is famous for outstanding performance in the processing of sequence-to-sequence data structure [31]. What is more, it also benefits speech emotion recognition as a result of the specific sequence framework of speech signal. Attention mechanism aims to choose interrelated encoded vectors according to weight value instead of achieving simply a mean or max pooling layer. It can take a series of high-level feature maps from BiLSTM layer to acquire further notable emotional information at the last stage of classification [32, 33], etc. In the framework of attention layers, we mark out the k that is the significant part to draw into the model shown in Eq. (11).

$$k = \sum_{t=1}^{T} \alpha_t Q_t, \tag{11}$$

where Q_t substantiates the n-thhidden condition message of encoding input from recurrent layer that is $Q_t = \left[\overrightarrow{Q_t}; \overleftarrow{Q_t}\right]$ at n-th time step. It has ability to integrate available hidden information amongst forward sequence and backward sequence. And it is a remarkable fact that a significant parameter α_t decides to gain the major sequence data in the layer of output shown in Eq. (12).

$$\alpha_t = \frac{\exp(W \cdot Q_t)}{\sum\limits_{t=1}^{T} \exp(W \cdot Q_t)}, \tag{12}$$

where α_t expresses the relationship weights between the n-th sequence and the whole sequence. Therefore, it is vital to select useful information according to the value of weights, and the bigger the weights, the more valuable emotional message is. Besides, k denotes the representations of utterance-level which will be laid into our proposed layer to obtain the further high-emotional explanation.

Next, we pick up further enhanced features by means of setting a layer of upsampling after putting sequence features into attention mechanism. First, we perform a layer of dense with one cell and the loss function is set to 'sigmoid', and then it is necessary to append an upsampling layer with 64 cells for the sake of amplifying the concentrated useful feature area. In this process, we make the shape of feature sequence more suitable for the input and output by adding reshape layer with 64 cells. Next, we combine the original feature from the output of attention layer and postprocessed feature set to perform a multiplication operation. The purpose of designing such an enhanced attention structure is to enlarge useful feature sequence as much as possible and narrow down useless feature area shown in Fig. 3.

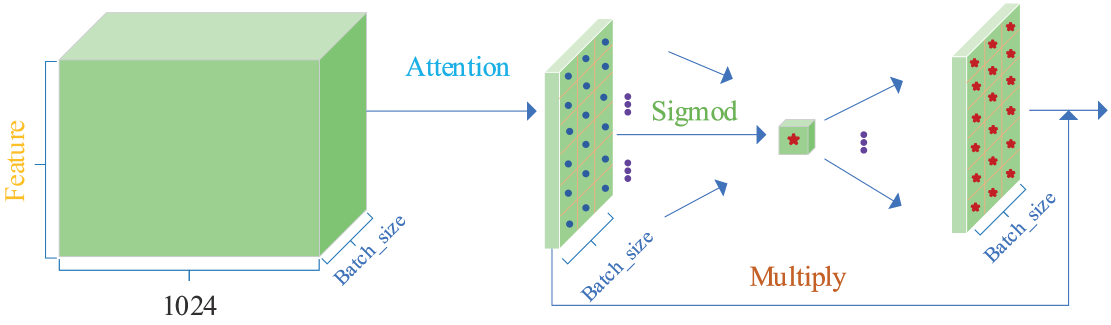

Fig. 3. Illustration of the whole structure to enhance attention mechanism based on the upsampling layer

3) Skip connection: For containing some frequency domain information, we employ 140 descriptions feature set from the second extraction layer as a skip connection structure to catenate postprocessed data from attention framework, and we decide to exploit two fully connected layers to extract further feature set with 128 cells and 64 cells respectively in the branch. The application of batch normalization has ability to help fully connected layer to accelerate the process of training and strengthens the generalization performance instead of the trick of dropout.

Finally, the refining feature map is adopted as a fully connected layer to get the classification result for SER with the loss function of ʻsoftmax' to better classify the utterance representations into N diverse areas, where N means the number of emotional categories.

4. Experiment

4.1. Experiment setup. In this paper, we utilize the famous interactive emotional dyadic motion capture corpus (IEMOCAP) [34] and the Berlin Database of Emotional Speech (EMODB) [35] to test and evaluate the effectiveness and robustness of features that are the aggregation of WPR and frame-level above the performance of our proposed framework. The IEMOCAP corpus was collected following the theatrical theory for the sake of simulating natural dyadic interactions between motion capture markers and audio data from five pairs of actors (one male and one female of each pair). It is utilized two types of emotional elicitation method, one is the theatrical scripts performances, another is the improvisations of emotional scenarios experiments in the dialog recording. There are 1150 utterances in the prototypical data (entire agreement on the emotional state from evaluators) of improvisations of emotional scenarios (neutral, angry, excited, happy, frustrated, sad and surprise). The EMODB database contains speech samples in German language spoken by 5 male and 5 female to express seven emotional state (angry, neutral, fear, disgust, joy, sadness and boredom) and display 535 sentences from life every day. The speech signals are sampled at 16 kHz sampling frequency and 18 bits in each sampling point.

In IEMOCAP corpus, we get ready for utilizing the utterances that are the emotional improvisational scenarios spontaneously and extract four affective state: neutral, happy, sad and angry. In addition, data from the whole EMODB database with seven emotions can be applied in our performances.

For both of two databases, we both employ a 10-fold cross-validation contain ten speakers. Specifically, we select 8 speakers as the training set, 1 speaker as the test set, and the other is chosen to be validation set for obtaining better model. Besides, we adopt five times diverse parameter initialization with different random seed for confirming the significance of proposed feature set and we report the unweighted accuracy with the average on the test data due to the imbalance property of data distribution. The destination of selecting above-mentioned method guarantees the results that are more genuine and believable.

Besides, we conducted extensive speaker-dependent experiments with our contribution to the selected databases respectively, and the division of each corpus is different from speaker-independent performance. In each experiment, it is necessary for us to split the whole set into two sets randomly that include 80% training set and 20% testing set of the data with diverse random seeds. Then, we also report the unweighted accuracy with the average value on the test data. The similar results of the performances illuminate that the designed model networks based on proposed feature extraction have a positive effect on speech emotion recognition shown as Table 6 and Table 7.

We start to extract the wavelet packet reconstruction feature map from the prepared training set after separating the databases. Then our proposed framework is responsible for disposing these features. For better helping the model classify, we set some appropriate hyperparameters to fit the modal include the number of min-batch, the value of learning rate and momentum from optimizer etc. Moreover, we also have some tricks like using the cross-entropy objective function, Adam optimizer [36], and batch normalization [37]. The experiment environment is implemented in the Keras toolkit.

4.2. Comparison of networks architectures. In this paper, we propose a novel algorithm with utilizing 2-D wavelet packet reconstruction feature and 1-D conventional LLDs feature set combined with suitable networks respectively. It is worth nothing that it has more superiorities in the process of feature extraction, because WPR signal set contains quite a bit of information from temporal domain and frequency domain. Besides, we also improve our framework of model with appending an upsampling layer in our contribution appropriately and achieve better performance in the experiments. Therefore, for confirming the effectiveness of our innovation, we compare our approach with several baselines that don't take account of something additional simulation conditions at the same time.

In the ablation study from speaker-independent experiment shown in Table 3, we set up the fundamental baseline with exploiting 140 1-D original LLDs feature singly as the comparison experiment firstly. It does not only perform the 58.41% and 72.56% recognition accuracy in two databases respectively, but we also employ the diverse configurations of layers to find more appropriate model structure relatively. The basic experiment illuminates that two FCN (fully convolutional networks) layers acquire better results and use less parameters with 128 and 64 cells based on the batch normalization compared with other comparisons of reference. Though we can see diverse dense layers that could lead to tiny difference in the initial stage of feature extraction, we also adopt two layers due to considering the complexity of time and space.

Next, we investigate another comparison experiment that demonstrates the influence of experiment results under different wavelet packet decomposition (WPD) layers. The results express that three WPD layers have better performance compared with others, because less WPD layers could lead to ignore some available frequency domain information and more WPD layers could cause lots of redundant parameters and waste train-

Table 3
The ablation study of two databases with SER accuracy results (%) based on different algorithm in speaker-independent (SI) and speaker-dependent (SD) [WP: wavelet packet; BN: batch normalization]

Features (size)	Methods	Configurations of layers	SI (IEMOCAP)	SD (IEMOCAP)	SI (EMODB)	SD (EMODB)
LLDs (140)	DNN	2*FCN(128,64) + dropout (0.5) 2*FCN(128,64) + BN 2*FCN(128,64,32) + BN	57.76 **58.41** 58.24	71.78 **72.56** 72.16	60.68 **62.17** 61.84	70.85 **72.85** 74.03
No WP layer (36) 1 WP layer (72) 2 WP layers(144) 3 WP layers(288) 4 WP layers (576)	DRNN-attention (no upsampling)	FCN(256) + BiLSTM + attention (64)	54.35 56.63 57.86 **58.05** 57.97	67.23 69.14 71.36 **72.62** 72.45	58.84 59.76 61.65 **63.24** 63.14	69.74 72.03 73.27 **74.43** 74.06
WPR with fused conventional feature	DRNN-attention (no upsampling)	FCN(256) + BiLSTM + Attention (64)	60.84	75.96	65.15	79.34
	DRNN-attention (no upsampling)	FCN (256) + BiLSTM + attention (64) + upsampling (64)	**62.52**	**77.57**	**66.90**	**82.26**

ing time under the similar performance. The experimental result demonstrates 58.05% and 72.62% recognition accuracy approximately in two databases respectively. Compare with no WPR signal set, we gain about 3.70% and 5.39% recognition accuracy improvement. However, it achieves better performance 60.84% and 75.96% accuracy when we combine two feature set between 2-D WPR and 140 1-D LLDs, since conventional LLDs features have ability to replenish some useful frequency domain information. After that, we also make a comparison experiment that only adopts 2-D WPR feature set to input into DRNN-attention model structure based on no upsampling layer with average of five times in speaker-independent experiment. Compared with no upsampling layer, the complete model structure has ability to obtain 1.68% and 1.61% accuracy majorization separately in final.

On the other hand, we also employ another speaker-dependent classification style to validate the effectiveness and generalization of our proposed contribution. From the ablation study of Table 3, we discover that the performance results gradually improve along with the improvement of model structure and feature extraction. The mutative tendency of experimental results in speaker-dependent is similar with above performance classification. In the aspect of feature extraction of WPR signal set, the selection of three wavelet packet layers acquires 63.24% and 74.93% in two corpora respectively. In addition, we achieve 66.90% and 82.26% accuracy results with our proposed model framework after combining the WPR signal feature set and conventional LLDs feature set at last.

Therefore, we can find that emotional classification of IEMOCAP and EMODB are recognized better based on our designed novel feature framework and layers architecture.

4.3. Experiment results. Furthermore, we also demonstrate the confusion matrix to further analyze the experimental results of our proposed algorithm in speaker-independent and speaker-dependent performance illustrated in Tables 4–7 respectively.

In the speaker-independent experiment, we can discover improved recognition rate in each emotional state compared with baselines. Like most studies, our research also achieves a high recognition accuracy in angry and sad emotion for IEMOCAP corpus, which can obtain 64.14% and 79.16% result respectively as indicated in Table 4. Note that, it has a notable improvement that we can find the accuracy of neutral and happy emotion which is corresponding to 58.10% and 48.68% respectively compared with some other researches [23]. It could explain the mixed feature maps that promote the distinguished ability between neutral and happy emotion. In addition, the performance of speaker-independent in the EMODB database is also satisfactory as shown in Table 5. We also make similar progress in exploiting proposed algorithm into this database, it obtains 89.76%, 93.55% and 86.08% recognition accuracy rate in angry, sad and neutral emotional states separately, while 53.52% of happy emotions are misclassified as angry emotion. It could attribute these mistakes that the distribution of two kinds of emotional state is so close in activation-valence space that it is hard to distinguish the difference in mapping frequency domain. Besides, the less data volume of database is another reason to cause lower accuracy result in SER.

In the speaker-dependent experiment of IEMOCAP corpus, we acquire 73.31% and 82.37% accuracy results on average

Table 4
Confusion matrix with speaker-independent in the IEMOCAP corpus

		Predicated class			
		Angry	Sad	Happy	Neutral
True class	Angry	**64.14**	2.95	15.19	17.72
	Sad	1.40	**79.16**	4.38	15.06
	Happy	10.67	03.48	**48.68**	37.17
	Neutral	3.29	19.49	19.12	**58.10**

Table 5
Confusion matrix with speaker-independent in the EMODB database

		Predicated class						
		Angry	Boredom	Disgust	Fear	Happy	Sad	Neutral
True class	Angry	**89.76**	0.00	0.79	1.57	7.87	0.00	0.00
	Boredom	0.00	**79.16**	0.00	1.23	0.00	8.64	11.11
	Disgust	2.27	4.55	**70.45**	11.36	4.55	2.27	4.55
	Fear	2.94	2.94	1.47	**70.59**	2.94	1.47	17.65
	Happy	29.58	1.41	1.41	11.27	**53.32**	0.00	2.82
	Sad	0.00	1.61	0.00	1.61	0.00	**93.55**	3.23
	Neutral	0.00	10.13	0.00	1.27	0.00	2.53	**86.08**

corresponding to angry and sad emotional state respectively, and it has great improvement in angry emotion compared with speaker-independent especially as shown in Table 6. Besides, happy emotion and neutral emotion only have a little improvement in accuracy results, because two kinds of emotional state own some analogical emotion information in frequency domain, which causes some confusion in speech emotion recognition. After that, from speaker-dependent experiment in the EMODB database, there is an upward trend overall compared with another classification style shown in Table 7. It is a remarkable fact that our suggested method makes progress in the emotional state of angry, sad and neutral. These kinds of emotional states are corresponding to 91.34%, 93.55% and 88.61% accuracy rate respectively. However, like the comparability of analysis in speaker-independent experiment, happy and disgust emotions generate less accuracy rate results due to the lack of data in the EMODB database. In addition, the classification tests for other people outside the training set also give low recognition accuracy. The reason may be the uneven distribution of test set data and the small size of test data volume.

Table 6
Confusion matrix with speaker-dependent in the IEMOCAP corpus

		Predicated class			
		Angry	Sad	Happy	Neutral
True class	Angry	**71.31**	2.53	13.50	12.66
	Sad	1.75	**82.37**	3.49	12.39
	Happy	9.62	3.28	**51.80**	35.31
	Neutral	3.20	17.84	16.83	**62.12**

The two kinds of experimental results both reveal that the feature map chosen and the best SER architecture heavily depend on the type and size of the database, and this finding is a great significance for the development of SER syste ms on new datasets.

5. Conclusions

In this paper, we propose an original feature set comprised of wavelet features from wavelet packet reconstruction signals combined connected with conventional LLDs feature set to constitute mixed feature set, and it has ability to retain effective and advantageous information in temporal and frequency domain. Next, we utilize the novel model framework to further extract feature and acquire better recognition results in the

Table 7
Confusion matrix with speaker-dependent in the EMODB database

		Predicated class						
		Angry	Boredom	Disgust	Fear	Happy	Sad	Neutral
True class	Angry	**91.34**	0.00	0.79	1.57	6.30	0.00	0.00
	Boredom	0.00	**83.95**	0.00	1.23	0.00	6.17	8.64
	Disgust	2.27	4.55	**77.27**	6.82	4.55	2.27	2.27
	Fear	2.99	1.49	1.49	**79.10**	2.99	1.49	10.45
	Happy	19.72	1.41	1.41	11.27	**61.97**	0.00	4.23
	Sad	0.00	1.61	0.00	1.61	0.00	**93.55**	3.23
	Neutral	0.00	10.13	0.00	1.27	0.00	2.53	**88.61**

speaker-independent and speaker-dependent experiment from two databases respectively. The novel model includes FCN networks, BiLSTM based on the attention mechanism with UpSampling1D layer to deal with WPR feature set, and then FCN networks with batch normalization to process conventional feature set as another skip connection structure. Hence, it contains more temporal-frequency information and contextual dependencies from the features learned by our proposed algorithm. We also adopt the skill of voice activity detection to reduce noise and redundancy as much as possible so that we get a better performance. The experiment results demonstrate our proposed approach has more superiority in ter ms of recognition accuracy rate.

Although our proposed algorithm has made progress in the aspect of SER, we still need to accomplish other research about how to promote further the performance of recognition. In the future, we will apply ourselves to the model robustness and research more flexible framework that has the characteristic of stronger generalization in other speech databases. Besides, it is necessary to be conscious of how to accomplish better feature extraction. Because of the "black box" characteristics of deep learning network model, it is hard to describe which specific feature is useful for specific emotions when exploiting network model to extract features. We will also focus on this research to explore the role of specific features in deep learning methods under the SER.

Acknowledgements. Development Project of Ship Situational Intelligent Awareness System, Grant MC-201920-X01; National Natural Science Foundation of China, Grant 61673129.

References

[1] M. Gupta, et al., "Emotion recognition from speech using wavelet packet transform and prosodic features", *J. Intell. Fuzzy Syst.* 35, 1541–1553 (2018).

[2] M. El Ayadi, et al., "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognit.* 44, 572–587 (2011).

[3] P. Tzirakis, et al., "End-to-end speech emotion recognition using deep neural networks", *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5089–5093, doi: 10.1109/ICASSP.2018.8462677.

[4] J.M Liu, et al., "Learning Salient Features for Speech Emotion Recognition Using CNN", *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, Beijing, China, 2018, pp. 1–5, doi: 10.1109/ACIIAsia.2018.8470393.

[5] J. Kim, et al., "Learning spectro-temporal features with 3D CNNs for speech emotion recognition", *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, San Antonio, USA, 2017, pp. 383–388, doi: 10.1109/ACII.2017.8273628.

[6] M.Y Chen, X.J He, et al., "3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition", *IEEE Signal Process Lett.* 25(10), 1440–1444 (2018), doi: 10.1109/LSP.2018.2860246.

[7] V.N. Degaonkar and S.D. Apte, "Emotion modeling from speech signal based on wavelet packet transform", *Int. J. Speech Technol.* 16, 1–5 (2013).

[8] T. Feng and S. Yang, "Speech Emotion Recognition Based on LSTM and Mel Scale Wavelet Packet Decomposition", *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 2018)*, New York, USA, 2018, art. 38.

[9] P. Yenigalla, A. Kumar, et. al", Speech Emotion Recognition Using Spectrogram & Phoneme Embedding Promod", *Proc. Interspeech 2018*, 2018, pp. 3688–3692, doi: 10.21437/Interspeech.2018-1811.

[10] J. Kim, K.P. Truong, G. Englebienne, and V. Evers, "Learning spectro-temporal features with 3D CNNs for speech emotion recognition", *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, San Antonio, USA, 2017, pp. 383–388, doi: 10.1109/ACII.2017.8273628.

[11] S. Jing, X. Mao, and L. Chen, "Prominence features: Effective emotional features for speech emotion recognition", *Digital Signal Process.* 72, 216–231 (2018).

[12] L. Chen, X. Mao, P. Wei, and A. Compare, "Speech emotional features extraction based on electroglottograph", *Neural Comput.* 25(12), 3294–3317 (2013).

[13] J. Hook, et al., "Automatic speech based emotion recognition using paralinguistics features", *Bull. Pol. Ac.: Tech.* 67(3), 479–488, 2019.

[14] A. Mencattini, E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, and C. Di Natale, "Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure", *Knowl.-Based Syst.* 63, 68–81 (2014).

[15] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics", *Speech Commun.* 53(1), 36–50 (2011).

[16] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge", *Comput. Speech Lang.* 27(1), 4–39 (2013).

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

[18] G.Trigeorgis et.al, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network", *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5200–5204, doi: 10.1109/ICASSP.2016.7472669.

[19] Y. Xie et.al, "Attention-based dense LSTM for speech emotion recognition", *IEICE Trans. Inf. Syst.* E102.D, 1426–1429 (2019).

[20] F. Tao and G.Liu, "Advanced LSTM: A Study about Better Time Dependency Modeling in Emotion Recognition", *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 2906–2910, doi: 10.1109/ICASSP.2018.8461750.

[21] Y.M. Huang and W. Ao, "Novel Sub-band Spectral Centroid Weighted Wavelet Packet Features with Importance-Weighted Support Vector Machines for Robust Speech Emotion Recognition", *Wireless Personal Commun.* 95, 2223–2238 (2017).

[22] Firoz Shah A. and Babu Anto P., "Wavelet Packets for Speech Emotion Recognition", *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Chennai, 2017, pp. 479–481, doi: 10.1109/AEEICB.2017.7972358.

[23] K.Wang, N. An, and L. Li, "Speech Emotion Recognition Based on Wavelet Packet Coefficient Model", *The 9th International Symposium on Chinese Spoken Language Processing*, Singapore, China, 2014, pp. 478–482, doi: 10.1109/ISCSLP.2014.6936710.

[24] S. Sekkate, et al., "An Investigation of a Feature-Level Fusion for Noisy Speech Emotion Recognition", *Computers* 8, 91 (2019).

[25] Varsha N. Degaonkar and Shaila D. Apte, "Emotion Modeling from Speech Signal based on Wavelet Packet Transform", *Int. J. Speech Technol*. 16, 1–5 (2013).

[26] F. Eyben, et al., "Opensmile: the munich versatile and fast open-source audio feature extractor", *MM '10: Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[27] Ch.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artif. Intell*. 43(2), 155–177 (2015).

[28] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech Emotion Recognition From 3D Log-Mel SpectrogramsWith Deep Learning Network", *IEEE Access* 7, 125868–125881 (2019).

[29] Keren, Gil and B. Schuller. "Convolutional RNN: An enhanced model for extracting features from sequential data," *International Joint Conference on Neural Networks*, 2016, pp. 3412–3419.

[30] C.W. Huang and S.S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition", *IEEE International Conference on Multimedia and Expo (ICME)*, Hong Kong, 2017, pp. 583–588, doi: 10.1109/ICME.2017.8019296.

[31] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic Speech Emotion Recognition using Recurrent Neural Networks with Local Attention", *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 2227- 2231, doi: 10.1109/ICASSP.2017.7952552.

[32] Ashish Vaswani, et al., "Attention Is All You Need", *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, USA, 2017.

[33] X.J Wang, et al., "Dynamic Attention Deep Model for Article Recommendation by Learning Human Editors' Demonstration", *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017.

[34] C. Busso, et al., "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources & Evaluation* 42(4), 335 (2008).

[35] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B.Weiss, "A database of German emotional speech," *INTERSPEECH 2005 – Eurospeech*, Lisbon, Portugal, 2005, pp. 1517–1520.

[36] D. Kingma and J. Ba, "International Conference on Learning Representations (ICLR)", *ICLR*, San Diego, USA, 2015.

[37] F. Vuckovic, G. Lauc, and Y. Aulchenko. "Normalization and batch correction methods for high-throughput glycomics", *Joint Meeting of the Society-For-Glycobiology* 2016, pp. 1160–1161.