# Towards Discriminative Representation Learning for Speech Emotion Recognition

**Runnan Li**[1,2], **Zhiyong Wu**[1,2], **Jia Jia**[1,2,*], **Yaohua Bu**[2], **Sheng Zhao**[3] and **Helen Meng**[4]
[1]Graduate School at Shenzhen, Tsinghua University
[2]Dept. of Computer Science and Technology, Tsinghua University
[3]Search Technology Center Asia (STCA), Microsoft
[4]Dept. of Systems Engineering and Engineering Management, The Chinese University of Hong Kong
lirn15@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, jjia@tsinghua.edu.cn

## Abstract

In intelligent speech interaction, automatic speech emotion recognition (SER) plays an important role in understanding user intention. While sentimental speech has different speaker characteristics but similar acoustic attributes, one vital challenge in SER is how to learn robust and discriminative representations for emotion inferring. In this paper, inspired by human emotion perception, we propose a novel representation learning component (RLC) for SER system, which is constructed with Multi-head Self-attention and Global Context-aware Attention Long Short-Term Memory Recurrent Neutral Network (GCA-LSTM). With the ability of Multi-head Self-attention mechanism in modeling the element-wise correlative dependencies, RLC can exploit the common patterns of sentimental speech features to enhance emotion-salient information importing in representation learning. By employing GCA-LSTM, RLC can selectively focus on emotion-salient factors with the consideration of entire utterance context, and gradually produce discriminative representation for emotion inferring. Experiments on public emotional benchmark database IEMOCAP and a tremendous realistic interaction database demonstrate the outperformance of the proposed SER framework, with 6.6% to 26.7% relative improvement on unweighted accuracy compared to state-of-the-art techniques.

## 1 Introduction

Human speech is highly expressive, people in communication are willing to use emotions, intonations and styles to convey the underlying intent of messages. For intelligent speech interaction systems, recognizing such paralinguistic information, especially the emotion, can enhance the understanding of user intention and improve user experience. Speech emotion recognition (SER), aiming to detect emotions from speech, is thus becoming an increasing interest in the human-computer interaction research field.

In spoken human-computer interaction, the input sentimental speech may have different speaker characteristics while sharing similar acoustic attributes. Therefore, one vital challenge in SER is how to learn robust and discriminative representations from speech [Strapparava and Mihalcea, 2008]. Traditionally, a long statistical feature vector produced by a series of statistical aggregation functions (such as mean, max, variance, etc) on frame-level Low-Level Descriptors (LLDs) extracted from speech is used as the utterance-level representation [Schuller *et al.*, 2017]. The statistical feature vector can roughly describe the temporal variations and contours of LLDs, which are assumed to be highly related to speech emotion. The learned inference model, such as Deep Neural Network (DNN) [Stuhlsatz *et al.*, 2011], is then applied to infer emotion from the produced statistical vector.

With the strong ability in informative feature extraction and temporal aggregation, automatic feature learning algorithm on frame-level LLDs is also proposed, such as Convolutional Neural Networks (CNNs) [Poria *et al.*, 2016], Recurrent Neural Networks (RNNs) [Lee and Tashev, 2015] and its memory enhanced Long Short-Term Memory (LSTM) variants [Poria *et al.*, 2017], and has achieved significant improvement.

However, limitations still exist in these state-of-the-art approaches: the lack of attention ability and underutilization of contextual information. In human speech emotion perception, as reported in [Schirmer and Adolphs, 2017], incorporating context from entire speech is routine, efficient and somehow automatic, and attention-capturing vocalizations can produce greater activation to cortex than neutral vocalizations. This mechanism indicates the importance of introducing attention ability in developing SER systems. The challenge lies in two aspects while human perception is mostly based on word-level vocalizations, one is how to generate informative features with suitable resolution for emotion perception, and the other one is how to generate representation with selective attention under the understanding of the entire utterance.

In this paper, to address above challenges, we propose the combination use of Multi-head Self-attention [Vaswani *et al.*, 2017] and Global Context-aware Attention Long Short-Term Memory recurrent neutral network (GCA-LSTM) [Liu *et al.*, 2017] to construct a novel Representation Learning Component (RLC), aiming to learn robust and discriminative representations from speech for emotion inferring.

Developed from self-attention mechanism, Multi-head

*Corresponding author.

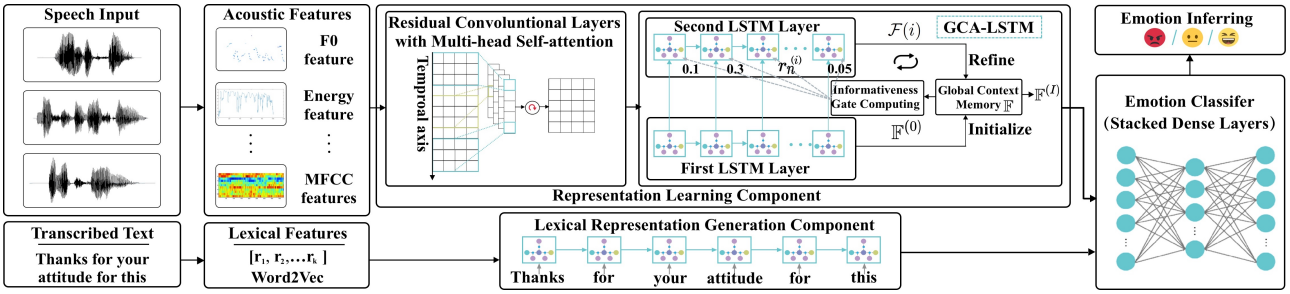


Figure 1: Overview of the proposed speech emotion recognition framework. Red recurrences represent the Multi-head Self-attention blocks.

Self-attention inherits the strong ability in modeling the element-wise correlative dependencies of input sequence, and can further jointly attend information from different representation subspaces to improve the modeling performance. By employing Multi-head Self-attention, RLC can model the common patterns of sentimental speech features and the relative dependencies between vocalizations, to focus on the attention-capturing parts of speech and enhance emotion-relevant information importing in feature learning.

Developed from LSTM network, GCA-LSTM inherits the ability to generate utterance-level representation from speech with arbitrary length while considering the context information. Furthermore, the additional global context memory is applied to alleviate the weakness in utilizing the utterance-level global contextual information of LSTM network. In GCA-LSTM, the global context information is fed to the network to compute informativeness scores of each step's input, and accordingly adjust the attention weights for them. For the input with higher informativeness score to the sentimental representation, the network learns more from it; on the contrary, the network blocks it. By employing GCA-LSTM, RLC can selectively focus on emotion-salient factors with the consideration of entire utterance context, and gradually produce discriminative representation for emotion inferring.

As depicted in Fig.1, in inference, LLDs are extracted from speech signal as the input acoustic features. Residual convolutional layers with Multi-head Self-attention are employed to extract emotion-salient features, and merge semantically similar hidden outputs to produce suprasegmental features with lower temporal resolution. The GCA-LSTM block is employed to produce utterance-level representation from suprasegmental features, which consists of two LSTM layers and one global context memory embodying the global contextual representation. For each input, the first LSTM layer encodes the input sequence and then initializes the global context memory. Then the global context is fed to calculate the relevance scores of inputs, and help the model to selectively focus on the informative factors in the second LSTM layer to produce the attention representation. The learned attention representation is further fed to refine the global context memory. Such iterative operation is repeated for several times to progressively produce robust and discriminative acoustic representation of speech, embedded in the global context memory. The proposed framework can also benefit from multimodal inputs. For auxiliary transcribed text, one LSTM layer is employed to encode input sequence into utterance-level representation. The acoustic representation and textual representation are then fed to classifier for inferring emotion.

The proposed framework is evaluated on two different databases comparing to state-of-the-art techniques: Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [Busso *et al.*, 2008] and a tremendous real scene interaction database (RID). The experimental results demonstrate the superior performance of the proposed framework, achieving 14.3% relative improvement (from 60.7% to 69.4% on unweighted accuracy) on IEMOCAP with acoustic input only, and 6.6% relative improvement (from 74.3% to 79.2%) using both acoustic and lexical features, and 26.7% relative improvement (from 71.2% to 90.2%) on RID with acoustic input only. The main contributions are summarized as follows.

- **A novel representation learning component.** Constructed with Multi-head Self-attention and GCA-LSTM, the proposed RLC inherits the feature learning ability of conventional CNNs, meanwhile promotes its attention ability in utilizing emotion-salient factors of speech to produce discriminative presentation.
- **Superior-performance in speech emotion recognition.** The experimental results demonstrate the proposed RLC can effectively produce robust and discriminative representation from speech, and significantly outperform state-of-the-art approaches in experiments.

## 2 Related Work

**Speech Emotion Recognition.** State-of-the-art SER techniques are mainly developed with neutral networks. [Poria *et al.*, 2016] proposed a CNN based feature learning approach to extract emotion-related features from frame-level LLDs. [Lee and Tashev, 2015; Poria *et al.*, 2017] proposed the use of RNN and its LSTM variants to model contextual information. [Trigeorgis *et al.*, 2016] proposed an end-to-end learning approach to reduce hassle and cost in developing SER model.

**Attention Mechanism.** The successful employment of attention mechanism, as in automatic speech recognition (ASR) [Chorowski *et al.*, 2015] and machine translation (MT) [Bahdanau *et al.*, 2014], has empirically demonstrated the effectiveness of attention mechanism in selectively focusing on specific information. In developing SER systems, [Mirsamadi *et al.*, 2017] proposed the use of local attention with RNN, [Lian *et al.*, 2018] proposed the use of transformer structure in representation generation, both can significantly improve the emotion inferring performance.

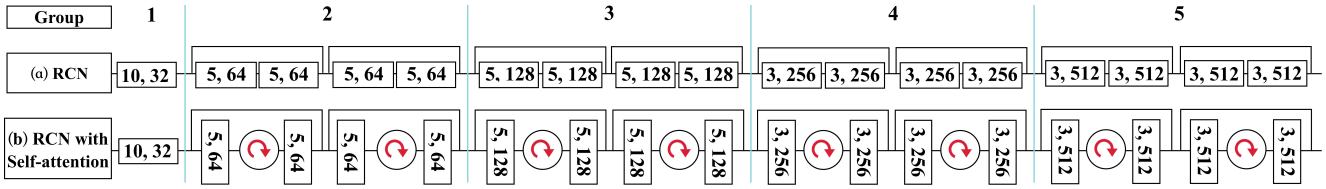


Figure 2: Group of layers are separated by the bold lines, and striding downsampling with a factor of 2 is processed after the blue lines. Red recurrences represent the Multi-head Self-attention blocks.

# 3 Methodology

## 3.1 Convolutional Feature Learning

Residual convolutional network (RCN) is proposed for informative feature learning with 1-D temporal convolutional layers. Developed from [He *et al.*, 2016], the stacked layers keep the original temporal structure of input signals through residual structure in the progressively feature extracting process, thus to retain the important temporal information of speech.

The structure of RCN is depicted in Fig.2(a). For the $l^{th}$ group $\mathcal{G}_i^l$, the first convolutional layer employs 2-stride length, increasing receptive field and reducing the temporal resolution; for the rest $i^{th}$ layer $\mathcal{G}_i^l$, the output is calculated as

$$(\mathcal{G}_i^l * f_i^l)(\mathbf{p}) = \sum_{\mathbf{a}+\mathbf{b}=\mathbf{p}} \mathcal{G}_i^l(\mathbf{a})\, f_i^l(\mathbf{b}) \tag{1}$$

where $f_i^l$ is the filter of $\mathcal{G}^l$, and the domain of $\mathbf{p}$ is the feature map in $\mathcal{G}_i^l$. Specially, to optimize the training efficiency, batch normalization [Ioffe and Szegedy, 2015] is utilized in the convolutional layers. Furthermore, human vocal perception is based on consonant-vowel syllables with duration from 150 msec to 200 msec [Steinschneider *et al.*, 2013]. In this work, for input acoustic features with 5 msec shift length, we employ $l = 5$ groups convolutional layers for feature learning, producing learned suprasegmental features with granularity of 160 msec, which is close to the granularity in human speech perception.

## 3.2 Multi-head Self-attention

Attention-capturing vocalizations in speech can contribute more to human emotion perception, and significantly affect the perception to other vocalizations. In this work, Multi-head Self-attention mechanism [Vaswani *et al.*, 2017] is proposed to model the relative dependencies between elements, meanwhile focusing on attention-capturing factors in speech. With the enhanced attention ability, the network can further import emotion-salient information from input, providing discriminative feature learning result for the representation learning component.

Developed from self-attention, Multi-head Self-attention compute a weighted output with input and key-value pairs, where the weights assigned are computed by a compatibility function using the input and corresponding keys.

For the given input sequence $H$, self-attention computes $d_k$, $d_k$, $d_v$ dimensional queries, keys, values $Q, K, V$ with linear projections. The attention output is then calculated as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

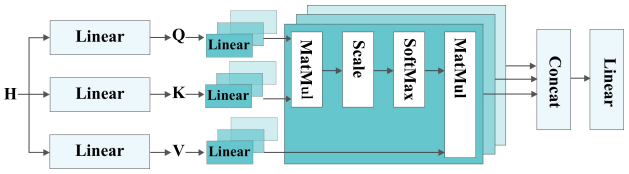


Figure 3: The parallel structure of Multi-head Self-attention.

To jointly attend to information from different representation subspaces at different positions, multi-head attention is proposed. Compared to single attention, multi-head attention performs $r$ times different linear projections from queries, keys, values $Q_i, K_i, V_i$, where $i = 1, ..., r$, and then performs the attention function in parallel, yielding $(d_v/r)$-dimensional output values. These values are concatenated and projected again to produce the final values, resulting in higher effectiveness in producing attention representation [Vaswani *et al.*, 2017]. As depicted in Fig.3, the Multi-head Self-attention is calculated as

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, ..., head_r)W^O \\ where\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{3}$$

where $W_i^Q$ $W_i^K$, $W_i^V$ are the weight matrices in parallel attentions, with the dimension of $d_k/r$, $d_k/r$ , $d_v/r$ respectively. $W^O$ is the weight matrix in linear output function. In this work, as depicted in Fig.2(b), Multi-head Self-attention is applied in residual blocks, each with $r = 4$ parallel heads.

## 3.3 Global Context-aware Attention LSTM

In communication, participants perceived each other's emotion through the understanding of the whole sentence. Meanwhile, vocalizations in speech contribute different to the final emotion perception. To model the contribution differences, GCA-LSTM [Liu *et al.*, 2017] is employed in this work, which exploits the global context information to measure the relevance of vocalizations, and selectively imports information for utterance-level representation generation.

As depicted in Fig.4, GCA-LSTM contains two LSTM layers and one global context memory. The representation $\mathbb{F}$ is maintained in the global context memory, and gradually refined. In process, the first LSTM layer encodes the learned suprasegmental features, and initializes the global context memory $\mathbb{F}^{(0)}$. The second LSTM layer performs attention over the hiddens to compute the attention representation $\mathcal{F}$, which is used to refine $\mathbb{F}$. The refining of $\mathbb{F}$ is iteratively processed $I$ times to progressively learn discriminative rep-

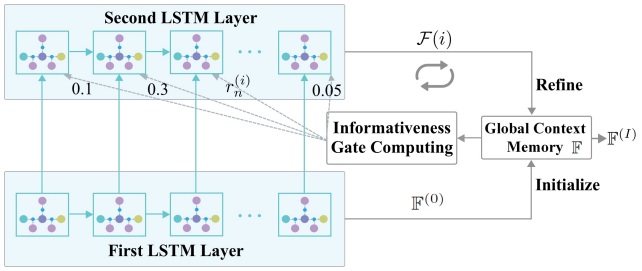


Figure 4: Global Context-aware Attention LSTM. The output $\mathbb{F}^{(I)}$ is used as the generated representation for emotion inferring.

resentation. After $I$-th refining iteration, the utterance-level representation $\mathbb{F}^{(I)}$ is used as the learned representation.

**Global Context Memory Initialization**

For the suprasegmental feature sequence $M = (m_1, ..., m_n, ..., m_N)$, the first LSTM layer computes the hidden vector sequence $h = (h_1, ..., h_n, ..., h_N)$ from $n = 1$ to $N$ as following equations:

$$f_n = \sigma(W_f m_n + U_f h_{n-1} + b_f) \quad (4)$$
$$i_n = \sigma(W_i m_n + U_i h_{n-1} + b_i) \quad (5)$$
$$o_n = \sigma(W_o m_n + U_o h_{n-1} + b_o) \quad (6)$$
$$c_n = f_n \circ c_{n-1} + i_n \circ tanh(W_c m_n + U_c h_{n-1} + b_c) \quad (7)$$
$$h_n = o_n \circ tanh(c_n) \quad (8)$$

where $\sigma$ is the Sigmoid activation function, $f$, $i$, $o$ and $c$ are the *input gate*, *forget gate*, *output gate* and *memory cell* activation vectors respectively, $W$, $U$ and $b$ items are the weight matrices and bias vectors of each gate. The last hidden output $h_N$ is used as the initialization value $\mathbb{F}^{(0)}$ of global context memory $\mathbb{F}$. To facilitate description, the hidden output $h_n$ of the first LSTM layer is written as $\mathbb{h}_n$ in the following.

**Global Context Memory Refining**

The informativeness degree of the input is firstly assessed at each time step. In the $i$-th iteration, the network learns an informativeness gate $r_n^{(i)}$ using both $\mathbb{h} = (\mathbb{h}_1, ..., \mathbb{h}_n, ..., \mathbb{h}_N)$ and $\mathbb{F}^{(i-1)}$ from the previous iteration

$$e_n^{(i)} = W_{e_1}(tanh(W_{e_2}\begin{pmatrix}\mathbb{h}_n \\ \mathbb{F}^{(i-1)}\end{pmatrix})) \quad (9)$$

$$r_n^{(i)} = \frac{exp(e_n^{(i)})}{\sum_{n=1}^{N} exp(e_n^{(i)})} \quad (10)$$

The learned informativeness gate $r_n^{(i)}$ is normalized by min-max normalization to range $[0, 1]$, representing the relevance degree of input $\mathbb{h}$ to the global context $\mathbb{F}^{(i-1)}$. The calculation of memory cell $c_n$ is modified to:

$$\begin{aligned} c_n =& f_n \circ c_{n-1} \circ (1 - r_n^{(i)}) \\ &+ i_n \circ tanh(W_c \mathbb{h}_n + U_c h_{n-1} + b_c) \circ r_n^{(i)} \end{aligned} \quad (11)$$

With the use of informativeness gate $r_n^{(i)}$, the cell state $c_n$ is updated according to the relevance of the input $\mathbb{h}_n$ to the global context, i.e., the cell state will import more information from the input with higher relevance score. With this mechanism, the utterance-level representation learning will selectively focus on the informative parts of speech with consideration of global contextual information.

The hidden output $h_n$ of the second LSTM layer at each step is calculated following Eq.4 to Eq.8, with the memory cell state updating scheme replaced from Eq.7 to Eq.11, using $\mathbb{h} = (\mathbb{h}_1, ..., \mathbb{h}_n, ..., \mathbb{h}_N)$ from the first LSTM layer as input. For the $i$-th iteration, the last hidden output $h_N^{(i)}$ is used as the attention representation $\mathcal{F}(i)$ to refine the representation $\mathbb{F}$.

$$\mathbb{F}^{(i)} = ReLU(W_F\begin{pmatrix}\mathcal{F}(i) \\ \mathbb{F}^{(i-1)}\end{pmatrix}) \quad (12)$$

After $I$-th refining iteration, $\mathbb{F}^{(I)}$ is fed to the following classifier to infer speech emotion. With the iteratively refining, the relevance score is progressively enhanced and benefit the representation learning component in emotion-salient information importing, providing more robust and discriminative representation for emotion inferring and thus to improve the overall performance of the proposed SER framework.

### 3.4 Lexical Representation Learning

The proposed framework can also benefit from employing multi-modal inputs. For the transcribed text with $K$ words, Word2Vec [Mikolov *et al.*, 2013] is utilized to generate the lexical features $R = (r_1, ..., r_k, ..., r_K)$, which are fed to an LSTM layer to calculate hidden $hr = (hr_1, ..., hr_k, ..., hr_K)$. The last hidden output $hr_K$ is employed as the utterance-level representation $RT$ of the sentence, which is concatenated with $\mathbb{F}^{(I)}$ as the input to the emotion classifier.

## 4 Experiments

The public emotion benchmark database IEMOCAP [Busso *et al.*, 2008] and real scene database RID are used in the experiments for performance evaluation. IEMOCAP is employed to compare the performance of the proposed SER framework with state-of-the-art approaches, and RID is employed to assess the robustness and effectiveness of the proposed framework in realistic interaction scenarios. The implementations of this work are shared on the public website[1].

### 4.1 Experimental Setup

**Database.** The IEMOCAP database contains 12 hours of conversations in English, segmented into 5,531 utterances and categorized with 9 emotion classes: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and others. In experiment, to compare with state-of-the-art approaches, the utterances labeled 'excitement' are combined with the 'happy' class, forming a four-class database labeled $\{happy, angry, sad, neutral\}$ with each class containing $\{1636, 1103, 1084, 1708\}$ utterances respectively. RID is collected from realistic human-computer interactions by Microsoft, authored and labeled by users. The database contains 358,024 utterances, categorized into 5 emotional classes: angry, neutral, happy, sad and surprise, each containing $\{82952, 61629, 67499, 82125, 63819\}$ utterances respectively. Both IEMOCAP and RID are randomly shifted and divided into three partitions with a proportion of 8:1:1 for training, validation and testing.

[1]https://github.com/thuhcsi/IJCAI2019-DRL4SER/

| | Method | IEMOCAP | | | | IEMOCAP | | | | RID | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Input | Reported UA* | UA | F1 | Input | Reported UA* | UA | F1 | Input | UA | F1 |
| [Xia and Liu, 2017] | DNN | A | 60.1% | 60.4% | 0.597 | A+L | - | 72.8% | 0.731 | A | 68.7% | 0.691 |
| [Poria *et al.*, 2016] | CNN | A | 61.3% | 60.7% | 0.608 | A+L | 65.1% | 69.7% | 0.702 | A | 71.2% | 0.719 |
| [Poria *et al.*, 2017] | LSTM | A | 57.1% | 55.8% | 0.563 | A+L | 74.5% | 73.9% | 0.740 | A | 62.1% | 0.620 |
| [Mirsamadi *et al.*, 2017] | RNN & Attention | A | 58.8% | 59.6% | 0.594 | A+L | - | 74.3% | 0.745 | A | 69.1% | 0.687 |
| Our approach | The proposed RLC | A | - | **69.4%** | **0.693** | A+L | - | **79.2%** | **0.791** | A | **90.2%** | **0.901** |

Table 1: The performances of state-of-the-art approaches and the proposed framework on IEMOCAP and RID. Unweighted Accuracy (UA) and F1-measure score (F1) are the higher the better. A: acoustic features, L: lexical features. (*: the original performance reported in paper.)

**Features.** For better comparison to state-of-the-art approaches, acoustic features and textual features are extracted from speech and corresponding transcribed text. As suggested in computational paralinguistic challenges (ComParE) [Schuller *et al.*, 2017], 17-dimensional LLD acoustic features are extracted as the input: 12-dimensional Mel-frequency cepstral coefficients (MFCCs) and 1-dimensional logarithmic energy, voicing probability, harmonic-to-noise ratio (HNR), logarithmic fundamental frequency (LF0) and zero-crossing rate, with 25 msec frame window length and 5 msec intervals. Lexical features are extracted using a well-trained Word2Vec model proposed in [Mikolov *et al.*, 2013], resulting in 300-dimensional vector for each word of input utterances.

**Hyper-parameters.** In the proposed framework, filters employed in the residual convolutional layers are depicted in Fig.2, each Multi-head Self-attention block has 4 parallel heads, and each LSTM contains 256 units. The iteration times $I$ in GCA-LSTM is empirically set at 3. The emotion classifier is constructed with three stacked dense layers, each contains 256 units. The initial learning rate of training is $10^{-3}$.

**Implementation and Training.** The proposed framework and state-of-the-art comparisons are implemented using TensorFlow [Abadi *et al.*, 2015] deep learning framework, trained by stochastic optimization with 128 samples per batch. Cross-entropy loss is employed as the loss function to measure the performance of emotion recognition, and Adam [Kingma and Ba, 2014] algorithm is employed as the optimizer in training. Specially, back-propagation through time (BPTT) is employed to train the LSTM oriented models.

**Evaluation Metrics.** In this work, unweighted accuracy (UA) [Rozgic *et al.*, 2012] and F1-measure [Powers, 2011] are employed to measure the performance of the proposed framework and the comparisons. UA is defined as the mean of accuracies for different emotion categories. All the experimental results reported are based on 10-fold cross-validation.

## 4.2 Comparison to State-of-the-art

Four representative state-of-the-art SER approaches are implemented for comparison with the proposed framework on IEMOCAP and RID. Specially, the numbers of filters/units of convolution/dense layers in comparisons are balanced to ensure the parameters consistency.

1) [Xia and Liu, 2017] employs a series of statistics functions on LLDs to generate utterance-level representation, and a DNN based classifier to infer emotion.
2) [Poria *et al.*, 2016] employs CNN and multiple kernel learning classifier to extract features from multimodalities and infer emotion.
3) [Poria *et al.*, 2017] the conventional LSTM based model is used to capture inner contextual information of utterances for speech emotion recognition. Specially, to compare with the proposed framework, the cross-utterance contextual information is not considered.
4) [Mirsamadi *et al.*, 2017] employs the local attention-based RNN for emotional relevant feature production and emotion recognition.

**Experimental Result.** As shown in Table.1, our implementations of state-of-the-art approaches have similar performance to the reported results of original papers. On IEMOCAP, when using acoustic features only, the proposed SER system achieves significant improvement on UA, +14.9% relative improvement compared to [Xia and Liu, 2017], +14.3% compared to [Poria *et al.*, 2016], +24.3% compared to [Poria *et al.*, 2017], +16.4% compared to [Mirsamadi *et al.*, 2017].When evaluated with larger data scale on RID, all the implementations have achieved better performance, and the proposed framework gains +26.7% to +45.2% relative improvement on UA compared to state-of-the-art approaches. When considering both audio and lexical features, the proposed system also outperforms other comparisons with +6.6% to +13.6% relative improvements on UA.

## 4.3 Component Contribution Research

To figure out the contribution of individual components, systems with different combining of measures are implemented:

1) Baseline system, conventional LSTM based approach [Poria *et al.*, 2017], employing stacked LSTM layers to infer emotion using input features directly.
2) System 1 (S1), residual convolutional network is employed for feature learning, and stacked LSTM layers are used to produce the utterance-level representation.
3) System 2 (S2), Multi-head Self-attention blocks in residual convolutional network are employed to enhance the effectiveness of feature learning.
4) System 3 (S3), GCA-LSTM is employed to replace the stacked LSTM layers for representation generation.

**Experimental Result.** As shown in Table.2, when evaluated on IEMOCAP with acoustic features, applying RCN gains +9.6% relative improvement on UA compared to the baseline system. Applying Multi-head Self-attention can further gain +6.9% relative improvement on UA. Compared with S2, with GCA-LSTM for representation generation, the proposed S3 system gains +5.0% relative improvement on UA. Similar improvements can be also observed in experiments on RID, indicating the effectiveness of the used components.

| | Parameters | Residual CNN | Multi-head Self-attention | RNN Cell | IEMOCAP | | | | | | RID | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Input | UA (%) | F1 | Input | UA (%) | F1 | Input | UA (%) | F1 |
| Baseline | 9.27M | NO | NO | LSTM | A | 56.4% | 0.565 | A+L | 72.9% | 0.731 | A | 62.1% | 0.620 |
| S1 | 9.15M | YES | NO | LSTM | A | 61.8% | 0.621 | A+L | 74.3% | 0.745 | A | 74.6% | 0.744 |
| S2 | 9.07M | YES | YES | LSTM | A | 66.1% | 0.667 | A+L | 77.1% | 0.770 | A | 85.3% | 0.849 |
| S3 | 9.11M | YES | YES | GCA-LSTM | A | **69.4%** | **0.693** | A+L | **79.2%** | **0.791** | A | **90.2%** | **0.901** |

Table 2: Experimental results for component contribution evaluation. A: acoustic features. L: lexical features. Unweighted Accuracy (UA) and F1-measure score (F1) are the higher the better. The units employed in comparison systems are balanced to ensure parameters consistency.

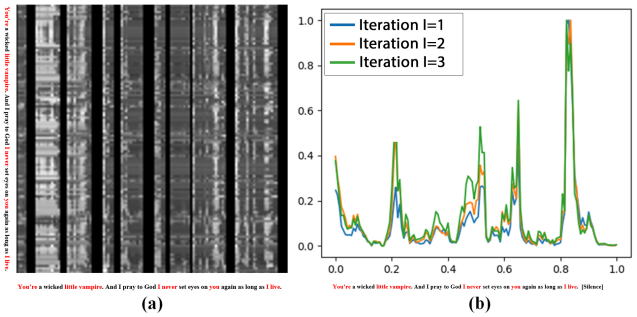


Figure 5: The intermediate result in processing an angry speech. Red words present the attention-capturing vocalizations for human annotator. (a): Learned correlation matrix in a head of Multi-head Self-attention block, the brighter the pixel, the higher the attention value is. (b): The informativeness degree computed by GCA-LSTM.

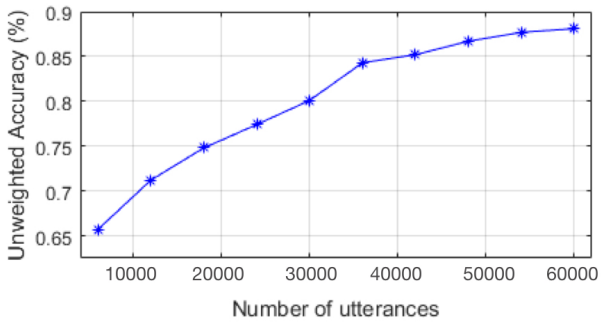


Figure 6: Data scale evaluation, the value of horizontal axis represents the number of utterances from each emotion class in RID.

### 4.4 Analysis and Discussion

**Learned Suprasegmental Features Analysis.** As depicted in Fig. 5(a), in the learned correlation matrix, the attention-capturing vocalizations are presented with higher importance weights than other elements. As the correlation matrix is employed to compute the final value of output, this mechanism can help the model to selectively focus on emotion-salient informative elements in feature learning.

**Informativeness Degree Analysis.** As shown in Fig. 5(b), in GCA-LSTM, emotion-salient factors are presented with significantly higher informative scores; and the informative score difference can be further enhanced in the iteratively learning. The informative scores can directly determine the information importing, hence the generated representation will focus more on emotion-salient factors in speech, and meanwhile ignore the stochastic disturbance. In this way, the learned representation can be more robust and discriminative.

**Data Scale Evaluation.** The training data scale can significantly affect the performance of the proposed framework. As shown in Fig.6, started with restricted training data scale, the framework performance is limited and gradually improved with increasing training data, and achieves relative stability after using 36,000 utterances from each emotion class.

From the results and analysis, conclusions are summarized:

1) The comparison between [Poria *et al.*, 2017] and other implementations has stated the effectiveness of LSTM in representation generation; however, for input with long temporal steps, the performance is limited.
2) Residual convolutional network is effective in learning paralinguistic features, meanwhile providing efficient time-resolution reduction.
3) Using Multi-head Self-attention blocks in residual convolutional network provides strong ability in modeling the element-wise relative dependencies across the learned suprasegmental features, and further enhances the ability in learning emotion-salient information.
4) By employing GCA-LSTM, with the strong global context-aware attention ability, the RLC can selectively import information from emotion-salient factors, providing more discriminative representation for SER.
5) Data scale is important in training an effective SER system. The performance of SER system is limited when trained with limited data; with the increase of data scale, the performance will improve significantly.

## 5 Conclusion

In this paper, we proposed a novel representation learning component (RLC) for speech emotion recognition. Constructed with Multi-head Self-attention and GCA-LSTM, the RLC can extract informative suprasegmental features from acoustic features, and produce robust and discriminative representation with selective attention. Experiments on IEMOCAP and a real scene interaction database demonstrate the outperformance of the proposed SER framework, with significant improvement comparing to state-of-the-art approaches.

## Acknowledgments

This work was conducted when the first author was an intern at Microsoft, and is supported by joint research fund of National Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) (61531166002, N_CUHK404/15), National Natural Science Foundation of China (61831022, 61521002, 61433018, 61375027) and National Social Science Foundation of China (13&ZD189).

## References


[Abadi *et al.*, 2015] Martín Abadi, Ashish Agarwal, and et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Busso *et al.*, 2008] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, et al. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[Chorowski *et al.*, 2015] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lee and Tashev, 2015] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. 2015.

[Lian *et al.*, 2018] Zheng Lian, Ya Li, Jianhua Tao, and Jian Huang. Improving speech emotion recognition via transformer-based predictive coding through transfer learning. *CoRR*, abs/1811.07691, 2018.

[Liu *et al.*, 2017] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 7, page 43, 2017.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mirsamadi *et al.*, 2017] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2227–2231. IEEE, 2017.

[Poria *et al.*, 2016] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Proceedings of the IEEE International Conference on Data Mining*, pages 439–448. IEEE, 2016.

[Poria *et al.*, 2017] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 873–883, 2017.

[Powers, 2011] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[Rozgic *et al.*, 2012] Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad. Ensemble of svm trees for multimodal emotion recognition. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–4. IEEE, 2012.

[Schirmer and Adolphs, 2017] Annett Schirmer and Ralph Adolphs. Emotion perception from face, voice, and touch: comparisons and convergence. *Trends in cognitive sciences*, 21(3):216–228, 2017.

[Schuller *et al.*, 2017] Björn Schuller, Stefan Steidl, Anton Batliner, Elika Bergelson, et al. The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, pages 3442–3446, 2017.

[Steinschneider *et al.*, 2013] Mitchell Steinschneider, Kirill V Nourski, and Yonatan I Fishman. Representation of speech in human auditory cortex: is it special? *Hearing research*, 305:57–73, 2013.

[Strapparava and Mihalcea, 2008] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

[Stuhlsatz *et al.*, 2011] André Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Günter Meier, and Björn Schuller. Deep neural networks for acoustic emotion recognition: raising the benchmarks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5688–5691. IEEE, 2011.

[Trigeorgis *et al.*, 2016] George Trigeorgis, Fabien Ringeval, et al. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5200–5204. IEEE, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[Xia and Liu, 2017] Rui Xia and Yang Liu. A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on Affective Computing*, (1):3–14, 2017.