Speech Emotion Recognition Using Fusion of Three Multi-Task Learning-based Classifiers: HSF-DNN, MS-CNN and LLD-RNN

Zengwei Yao, Zihao Wang, Weihuang Liu, Yaqian Liu, Jiahui Pan

PII: S0167-6393(19)30032-9
DOI: https://doi.org/10.1016/j.specom.2020.03.005
Reference: SPECOM 2698

To appear in: *Speech Communication*

Received date: 6 April 2019
Revised date: 5 February 2020
Accepted date: 20 March 2020

Highlights:

- Attention-based weighted-pooling method for efficient utterance-level aggregation
- Learning generalized representation between emotion classification task and regression task
- Confidence-based decision-level fusion of distinctive classifiers

Speech Emotion Recognition Using Fusion of Three Multi-Task Learning-based Classifiers: HSF-DNN, MS-CNN and LLD-RNN

Zengwei Yao[a,1], Zihao Wang[a,1], Weihuang Liu[a], Yaqian Liu[a], Jiahui Pan[a,*]

[a]*School of Software, South China Normal University, Guangzhou 510641, China*

Abstract

Speech emotion recognition plays an increasingly important role in emotional computing and is still a challenging task due to its complexity. In this study, we developed a framework integrating three distinctive classifiers: a deep neural network (DNN), a convolution neural network (CNN), and a recurrent neural network (RNN). The framework was used for categorical recognition of four discrete emotions (i.e., angry, happy, neutral and sad). Frame-level low-level descriptors (LLDs), segment-level mel-spectrograms (MS), and utterance-level outputs of high-level statistical functions (HSFs) on LLDs were passed to RNN, CNN, and DNN, separately. Three individual models of LLD-RNN, MS-CNN, and HSF-DNN were obtained. In the models of MS-CNN and LLD-RNN, the attention mechanism based weighted-pooling method was utilized to aggregate the CNN and RNN outputs. To effectively utilize the interdependencies between the two approaches of emotion description (discrete emotion categories and continuous emotion attributes), a multi-task learning strategy was implemented in these three models to acquire generalized features by simultaneously operating classification of discrete categories and regression of continuous attributes. Finally, a confidence-based fusion strategy was developed to integrate the power of different classifiers in recognizing different emotional states. Three experiments on emotion recognition based on the IEMOCAP corpus were conducted. Our experimental results show that the weighted pooling method based on attention mechanism endowed the neural networks with the capability to focus on emotionally salient parts. The generalized features learned in the multi-task learning helped the neural networks to achieve higher accuracies in the tasks of emotion classification. Furthermore, our proposed fusion system achieved weighted accuracy of 57.1% and unweighted accuracy of 58.3%, which were significantly higher than those of each individual classifier. The effectiveness of the proposed approach based on classifier fusion was thus validated.

Keywords: speech emotion recognition, attention mechanism, multi-task learning, classifier fusion

[*] Corresponding author
Email address: panjh@qq.com (Jiahui Pan)
[1] These authors contributed equally to the manuscript.

1. Introduction

Affective computing and sentiment analysis have been continuously attracting the interest of many researchers ever since their proposal [1]. In the area of affective computing, one of the key components is emotion recognition, which aims to recognize meaningful patterns extracted from gathered data including facial expression, human language, and speech segments. Among all types of data, speech segments are a valuable source of emotional information. The technology of speech emotion recognition has gained increasing popularity in fields of human-computer interaction such as artificial customer service [2], car driving [3], distance education [4], and medical assistance [5]. However, this technology remains a challenging issue due to the complexity of emotional expressions, arising from speaker variability, gender differentiation, and age-related differences [6, 7, 8].

Currently, speech emotion recognition tasks aim to make a decision on the utterance-level, and most related works focus on obtaining an effective representation that is relevant to its emotional state. Almost all studies consist of two stages: first, an utterance is divided into multiple short parts and features are extracted at these short-term levels; then, these short-term level features are aggregated into an utterance-level vector through an efficient method. Some acoustic features are considered to represent short-term emotional information, such as pitch, voicing probability, energy, zero-crossing rate, mel-filterbank features, and mel-frequency cepstral coefficients (MFCCs). These features are extracted from short frames whose length varies from 20 ms to 50 ms and are referred to as frame-level low-level descriptors (LLDs). However, the perception of emotion usually depends on the emotional information expressed in the speech over a certain period of time. It is often necessary to calculate the global characteristics of the whole utterance. Therefore, the extracted LLDs are concatenated into an utterance-level vector by applying numerous of high-level statistical functions (HSFs), which describe temporal variations over multiple frames [9]. A variety of models have been applied for classification based on HSFs, in which support vector machines (SVM) still perform competitively and act as baseline systems [10, 11]. Recently, a number of researchers applied the deep neural network (DNN) to automatically learn relevant features for speech emotion recognition. Stuhlsatz *et al.* used a DNN on the top of utterance-level HSFs to learn deeper representation [12]. Many researchers have used recurrent neural network (RNN) to learn long-time context from multiple frame-level LLDs [13, 14, 15, 16]. Meanwhile, attention mechanisms have been applied to focus on the emotionally-relevant parts instead of the whole utterance [14, 15, 17, 18]. For instance, Mirsamadi *et al.* utilized RNN for temporal modeling and investigated the performance of different pooling methods, such as mean pooling method and weighted pooling method, for aggregating the RNN outputs over all frames into utterance-level representation [14]. They proposed that the weighted pooling method based on attention mechanisms was the most efficacious because it automatically assigned more weights on emotionally salient parts and overcame the weakness of handling emotionally-irrelevant frames. Previous studies have begun to utilize Convolution neural network (CNN) to directly learn deep features from mel-spectrograms (MS) instead of acoustic features [19, 20, 21, 22]. To obtain MS, the short frames are converted into spectrograms by Short Term Fourier Transform (STFT) and the computed magnitude spectrograms are mapped to the mel-scale. The obtained MS images were fed into CNN to learn frequency and time domain representations. However, in studies using CNN based methods, there is a lack of in-depth investigations of implementing aggregation approaches over different time steps similar to those applied in RNN with LLDs [14]. The frame-level LLDs are hand-crafted acoustic features from short parts, and segment-level spectrograms are visual representations of a spectrum of frequencies from middle-length parts as it varies with time. The utterance-level HSFs are certain statistics based on LLDs over all frames of an utterance. These three are different forms of commonly used input feature maps, which need to be transformed to characterize the emotional state of an utterance more relevantly through respective models.

The multi-task strategy is known for its capability to learn inner relationships and improve generalization due to its sharing representations among related tasks, leading to its wide applications in speech processing area [23, 24, 25, 26]. Emotional states can be described either by commonly used discrete categories, such as angry and happy, or through continuous attributes, in which each emotional state is represented as a point in the 2-dimensional space defined by valence and activation. Previous research has shown that applying the strategy of multi-task learning to simultaneously generate a classification task on discrete categories and a regression task on continuous attributes could obtain more generalized representation through neural networks [27]. Considering continuous attribute information, to a great extent, helps improve the classification task. This multi-task learning method enables neural networks to acquire the interdependencies between different emotion representation spaces. Additionally, this method is efficient in using data, and thus it can be an appropriate candidate to further improve the performance of the above mentioned DNN models.

Different modeling methods will obtain different types of speech emotion information. Rather than using only a single type of information, we implemented a fusion strategy on the various sources of information to make a decision in classification, and a more sensible conclusion could be reached. The fusion methods of speech emotion features can be roughly separated into two categories: feature-level fusion [28, 29] and decision-level fusion[30, 31]. In feature-level fusion, emotion features extracted by different models are combined to generate a more informative representation for classification. For example, Zhao *et al.* established an end-to-end framework for speech emotion recognition, which directly concatenated the outputs from CNN and RNN and generated a spatio-temporal representation [28]. However, the method which straightforwardly concatenated features failed to take into account the disparity between output features generated by CNN and RNN, and a large scale of training was required for an acceptable outcome. In decision-level fusion, multiple classifiers are trained to make predictions, and their results are jointly analyzed to make final decisions. For example, Su *et al.* constructed a framework including an RNN model with multiple frame-level acoustic LLDs as inputs and a SVM model with HSFs as inputs. Both RNN and SVM generated the probabilities of concerned classes, which were used as the confidence score for these two models. The confidence scores of RNN and SVM were averaged for each concerned class as the fusion confidence score for the integrated model [30]. Considering the insufficiency of speech emotion corpus, in this study, we adopted decision-level fusion to achieve high performance on the emotion recognition task.

In this study, we proposed an efficient speech emotion recognition framework that integrated three distinctive classifiers to discriminate four common emotional states: angry, happy, neutral, and sad. First, we used DNN to obtain high-level representation from HSFs, CNN to model time and frequency domain from spectrograms, and RNN to learn long-context information from LLDs. To enable the neural network to automatically handle the problem of uneven distribution of emotion information across utterance parts, we employed an attention mechanism based weighted pooling method to generate a more efficient utterance-level representation. In addition, to achieve better performance with currently limited data volume, we applied multi-task learning to simultaneously operate classification of discrete categories and regression of continuous attributes so that the generalized features among different tasks could be acquired. Finally, we developed a confidence-based fusion method to integrate the varying capabilities of different models for handling different emotional states. Therefore, a more comprehensive classification system could be obtained by performing a fused recognition. We conducted experiments with the IEMOCAP dataset [32]. The experimental results show that the weighted-pooling method could efficiently aggregate the outputs of RNN and CNN to an utterance-level representation. The multi-task learning strategy enabled the neural networks to capture generalized features to achieve improvements in the recognition performance. Moreover, the strategy of confidence-based fusion was capable of integrating the power of each sub-classifier to achieve a significantly higher performance in emotion classification.

The major contributions of this study are summarized as follows. (1) An attention-based weighted pooling method for aggregating outputs of CNN and RNN respectively was proposed to enable the neural network to focus automatically on the emotionally obvious parts. (2) A multi-task learning strategy was proposed to leverage the information of continuous attributes. (3) A confidence-based fusion method was proposed to integrate the power of different classifiers (DNN, CNN, and RNN) in recognizing different emotional states.

2. Data

We used the interactive emotional dyadic motion capture (IEMOCAP) dataset [32] to evaluate the performance of our proposed method. IEMOCAP is widely used for speech emotion recognition, and it is recorded by 10 actors and contains 5 sessions. The speech of each session came from a man and a woman and the total duration was 12 hours. Each utterance from either of the actors in the interaction was evaluated categorically over the previously defined emotion set (angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and other) by at least three different annotators, and dimensionally over the axes of valence, activation, and dominance by at least two different annotators. The affective dimension ratings range from 1 to 5. The most commonly used utterances for emotion category classification fall into four categories (i.e., 1103 in angry, 1636 in happy, 1708 in neutral and 1084 in sad). We performed experiments with a total of 5531 utterances from these four categorical emotional states.

3. Methods

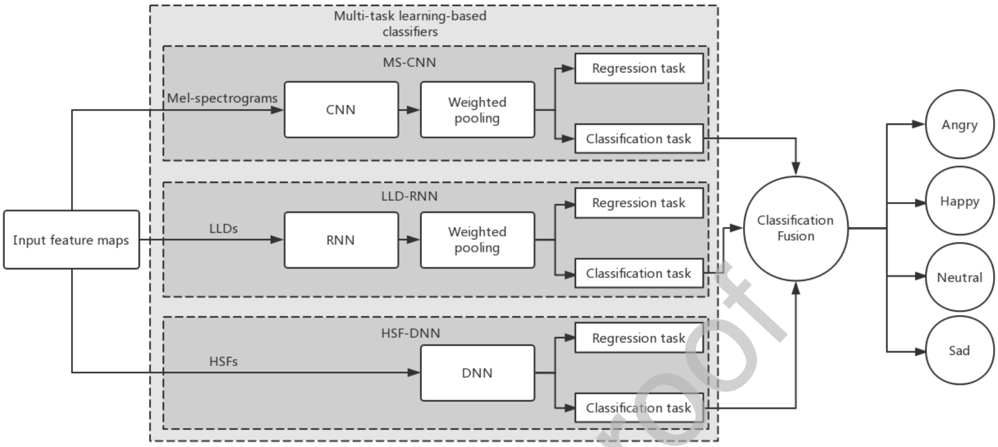


Figure 1: Framework of the proposed method using fusion of three classifiers: MS-CNN, LLD-RNN and HSF-DNN

As shown in Figure 1, we separately trained three classifiers using different types of inputs, and then applied the fusion strategy at the decision level to reach the final conclusion. Specifically, we passed HSFs to DNN (denoted as HSF-DNN), MS to CNN (denoted as MS-CNN), and LLDs to RNN (denoted as LLD-RNN). We implemented a multi-task learning framework for each individual classifier. The involved components are described in detail in the following sections.

3.1. Input feature maps

The speech signal was sampled at 16000Hz sampling rate, and windowed into short frames. For MS extraction, the length of each frame was 32-ms without overlap and the librosa toolkit [33] was utilized. For LLD and HSF extraction, the length of each frame was 25-ms with 15-ms overlap between neighboring frames and the openSMILE toolkit [34] was utilized.

To extract MS, each frame x was converted into magnitude spectrum X by applying Discrete Fourier Transform (DFT):

$$X[k] = \sum_{i=0}^{N-1} x[i] \exp(\tfrac{-j2\pi ik}{N}); \quad (0 \leq k \leq N-1) \tag{1}$$

where N was the number of points within each frame. Then the mel-spectrum s was computed with the following formula:

$$s[m] = \sum_{k=0}^{N-1} (|X[k]|^2 H_m(k)); \quad (0 \leq m \leq M-1) \tag{2}$$

where M was the number of mel-scale filters and $H_m(k)$ denoted the *m-th* triangular filter output on *k-th* spectrum bin. In this study, 40 mel-scale filters were used here. Each frame, surrounded by 7 neighboring left contextual vectors and 8 neighboring right contextual vectors, was formed as a 16 × 40-dimensional segment-level MS image.

To extract LLD of each frame x, the extracted acoustic features consisted of the root-mean-square signal frame energy (RMSE), zero-crossing rate of time signal (ZCR), fundamental frequency (F_0), harmonics-to-noise ratio (HNR) and 12 MFCCs. The RMSE was defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}|x[i]|^2} \tag{3}$$

The ZCR was defined as:

$$ZCR = \frac{1}{N-1}\sum_{i=1}^{N-1} I_{<0}(x_i x_{i-1}) \tag{4}$$

where $I_{<0}(x_i x_{i-1})$ indicated whether $x_i x_{i-1}$ was negative. The F_0 was defined as:

$$F_0 = \frac{1}{T} \tag{5}$$

where T denoted the smallest value of period. The HNR was defined as:

$$HNR = \frac{\sum_{i=0}^{N-1} h[i]^2}{\sum_{i=0}^{N-1} n[i]^2} \tag{6}$$

where $x[i] = h[i] + n[i]$, $h[i]$ denoted the harmonic component, and $n[i]$ denoted the noise component. To compute MFCCs c, the mel-spectrum s was transformed into log scale and then the Discrete cosine transform (DCT) was applied to obtain cepstral coefficients with the following formula:

$$c[i] = \sum_{m=0}^{M-1} log_{10}(s[m]))cos(\frac{\pi i(m-0.5)}{M}); \quad (i = 0, 1, 2, ..., C-1) \tag{7}$$

where C denoted the number of MFCCs. In addition, for each of above 16 acoustic features, the delta coefficients were computed. Thus, the 32-dimensional LLD of each frame was obtained.

To extract HSF, twelve statistical aggregation functions were applied to the extracted LLDs of each frame over an entire utterance. The statistical functions consisted of mean, range, standard deviation, kurtosis, skewness, minimum and maximum values, absolute positions of the maximum and minimum values, slope and offset of the linear regression line, quadratic error between contour and linear regression line. For each function F_i, a 32-dimensional vector HSF_i could be calculated as the following formula:

$$HSF_i = F_i(LLD_1, LLD_2, ..., LLD_T); \quad (0 \le i \le 11) \tag{8}$$

where T denoted the number of frames of the current utterance. Thus, the results were concatenated into an utterance-level 384-dimensional vector.

The contiguous 32-dimensional frame-level LLDs, 16 × 40-dimensional segment-level MS, and 384-dimensional HSFs were fed into RNN, CNN, and DNN, separately.

3.2. Multi-task learning-based classifiers

We developed three multi-task learning-based classifiers HSF-DNN, MS-CNN, and LLD-RNN. They were all trained for emotion category classification tasks and attribute regression tasks, while the weighted pooling layer was applied to both MS-CNN and LLD-RNN.

3.2.1. DNN

DNN was capable of learning high-level representation with the non-linear transformation operations, and used to learn a more emotionally discriminative vector from the utterance-level HSF vector [12]. In order to transform the utterance-level HSF vector into a deeper representation, we utilized the non-linear transformation operations of DNN to obtain a more emotionally discriminative vector. As shown in Figure 2, in this study, we used 2 fully connected layers with a *ReLU* activation function and each layer consisted of 128 units.

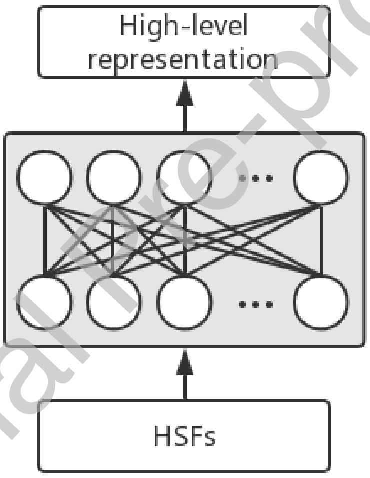


Figure 2: Framework of HSF-DNN

3.2.2. CNN

CNN was promising in learning features from raw images with local receptive field and parameter sharing [35]. Figure 3 shows the framework of our proposed MS-CNN. In this study, 2D-CNN was applied to learn time-frequency information from the feature context of the MS. The mel-spectrograms consisted of the time dimension and frequency dimension. In each convolution layer, the 2D time-frequency kernels convolved along the time dimension and frequency dimension. The pooling layer conducted pooling operation based on the outcomes of the prior convolution layer and reduced the feature dimensions. Specifically, we used three convolutional layers with sixteen 3×3 time-frequency kernels in each layer. The second and third convolution layers were followed by one 2×2 max-pooling layer separately. After the convolution operation, we flattened the CNN outputs of each mel-spectrogram and obtained a sequence of time-frequency representations.

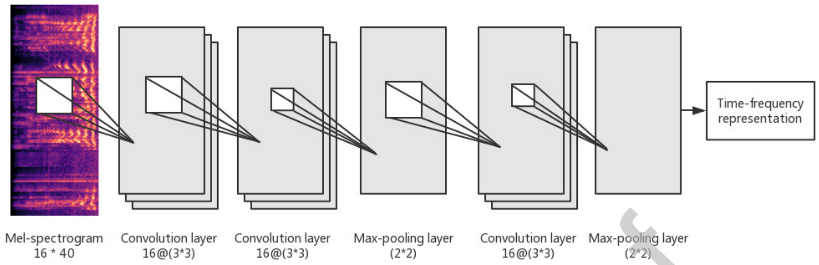

Time-frequency representation
Mel-spectrogram 16 * 40
Convolution layer 16@(3*3)
Convolution layer 16@(3*3)
Max-pooling layer (2*2)
Convolution layer 16@(3*3)
Max-pooling layer (2*2)

Figure 3: Framework of MS-CNN for each segment

3.2.3. RNN

RNN was designed to handle data sequence with its recursive connections between hidden layer activations at neighboring time steps [36]. In this study, RNN was applied to learn long-term dependency on sequential frame-level LLDs. Over continuous LLDs, RNN outputs represented long-term integrations of different time steps [14]. As shown in Figure 4, for a better representation of an utterance, we captured forward and backward information varieties by employing bidirectional long short term memory (BLSTM) neural network [36] with multiple frame-level LLDs as input. Within the BLSTM, the forward layer generates hidden states $(\overrightarrow{h}_1, ... \overrightarrow{h}_T)$ (T is the length of the input sequence) which represent forward varieties and the backward layer generates hidden states $(\overleftarrow{h}_1, ... \overleftarrow{h}_T)$ which represent backward feature varieties. The concatenated hidden states for each time step $h_i = [\overrightarrow{h}_i \oplus \overleftarrow{h}_i]$ allow us to preserve both forward and backward information. Specifically, we applied one BLSTM layer with 128 memory cells (64 memory cells in each direction). After processing through the temporal model, we obtained a sequence of long-term representations corresponding to different time steps.

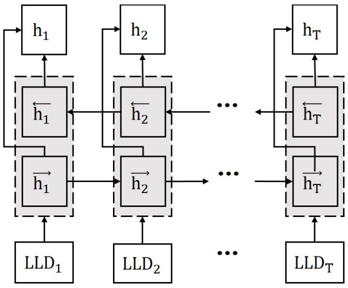


Figure 4: Framework of LLD-RNN

3.2.4. Weighted pooling

In this study, we applied the attention mechanism-based weighted pooling for both CNN and RNN, where CNN-outputs and RNN-outputs at various time steps separately were aggregated to generate an utterance-level representation. Not all time steps contribute equally to the emotion representation of an utterance. Based on this fact, we employed a weighted pooling strategy to enable neural networks to assign appropriate attention on different parts according to their contribution to utterance-level emotion recognition.

The attention mechanism was applied to obtain an aggregated vector by computing a weighted sum over time. Specifically, we first passed the RNN output h_i (or the CNN-output) to a 128-unit fully connected layer with a $tanh$ activation function to obtain the hidden representation u_i for h_i. Then we measured the importance of a time step by the similarity of u_i with a time step-level context vector U (inner product of u_i and U). A $softmax$ function was applied to the inner products to obtain the normalized weight a_i for each time step. Finally, we obtained the weighted sum H based on the computed weights as the aggregated utterance-level representation (defined in (9)). All the parameters mentioned above were optimized together during the network training process.

$$\begin{aligned} u_i &= tanh(Wh_i + b) \\ a_i &= \exp(u_i^\top U)/\sum_i \exp(u_i^\top U) \\ H &= \sum_i^T a_i h_i \end{aligned} \tag{9}$$

3.2.5. Classification task and Regression task

As shown in Figure 5, we built a multi-task learning framework on the MS-CNN, LLD-RNN and HSF-DNN components to improve their recognition performance by learning inter-relationships between different tasks.

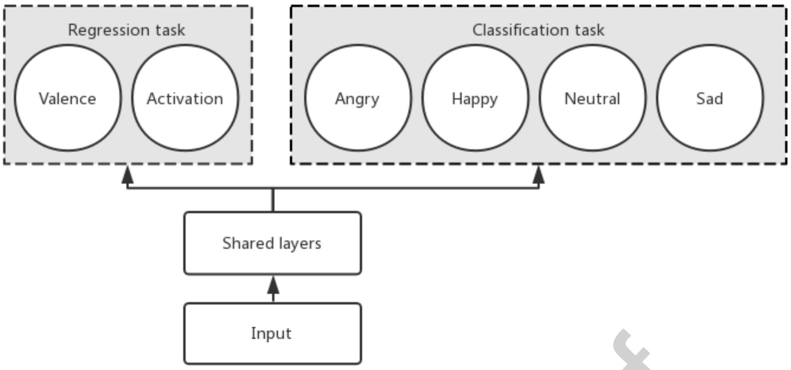


Figure 5: Multi-task learning framework for individual classifiers

Multi-task learning with both discrete categories and continuous attributes value helps improve the performance of the classification task [27]. To capture the inner relationship between two different descriptions of emotion (discrete categories and continuous attributes) and to improve the generalization of the neural network model, we designed a multi-task learning framework. We considered categorical emotion classification as the primary task and attribute regression as the secondary task.

In each model, the two tasks shared the same layers. Specifically, in MS-CNN, the tasks shared the CNN and weighted-pooling layers; in LLD-RNN, they shared the BLSTM and weighted-pooling layers; in HSF-DNN, they shared the DNN layers. We simultaneously fed the same aggregated utterance-level representation into a $softmax$ layer to obtain four-class probability distribution and two non-linear transformation units to predict two attribute values. With reference to the study [27], we assumed that the values of the attributes are in the range $[-1, 1]$, and we utilized the non-linear $tanh$ as the activation function for the two units because its output is in the range of $[-1, 1]$. During the training process, we used categorical cross-entropy as a loss function for the primary task and mean squared error for the secondary task. We defined the loss function for the whole system as

$$J = L_{ce} + \alpha \cdot (L_{mse-v} + L_{mse-a}) \tag{10}$$

where L_{ce} refers to the loss function for classification, L_{mse-v} and L_{mse-a} refer to the loss function in the prediction of two attributes - valence and activation, and α is a hyper-parameter used to control the contribution of the secondary task to the whole system. In this work, we set $\alpha = 2.0$ for MS-CNN and HSF-DNN and $\alpha = 0.8$ for LLD-RNN. In the corpus we used in this study, the continuous attributes are valued in $[1, 5]$. Prior to the process of multi-task learning, we mapped the continuous attribute values to the range of $[-1, 1]$ with the following formula:

$$v_{new} = v_{raw}/2 - 1.5 \tag{11}$$

where v_{raw} refers to the original value and v_{new} refers to the transformed value.

3.3. Classification Fusion

After we obtained the outputs from the three different classifiers - each used a different form of feature for certain speech utterances as input, we incorporated the three models to improve the ultimate recognition performance. Specifically, we developed confidence-based decision-level fusion using the sum of confidence scores referred to the study [30]. The confidence scores were separately generated from the $softmax$ layer in three individual classifiers. Let $\mathbb{I}$ denote the set of the three involved classifiers (HSF-DNN, MS-CNN, and LLD-RNN), $\mathbb{C}$ denote the set of four emotional states (angry, happy, neutral, and sad), and $s_i(c)$ denote the confidence score of the emotional state c in the classifier i. Let $p(c)$ denote the fusion confidence score of a given emotional state c, which can be calculated with the following formula:

$$p(c) = \sum_{i} s_i(c) \tag{12}$$

The emotional state with the highest fusion confidence score would then be chosen as the ultimate predicted class r:

$$r = \arg\max p(c) \tag{13}$$

4. Experiments and Results

To evaluate the performance of our proposed approach in the speaker-independent environment, we performed three experiments on the IEMOCAP dataset [32] using leave-one-session-out strategy. The weighted accuracy (WA) and unweighted accuracy (UA) were used as the metric for evaluation. WA corresponded to the overall accuracy which was calculated by dividing the number of true positives by the total number of samples. For each class, the recall value was calculated by dividing the number of true positives by the sum of the number of true positives and the number of false negatives. UA was given by the average of the recall values of four classes. Furthermore, we performed paired t-test to compare the accuracies between different methods. The results were considered significant when the p values were below 0.05. Specifically, we obtained the performance for each actor and performed the t-test based on these actor-based samples.

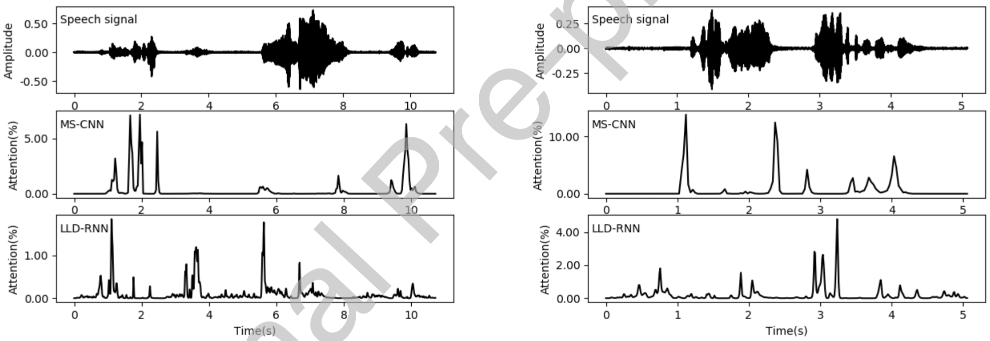
4.1. Experiment I: Different pooling methods for MS-CNN and LLD-RNN

In the Experiment I, we investigated the performance of the mean pooling and weighted pooling methods with reference to the study [14]. These two pooling methods were used for aggregating the sequential outputs of CNN and RNN into an utterance-level representation. Specifically, for the mean pooling method, we obtained the mean value of all sequential hidden outputs. For the weighted pooling method, we obtained the weighted sum over all sequential hidden outputs, in which the weights were obtained with the attention mechanism described above.

Table 1: Results of different pooling methods for MS-CNN and LLD-RNN

Classifiers	Pooling methods	WA (%)	UA (%)
MS-CNN	Mean pooling	43.0	38.1
	Weighted pooling	48.0	48.3
LLD-RNN	Mean pooling	43.9	41.9
	Weighted pooling	52.6	54.1

As shown in Table 1, the mean pooling method obtained the WA of 43.0% and the UA of 38.1% for MS-CNN, and the WA of 43.9% and the UA of 41.9% for LLD-RNN. The weighted pooling method obtained the WA of 48.0% and the UA of 48.3% for MS-CNN, and the WA of 52.6% and the UA of 54.1% for LLD-RNN. For both MS-CNN and LLD-RNN classifiers, the WA and UA of the weighted pooling method were significantly higher than those of the mean pooling method (paired t-test, p values < 0.05).



(a) (b)

Figure 6: Obtained weights for two speech samples. Top: raw speech signal waveform; middle: obtained weights in MS-CNN; bottom: obtained weights in LLD-RNN

Figure 6 shows the weight assignment on two samples in MS-CNN and LLD-RNN, in which the original speech signal waveforms of the two samples were passed and the attention mechanism was applied. The weight distribution has similar patterns in both systems: the neural network assigned a small proportion of weight on silent segments while a large proportion on segments with relatively greater amplitude. However, the larger amplitude would not guarantee that a higher proportion of weight was assigned to the corresponding segment. In fact, weights were assigned based on the extent to which certain speech segments could contribute to the utterance-level emotional state.

4.2. Experiment II: Different multi-task learning frameworks

In the Experiment II, we investigated the effect of applying different multi-task learning frameworks on emotion category classification accuracies. To explore the contribution of emotional attributes to classification in multi-task learning, we evaluated the performance of the frameworks not only when jointly predicting both attributes valence (V) and activation (A), but also when predicting only one of them. Specifically, both the MS-CNN and LLD-RNN classifiers were applied with the weighted pooling method.

As shown in Table 2, with our proposed frameworks (V&A), HSF-DNN achieved the WA of 54.4% and UA of 55.6%, MS-CNN achieved WA of 49.5% and UA of 50.1%, and LLD-RNN achieved WA of 54.2% and UA of 55.6%. Compared with single-task learning-based methods, the multi-task learning-based methods (V&A) achieved better performance in terms of both WA and UA. For LLD-RNN, the improvements in both WA and UA were significant at the $p = 0.05$ level. However, for HSF-DNN and MS-CNN, none of improvements in WA or UA were statistically significant. Furthermore, the multi-task learning-based methods jointly predicting valence and activation attributes also obtained the higher WA and UA than those of the methods predicting one attribute individually.

Table 2: Results of different multi-task frameworks for individual classifiers

Classifiers	Frameworks	WA (%)	UA (%)
HSF-DNN	Single-task	52.5	53.6
	Multi-task (V)	53.0	53.7
	Multi-task (A)	53.8	54.9
	Multi-task (V&A)	54.4	55.6
MS-CNN	Single-task	48.0	48.3
	Multi-task (V)	48.2	49.1
	Multi-task (A)	48.8	49.4
	Multi-task (V&A)	49.5	50.1
LLD-RNN	Single-task	52.6	54.1
	Multi-task (V)	53.1	54.3
	Multi-task (A)	53.2	54.4
	Multi-task (V&A)	54.2	55.6

4.3. Experiment III: Different methods

Table 3: Results of different methods

Methods	WA (%)	UA (%)	Angry (%)	Happy (%)	Neutral (%)	Sad (%)
Baseline	54.0	55.3	62.7	48.6	49.4	60.4
HSF-DNN	54.4	55.6	61.5	50.6	48.9	61.5
MS-CNN	49.5	50.1	50.1	42.3	52.4	55.4
LLD-RNN	54.2	55.6	59.6	50.6	47.5	64.7
Fusion	57.1	58.3	63.0	52.7	52.5	64.9

In the Experiment III, we compared our fusion method with a baseline method and three individual classifiers to validate our approach and system. In this study, a linear SVM (cost parameter $c = 0.002$) which used utterance-level HSFs representation as input was adopted as a baseline method [10]. Table 3 shows the WA, UA, and averaged accuracies of each of the four categories (angry, happy, neutral, and sad) using the baseline method, the fusion method and the individual classifiers. Our fusion method significantly increased WA by 3.1% and UA by 3.0%, compared with the baseline method. Furthermore, the fusion method achieved WA of 57.1%, UA of 58.3% and averaged accuracies in angry, happy, neutral and sad states of 63.0%, 52.7%, 52.5% and 64.9%, respectively. Note that the individual classifiers demonstrated varied capabilities in recognizing different emotional states. Specifically, the HSF-DNN and LLD-RNN classifiers were powerful in recognizing the angry, happy and sad states, whereas the MS-CNN classifier was strong in recognizing the neutral state.

Furthermore, we performed paired t-test to compare the WA, UA, and averaged accuracies in four categories between the fusion method and each of the four methods including baseline, HSF-DNN, MS-CNN, and LLD-RNN. As shown in Table 4, the statistical test results indicate that significant differences existed between the fusion method and each of the individual classifiers in both WA and UA.

Table 4: Results of paired t-test (p values) between fusion method and different non-fusion methods

Methods	WA	UA	Angry	Happy	Neutral	Sad
Baseline	<0.05	<0.05	0.77	0.17	0.32	0.23
HSF-DNN	<0.05	<0.05	0.45	0.24	0.25	0.06
MS-CNN	<0.05	<0.05	<0.05	<0.05	0.66	<0.05
LLD-RNN	<0.05	<0.05	0.18	0.15	<0.05	0.71

5. Discussion

In speech emotion recognition, an important task is to obtain effective features that represent the emotional state of an utterance. When designing an algorithm to achieve emotion category classification, we need to make efficient use of a limited amount of data and solve the problems of silent and emotionless parts. In this paper, we developed a speech emotion recognition framework that integrates three distinctive classifiers (HSF-DNN, MS-CNN, and LLD-RNN), in which high-level statistical function outputs, mel-spectrograms, and LLDs were passed as their respective inputs. First, we applied the attention mechanism based weighted pooling method to aggregate the sequence of hidden representations in MS-CNN and LLD-RNN. Then, we implemented a multi-task learning method for HSF-DNN, MS-CNN, and LLD-RNN, which was trained for jointly category classification and attribute regression. Finally, we designed a decision-level fusion strategy using the sum of confidence scores of three individual classifiers. Experimental results on the IEMOCAP corpus with a leave-one-session-out strategy show that our framework achieved significant improvements in WA and UA compared with the baseline method and each individual classifier.

Experimental results (as shown in Table 1) demonstrate the high efficiency of the proposed weighted pooling method in utterance-level aggregation. Emotion intensity is uneven across a given speech utterance which usually includes some silent or weak emotional parts. Thus, any possible impact caused by this phenomenon should be considered in aggregating multiple short-term features. Mirsamadi *et al.* proposed that while pooling the RNN outputs, desired representation would be distorted by silent and non-emotional speech parts with the mean pooling method [14]. With the weighted pooling method, the attention mechanism could automatically assign higher weights on emotionally obvious parts and lower weights on weak emotions or silent parts according to the class information gained from supervised learning. We further applied the weighted pooling method on the CNN outputs which represent short-term and frequency features and obtained an improved utterance-level representation. However, our weighted summation approach in the attention mechanism was linear, whereas a non-linear combination of the attention weights would be more appealing.

Multi-task learning is able to improve the generalization of neural networks through shared layers. Rui Xia *et al.* developed a multi-task learning method that utilized the activation and valence information of emotion to improve the recognition capability of a deep belief network [27]. In this study, we implemented a multi-task learning strategy in our HSF-DNN, MS-CNN and LLD-RNN models, aiming to further promote their respective performance. As shown in Table 2, the single-task model LLD-RNN with a weighted pooling method [14] achieved WA of 52.6% and UA of 54.1%, whereas our multi-task model LLD-RNN (V&A) obtained 1.6% and 1.5% improvements in WA and UA, respectively ($p < 0.05$). The experimental result shows that when simultaneously processing the classification task of discrete categories and the regression task of continuous attributes, the shared layer acquired interdependence between different tasks, assisting the model in promoting its distinguishing ability towards emotion categories. However, with the multi-task learning strategy, the improvements in WA and UA for HSF-DNN and MS-CNN were not significant at $p = 0.05$ level. The reason could be depicted in two-fold as follow. First, the valence and activation attributes were considered equally important and assigned the same weights in the loss function. Assigning different weights for different attributes might be a better strategy, which could take into account the difference between different attributes. Second, generalized representation learning could be limited by the lack of sufficient training data.

In this study, we developed a confidence-based fusion strategy for the HSF-DNN, MS-CNN, and LLD-RNN classifiers at the decision level. Each sub-classifier have their own capabilities and limitations in recognizing different emotional states. For instance, DNN classifier obtained high-level representations from static statistics, CNN classifier captured time-frequency variation, and RNN classifier learned long-time context. Information could be integrated by combining multiple estimators which were based on different acoustic cues or different processing principles [37] to optimize the detection performance. This was demonstrated by our experimental results in Tables 3 and 4. Our fusion method has significantly increased WA by 3.1% and UA by 3.0% compared with the baseline SVM method [10] ($p < 0.05$).

However, there were no significant differences between the fusion method and each sub-classifier for averaged accuracies in each specific emotional state. The reason could be depicted in two-fold as follow. First, the lack of sufficient training data might limit the parameters learning of the neural networks in our frameworks. Second, there was a limitation of our confidence-based fusion strategy, which simply added up the confidences and could not efficiently utilize the discrepancies of different models in distinguishing different emotional states. A more effective fusion method could be conducted with sufficient training corpus. Future studies should include the following two folds. First, a non-linear weighted fusion method in the decision level could be implemented by training a non-linear neural network to fit the fusion weights. Second, a feature-level fusion framework could be designed based on jointly optimizing sub-classifiers.

6. Conclusion

In this study, we presented a confidence-based fusion method consisting of three multi-task learning-based sub-classifiers: DNN with utterance-level HSFs, CNN with multiple segment-level MS, and RNN with multiple frame-level LLDs, which was used for categorical recognition of discrete emotions (angry, happy, neutral and sad). We applied a weighted pooling method based on the attention mechanism to aggregate multiple outputs of CNN or RNN. The attention mechanism endowed the neural network with the capability to focus on emotionally salient parts. The multi-task learning-based methods for simultaneously emotion category classification and attribute regression could learn generalized representations which improved the classification accuracies. Our proposed fusion method could integrate different recognition powers of the three sub-classifiers, and achieved WA of 57.1% and UA of 58.3%, which were significantly higher than those of each individual classifier. The effectiveness of the proposed approach based on classifier fusion is thus validated.

Author contributions

Jiahui Pan, Zengwei Yao and Zihao Wang conceived of and designed the experiments. Zengwei Yao, Zihao Wang, Weihuang Liu, and Yaqian Liu performed the experiments and analyzed the data. Jiahui Pan, Zengwei Yao and Zihao Wang wrote and reviewed the manuscript. Jiahui Pan supervised the project.

Funding

This study was supported by the National Natural Science Foundation of China under Grants 61876067 and 61503143; the Guangdong Natural Science Foundation under Grant 2019A1515011375; the Pearl River S&T Nova Program of Guangzhou under Grant 201710010038; and the Guangdong Colleges and Universities Special Project Foundation in Key Areas of Artificial Intelligence under Grant 2019KZDZX1033. .

Competing interests

None of the authors have potential conflicts of interest to be disclosed.

References

[1] R. W. Picard, R. Picard, Affective computing. vol. 252, MIT press Cambridge.EEG-detected olfactory imagery to reveal covert consciousness in minimally conscious state. Brain injury 29 (13-14) (1997) 1729–1735.

[2] C. M. Lee, S. S. Narayanan, et al., Toward detecting emotions in spoken dialogs, IEEE transactions on speech and audio processing 13 (2) (2005) 293–303.

[3] B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, IEEE, 2004, pp. I–577.

[4] Q. Luo, H. Tan, Facial and speech recognition emotion in distance education system, in: The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007), IEEE, 2007, pp. 483–486.

[5] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk, IEEE transactions on Biomedical Engineering 47 (7) (2000) 829–837.

[6] N. Ding, V. Sethu, J. Epps, E. Ambikairajah, Speaker variability in emotion recognition-an adaptation based approach, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2012, pp. 5101–5104.

[7] T. Vogt, E. André, Improving automatic emotion recognition from speech via gender differentiaion., in: LREC, 2006, pp. 1123–1126.

[8] A. Mill, J. Allik, A. Realo, R. Valk, Age-related differences in emotion recognition ability: A cross-sectional study., Emotion 9 (5) (2009) 619.

[9] B. Schuller, S. Steidl, A. Batliner, The interspeech 2009 emotion challenge, in: Tenth Annual Conference of the International Speech Communication Association, 2009.

[10] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, Acoustic emotion recognition: A benchmark comparison of performances, in: 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, IEEE, 2009, pp. 552–557.

[11] S. G. Koolagudi, K. S. Rao, Emotion recognition from speech: a review, International journal of speech technology 15 (2) (2012) 99–117.

[12] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, B. Schuller, Deep neural networks for acoustic emotion recognition: raising the benchmarks, in: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2011, pp. 5688–5691.

[13] J. Lee, I. Tashev, High-level feature representation using recurrent neural network for speech emotion recognition, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[14] S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 2227–2231.

[15] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, B. Schuller, Speech emotion classification using attention-based lstm, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (11) (2019) 1675–1685.

[16] B. T. Atmaja, M. Akagi, Speech emotion recognition based on speech segment using lstm with attention model, in: 2019 IEEE International Conference on Signals and Systems (ICSigSys), IEEE, 2019, pp. 40–44.

[17] L. Tarantino, P. N. Garner, A. Lazaridis, Self-attention for speech emotion recognition, Proc. Interspeech 2019 (2019) 2578–2582.

[18] R. Li, Z. Wu, J. Jia, Y. Bu, S. Zhao, H. Meng, Towards discriminative representation learning for speech emotion recognition, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 5060–5066.

[19] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Transactions on Multimedia 20 (6) (2017) 1576–1590.

[20] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, J. Vepa, Speech emotion recognition using spectrogram & phoneme embedding., in: Interspeech, 2018, pp. 3688–3692.

[21] H. Meng, T. Yan, F. Yuan, H. Wei, Speech emotion recognition from 3d log-mel spectrograms with deep learning network, IEEE Access 7 (2019) 125868–125881.

[22] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, H. Meng, Learning discriminative features from spectrograms using center loss for speech emotion recognition, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 7405–7409.

[23] X. Li, Y. Y. Wang, G. Tr, Multi-task learning for spoken language understanding with shared slots, in: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011, 2011.

[24] M. L. Seltzer, J. Droppo, Multi-task learning in deep neural networks for improved phoneme recognition, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 6965–6969.

[25] Z. Wu, C. Valentini-Botinhao, O. Watts, S. King, Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 4460–4464.

[26] S. Kim, T. Hori, S. Watanabe, Joint ctc-attention based end-to-end speech recognition using multi-task learning, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 4835–4839.

[27] R. Xia, Y. Liu, A multi-task learning framework for emotion recognition using 2d continuous space, IEEE Transactions on Affective Computing 8 (1) (2017) 3–14.

[28] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, C. Li, Exploring spatio-temporal representations by integrating attention-based bidirectional-lstm-rnns and fcns for speech emotion recognition, Proc. Interspeech 2018 (2018) 272–276.

[29] C. Zheng, C. Wang, W. Sun, N. Jia, Research on speech emotional feature extraction based on multidimensional feature fusion, in: International Conference on Advanced Data Mining and Applications, Springer, 2019, pp. 535–547.

[30] B.-H. Su, S.-L. Yeh, M.-Y. Ko, H.-Y. Chen, S.-C. Zhong, J.-L. Li, C.-C. Lee, Self-assessed affect recognition using fusion of attentional blstm and static acoustic features, 2018, pp. 536–540. doi:10.21437/Interspeech.2018-2261.

[31] J. Sebastian, P. Pierucci, Fusion techniques for utterance-level emotion recognition combining speech and transcripts, in: Proc. Interspeech, 2019, pp. 51–55.

[32] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (4) (2008) 335.

[33] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, Oriol Nieto, librosa: Audio and Music Signal Analysis in Python, in: Kathryn Huff, James Bergstra (Eds.), Proceedings of the 14th Python in Science Conference, 2015, pp. 18 – 24. doi:10.25080/Majora-7b98e3ed-003.

[34] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, ACM, 2010, pp. 1459–1462.

[35] T. N. Sainath, O. Vinyals, A. Senior, H. Sak, Convolutional, long short-term memory, fully connected deep neural networks, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 4580–4584.

[36] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (11) (1997) 2673–2681.

[37] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, A. Klapuri, Automatic music transcription: challenges and future directions, Journal of Intelligent Information Systems 41 (3) (2013) 407–434.

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Author statement

Speech Emotion Recognition Using Fusion of Three Multi-Task Learning-based Classifiers: HSF-DNN, MS-CNN and LLD-RNN:

Jiahui Pan, Zengwei Yao and Zihao Wang conceived of and designed the experiments. Zengwei Yao, Zihao Wang, Weihuang Liu, and Yaqian Liu performed the experiments and analyzed the data. Jiahui Pan, Zengwei Yao and Zihao Wang wrote and reviewed the manuscript. Jiahui Pan supervised the project.