A Study on Speech Emotion Recognition Using Self-Attention Transformer and Cross-Attention Transformer

(自己注意型トランスフォーマーと相互注意型トランスフォーマーを用いた音声感情認識に関する研究)

何 雨潤

He Yurun

ID Number: 37-206536

Supervisor: Prof. Nobuaki Minematsu

Department of Electrical Engineering and Information Systems,
Graduate School of Engineering,
The University of Tokyo

Master Thesis
July 2022

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

He Yurun
July 2022

Acknowledgements

I would like to express my deep gratitude to my supervisor Prof. Nobuaki Minematsu, who has offered me valuable guidance in my academic studies and whose illuminating suggestions and useful comments have contributed greatly to the completion of this thesis. Also, I would like to thank Associate Prof. Daisuke Saito for giving me much advice and maintaining the laboratory server so that I can have the best environment to conduct these experiments, even if it is impossible for me to come to Japan due to the COVID-19 pandemic. I want to thank all the members in the laboratory who have always been kind to me, the two years of research life is the cherished and indelible memory in my life. Finally, I would like to dedicate this thesis to my loving family, who have kept giving me great spiritual and material support.

Abstract

Speech emotion recognition (SER) is a fundamental task to detect the implied emotions from speech signals. It is of great challenge in learning suitable affect-salient features for achieving good performance. With the advent of deep learning (DL), many pieces of research conducted on DL-based SER systems have shown extreme advantages. Among them, transformer exhibited outstanding qualities in learning relevant representations associated with this task. However, there are still several problems to deal with: 1) a normal transformer is only able to process the uni-source input, 2) interaction between transformer and other DL structures — convolutional neural network (CNN) and Long short-term memory (LSTM) are still needed to be investigated, 3) there is often only one kind of input features in a transformer-based SER system, which may cause limited knowledge. For the first problem, we attempt to use the cross-attention transformer (CAT) to handle bi-source input. For better differentiation, we alias the normal transformer model as self-attention transformer (SAT). Towards the second problem, we propose two SER systems: integration of LSTM with SAT (ILSAT) and CAT for CNN and LSTM sources (CL-CAT). While the former is to replace the positional encoding in SAT by LSTM, the latter is for joint encoding of CNN and LSTM sources using CAT. Experimental results conducted on the IEMOCAP dataset show that both our proposals can make a promising improvement relative to the baseline system. To solve the third problem, additional acoustic (multiple acoustic features) and textual features (multimodal features) are considered and fused with previous input features by CAT, and corresponding SER systems SMW_CAT (S, M, W are the abbreviation of log mel spectrogram, MFCC, and raw waveform data) and AT_CAT-SAT (A, T are the abbreviation of audio and text) are proposed respectively. SMW_CAT has achieved a 73.80% WA and 74.25% UA, which outperforms existing state-of-art approaches. AT_CAT-SAT has achieved a 73.64% weighted accuracy (WA) and 75.05% unweighted accuracy (UA), where substantial improvement can be observed compared with SER systems with single modal input. In addition, we exploit t-SNE to visualize the process of learning relative representations in our systems.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Research background

The speech signal gives us the most natural, intuitive, and speedy way to express ourselves in daily life. This fact has motivated researchers to think of speech as an efficient method for human-computer interaction (HCI). Although there have been tremendous efforts on speech recognition which refers to the process of converting human speech into word sequences, it is still difficult for machines to fully understand a person's voice, since it carries not only explicit linguistic information but also implicit paralinguistic information such as emotion. Emotion is indispensable in the speech that can help the machine to better catch the intention of the speaker. This has introduced a related research field, called speech emotion recognition (SER), which is defined as detecting the emotional state of a speaker from his/her speech. SER is particularly functional for a wide range of applications. For those systems of web movies and computer tutorials, the response to the user should depend on the detected emotion. It is also useful for in-car board systems where information on the mental state of the driver may be provided to the system to initiate his/her safety. In call center applications and mobile communication, the main objective of employing SER is to adapt the system response upon detecting frustration or annoyance in the speaker's voice [9].

Although SER has received substantial attention from both academia and industry because of its practical importance, it is a challenging task, because emotions are subjective. There is no common consensus on how to measure or categorize them, i.e. there exists no unanimously accepted emotion model [1]. However, generic emotion models can be divided into two types, namely discrete emotion models and dimensional emotion models. While the former mainly defines emotions into limited categories, such as happy, angry, sad, and so on, the latter uses two dimensions (valence-arousal) or three dimensions (valence-arousal-dominance) to describe

emotion. Because most of the existing SER systems focus on the former, in this thesis, we will build our SER systems based on the discrete emotion model.

SER aims to identify the high-level affective status of an utterance from the low-level features. It can be treated as a classification problem on sequences [13]. In the past, many different methods were proposed, most of which extracted a large amount of complex low-level handcrafted features (such as pitch, energy, etc) out of the utterance and then applied conventional classification algorithms like Hidden Markov Model (HMM) [30, 17] and support vector machines (SVM) [15]. In recent years, the appearance of deep learning has changed this field in ways of extracting discriminative features. [13] proposed to use the segments with the highest energy to train a deep neural network (DNN) model to extract effective emotional information. [24] first used convolutional neural networks (CNN) to learn robust features for SER and showed excellent performances on several benchmark datasets. [18] applied a long short-term memory (LSTM) to learn long-range temporal relationships for SER. In [41], they directly used raw audio samples to train a convolutional recurrent neural network (CRNN) to build continuous arousal and valence space.

More recently, the application of attention-based deep-learning approaches has sprung up in this task. In [6], attention layers were used to focus on the emotionally relevant parts and produced utterance-level affective-salient features for SER. The authors of other researches showed the efficiency of a neoteric deep learning model — transformer on the SER task [38, 27, 36].

1.2 Research objective

The above works have promoted the progress in the research of SER, and the successful application of the transformer has inspired us to extract related features from the speech signal. However, some issues still need to be addressed. First, the interaction between the transformer and two widely used deep learning models — CNN and LSTM lacks investigation. Second, most current SER systems utilize only one kind of acoustic feature as input embeddings, which may lead to limited knowledge. Third, emotion can be technologically captured and assessed in a multimodal way, including not only speech signals but also facial expressions, physiological signals, word recognition, brain signals, and so on [1].

To solve the three problems, the transformer model must be expanded, as it can only process the uni-source input. Considering its peculiarity, we modify the transformer to deal with bi-source input, named cross-attention transformer (CAT). For better differentiation, we alias the normal transformer model as self-attention transformer (SAT).

With SAT and CAT, we can deal with the issues. Towards the first issue, we propose integration of LSTM with SAT (ILSAT) and CAT for CNN and LSTM sources (CL-CAT), where the former is for replacing the positional encodings in SAT with LSTM by integrating them in a parallel manner, and the latter is to use CAT to joint-encode CNN and LSTM. By introducing three acoustic features — raw waveform data, log mel spectrogram, and MFCC, we propose SMW_CAT (S, M, W are the abbreviation of log mel spectrogram, MFCC, and raw waveform data) to solve the second problem. Eventually, for addressing the third issue, AT_CAT-SAT (A, T are the abbreviation of audio and text) is proposed to take acoustic and textual features as input embeddings. Experiments conducted on the IEMOCAP dataset [5] show the effectiveness of these SER systems.

1.3 Thesis organization

This thesis is organized as follows: Chapter 2 introduces components in a traditional SER system. Chapter 3 gives a review of some exemplary deep learning-based SER systems. Chapter 4 describes the model architecture of SAT and CAT and the design of our proposals in detail. Chapter 5 presents the experimental details, results, and analysis of the proposed networks on the IEMOCAP dataset. Finally, Chapter 6 concludes the whole paper and points out directions for some future works.

Chapter 2

Components of Traditional Speech Emotion Recognition System

2.1 Overview

As illustrated in Figure 2.1, a traditional SER system mainly consists of two parts: a front-end processing module that extracts the appropriate features from the input speech data, and a back-end classifier that predicts the underlying emotion. During the feature extraction stage, a speech signal is converted into numerical values using various front-end signal processing techniques. Extracted feature vectors have a compact form and ideally should capture essential information from the signal. In the back-end, an appropriate classifier is selected according to the task to be performed [21]. Besides, the definition of emotion is essential to construct a criterion for SER. In this chapter, initially emotion models are introduced. Then feature extraction and classification, the two core modules in conventional SER system, will be addressed.

2.2 Emotion Models

To successfully implement an SER system, emotion must be carefully defined and modeled. From the psychological point of view, human emotions can be identified and grouped based on emotion type, emotion intensity, and many other parameters, which can be all combined and realized into emotion models. However, there are no unanimously accepted emotion models, and it is still an open question in psychology. Based on different emotion theories, existing emotion models can be divided into two classes: discrete and dimensional.

Discrete emotion models, also known as categorical emotion models, define emotions into limited categories. Depicted in Figure 2.2 (a), the most widely used one is Ekman's basic

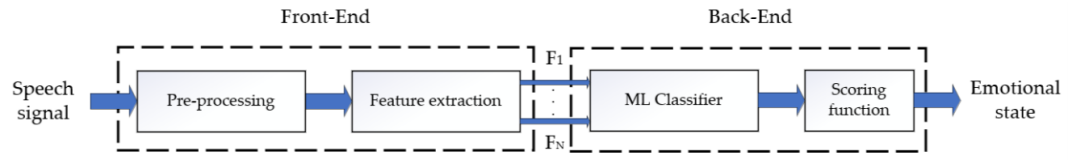
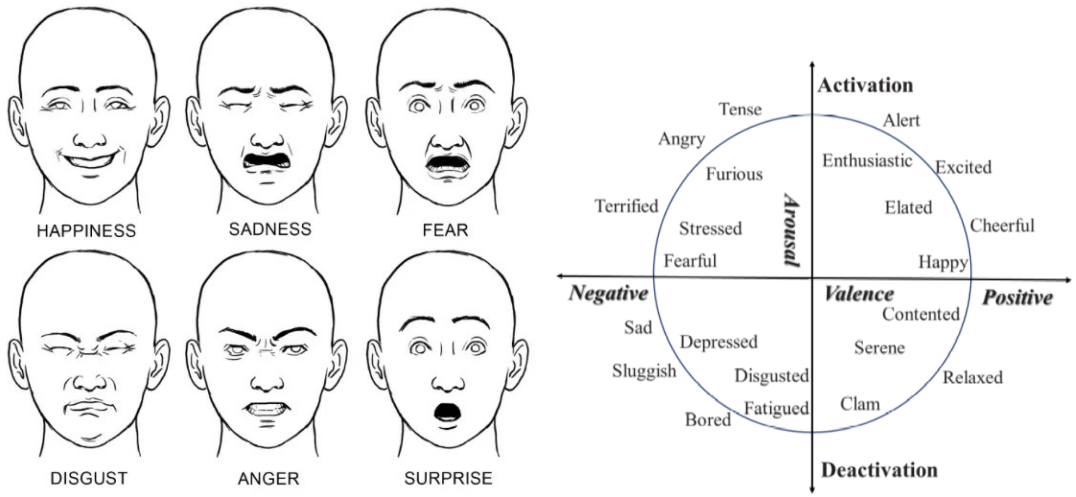


Fig. 2.1 Block diagram of traditional SER system [21].



(a) Discrete emotion model (b) Dimensional emotion model

Fig. 2.2 Two kinds of emotion models.

emotion model [8], which is based on the six basic emotions: happiness, sadness, fear, disgust, anger, and surprise. Other emotions are obtained by the combination of the basic ones. On the other hand, dimensional emotion models define a few dimensions with some parameters and specify emotions according to those dimensions. Two or three dimensions are used in most dimensional emotion models — 'valence' (indicates the positivity or negativity of an emotion), 'arousal' (indicates the excitement level of an emotion) and 'dominance' (indicates the level of control over an emotion) [33, 4]. A valence-arousal based dimensional emotion model is shown in Figure 2.2 (b).

Since most of the existing SER systems focus on the discrete emotion model, the rest of this paper will only be discussed based on it.

2.3 Feature extraction

In SER, one of the central research issues is how to extract discriminative, affect-salient features from speech signals, i.e. features that are sensitive to emotion and invariant to nuisance factors such as speakers and contents [9]. The performance of SER systems significantly relies on

Table 2.1 Four categories of speech features.

Prosodic features	pitch, energy, duration
Spectral features	Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients(LPCC), Log-Frequency Power Coefficients (LFPC), Gammatone Frequency Cepstral Coefficients (GFCC), formants
Voice quality features	jitter, shimmer, harmonics-to-noise ratio (HNR), Normalized Amplitude Quotient (NAQ), Quasi Open Quotient (QOQ)
Teager Energy Operator (TEO) based features	TEO-decomposed frequency modulation variation (TEO-FM-Var), TEO auto-correlation envelope area (TEO-Auto-Env), critical band based TEO auto-correlation envelope area (TEO-CB-Auto-Env)

the selection of suitable features. This section first introduces types of acoustic features. Then the extraction procedure of log mel spectrogram and MFCC features are explained in-depth, as they are about to be used as input features in this paper. After that, the definition of global features and local features will be discussed.

2.3.1 Categories of acoustic features

Generally, acoustic features in SER are grouped in four categories: prosodic features, spectral features, voice quality features, and Teager Energy Operator (TEO) based features [1]. Table 2.1 shows examples of features belonging to each type.

It is common in SER to combine features that belong to different categories to obtain better results. Over the many years of research, the focus has been placed on the selection of the ideal set of descriptors for emotional speech. Some hand-crafted features sets such as eGeMAPS [10], ComParE [44] were proposed. However, despite great research efforts, until now there is still no consensus on the most appropriate features for precise and distinctive classification.

2.3.2 The feature extraction of log mel spectrogram and MFCC

As shown in Figure 2.3, the extraction of log mel spectrogram and MFCC basically includes pre-emphasis, framing, windowing, Fast Fourier Transform (FFT), Mel-filter bank, and Discrete cosine transform (DCT). The detailed description of each step is explained below.

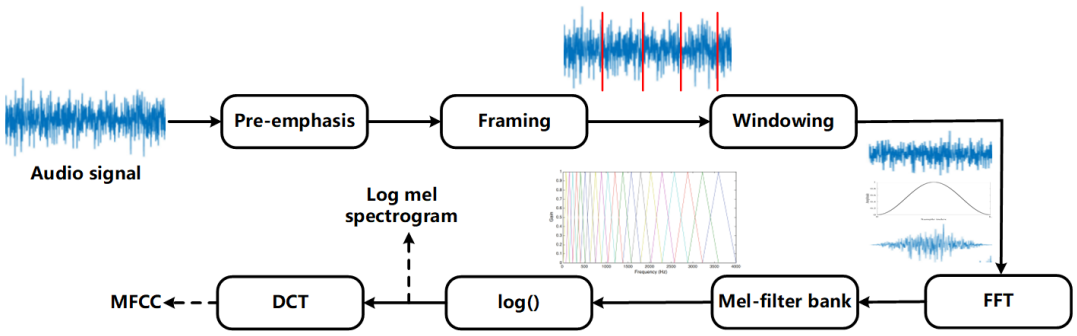


Fig. 2.3 Process of extracting log mel spectrogram and MFCC.

Pre-emphasis

When human speech is converted into a spectrum and observed, the energy of the high-frequency component tends to be lower than that of the low frequency, thus more difficult to be caught. The purpose of pre-emphasis is to reinforce the high-frequency part so that the raw speech signal will have a relatively flattened energy distribution over the entire frequency range. Let $s(n)$ be the digitalized speech signal, pre-emphasis is calculated as follows. Usually coefficient α is set to 0.95 or 0.97.

$$y(n) = s(n) - \alpha s(n-1) \tag{2.1}$$

Framing and windowing

Although the speech signal varies very fast over time, in a sufficiently short interval (generally considered as 20 to 30 ms), it can be relatively stable, i.e. the speech signal has short-term stability. These small intervals are called frames. Besides, in order to smooth the transition between frames, there will be an overlap between the previous frame and the next frame, which is known as frame shift, often being set to half of the frame length.

After framing, the next step is generally applying a window function to frames. While breaking the signal into frames, if we directly chop it off at the edges of the signal, the sudden fall in amplitude at the edges will produce noise in the high-frequency domain. Ordinarily, a Hamming window is utilized to solve this problem, which can be formulated as:

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{M-1}), \quad 0 \leq n \leq M-1 \tag{2.2}$$

where M is the window size.

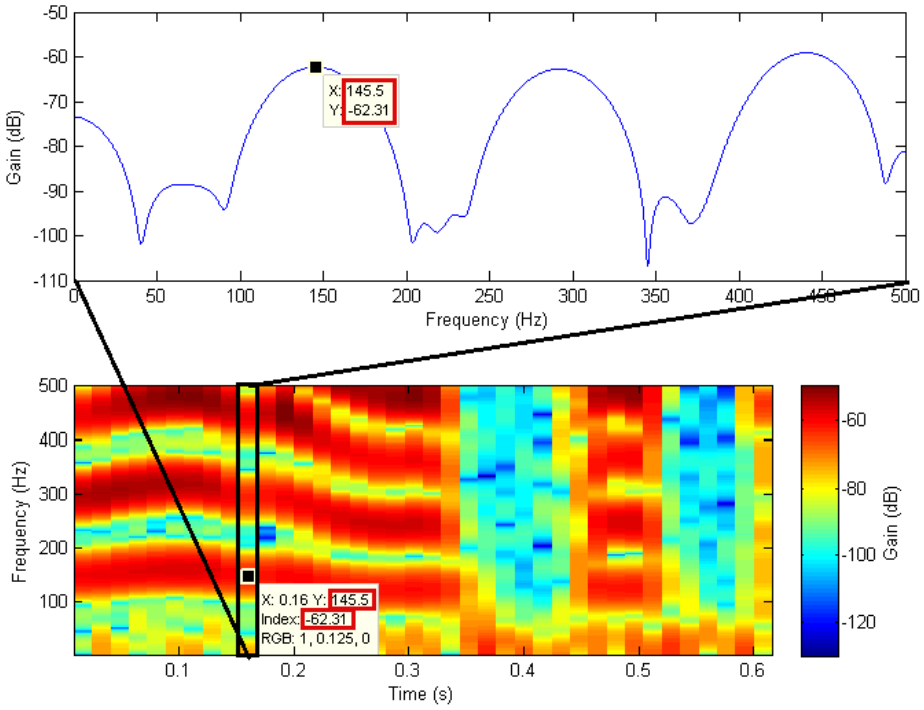


Fig. 2.4 Relationship between spectrum (top) and spectrogram (bottom) [40].

Then we can add it to each frame:

$$x(n) = y(n)w(n) \tag{2.3}$$

Fast Fourier Transform (FFT)

We will convert the signal from the time domain to the frequency domain by applying the Discrete Fourier Transform (DFT). For audio signals, analyzing in the frequency domain is easier than in the time domain.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}}, \quad 0 \leq k \leq N-1 \tag{2.4}$$

where N is the number of points used to compute the DFT. $X(k)$ is called spectrum, and $X(k)$ of all frames is known as spectrogram. The relation between them is depicted in Figure 2.4. In addition, normally Fast Fourier Transform (FFT) will be exploited to substitute DFT for improving the calculation speed.

Mel-filter banks

The way our ears will perceive the sound is different from how the machines will perceive the sound. Our ears have higher resolution at a lower frequency than at a higher frequency. So if we hear sound at 200Hz and 300Hz we can differentiate it easily when compared to the sounds at 1500Hz and 1600Hz even though both have a difference of 100Hz between them. Whereas for the machine, the resolution is the same at all the frequencies. It is noticed that modeling the human hearing property at the feature extraction stage will improve the performance of the model. So we will use the mel-scale to map the actual frequency to the frequency that human beings will perceive.

$$f_{Mel} = 1127 \ln(1 + \frac{f}{700}) \tag{2.5}$$

where f denotes the physical frequency in Hz, and f_{Mel} denotes the perceived frequency.

To implement this, the spectrum $X(k)$ will pass through a set of band-pass filters $H_m(k)$ known as Mel-filter bank, resulting in the mel spectrum $S_{Mel}(m)$. For the reason that humans are less sensitive to change in audio signal energy at higher energy compared to lower energy, and log function also has a similar property: at a low value of the input, the gradient of the log function will be higher, but at high value of the input, gradient value is less, mel spectrum is always represented on a log scale to mimic the human hearing system. Eventually, we can obtain the log mel spectrum $S(m)$.

$$S(m) = \ln(S_{Mel}(m)) = \ln\left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right), \quad 0 \leq m \leq M-1 \tag{2.6}$$

$$H_m(k) = \begin{cases} \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k \leq f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & others \end{cases} \tag{2.7}$$

where M is the total number of Mel-filter banks, $f()$ is the list of mel-spaced frequencies.

Discrete cosine transform (DCT)

Since the Mel-filter banks are all overlapping, the filter bank energies are quite correlated with each other. Therefore DCT is applied to the log mel spectrum $S(m)$ for decorrelation and

produces a set of coefficients $c(n)$, called Mel-Frequency Cepstral Coefficients (MFCC).

$$c(n) = \sum_{m=0}^{M-1} S(m) \cos\left(\frac{\pi n(m-0.5)}{M}\right), \quad n = 0, 1, 2, \cdots, C-1 \tag{2.8}$$

where C is the dimension of MFCCs. As most of the signal information is represented by the first few cepstral coefficients, the system can be made robust by extracting only those coefficients ignoring or truncating higher order DCT components. In the task SER, the dimension of MFCC is usually 13, 20, or 40.

2.3.3 Local features and global features

Based on the region of analysis used for feature extraction, features can be separated into two types. Local features, also called short-term or frame-level features, are extracted from frames. On the other hand, different statistical aggregation functions (such as mean, max, variance, linear regression coefficients, etc.) are applied to every local feature over the duration of the utterance, and results are concatenated into a long feature vector at the utterance level, which is named global features, also known as long-term or utterance-level features. The role of these global features is to roughly describe the temporal variations and contours of the different local features during the utterance [26]. An example of obtaining global features from local features will be described in Section 3.2.2.

There has been a disagreement on which local and global features are more suitable for SER. The two kinds of features have their own advantages and disadvantages, it is hard to say which is better. For instance, the number of global features is less than local ones, meaning the application of cross-validation and feature selection algorithms to them are executed much faster. However, temporal information present in speech signals is completely lost in global features [9].

2.4 Classification

Once the features are extracted from speech data, models are trained on the extracted feature set so that new instances can be classified based on the emotions they portray. There are many different classifiers used as models to analyze emotions from data. Basically, traditional SER systems can be realized at the frame level or at the utterance level.

In the frame approach, generative models like Hidden Markov Model (HMM) [30, 17] and Gaussian Mixture Model (GMM) [35] are used to learn the underlying probability distribution of local features of each emotional state and then a Bayesian classifier is trained using maximum

likelihood principle. HMM is a Markov process with the incapacity of observing the process that generates states directly. Each state has a probability distribution over the possible output tokens, and the current state at time t only depends on the previous state at time $t-1$. The internal behavior of HMM refers to the state sequence through which the model passes. GMM is another probabilistic model which can be considered as a special continuous HMM that contains only one state. The idea behind the mixture models is to model the data in terms of a mixture of several components, where each component has a simple parametric form, such as a Gaussian.

In the utterance approach, global features are calculated and utilized for training discriminative classifiers such as Support Vector Machines (SVM) [15] and K Nearest Neighbours (KNN) [31]. SVM is a supervised classifier that finds an optimal hyperplane for linearly separable patterns which has the maximum margin between data points of binary classes, i.e. to classify multiple emotions using SVM, the problem will convert into a binary classification problem in a high-dimensional space. KNN is another versatile classifier for estimating the likelihood that a data point will become a member of one group or another based on what group the data points nearest to it belong to. It is a lazy learning algorithm because it doesn't perform any training when supplying the training data.

Chapter 3

Speech Emotion Recognition Systems based on Deep Learning

3.1 Overview

In recent years, deep learning has achieved tremendous success in various domains such as object detection [32] and dialogue systems [43]. It is part of a broader family of machine learning methods that are characterized by a graded multi-layer structure. As mentioned in Section 2.3, it is of importance to select efficient hand-crafted features for SER, but often requires professional knowledge. The emergence of deep learning can help this task to learn adaptive low-level features from raw data and high-level features from low-level ones in a hierarchical manner nullifying the over-dependence of traditional SER models on the choice of features [11].

In the early period of deep learning, generally the SER systems such as DNN-ELM [13] and CNN-SVM [24] utilized neural networks to obtain representations from raw features, but still used the classifier part like conventional ones. That means, we must train the deep learning structures at the frontend for feature extraction in the beginning, and then train the classifiers at the backend for final emotion categorization. This is called two-stage SER systems. As time goes on, many end-to-end systems came forth with the advantage of training only one time. Among them, the most widely used one is CRNN [22, 14, 23]. Meantime, some enhancement techniques were proposed for further performance improvement, including attention mechanism [6] and transfer learning. More recently, a novel deep learning model named transformer [42] has caused the second ripples in the field. This chapter will briefly review exemplary systems based on deep learning and describe the architecture of CRNN, attention mechanism, and transformer in detail.

3.2 Representative systems in early deep learning era

3.2.1 Two-stage systems

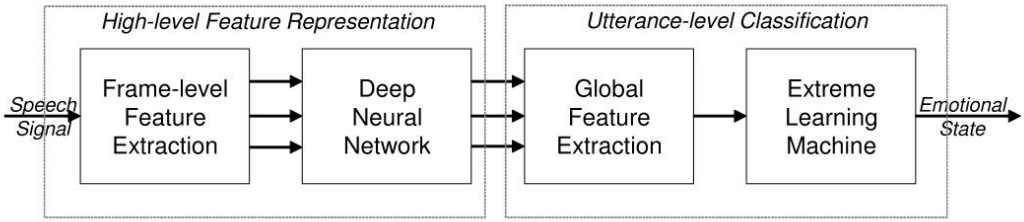
In two-stage SER systems, deep learning architectures like deep neural network (DNN), recurrent neural network (RNN), and convolutional neural network (CNN) are trained for frontend feature extraction, followed by a backend emotion recognizer such as SVM and extreme learning machine (ELM).

[13] proposed a DNN-ELM system for emotion recognition, as shown in Figure 3.1 (a). Their idea is to split each utterance into frames and calculate low-level features at first. Because emotions manifest in speech in a slow manner, i.e. they always exist in a long-range striding over multiple frames, the input features use the current frame concatenated with few context frames. Then DNN with three hidden layers is used to transform this sequence of features into the sequence of probability distributions over the target emotion labels.

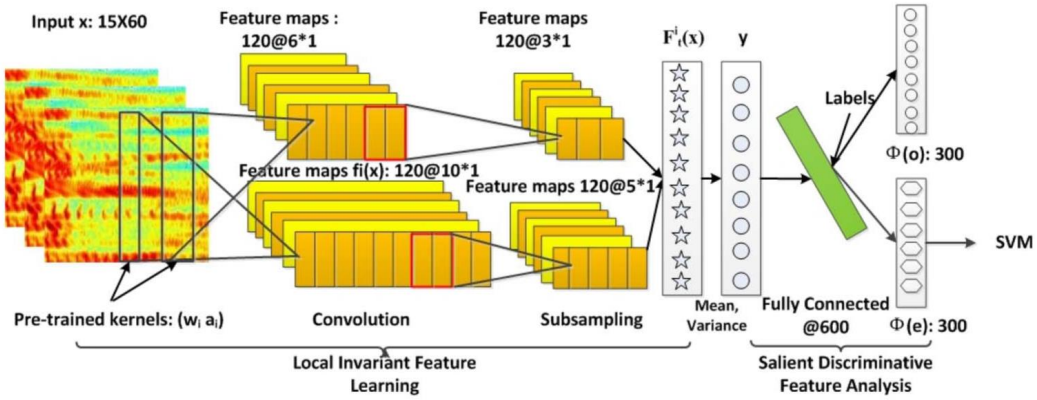
The system is trained by the following steps:

1. Train DNN by back propagation (BP) algorithm using cross-entropy loss. Then fix the parameters of DNN.
2. Aggregate the output probabilities of each frame into utterance-level features using simple statistics like maximum, minimum, average, percentiles, etc.
3. Train ELM is trained to classify utterances by emotional state.

To solve the shortcoming in [13] that the estimation of high-level features for the current frame uses few context frames are not sufficient to cover the long-range contextual effect in emotional speech, [18] proposed to replace DNN with RNN, thereupon they obtained the RNN-ELM model. RNN extends the notion of typical feed-forward architecture by adding inter-layer and self connections to units in the recurrent layer, it can effectively remember relevant long-term context from the input features. In [24], the authors utilized CNN to learn salient features to be used by an SVM for classification, where CNN is able to learn features that are insensitive to small variations in the input speech which can help in disentangling speaker-dependent variations as well as other sources of distortion. As depicted in Figure 3.1 (b), initially they used sparse auto-encoders to learn filters from spectrogram segments. Then they convolved the learned filters with spectrogram fragments to produce feature vectors. The feature vectors are mapped into two smaller feature vectors using a semi-supervised objective function, which disentangled affect-salient features from other non-salient features. Finally, the affect-salient features were used to train SVMs.



(a) DNN-ELM [13]



(b) CNN-SVM [24]

Fig. 3.1 Block diagram of two representative two-stage SER systems.

3.2.2 End-to-end systems

As time goes on, some end-to-end approaches based on deep learning have already arisen [22, 12, 29, 46]. Compared with two-stage SER systems, one of their advantages is that different modules are trained together instead of sequentially. Besides, they can avoid greedily enforcing the distribution of intermediate layers to approximate that of labels, and more proper representations can be obtained from the final layer [23]. Among these end-to-end systems, the most typical one is convolutional recurrent neural network (CRNN) [22, 14, 23], which is mainly a combination of CNN and RNN. The motivation is that they are complementary in their modeling capabilities, as CNN makes a good fist at reducing frequency variations (i.e. frequency modeling), and RNN is skilled at learning characteristics of data over long periods of time (i.e. temporal modeling). Therefore, the idea of combining them to utilize the merits of both comes naturally. A standard CRNN SER system is depicted in Figure 3.2.

Let $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_T]^T$ be the input features from a speech utterance split into segments in advance, where T represents the number of segments. The features are fed into a CNN being convolved and pooled in several times and eventually we can obtain a flattened feature

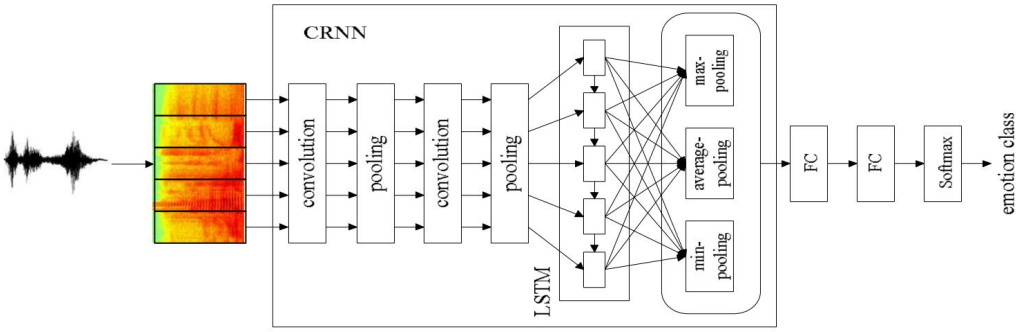


Fig. 3.2 Block Diagram of a standard CRNN SER system [23].

map $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_T]^T$. Next, a special RNN cell named Long short-term memory (LSTM) is applied to learn long-term dependencies and contextual information by introducing the gating mechanism. In RNN, every time step corresponds to a segment of the original audio utterance. After getting the output $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_T]^T$, three pooling functions maximum, average, minimum are applied to integrate over the time dimension resulting three global features $\mathbf{O}_{max}, \mathbf{O}_{avg}, \mathbf{O}_{min}$. Finally, they are concatenated together and pass through the fully connected layer and softmax layer to predict the final probabilities of each emotion class.

$$\mathbf{C} = \text{CNN}([\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_T]^T) \tag{3.1}$$

$$\mathbf{r}_t = \text{LSTM}(\mathbf{r}_{t-1}, \mathbf{c}_t), \text{ where } t \in 1, \cdots, T \tag{3.2}$$

$$\mathbf{O}_{max} = \max_{1 \le t \le T} \mathbf{r}_t \tag{3.3}$$

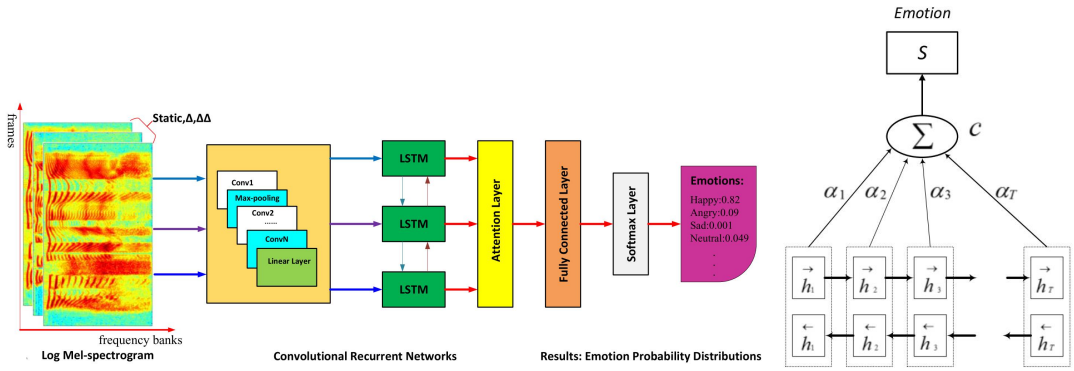
$$\mathbf{O}_{avg} = \frac{\sum_{t=1}^{T} \mathbf{r}_t}{T} \tag{3.4}$$

$$\mathbf{O}_{min} = \min_{1 \le t \le T} \mathbf{r}_t \tag{3.5}$$

3.3 Enhancement Techniques

3.3.1 Attention mechanism

One issue that appears to still puzzle researchers applying deep learning framework in SER, is how to effectively balance the short-term characterization at the frame level and long-term aggregation at the utterance level [26]. To get the global features, usually very simple and naive aggregation functions such as mean and max are applied to each of the local features over the duration of the utterance, which has shown in Figure 3.2. However, in fact, SER is related to



(a) Overview of the whole system (b) Attention layer

Fig. 3.3 Block diagram of Attention-CRNN SER system [6].

utterance classification with emotional content being differently distributed over the signal, and the emotion of the whole signal is a composition of emotions from different parts of the signal [28]. Therefore, we need a technique to focus more on the emotional part and focus less on the emotionless part of the whole utterance.

The emergence of attention mechanism [3] fulfills such a requirement. It ensures that the classifier pays attention to specific locations of the given utterance based on attention weights in each portion of the local features. A CRNN with attention layer [6] is depicted in Figure 3.3.

As shown in Figure 3.3 (b), at each time frame t, a softmax function is applied to obtain the normalized importance weight α_t which sum to unity:

$$\alpha_t = \frac{\exp(\mathbf{h}_t \mathbf{W})}{\sum_{\tau=1}^{T} \exp(\mathbf{h}_\tau \mathbf{W})} \tag{3.6}$$

where $\mathbf{W}$ is a trainable weighted matrix, $\mathbf{h}_t$ is the RNN output at time step t, $\mathbf{h}_t \mathbf{W}$ indicates a score for the contribution of frame t to the final utterance-level representation of the emotion.

The obtained weights are then used in a weighted average in time to obtain the utterance-level representation:

$$\mathbf{c} = \sum_{t=1}^{T} \alpha_t \mathbf{h}_t \tag{3.7}$$

3.3.2 Transfer learning

SER extremely lacks training data. Currently, the most widely used SER database is IEMOCAP [5] only comprises 10039 utterances, approximately a total of 12h. For those complicated SER systems, a small amount of training data means easily overfitting, resulting in reduced performance. One way to alleviate this data lacking issue is to transfer the knowledge learned from data in other related tasks (source tasks) to the task at hand (target task), which is known as transfer learning. In order to use transfer learning, the source model needs to be general enough. The most common approach for transfer learning is to train a source model with a set of source data [7] or use a pre-trained model [37], then use the learned knowledge as a starting point on a related task.

3.4 Transformer

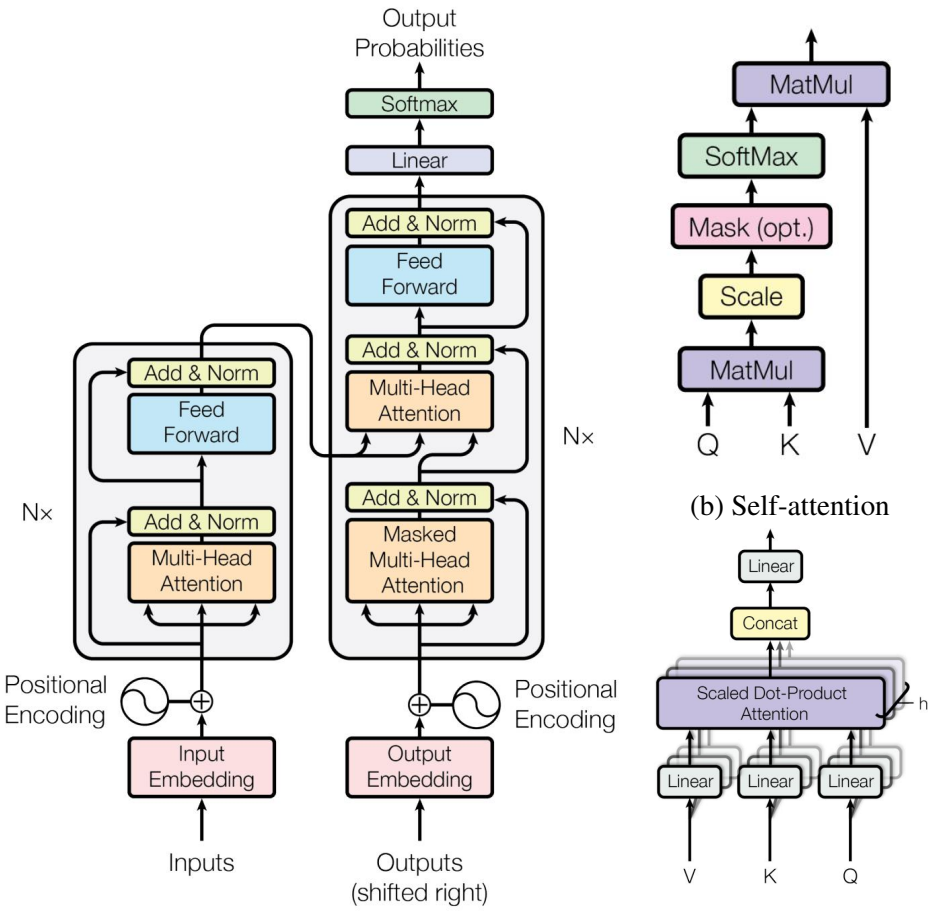
Although the appearance of the attention mechanism introduced in previous Section 3.3.1 can let the model look at different parts of speech with different weights, there still exists a problem: RNN families' calculation is limited to be sequential, i.e. RNN related algorithms can only be calculated from left to right or from right to left, causing

- The computation of time slice t depends on the result of $t-1$, which limits the parallel ability of the model.
- In the process of sequential computing, information will be lost. Although gate structures such as LSTM can alleviate the problem of long-term dependence to some extent, they still do nothing to solve the phenomenon of special long-term dependence.

To solve the two issues above, [42] proposed transformer, which exploits the self-attention mechanism to reduce the distance between any two positions in the sequence to a constant, as shown in Figure 3.4 (b). In addition, it is not an RNN-like sequential structure, which means better parallelism.

The calculation of self-attention can be divided into the following steps:

1. Obtain queries, keys, and values vectors through input sequential vector.
2. Compute the dot products of the queries of time t with all keys, divide each by $\sqrt{d_k}$, and apply a softmax function to obtain the weights on the values, where $\sqrt{d_k}$ represents the dimension of queries vector.
3. Multiply the weights and values and sum them, producing the output vector of time t.



(a) Overview of the whole transformer

(b) Self-attention

(c) Multi-head attention

Fig. 3.4 Block diagram of transformer [42].

4. Calculate the output vector of other times along the steps above.

In practice, the calculation above can be done in matrix form for faster processing. Given an input sequential matrix as $\mathbf{X} \in \mathbb{R}^{d_t \times d_f}$, by multiplying with three different trainable weight matrix $\mathbf{W}^Q \in \mathbb{R}^{d_f \times d_k}, \mathbf{W}^K \in \mathbb{R}^{d_f \times d_k}, \mathbf{W}^V \in \mathbb{R}^{d_f \times d_v}$, we can obtain the set of queries $\mathbf{Q} \in \mathbb{R}^{d_t \times d_k}$, the set of keys $\mathbf{K} \in \mathbb{R}^{d_t \times d_k}$, and the set of values $\mathbf{V} \in \mathbb{R}^{d_t \times d_v}$. Then the self-attention can be calculated as:

$$\mathbf{Z} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{3.8}$$

where variable $\mathbf{Z} \in \mathbb{R}^{d_t \times d_v}$ represents the attentional matrix.

Researchers found it beneficial to linearly project the queries, keys and values h times with different weight matrix $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ respectively. Each of these projected versions

are then performed the attention function in parallel, yielding d_v dimensional output values. These are concatenated and once again projected by multiplying with another weight matrix $\mathbf{W}^O \in \mathbb{R}^{hd_v \times d_f}$ to obtain the final output $\mathbf{O}_{MHA} \in \mathbb{R}^{d_t \times d_f}$:

$$\begin{aligned} \mathbf{O}_{MHA} = \text{MHA}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \text{Concat}(\mathbf{Z}_1,\ldots,\mathbf{Z}_h)\mathbf{W}^O \\ \text{where } \mathbf{Z}_i &= \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \end{aligned} \tag{3.9}$$

By doing so, the self-attention can be further refined into another mechanism called multi-head attention (MHA), illustrated in Figure 3.4 (c). It expands the model's ability to focus on different positions and gives the self-attention layer multiple representation subspaces.

Afterward, the output of the MHA layer is fed to a feed-forward neural network (FFN) to further increase the representation capacity. In addition, a residual connection is added around both MHA and FFN layer for solving the vanishing gradient problem in deep learning, followed by a layer-normalization step:

$$\mathbf{O}_{mid} = \text{LayerNorm}(\mathbf{X} + \mathbf{O}_{MHA}) \tag{3.10}$$

$$\mathbf{O}_{final} = \text{LayerNorm}(\mathbf{O}_{mid} + \text{FFN}(\mathbf{O}_{mid})) \tag{3.11}$$

To address the crucial problem that transformer has little ability to capture sequential sentences, it is required to add a positional encoding (PE). This means summing a sinusoid function with a large period over the input before feeding it to the first layer. The intuition here is that for any fixed offset k, PE_{pos+k} can be represented as a linear function of PE_{pos}, which provides great convenience for the model to capture the relative position relationship between sequential data:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d_f}}}\right) \tag{3.12}$$

$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d_f}}}\right) \tag{3.13}$$

where pos is the position of each sequential data in input features, i represents the i^{th} dimension of the input embedding of each data.

The whole transformer architecture is depicted in Figure 3.4 (a), which has an encoder part and a decoder part. Note that in the field SER, we only make use of the encoder component to get a better feature representation of the speech signal.

Chapter 4

Proposed Models

4.1 Overview

Currently, many pieces of research have shown transformer's outstanding qualities in learning relevant representations associated with the SER task [38, 27, 36]. However, there are still several problems to deal with.

The first problem is that a normal transformer is only able to process the uni-source input. Considering its peculiarity, transformer can be modified for interacting with bi-source input. We call this cross-attention transformer (CAT), which calculates queries from one source, and keys and values from another source. For better differentiation, the ordinary transformer model is aliased as self-attention transformer (SAT) in the rest of the paper. The detailed definition of CAT and SAT will be introduced in Section 4.2.

The second problem is that the interaction between transformer and other deep learning structures — CNN and LSTM is still needed to be investigated.

1. Although the appearance of transformer surpasses LSTM in most aspects, the lost location information can only be compensated by constant coding, i.e. positional encoding (PE) mentioned in Section 3.4. Since LSTM has a strong advantage for position relation, it is reasonable to use it in place of PE. Towards this direction, we propose an SER system called integration of LSTM with SAT (ILSAT), which will be mentioned in Section 4.3.1.
2. CNN and LSTM are skilled at frequency modeling and temporal modeling respectively and two sources can be exactly obtained from them, so we utilize CAT to interact and combine information from CNN and LSTM. The corresponding SER system is named CAT for CNN and LSTM sources (CL-CAT), which is about to be introduced in Section 4.3.2.

The third problem is that there is often only one kind of input feature in an SER system (in the case of our previous system, we choose log mel spectrogram), which may cause limited knowledge. Therefore, we attempt to alleviate this problem in the following directions:

1. Besides log mel spectrogram (abbreviated as S), add MFCC (abbreviated as M) and raw waveform data (abbreviated as W) as supplementary acoustic features, i.e. multiple acoustic features. Based on this, a novel SER system named SMW_CAT is presented, which will be explained in Section 4.4.
2. Besides audio features (abbreviated as A), add text (abbreviated as T) as additional modal features, i.e. multimodal features. Based on this, another new SER system named AT_CAT-SAT is presented, which will be explained in Section 4.5.

4.2 Cross-attention transformer (CAT) and self-attention transformer (SAT)

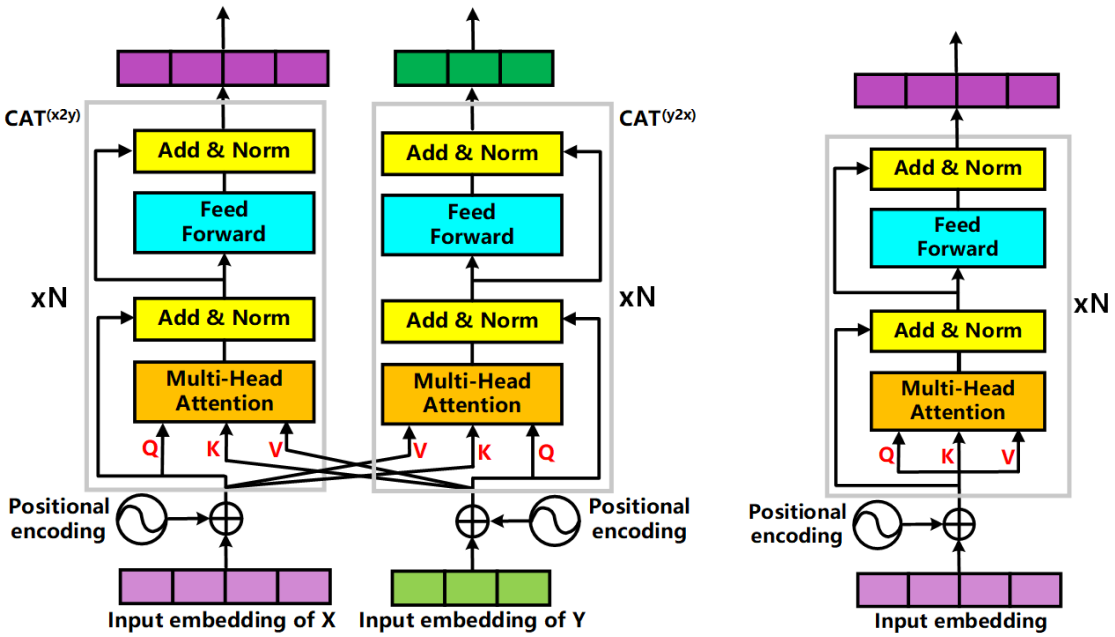
The purpose of cross-attention transformer (CAT) is to find interactive information from two different sources and associate the relevant and helpful part for emotion recognition in both sources. The word "cross attention" represents transmitting information from one source to another source according to the self-attention mechanism in transformer. Depicted in Figure 4.1 (a), considering two different sources X and Y with input embeddings $\mathbf{H}^{(x)} \in \mathbb{R}^{T_x \times f_x}$ and $\mathbf{H}^{(y)} \in \mathbb{R}^{T_y \times f_y}$ respectively, we can obtain queries, keys, values $\mathbf{Q}^{(x)} \in \mathbb{R}^{T_x \times d_k}, \mathbf{K}^{(x)} \in \mathbb{R}^{T_x \times d_k}, \mathbf{V}^{(x)} \in \mathbb{R}^{T_x \times d_v}, \mathbf{Q}^{(y)} \in \mathbb{R}^{T_y \times d_k}, \mathbf{K}^{(y)} \in \mathbb{R}^{T_y \times d_k}, \mathbf{V}^{(y)} \in \mathbb{R}^{T_y \times d_v}$ from them through relevant mapping matrices $\mathbf{W}^{Q(x)} \in \mathbb{R}^{f_x \times d_k}, \mathbf{W}^{K(x)} \in \mathbb{R}^{f_x \times d_k}, \mathbf{W}^{V(x)} \in \mathbb{R}^{f_x \times d_v}, \mathbf{W}^{Q(y)} \in \mathbb{R}^{f_y \times d_k}, \mathbf{W}^{K(y)} \in \mathbb{R}^{f_y \times d_k}, \mathbf{W}^{V(y)} \in \mathbb{R}^{f_y \times d_v}$:

$$\mathbf{Q}^{(x)} = \mathbf{H}^{(x)}\mathbf{W}^{Q(x)}, \qquad \mathbf{Q}^{(y)} = \mathbf{H}^{(y)}\mathbf{W}^{Q(y)} \tag{4.1}$$

$$\mathbf{K}^{(x)} = \mathbf{H}^{(x)}\mathbf{W}^{K(x)}, \qquad \mathbf{K}^{(y)} = \mathbf{H}^{(y)}\mathbf{W}^{K(y)} \tag{4.2}$$

$$\mathbf{V}^{(x)} = \mathbf{H}^{(x)}\mathbf{W}^{V(x)}, \qquad \mathbf{V}^{(y)} = \mathbf{H}^{(y)}\mathbf{W}^{V(y)} \tag{4.3}$$

When $\mathbf{Q}^{(x)}, \mathbf{K}^{(y)}, \mathbf{V}^{(y)}$ are fed into a transformer, it is responsible for learning the latent representation from source X to source Y, denoted as $\mathrm{CAT}^{(x2y)}$:



(a) Cross-attention transformer (CAT) (b) Self-attention transformer (SAT)

Fig. 4.1 Block diagram of CAT and SAT.

$$\mathbf{O}_{MHA}^{(x2y)} = \text{MHA}(\mathbf{Q}^{(x)}, \mathbf{K}^{(y)}, \mathbf{V}^{(y)}) \tag{4.4}$$

$$\mathbf{O}_{mid}^{(x2y)} = \text{LayerNorm}(\mathbf{H}^{(x)} + \mathbf{O}_{MHA}^{(x2y)}) \tag{4.5}$$

$$\mathbf{O}_{final}^{(x2y)} = \text{LayerNorm}(\mathbf{O}_{mid}^{(x2y)} + \text{FFN}(\mathbf{O}_{mid}^{(x2y)})) \tag{4.6}$$

Similarly, when $\mathbf{Q}^{(y)}, \mathbf{K}^{(x)}, \mathbf{V}^{(x)}$ are fed into another transformer, it is responsible for learning the latent representation from source Y to source X, denoted as $\text{CAT}^{(y2x)}$:

$$\mathbf{O}_{MHA}^{(y2x)} = \text{MHA}(\mathbf{Q}^{(y)}, \mathbf{K}^{(x)}, \mathbf{V}^{(x)}) \tag{4.7}$$

$$\mathbf{O}_{mid}^{(y2x)} = \text{LayerNorm}(\mathbf{H}^{(y)} + \mathbf{O}_{MHA}^{(y2x)}) \tag{4.8}$$

$$\mathbf{O}_{final}^{(y2x)} = \text{LayerNorm}(\mathbf{O}_{mid}^{(y2x)} + \text{FFN}(\mathbf{O}_{mid}^{(y2x)})) \tag{4.9}$$

In summary, the whole CAT can be formulated as:

$$\mathbf{O}_{final}^{(x2y)}, \mathbf{O}_{final}^{(y2x)} = \text{CAT}(\mathbf{H}^{(x)}, \mathbf{H}^{(y)}) \tag{4.10}$$

On the other hand, in the normal transformer model, queries, keys, and values are gained from the same source, for better differentiation, it will be aliased as self-attention transformer (SAT) in the rest of the paper. The architecture of SAT is shown in Figure 4.1 (b).

4.3 Fusion between transformer, CNN and LSTM

In this section, two approaches to utilizing network-based structures aggregated with transformer are proposed. We first integrate LSTM with SAT (ILSAT), trying to replace the function of PE in the transformer with LSTM. Then we use CAT to combine the information obtained from CNN and LSTM (CL-CAT).

4.3.1 Integration of LSTM with SAT (ILSAT)

As the PE in SAT is just a fixed positional representation of input features, initially we replaced it by inserting LSTM between CNN and SAT. However, in our preliminary experiment, we found this approach cannot work well. Therefore we propose to use a parallel combination of LSTM and SAT instead of their cascaded structure, illustrated in Figure 4.2 (a). Suppose that the input features (log mel spectrogram) are represented as a sequence of vectors $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times f}$, where N is the number of frames. Because of the image-like character of log mel spectrogram, a designed CNN structure called Light CNN is used to convert it into a higher-level abstraction. The detailed architecture of light CNN will be introduced in Section 5.2.2. After passing through CNN, the output is flattened and reshaped into $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{N'}]^T \in \mathbb{R}^{N' \times f'}$ and then split into two flows to the subsequent models respectively. The first flow is to LSTM, which is applied to make good use of the contextual information. We utilize the bidirectional LSTM (BiLSTM), whose output at the current time step can be both learned from the previous and next states. Also, the LSTM output can be regarded as additional position information. After obtaining the sequential representation $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \cdots, \mathbf{l}_{N'}]^T$ from LSTM, we take average of them for eliminating the time dimension, resulting $\mathbf{O}_{LSTM} \in \mathbb{R}^{f'}$. The second flow is to SAT, in which the PE will be opened or closed to verify how much position information can be obtained from BiLSTM. For the output from this flow, an average operation is conducted for temporal aggregation, resulting $\mathbf{O}_{SAT} \in \mathbb{R}^{f'}$. After that, two outputs are integrated together. We try to utilize the following three strategies for fusion:

concatnation

$$\mathbf{O}_{final} = [\mathbf{O}_{LSTM}; \mathbf{O}_{SAT}] \tag{4.11}$$

plus

$$\mathbf{O}_{final} = \mathbf{O}_{LSTM} + \mathbf{O}_{SAT} \tag{4.12}$$

trainable plus

$$\begin{aligned} \mathbf{O}_{final} &= \alpha_A \mathbf{O}_{LSTM} + \alpha_B \mathbf{O}_{SAT} \\ \alpha_A &= \frac{\exp(\mathbf{W}_A)}{\exp(\mathbf{W}_A) + \exp(\mathbf{W}_B)} \\ \alpha_B &= \frac{\exp(\mathbf{W}_B)}{\exp(\mathbf{W}_A) + \exp(\mathbf{W}_B)} \end{aligned} \tag{4.13}$$

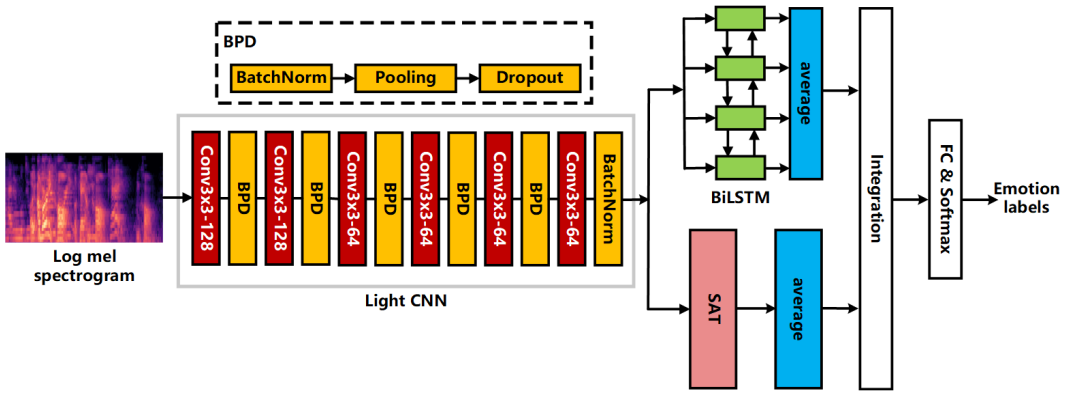
where $\mathbf{W}_A$ and $\mathbf{W}_B$ are trainable weight matrices.

4.3.2 CAT for CNN and LSTM sources (CL-CAT)

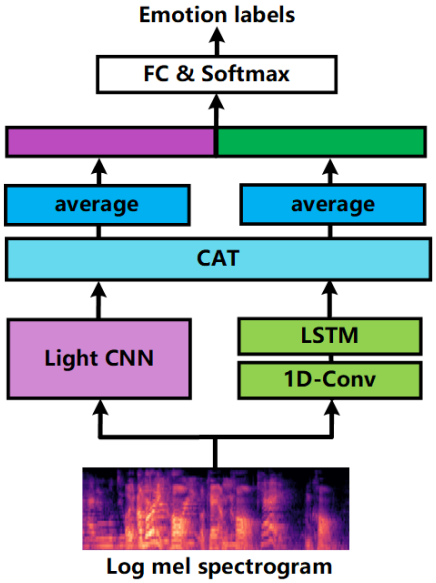
In Section 3.2.2, it is said that CNN and LSTM are complementary in their modeling capabilities. Inspired by this, we propose to utilize CAT to joint-encoding these two sources. Depicted in Figure 4.2 (b), the input features are fed into CNN and LSTM simultaneously. Considering that a tremendous length of input features will affect the function of LSTM and bring more parameters for training, we add an 1D temporal convolutional layer before LSTM not only for ensuring that each element of the input sequence has sufficient awareness of its neighborhood elements, but also for reducing the length of it. Let the source from CNN and LSTM be $\mathbf{H}^{(C)} = [\mathbf{h}_1^{(C)}, \mathbf{h}_2^{(C)}, \cdots, \mathbf{h}_J^{(C)}]^T$ and $\mathbf{H}^{(L)} = [\mathbf{h}_1^{(L)}, \mathbf{h}_2^{(L)}, \cdots, \mathbf{h}_K^{(L)}]^T$ respectively. According to CAT introduced in Section 4.2, we can obtain $\text{CAT}^{(c2l)}$ by calculating queries from $\mathbf{H}^{(C)}$, keys and values from $\mathbf{H}^{(L)}$, and $\text{CAT}^{(l2c)}$ by calculating queries from $\mathbf{H}^{(L)}$, keys and values from $\mathbf{H}^{(C)}$. Similar as ILSAT, outputs from $\text{CAT}^{(c2l)}$ and $\text{CAT}^{(l2c)}$ are averaged over time respectively. Eventually, they are concatenated together to and pass through the fully connected layer and softmax layer to obtain the final posterior probabilities of each emotion.

4.4 CAT with multiple acoustic features

Up to now, approaches to the fusion between transformer, CNN, and LSTM has been done, but they still exist a critical issue: log mel spectrogram is selected as the only input feature, which may cause limited acoustic knowledge. Towards this problem, in this section, we attempt to append two other acoustic features — raw waveform data and MFCC. Since CAT is suitable for joint-encoding outputs from different sources, it can be used for uniting these features. Hence



(a) Integrate LSTM with SAT (ILSAT)



(b) CAT for CNN and LSTM sources (CL-CAT)

Fig. 4.2 Block diagram of ILSAT and CL-CAT.

we propose a novel SER system named SMW_CAT, as shown in Figure 4.3, where S, M, W are the abbreviation of log mel spectrogram, MFCC, and raw waveform data.

Encoder module

Given an utterance $\mathbf{X}^{(w)} \in \mathbb{R}^{T^{(w)} \times 1}$, we can calculate the log mel spectrogram $\mathbf{X}^{(s)} \in \mathbb{R}^{T^{(s)} \times f^{(s)}}$ and MFCC $\mathbf{X}^{(m)} \in \mathbb{R}^{T^{(m)} \times f^{(m)}}$ according to Section 2.3.2.

For raw waveform data, a pre-trained wav2vec2 [2] model on the Automatic Speech Recognition (ASR) task are adopted, which can be viewed as a case of transfer learning.

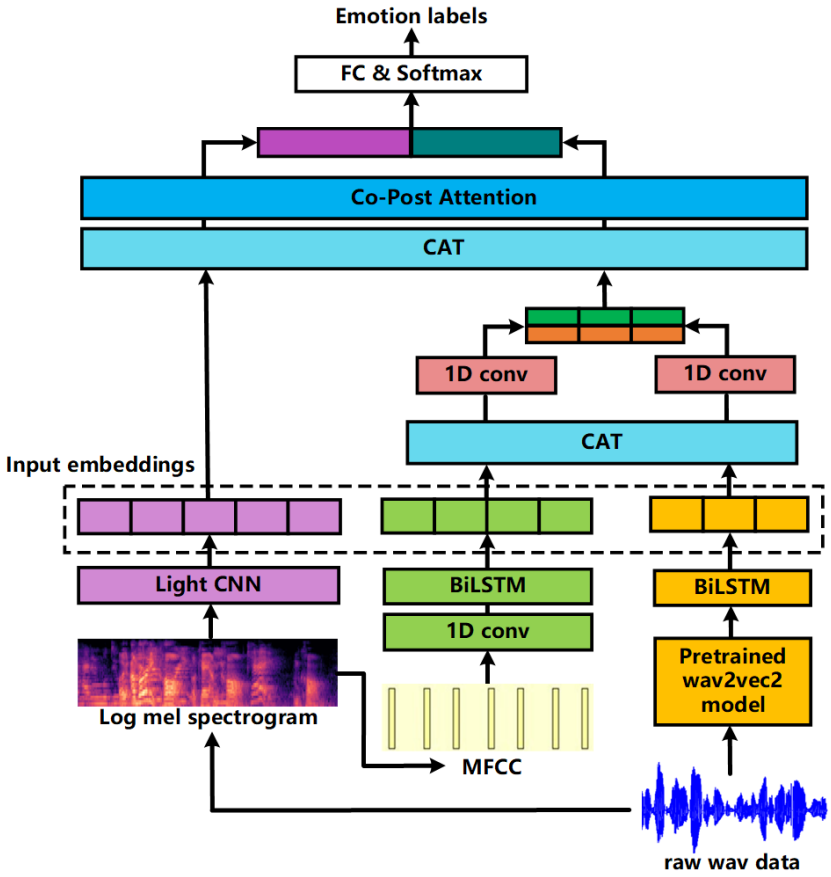


Fig. 4.3 Block Diagram of SMW_CAT.

Then BiLSTM layer is exploited for capturing the temporal information within the sequence, resulting input embedding $\mathbf{H}^{(w)} = [\mathbf{h}_1^{(w)}, \mathbf{h}_2^{(w)}, \cdots, \mathbf{h}_{T^{(w)\prime}}^{(w)}]^T \in \mathbb{R}^{T^{(w)\prime} \times f}$:

$$\mathbf{X}^{(w)\prime} = \text{Wav2vec2}(\mathbf{X}^{(w)}) \tag{4.14}$$

$$\mathbf{h}_t^{(w)} = [\overrightarrow{\mathbf{h}_t^{(w)}}; \overleftarrow{\mathbf{h}_t^{(w)}}] = [\overrightarrow{\text{LSTM}}(\mathbf{h}_{t-1}^{(w)}, \mathbf{x}_t^{(w)\prime}); \overleftarrow{\text{LSTM}}(\mathbf{h}_{t+1}^{(w)}, \mathbf{x}_t^{(w)\prime})],\ t \in 1, \cdots, T^{(w)\prime} \tag{4.15}$$

The log mel spectrogram is fed into Light CNN for embedding extraction:

$$\mathbf{H}^{(s)} = \text{LightCNN}(\mathbf{X}^{(s)}) \tag{4.16}$$

where $\mathbf{H}^{(s)} \in \mathbb{R}^{T^{(s)\prime} \times f}$

In the case of MFCC, we first apply a 1D temporal convolutional layer for capturing local patterns and reducing length in time dimension, and then use BiLSTM layer to obtain the final embedding $\mathbf{H}^{(m)} = [\mathbf{h}_1^{(m)}, \mathbf{h}_2^{(m)}, \cdots, \mathbf{h}_{T^{(m)\prime}}^{(m)}]^T \in \mathbb{R}^{T^{(m)\prime} \times f}$:

$$\mathbf{X}^{(m)\prime} = \text{1DConv}(\mathbf{X}^{(m)}) \tag{4.17}$$

$$\mathbf{h}_t^{(m)} = [\overrightarrow{\mathbf{h}_t^{(m)}}; \overleftarrow{\mathbf{h}_t^{(m)}}] = [\overrightarrow{\text{LSTM}}(\mathbf{h}_{t-1}^{(m)}, \mathbf{x}_t^{(m)\prime}); \overleftarrow{\text{LSTM}}(\mathbf{h}_{t+1}^{(m)}, \mathbf{x}_t^{(m)\prime})],\ t \in 1, \cdots, T^{(m)\prime} \tag{4.18}$$

CAT based fusion module

In our preliminary experiment, among three features, input embedding of raw waveform data achieved the highest accuracy, followed by MFCC, and log mel spectrogram was the lowest one. Consequently, we propose to fuse from the best to the worst, i.e. utilize the fusion result from the top two to increase the generalization performance of the last. Specifically, at first the input embeddings of raw waveform data $\mathbf{H}^{(w)}$ are combined with MFCC $\mathbf{H}^{(m)}$ based on CAT, then two outputs are normalized into the same length by 1D temporal convolutional layer and concatenated together to be ready for the next fusion.

$$\mathbf{H}^{(m2w)}, \mathbf{H}^{(w2m)} = \text{CAT}(\mathbf{H}^{(m)}, \mathbf{H}^{(w)}) \tag{4.19}$$

$$\mathbf{H}^{(mw)} = [\text{1DConv}(\mathbf{H}^{(m2w)}); \text{1DConv}(\mathbf{H}^{(w2m)})] \tag{4.20}$$

where $\mathbf{H}^{(mw)} \in \mathbb{R}^{T^{(mw)} \times f}$

Next is to adopt CAT to fuse $\mathbf{H}^{(mw)}$ with the input embeddings of log mel spectrogram. The two outputs are further interacted by two attention layers mentioned in Section 3.3.1 but share the same projection matrix $\mathbf{W}^{(smw)} \in \mathbb{R}^{f \times 1}$, which can be regarded as a co-post attention layer relative to the pre attention layer composed by CAT. Finally, we can get the representation $\mathbf{O}^{(smw)} \in \mathbb{R}^{2f}$ before fully connected layer.

$$\mathbf{H}^{(s2mw)}, \mathbf{H}^{(mw2s)} = \text{CAT}(\mathbf{H}^{(s)}, \mathbf{H}^{(mw)}) \tag{4.21}$$

$$\begin{aligned}
\mathbf{O}^{(s2mw)}, \mathbf{O}^{(mw2s)} &= \text{Co-PostAttention}(\mathbf{H}^{(s2mw)}, \mathbf{H}^{(mw2s)}, \mathbf{W}^{(smw)}) \\
&= \sum_{t=1}^{T^{(s)\prime}} \left(\frac{\exp(\mathbf{h}_t^{(s2mw)} \mathbf{W}^{(smw)})}{\sum_{\tau=1}^{T^{(s)\prime}} \exp(\mathbf{h}_\tau^{(s2mw)} \mathbf{W}^{(smw)})} \right) \mathbf{h}_t^{(s2mw)}, \\
&\quad \sum_{t=1}^{T^{(mw)}} \left(\frac{\exp(\mathbf{h}_t^{(mw2s)} \mathbf{W}^{(smw)})}{\sum_{\tau=1}^{T^{(mw)}} \exp(\mathbf{h}_\tau^{(mw2s)} \mathbf{W}^{(smw)})} \right) \mathbf{h}_t^{(mw2s)}
\end{aligned} \tag{4.22}$$

$$\mathbf{O}^{(smw)} = [\mathbf{O}^{(s2mw)}; \mathbf{O}^{(mw2s)}] \tag{4.23}$$

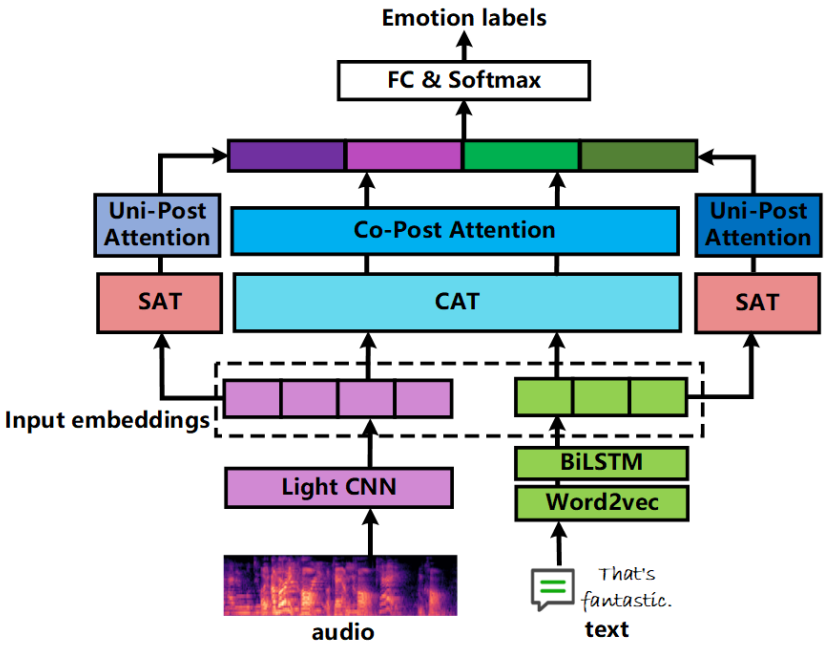


Fig. 4.4 Block Diagram of AT_CAT-SAT.

4.5 CAT with multimodal features

Usually, emotion is expressed in a multimodal way, including speech, text, facial expression, and so on. Considering an utterance "I really appreciate your kindness" labeled happiness but spoke in a calm tone, it will be predicted as neutral if speech is the only exploited modal. By joining the text modal, the word "appreciate" and "kindness" can be greatly helpful for recognizing the utterance correctly, i.e. in this case, emotion is almost hidden in the linguistic contents. Hence, it is evident that the performance of an SER system will be improved by taking multimodal features as input. Besides, when fusing multimodal information, two types of interaction should be thought over: the intra-modal interaction and the inter-modal interaction [45]. Due to the characteristics of SAT and CAT, the former can be responsible for capturing intra-modal dynamics, while the latter can learn inter-modal relations. Based on this theory, we propose another new SER system called AT_CAT-SAT for two kinds of multimodal input features — audio and text, as shown in Figure 4.4, where A, T are the abbreviation of audio and text.

Encoder module

Support that the input acoustic and textual features are denoted as $\mathbf{X}^{(a)} \in \mathbb{R}^{T^{(a)} \times f^{(a)}}$ and $\mathbf{X}^{(t)} \in \mathbb{R}^{T^{(t)} \times f^{(t)}}$, respectively. In our experiment, we use log mel spectrogram as acoustic

features. The textual features are just a preprocessed sequence of words split according to the blank space in the transcription of the utterance.

For audio data, the encoder for log mel spectrogram in Section 4.4 is conducted again, i.e.

$$\mathbf{H}^{(a)} = \text{LightCNN}(\mathbf{X}^{(a)}) \tag{4.24}$$

where $\mathbf{H}^{(a)} \in \mathbb{R}^{T^{(a)\prime} \times f}$

For text data, word2vec [25] is used to embed each word, then BiLSTM layer is applied to the embedded words, resulting input embedding $\mathbf{H}^{(t)} = [\mathbf{h}_1^{(t)}, \mathbf{h}_2^{(t)}, \cdots, \mathbf{h}_{T^{(t)\prime}}^{(t)}]^T \in \mathbb{R}^{T^{(t)\prime} \times f}$

$$\mathbf{X}^{(t)\prime} = \text{Word2vec}(\mathbf{X}^{(t)}) \tag{4.25}$$

$$\mathbf{h}_t^{(t)} = [\overrightarrow{\mathbf{h}_t^{(t)}}; \overleftarrow{\mathbf{h}_t^{(t)}}] = [\overrightarrow{\text{LSTM}}(\mathbf{h}_{t-1}^{(t)}, \mathbf{x}_t^{(t)\prime}); \overleftarrow{\text{LSTM}}(\mathbf{h}_{t+1}^{(t)}, \mathbf{x}_t^{(t)\prime})],\ t \in 1, \cdots, T^{(t)\prime} \tag{4.26}$$

CAT based fusion module

From the input embeddings of audio $\mathbf{H}^{(a)}$ and text $\mathbf{H}^{(t)}$, we can capture the inter-modal interaction by CAT and the intra-modal interactions by SAT. After that, the two outputs from CAT are further interacted by the co-post attention layer, and the two outputs from SAT are further interacted by the uni-post attention layer respectively. Eventually, representation $\mathbf{O}^{(at)} \in \mathbb{R}^{4f}$ before fully connected layer can be obtained.

$$\mathbf{H}^{(a2t)}, \mathbf{H}^{(t2a)} = \text{CAT}(\mathbf{H}^{(a)}, \mathbf{H}^{(t)}) \tag{4.27}$$

$$\mathbf{H}^{(a)\prime} = \text{SAT}(\mathbf{H}^{(a)}), \mathbf{H}^{(t)\prime} = \text{SAT}(\mathbf{H}^{(t)}) \tag{4.28}$$

$$\mathbf{O}^{(a2t)}, \mathbf{O}^{(t2a)} = \text{Co-PostAttention}(\mathbf{H}^{(a2t)}, \mathbf{H}^{(t2a)}, \mathbf{W}^{(at)}) \tag{4.29}$$

$$\begin{aligned} \mathbf{O}^{(a)\prime} &= \text{Uni-PostAttention}(\mathbf{H}^{(a)}, \mathbf{W}^{(a)}) \\ &= \sum_{t=1}^{T^{(a)\prime}} \left(\frac{\exp(\mathbf{h}_t^{(a)\prime} \mathbf{W}^{(a)})}{\sum_{\tau=1}^{T^{(a)\prime}} \exp(\mathbf{h}_\tau^{(a)\prime} \mathbf{W}^{(a)})} \right) \mathbf{h}_t^{(a)\prime} \end{aligned} \tag{4.30}$$

$$\mathbf{O}^{(t)\prime} = \text{Uni-PostAttention}(\mathbf{H}^{(t)}, \mathbf{W}^{(t)}) \tag{4.31}$$

$$\mathbf{O}^{(at)} = [\mathbf{O}^{(a2t)}; \mathbf{O}^{(t2a)}; \mathbf{O}^{(a)\prime}; \mathbf{O}^{(t)\prime}] \tag{4.32}$$

where $\mathbf{W}^{(at)}, \mathbf{W}^{(a)}, \mathbf{W}^{(t)} \in \mathbb{R}^{f \times 1}$.

Chapter 5

Experiments and Analysis

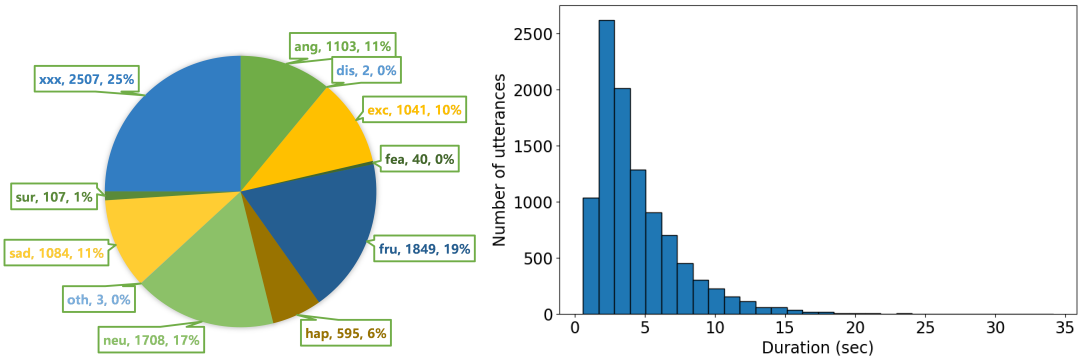
5.1 Overview

To evaluate the performance of our proposed models, we perform experiments on the IEMOCAP benchmark dataset. In this chapter, the experimental setup is firstly described. And then, the experimental results and analysis of the proposed models are presented.

5.2 Experimental setup

5.2.1 The IEMOCAP dataset

We evaluate all of our approaches on the interactive emotional dyadic motion capture (IEMOCAP) dataset [5], which is a standard benchmark dataset widely used for SER. It contains approximately 12 hours of speech. There are 10 actors (5 males and 5 females) to perform 5 dyadic sessions and two actors are grouped in a single session. All the conversations are separated into small utterances, labeled with the following emotions by at least three different annotators: *anger*, *disgust*, *excitement*, *fear*, *frustration*, *happiness*, *neutral*, *sadness*, *surprise*, and *other*, which have been evaluated by at least three different annotators. In our experiments, ground truth label are obtained by majority voting resulting the distribution shown in Figure 5.1 (a). To stay consistent with most previous researches on IEMOCAP, only the following five emotions were selected: *anger*, *happiness*, *excitement*, *sadness*, and *neutral*. Then *happiness* and *excitement* are merged into single happiness category because of the similarity between them. Therefore, totally 5531 utterances are exploited including 1103 *anger*, 1636 *happiness*, 1708 *sadness*, and 1084 *neutral*. We perform 10-fold leave-one-speaker-out cross-validation strategy, i.e. in each validation, the total dataset is split into 8:1:1 training set, validation set, and



(a) Distribution of emotion labels (b) Distribution of utterance durations

Fig. 5.1 IEMOCAP [5] overview (*ang*, anger state; *dis*, disgust; *exc*, excitement; *fea*, fear; *fru*, frustration; *hap*, happiness; *neu*, neutral; *oth*, other; *sad*, sadness; *sur*, surprise; and *xxx*, emotions that cannot be acquired by majority voting).

test set according to the difference of speaker, respectively. The final performance is calculated by taking the average of all test sets.

5.2.2 Implementation details

Due to the length distribution of IEMOCAP depicted in Figure 5.1 (b) and the input of our systems, utterances are sampled at 16*k*Hz, and then two splitting strategies are applied. For experiments on ILSAT, CL-CAT, and SMW_CAT, we spilt each utterance into segments with a length of 3 seconds. For experiments on AT_CAT-SAT, each utterance is normalized to 7.5 seconds. We use the modified utterances to obtain the log mel spectrograms and MFCC with the window size of 40 ms and the frame size of 10 ms as input features to our model. The number of filter banks in the log mel spectrogram is set to 128, and the number of MFCC is set to 40. For textual features, the off-the-shelf transcription in IEMOCAP is used, i.e. the word recognition error rate is regarded as zero. We first employ a tokenizer to split each utterance according to the blank space. Then these words are embedded into a 300-dimensional vector using the word2vec model. We assume a maximum of 30 words in each utterance.

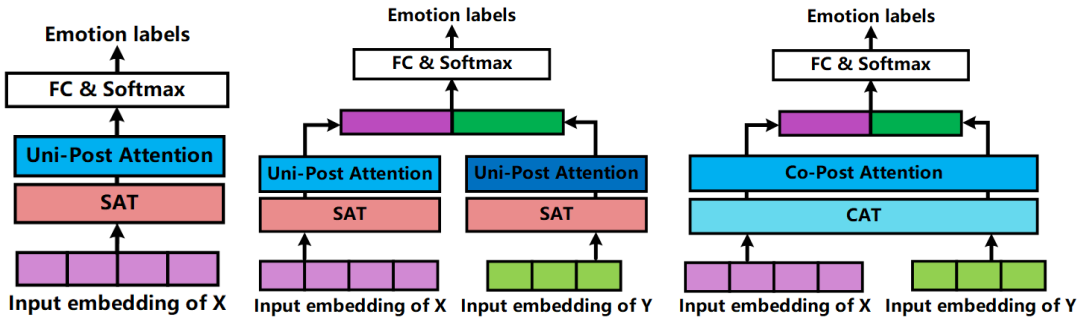
We implement our model within the PyTorch framework. The detailed parameters of light CNN are shown in Table 5.1. We use BiLSTMs with 64 hidden units followed by the dropout layer with 0.5 dropout probability. For transformer models, all of them have 64 embedding dimensions and 4 attention heads. The final fully connected layers consist of three linear layers with output dimensions 256, 256, and 4, respectively. To train the model, we choose cross-entropy loss as the loss function, and Adam with an initial learning rate of 0.00001 and 0.001 in the case with and without the pre-trained wav2vec2 model as optimizer. The learning rate decreases by one-tenth every 10 epochs. The models are trained for at most 100

Table 5.1 Layer parameters of the light CNN model. T denotes the time dimension, F denotes the frequency dimension, C is the number of input channels. Note that the BatchNorm and Dropout layers are omitted in this table.

Layers	Kernel size	Stride	Output shape
input	–	–	$T \times F \times C$
Convolution (1^{st})	3×3	1×1	$T \times F \times 128$
Pooling (1^{st})	2×2	2×2	$T/2 \times F/2 \times 128$
Convolution (2^{nd})	3×3	1×1	$T/2 \times F/2 \times 128$
Pooling (2^{nd})	2×2	2×2	$T/4 \times F/4 \times 128$
Convolution (3^{rd})	3×3	1×1	$T/4 \times F/4 \times 64$
Pooling (3^{rd})	2×2	2×2	$T/8 \times F/8 \times 64$
Convolution (4^{th})	3×3	1×1	$T/8 \times F/8 \times 64$
Pooling (4^{th})	2×2	2×2	$T/16 \times F/16 \times 64$
Convolution (5^{th})	3×3	1×1	$T/16 \times F/16 \times 64$
Pooling (5^{th})	2×2	2×2	$T/32 \times F/32 \times 64$
Convolution (6^{th})	3×3	1×1	$T/32 \times F/32 \times 64$

epochs with a batch size of 32. Besides, all the hyper parameters of our model are fine-tuned to maximize the sum of WA and UA to be introduced in the next section.

To confirm the dominance of our proposals, we design several baseline systems. For experiments on ILSAT and CL-CAT to be described in Section 5.3, we test the CNN connected with SAT sequentially (C-SAT), which is regarded as Figure 4.2 (b) with LSTM removed. Moreover, we insert LSTM into the middle of our baseline system to be utilized as another baseline system (C-L-SAT). For experiments on SMW_CAT and AT_CAT-SAT which will be discussed in Section 5.4 and Section 5.5, three types of baseline systems are designed: X_SAT, XY_SAT, and XY_CAT, as depicted in Figure 5.2. X_SAT uses SAT to process input embeddings from source X and applies an attention layer as uni-post attention to eliminate the time dimension. XY_SAT can be regarded as the simple combination of the output from X_SAT and Y_SAT. In XY_CAT, the interactive information from two different sources X and Y are captured by CAT, and then two outputs are further integrated by the co-post attention layer sharing weight, as mentioned in Section 4.4.



(a) X_SAT (b) XY_SAT (c) XY_CAT

Fig. 5.2 Block diagram of three baseline SER systems, where X and Y are two different sources.

5.2.3 Evaluation metrics

With respect to the evaluation metrics for measuring the model performance, confusion matrix, weighted accuracy (WA), and unweighted accuracy (UA) are employed in the experiments. Confusion matrix depicts how each class is classified and misclassified. WA (also known as overall accuracy) weighs each class according to the number of samples that belong to that class in the dataset, it represents the classification accuracy over the entire dataset. UA gives the same weight to each class, regardless of how many samples of that class the dataset contains, it can better reflect imbalance among classes. The distinction between these two measures is useful especially if there exist emotion classes that are under-represented by the samples. Suppose there are N emotion categories $c_1, c_2, \cdots, c_N$ in total, where number of samples and correctly predicted samples in each category is denoted as $|c_1|, |c_2|, \cdots, |c_N|$ and $TP_{c_1}, TP_{c_2}, \cdots, TP_{c_N}$ respectively. Based on this, WA and UA can be calculated as:

$$WA = \frac{\sum_{i=1}^{N} TP_{c_i}}{\sum_{i=1}^{N} |c_i|} \tag{5.1}$$

$$UA = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_{c_i}}{|c_i|} \tag{5.2}$$

5.3 Experiments on ILSAT and CL-CAT

Table 5.2 summarizes the property of ILSAT and our baselines in different conditions. In all systems, opening the PE achieves higher performance than closing, verifying the usefulness of it. However, far from obtaining a better behavior, simply inserting LSTM between CNN and SAT brings about a little bit worse accuracy. Among the three fusion ways tested (concatenation, plus, and trainable plus) in ILSAT, the first one shows the best result. In this case, the ILSAT

Table 5.2 Accuracy comparison among two baseline systems and our proposed ILSAT. PE represents positional encoding in the transformer.

Model	Fusion Mode	PE	WA(%)	UA(%)
C-SAT (baseline1)	–	False	65.89	65.60
		True	**66.29**	**65.54**
C-L-SAT (baseline2)	–	False	64.71	64.41
		True	66.35	64.82
ILSAT	concatenation	False	67.12	67.29
		True	**68.01**	**67.49**
	plus	False	65.02	64.62
		True	65.01	64.84
	trainable plus	False	64.87	64.42
		True	65.87	65.41

without PE exceeds the C-SAT with PE. This reveals that, by putting LSTM in the same position as SAT and concatenating them together, it has the ability to replace the function of PE, and we can focus more adequately on emotional parts of input utterances. Besides, when LSTM and PE exist at the same time, the best performance can be obtained, which exceeds the baseline models with 1.72% and 1.95% absolute improvements in WA and UA, respectively.

Depicted in Table 5.3, our CL-CAT model attains 67.32% WA and 67.26% UA, which has an increase of 1.03% and 1.72% over the baseline C-SAT system. It clearly indicates the efficacy of joint learning of the information from CNN and LSTM. Meanwhile, we explore an ablation study to examine the contributions of different parts in the CL-CAT. We first eliminate the CAT module, showing that it can bring 1.81% WA improvement and 1.27% UA improvement. By dropping the $\text{CAT}^{(l2c)}$ and $\text{CAT}^{(c2l)}$ in IL-CAT, a decrease of 0.79%/1.94% WA and 0.39%/2.34% UA can be observed respectively. This shows the importance of these two modules in CAT. In addition, the performance becomes worse in the absence of $\text{CAT}^{(c2l)}$, indicating it is more significant than $\text{CAT}^{(l2c)}$. Comparing the second row, the third row, and the last two rows in the table, we can verify that mapping into a different model can certainly bring elevation of performance.

Table 5.3 Ablation study on the CL-CAT.

Model	WA(%)	UA(%)
CL-CAT	**67.32**	**67.26**
C-SAT (baseline)	**66.29**	**65.54**
L-SAT	64.64	64.81
CL-CAT (w/o CAT)	65.51	65.64
CL-CAT (w/o CAT$^{(l2c)}$)	66.53	66.87
CL-CAT (w/o CAT$^{(c2l)}$)	65.38	64.92

5.4 Experiments on SMW_CAT

The experimental results of SMW_CAT and some comparing systems are summarized in Table 5.4. When exploiting only one kind of acoustic input feature, W_SAT achieves the highest accuracy, followed by M_SAT, and S_SAT is the lowest one. With MFCC as another additional feature, the performance of SM_SAT exceeded both S_SAT and M_SAT. The same conclusion can be drawn in the case of SW_SAT and MW_SAT, pointing to the effectiveness of XY_SAT. By replacing the architecture of XY_SAT with XY_CAT, further improvement in WA and UA can be observed. To better understand the contribution of different parts in XY_CAT, we select SW_CAT as an example and replace the CAT and co-post attention with the corresponding part in SW_SAT respectively. As can be seen from the results in the third and fourth rows from the bottom, both parts bring increasement and neither of them exceeds the performance of SW_CAT which exploits them concurrently. Utilizing all three features, our proposal SMW_CAT achieves the best performance of 73.80% in terms of WA and 74.25% in terms of UA.

To verify the effectiveness of the proposal, we further compare our system with other currently advanced approaches. Experimental results of different methods are listed in Table 5.5. SMW_CAT achieves the best WA and UA scores. It outperforms the best state-of-the-art method [47] by absolute 0.1%WA increase and absolute 0.35% UA increase. The result gives the credit to the excellent discrimination ability of CAT.

We also plot the confusion matrix and execute the t-SNE visualization on S_SAT, M_SAT, W_SAT, and SMW_CAT, depicted in Figure 5.3 and Figure 5.4. From W_SAT to S_SAT, *angry* is assembled worse and easier to overlap with other emotions, thus there exists a significant degradation of 19.6%. On the other side, *sad* reaches a higher probability of being correctly recognized due to the suppression in misclassifying with *neutral*. Considering when log mel spectrogram is calculated from raw waveform data the phase info will be lost, it can be

Table 5.4 Performance of our proposed SMW_CAT and several comparing systems. *S* denotes log mel spectrogram, *M* denotes MFCC, *W* denotes raw waveform data, *CPA* denotes co-post attention, *UPA* denotes uni-post attention.

Model	WA(%)	UA(%)
S_SAT	66.74	66.03
M_SAT	67.75	66.53
W_SAT	71.44	71.99
SM_SAT	68.31	68.15
SW_SAT	71.45	72.22
MW_SAT	72.93	73.78
SM_CAT	68.47	68.15
SW_CAT	72.43	73.55
SW_CAT (CAT $\rightarrow$ SAT)	71.68	72.45
SW_CAT (CPA $\rightarrow$ UPA)	71.99	72.97
MW_CAT	73.59	73.86
SMW_CAT	**73.80**	**74.25**

inferred that the absence of phase causes the above phenomena. From log mel spectrogram to MFCC, higher-order DCT components are ignored, which corresponds to the slight vibration in high quefrency. In M_SAT, *happy* achieves a performance improvement of 4.4% compared with S_SAT because it gathers into a cluster far from other emotions in t-SNE visualization. However, *sad* decreases to 66.9% as it becomes more scattered over the mapping space. In our proposed SMW_CAT, seen in Figure 5.3 (d) and Figure 5.4 (d), the performance looks like a complementarity and enhancement over S_SAT, M_SAT and W_SAT. In addition, the misclassification probability among *angry*, *happy* and *sad* is relatively low, meaning that the only bottleneck for our system is the confusion between *neutral* and them.

5.5 Experiments on AT_CAT-SAT

Table 5.6 shows the performance of our proposed AT_CAT-SAT and some comparing systems. From the first two rows in the table, SAT with audio-only achieved a better result than SAT with text-only, meaning that more emotion is implicated in acoustic features than in textual ones. When these two modalities are combined by SAT for capturing intra-modal dynamics and by CAT for learning inter-modal relations, corresponding to the third and fourth row in the

Table 5.5 Comparison of our proposed SMW_CAT and previous state-of-the-art approaches on the IEMOCAP dataset.

Model	WA(%)	UA(%)
CNN-BiLSTM [34]	68.8	59.4
Fusion-ConvBERT [19]	66.47	67.12
CNN_TF_Att.pooling [20]	71.75	68.06
Self-attention [39]	70.17	70.85
Co-attention [48]	71.64	72.70
MLT-Dnet [16]	-	73.01
GLAM [47]	73.70	73.90
SMW_CAT (proposed)	**73.80**	**74.25**

Table 5.6 Performance of our proposed AT_CAT-SAT and several comparing systems. *A* denotes audio, *T* denotes text.

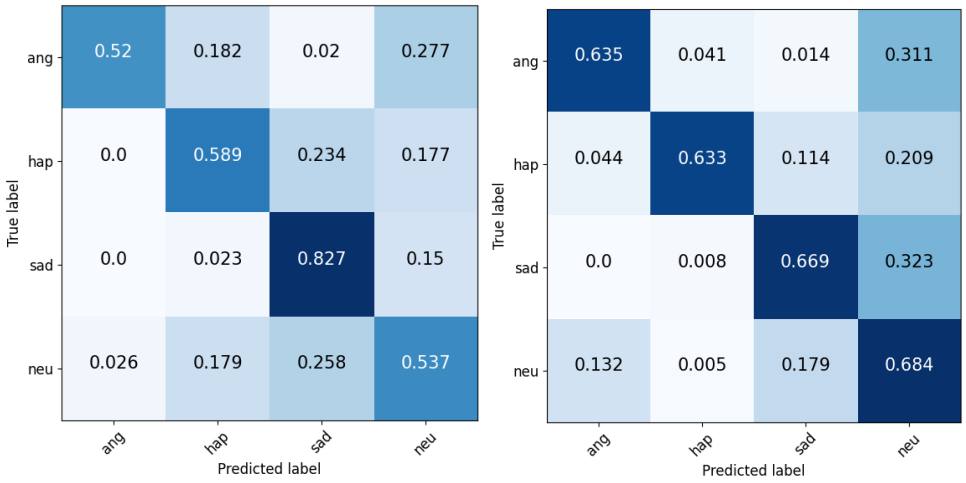
Model	WA(%)	UA(%)
A_SAT	65.26	65.96
T_SAT	61.48	62.81
AT_SAT	72.72	74.13
AT_CAT	73.22	74.50
AT_CAT-SAT	**73.64**	**75.05**

table respectively, performance has obviously improved compared with unimodality systems, i.e. A_SAT and T_SAT. It points out that both feature fusion strategies can make full use of the multimodal features and gain complementarity. Besides, AT_CAT has a 0.5% WA and 0.37% UA improvement than AT_SAT, indicating that more inter-modal interactions are modeled than intra-modal ones. Eventually, if we model these two interactions simultaneously, the best behavior of 73.64% WA and 75.05% UA can be obtained.

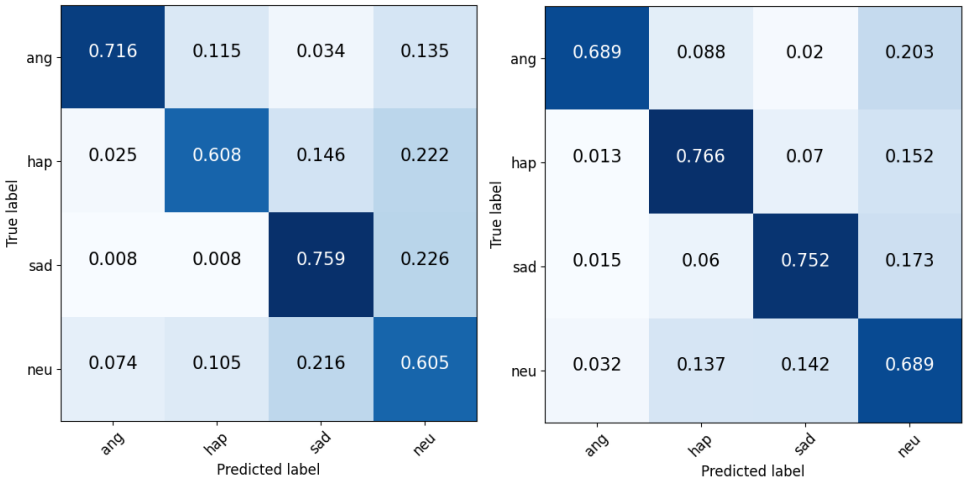
In detail, we compute confusion matrix to exploit accuracies for each emotional category, depicted in Figure 5.5. Except for *neutral*, AT_CAT has a higher correctly recognition probability in three other classes than both A_SAT and T_SAT. By introducing intra-modal interactions, AT_CAT-SAT restrains confusion sets of *angry* $\rightarrow$ *neutral* and *neutral* $\rightarrow$ *sad* so that 10.3% improvement on prediction of *angry* label and 4.1% improvement on prediction of *neutral* label can be observed. From the t-SNE visualization shown in Figure 5.6, data distribution of *happy* is more centralized, while *sad* and *neutral* is more scattered in T_SAT compared with A_SAT.

Moreover, AT_CAT is able to exploit advantages and suppress disadvantages from the two modalities, thus boundaries in the different emotions become clearer.

To understand the process of joint encoding information from audio and text more intuitively, we analyze the attention weights in CAT. Figure 5.7 (a) shows a rightly recognized *angry* utterance. We can see those words "vile", "ill-tempered", "wicked" that tend to hint the *angry* emotion have a higher attention. Among them, the word "wicked" takes the determined position. Besides, the word "wicked" is not only correlated to acoustic features when speaking itself but also stronger relations with acoustic features when speaking "You 're a vile" and "I never see you again" are detected, which cannot be learned by forced alignment. Another precisely predicted utterances labeled *happy* is depicted in Figure 5.7 (b). More glaring color is perceived in the words "That's amazing" and "Wow", from which the underlying *happy* emotion can be easily conjectured. While the words "That's amazing" has an association with acoustic features when speaking itself and "Wow", the word "Wow" only concentrated on acoustic features when speaking "That's amazing". These observations again indicate the superiority of alignment automatically obtained by CAT. However, for the words "what are going to", the uncanny focus is given, especially for the word "what". This eccentric phenomenon should be further researched.

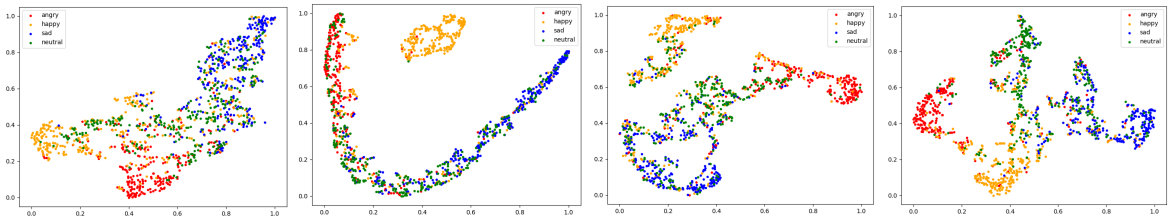


(a) S_SAT (b) M_SAT



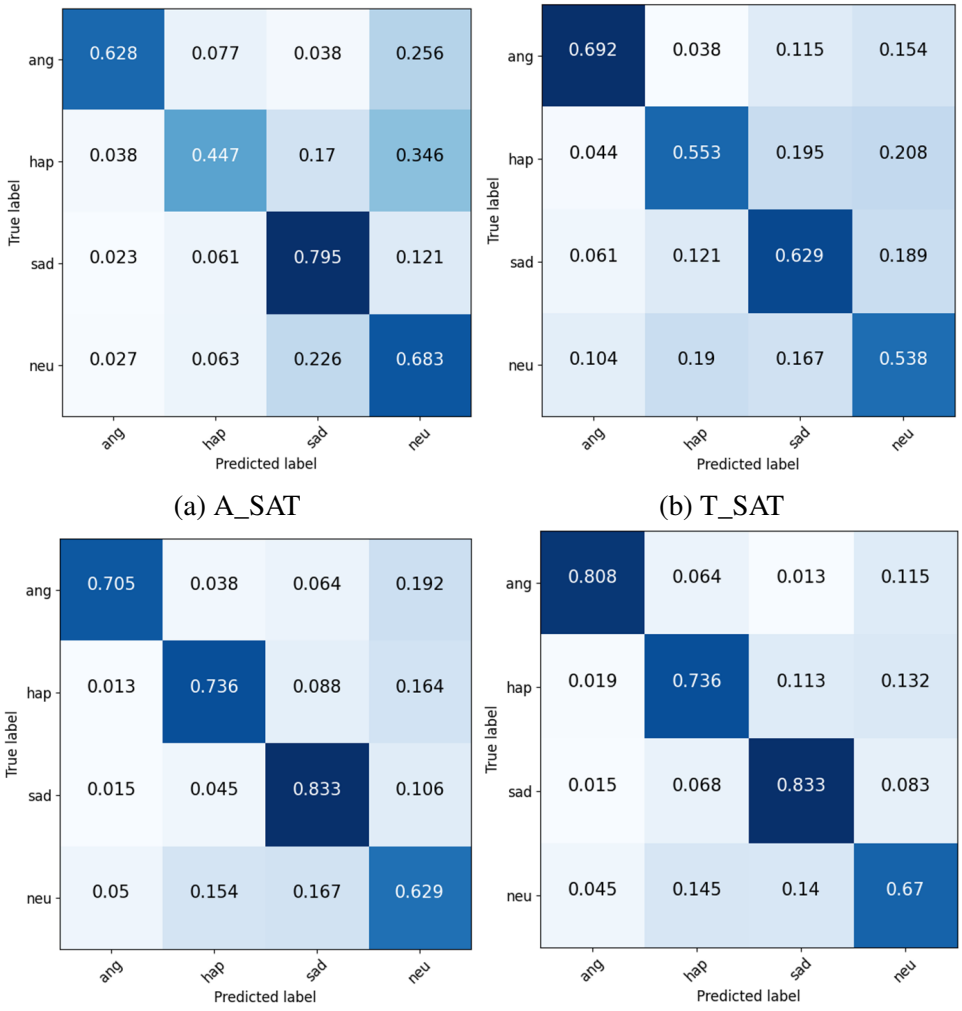
(c) W_SAT (d) SMW_CAT

Fig. 5.3 Confusion matrix of SER systems related to multiple acoustic features. (*ang* angry state; *hap* happy; *sad* sad; *neu* neutral)



(a) S_SAT (b) M_SAT (c) W_SAT (d) SMW_CAT

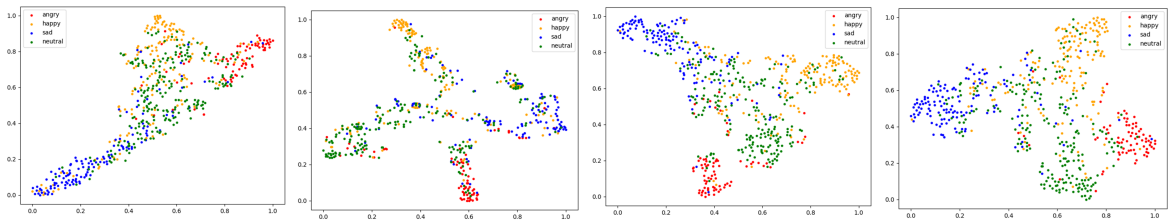
Fig. 5.4 The t-SNE visualization of the last hidden layer in SER systems related to multiple acoustic features.



(a) A_SAT (b) T_SAT

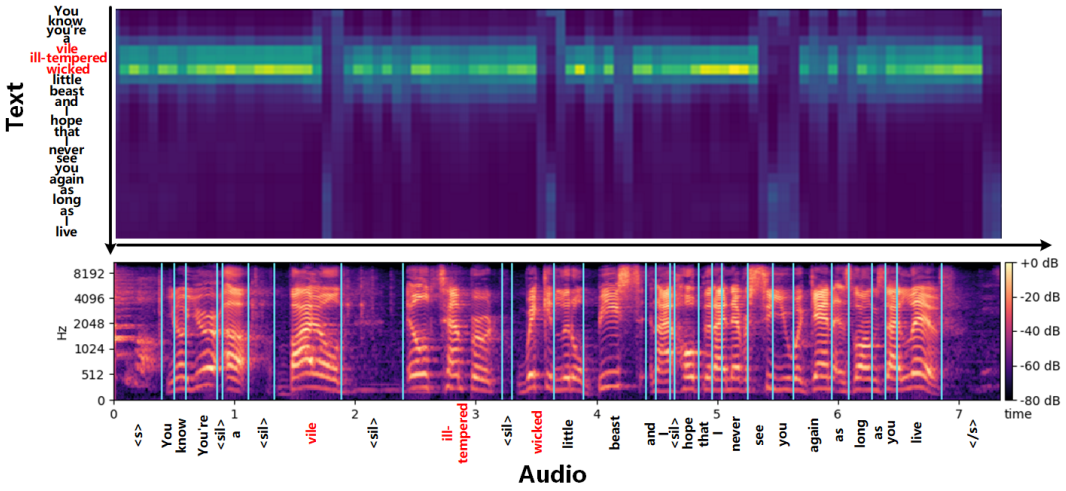
(c) AT_CAT (d) AT_CAT-SAT

Fig. 5.5 Confusion matrix of SER systems related to multimodal features. (*ang* angry state; *hap* happy; *sad* sad; *neu* neutral)

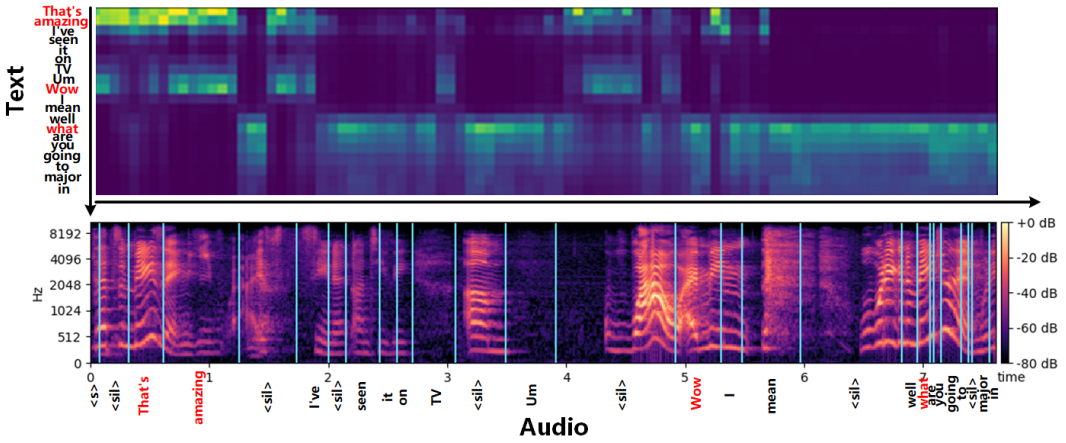


(a) A_SAT (b) T_SAT (c) AT_CAT (d) AT_CAT-SAT

Fig. 5.6 The t-SNE visualization of the last hidden layer in SER systems related to multimodal features.



(a) Attention weights of a sample utterance labeled *angry* and correctly predicted



(b) Attention weights of a sample utterance labeled *happy* and correctly predicted

Fig. 5.7 Visualization of attention weights from the final layer of CAT$^{(a2t)}$ in two sample utterances. In each figure, the transcription (i.e. textual features) is drawn on the left of attention weights, and the log mel spectrogram (i.e. acoustic features) is drawn at the bottom of attention weights. We also add the forced alignment data for comparison. Words with higher weight are emphasized in red color. Note that because of the pooling operation in light CNN, each block contains not only the representation of the corresponding acoustic feature but also its context elements.

Chapter 6

Conclusions and Future Works

6.1 Conclusions

In this paper, we propose several SER systems exploiting self-attention transformer (SAT) and cross-attention transformer (CAT), and then conduct experiments and analysis on them on the IEMOCAP benchmark dataset over 4-class emotion classification.

First, the definition of SAT and CAT are introduced. SAT is the normal transformer that can process the uni-source input by self-attention mechanism. While CAT, composed of two transformer modules, is able to deal with bi-source input by cross-attention mechanism.

Second, towards interaction between the transformer and other deep learning structures, we propose two kinds of SER systems. ILSAT, the model integrating LSTM with SAT in a parallel manner, aims to substitute the positional encodings in SAT with LSTM. CL-CAT, the model utilizing CAT to interact and combine information from CNN and LSTM, is presented grounded on the complementarity between them. Experiment results show that both our proposals can make a promising improvement relative to the C-SAT baseline system.

Third, considering the problem that there is often only one kind of input feature in previous SER systems, we attempt to import additional acoustic features and joint encode them by CAT. Based on this direction, a novel model called SMW_CAT is proposed, showing its effectiveness not only in achieving the best performance of 73.80% WA and 74.25% UA beyond all of comparing systems we designed but also in surpassing existing state-of-art approaches. We also analyze the t-SNE visualization for viewing differences among three acoustic features in an intuitive way.

Last, we discuss our AT_CAT-SAT, which exploits textual features from another modality. It can model the intra-modal interaction and inter-modal interaction simultaneously. Experiment results suggest its superiority over both uni-modal SER systems, achieving a 73.64% WA and

75.05% UA. Besides, visualization of the attention weights in CAT indicates it can correlate between affect-salient acoustic and textual features automatically.

6.2 Future works

For future works, we plan to:

- Find a dataset being collected in multi-languages and test the generalization ability of our proposed SER systems to the cross-linguistic situation.
- Use the results of recognized text data by automatic speech recognition (ASR) instead of zero word recognition error rate transcription to simulate the real-time case. Also, bring in the visual features as the third modality and fuse it with acoustic and lexical features according to the architecture of SMW_CAT.
- Train with more conditions such as continuous emotion labels and gender labels by multi-task learning (MTL) to further improve the performance.

References

[1] Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

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

[4] Iris Bakker, Theo Van der Voordt, Peter Vink, and Jan De Boon. Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33(3):405–421, 2014.

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359, 2008.

[6] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, 2018.

[7] Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction*, pages 511–516. IEEE, 2013.

[8] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987.

[9] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3): 572–587, 2011.

[10] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[11] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Towards real-time speech emotion recognition using deep neural networks. In *2015 9th international conference on signal processing and communication systems (ICSPCS)*, pages 1–5. IEEE, 2015.

[12] Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017.

[13] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Proc. Interspeech 2014*, pages 223–227, 2014.

[14] Che-Wei Huang and Shrikanth Shri Narayanan. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 583–588, 2017.

[15] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *Eighth European conference on speech communication and technology*, 2003.

[16] Soonil Kwon et al. Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. *Expert Systems with Applications*, 167:114177, 2021.

[17] Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition based on phoneme classes. In *Eighth international conference on spoken language processing*, 2004.

[18] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *Proc. Interspeech 2015*, pages 1537–1540, 2015.

[19] Sanghyun Lee, David K Han, and Hanseok Ko. Fusion-convbert: parallel convolution and bert fusion for speech emotion recognition. *Sensors*, 20(22):6688, 2020.

[20] Pengcheng Li, Yan Song, Ian McLoughlin, Wu Guo, and Lirong Dai. An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition. In *Proc. Interspeech 2018*, pages 3087–3091, 2018.

[21] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, 10(10), 2021.

[22] Wootaek Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*, pages 1–4. IEEE, 2016.

[23] Danqing Luo, Yuexian Zou, and Dongyan Huang. Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition. In *Proc. Interspeech 2018*, pages 152–156, 2018.

[24] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE transactions on multimedia*, 16(8):2203–2213, 2014.

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

[26] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2227–2231. IEEE, 2017.

[27] Anish Nediyanchath, Periyasamy Paramasivam, and Promod Yenigalla. Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7179–7183. IEEE, 2020.

[28] Michael Neumann and Ngoc Thang Vu. Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. In *Proc. Interspeech 2017*, pages 1263–1267, 2017.

[29] Yafeng Niu, Dongsheng Zou, Yadong Niu, Zhongshi He, and Hua Tan. A breakthrough in speech emotion recognition using deep retinal convolution neural networks. *arXiv preprint arXiv:1707.09917*, 2017.

[30] Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B Mariño. Speech emotion recognition using hidden markov models. In *Seventh European conference on speech communication and technology*, 2001.

[31] Tsang-Long Pao, Charles S Chien, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng, and Wen-Yuan Liao. Combination of multiple classifiers for improving emotion recognition in mandarin speech. In *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007)*, volume 1, pages 35–38, 2007.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[33] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[34] Aharon Satt, Shai Rozenberg, and Ron Hoory. Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech*, pages 1089–1093, 2017.

[35] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–1. Ieee, 2003.

[36] Guang Shen, Riwei Lai, Rui Chen, Yu Zhang, Kejia Zhang, Qilong Han, and Hongtao Song. Wise: Word-level interaction-based multimodal fusion for speech emotion recognition. In *INTERSPEECH*, pages 369–373, 2020.

[37] Melissa N Stolar, Margaret Lech, Robert S Bolia, and Michael Skinner. Real time speech emotion recognition using rgb image classification and transfer learning. In *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–8. IEEE, 2017.

[38] Lorenzo Tarantino, Philip N Garner, and Alexandros Lazaridis. Self-attention for speech emotion recognition. In *Interspeech*, pages 2578–2582, 2019.

[39] Lorenzo Tarantino, Philip N. Garner, and Alexandros Lazaridis. Self-Attention for Speech Emotion Recognition. In *Proc. Interspeech 2019*, pages 2578–2582, 2019.

[40] Noé Tits, Kevin El Haddad, and Thierry Dutoit. The theory behind controllable expressive speech synthesis: A cross-disciplinary approach. In *Human 4.0-From Biology to Cybernetic*. IntechOpen, 2019.

[41] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE, 2016.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[43] Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[44] Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer. On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology*, 4:292, 2013.

[45] Xueming Yan, Haiwei Xue, Shengyi Jiang, and Ziang Liu. Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling. *Applied Artificial Intelligence*, 36(1):2000688, 2022.

[46] Ziping Zhao, Yiqin Zhao, Zhongtian Bao, Haishuai Wang, Zixing Zhang, and Chao Li. Deep spectrum feature representations for speech emotion recognition. pages 27–33, 10 2018.

[47] Wenjing Zhu and Xiang Li. Speech emotion recognition with global-aware fusion on multi-scale feature representation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6437–6441. IEEE, 2022.

[48] Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng. Speech emotion recognition with co-attention based multi-level acoustic information. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7367–7371. IEEE, 2022.

Appendix A

Publications

National conferences and meetings

- Yurun He, Nobuaki Minematsu and Daisuke Saito, “Effective Integration of Transformer for Network-based Speech Emotion Recognition”. In *IPSJ SIG Technical report*, 2021-SLP-139(7),1-6 (2021-11-24).
- Yurun He, Nobuaki Minematsu and Daisuke Saito, “Transformer-based Speech Emotion Recognition with Multiple Acoustic Features”. In *Proc. Autumn Meeting of Acoustical Society of Japan*, 2-8-11, 2022.

International conferences and meetings

- Yurun He, Nobuaki Minematsu and Daisuke Saito, “Multiple Acoustic Features Speech Emotion Recognition using Cross-Attention Transformer”. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (to be submitted).