Study on speech emotion recognition based on classification and reconstruction for improved practicality

March 2023

Jennifer Santoso

Study on speech emotion recognition based on classification and reconstruction for improved practicality

Graduate School of Science and Technology
Degree Programs in Systems and Information Engineering
University of Tsukuba

March 2023

Jennifer Santoso

Acknowledgements

I want to express my sincerest gratitude to my advisor Professor Takeshi Yamada, for allowing me to pursue the Ph.D. degree and work under his guidance. Without his unwavering support, advice, and encouragement, this thesis would not have been possible.

I want to thank Professor Shoji Makino, my co-advisor during the first half of my Ph.D. and coauthor, who provides insights and critical questions that help me broaden my view on this research and other acoustic-related research.

I thank the thesis committee members: Professor Keisuke Kameyama, Professor Hotaka Takizawa, Professor Mikio Yamamoto, and Professor Naoto Wakatsuki. Their valuable insights help broaden my view on the research and clarify this thesis.

Thank you for the support to RevComm, Inc., especially honorable mentors and coauthors Dr. Kenkichi Ishizuka, Dr. Taiichi Hashimoto, and Mr. Takekatsu Hiramura. Thank you for your valuable cooperation for the three years of Doctoral studies, and may the cooperation benefit us all. I look forward to working together and hope to contribute the best to the company.

I cannot forget my former and current labmates in Multimedia Laboratory. Special thanks to Dr. Li Li, a friend, life mentor, and fellow female researcher. Her guidance and sharing about many aspects of life as a researcher have helped me through the various challenges in pursuing the Ph.D. degree.

I would also thank my family for their continuous support throughout my study period: my mother, Nina Gustina Punta, and my father, Eddy Santoso, for encouraging me with my studies. Thanks to my sister Jessica Santoso, for being cool and spending time in many interesting discussions. Thanks to my aunt Drg. Buddiwati Punta and grandma Ana Widjaja for supporting me financially throughout the early times of my Ph.D. study. Thanks to my aunties, uncles, and cousins for their continuous encouragement.

I want to thank my friends for their support and for providing not only interesting Ph.D. life but also continuous growth as a formidable researcher. I want to thank my friends, especially Candy Olivia Mawalim, for the friendship and for sharing

the fun time as well as valuable insights as a fellow female researcher in acoustics; Ivan Halim Parmonangan for exchanging ideas and for being a supportive research partner; Frederikus Hudi, for helping me stand up during one of my lowest points in life and reaffirm my values in not only pursuing the Ph.D. degree but also in making life decisions. I would also like to express my gratitude to Tsukuba Programming Circle for helping me get through life as an international Ph.D. student and allowing me to hone my mentoring skills early. I would also thank the Indonesian Student Association in Ibaraki (PPI Ibaraki), Tsukuba International Academy, the badminton circle, and many other friends and organizations I have encountered here and everywhere.

Thank you to JST and the University of Tsukuba for partially providing financial support for me to keep on living during part of my Ph.D. study period.

Abstract

Communication is one of the basic needs of human beings. To reach ideal goals for communication, an understanding of the condition of the communication partner, especially the emotional state, is essential. Within speech, as one of the most common means of communication, human also conveys emotion. Along with the advancement of technology and the widespread of communication systems, there is a need for computers to understand the emotions conveyed from speech and possibly with the combination of other types of information to realize a more affectively-aware system. This leads to the study of speech emotion recognition (SER), an essential component of affective computing.

This thesis explores the field of SER and aims to solve the problem related to its practicality. SER has gained a lot of benefits through improvements, enabling the SER to be adopted to practical use. One of these improvements is the combination with other types of information, such as text that can be obtained by transcribing speech using automatic speech recognition (ASR). Although the result from SER studies is promising, SER performance degrade due to the various conditions and limitations in practical use, such as ASR performance degradation due to emotion and the imbalanced training data to develop SER from situations such as business conversation.

In Chapter 3, we focus on improving SER performance by mitigating the effects of incorrect recognition from automatic speech recognition (ASR) to SER. SER is essential for understanding a speaker's intention. Recently, some groups have attempted to improve SER performance using a bidirectional long short-term memory (BLSTM) to extract features from speech sequences and a self-attention mechanism to focus on the important parts of the speech sequences. SER also benefits from combining the information in speech with text, which can be accomplished automatically using an ASR, further improving its performance. However, ASR performance deteriorates in the presence of emotion in speech. Although there is a method to improve ASR performance in the presence of emotional speech, it requires the fine-tuning of ASR, which has a high computational cost and leads to the loss of cues important for determining the presence of emotion in speech segments, which can be helpful in SER. We propose a BLSTM-and-self-attention-based SER method using self-attention weight correction (SAWC) with confidence measures to solve these problems. This method is applied to acoustic and text feature extractors in SER to adjust the importance weights of speech segments and words with a high possibility of ASR error. Our proposed SAWC reduces the importance of words with speech recognition errors in the text feature while emphasizing the importance of speech segments containing these words in acoustic features. Our experimental results on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset reveal that our proposed method outperforms other state-of-the-art methods.

In Chapter 4, we focus on the problem regarding the training of SER due to the conditions and availability of data in practical use. Although classification-based SER methods have achieved high overall performance, these methods tend to have lower performance for neutral speeches, which account for a large proportion in most practical situations. To solve the problem and improve the SER performance, we propose a neutral speech detector (NSD) based on the anomaly detection approach, which uses an autoencoder, the intermediate layer output of a pretrained SER classifier, and only neutral data for training. The intermediate layer output of a pretrained SER classifier enables the reconstruction of both acoustic and text features, which are optimized for SER tasks. We then propose the combination of the SER classifier and the NSD used as a screening mechanism for correcting the class probability of the incorrectly recognized neutral speeches. Results of our experiment using the IEMOCAP dataset indicate that the NSD can reconstruct both the acoustic and textual features, achieving a satisfactory performance for use as a reliable screening method. Furthermore, we evaluated the performance of our proposed screening mechanism, and our experiments show significant improvement in the F-score of the neutral class, and in the class-average weighted accuracy compared with state-of-the-art SER classifiers.

In Chapter 5, we focus on the problem of the imbalanced training data for SER in practical situations. Most of the existing SER methods are the classification-based method, which has some limitations, including maintaining the balance of the training data and the difficulty in handling additional emotional classes; it would be more difficult to add new emotion classes or to retrain the classifier from scratch. This chapter proposes a novel training strategy for an imbalanced dataset based on reconstruction error. We propose an SER method based on the reconstruction of acoustic and text features in latent space. The reconstructor for different emotion classes, including the neutral class, is used. The proposed method selects the emotion class with the lowest normalized reconstruction error as the SER result. Unlike the classifier approach, one reconstructor is dedicated to each emotion class and trained using only the data of the target emotion class. Therefore, the reconstructor can be trained without being affected by imbalanced training data and also facilitates the application of data augmentation to only a specific emotion class. Our experimental result obtained using the IEMOCAP dataset showed that the proposed method improved the class-average weighted accuracy compared with the state-of-the-art SER methods.

Contents

List of Figures

List of Tables

Abbreviations

ASR	Automatic speech recognition
BLSTM	Bidirectional long short-term memory
CM	Confidence measure
CNN	Convolutional neural network
CQT	Constant Q-transform
DNN	Deep neural network
F0	Fundamental frequency
FCN	Fully connected network
FFT	Fast fourier transform
GMM	Gaussian hidden Markov model
HMM	Hidden Markov model
LSTM	Long short-term memory
MFCC	Mel frequency cepstrum coefficient
MLP	Multilayer perceptron
MSE	Mean square error
NSD	Neutral speech detector
RNN	Recurrent neural network
SAWC	Self-attention weight correction
SER	Speech emotion recognition
STFT	Short-time fourier transform
SVM	Support vector machine
UA	Unweighted accuracy
WA	Weighted accuracy

Chapter 1

Introduction

1.1 Background

Communication is one of the essential needs to fulfill for human beings as social creatures. With communication, people can share different kinds of information within a specific context to provide mutual understanding. However, problems such as miscommunications and unpleasant situations become commonplace and obstacles to reaching this ideal condition. One of the main reasons is the lack of understanding of the communication partner's state of mind or emotion. Understanding the emotion can help people to be more aware of the condition of their communication partner and respond more appropriately. This would lead to ideal communication, which is smooth and natural, with all information conveyed as intended. This motivation to understand human emotions leads to the study of *emotion recognition*.

The emerging technology in recent years has provided support to realize emotion recognition and put it to practical use. Emotion recognition has been adopted into many real-life applications, such as call-center conversation analysis [1] and virtual assistants [2], and provides support for daily life communications. These applications use different kinds of information, such as gestures [3], facial expressions [4], biosignals such as EEG [5] and ECG [6], texts [7], and speech [8], which might influence the conveyed emotions. Among all this information, speech is one of the most prominent and readily available as it is the most common means of communication. Speech is rich in information, containing the speech content and the way it is conveyed, and therefore is often used as a cue for emotions. This leads to the task of classifying emotions through speeches or known as *speech emotion recognition (SER)*.

1.2 SER and its problems

SER has been conducted since the early days when speech inputs with their features were extracted and classified into emotional classes. SER was done by using hand-crafted input features, and simple statistics or simple machine learning methods, such as a Gaussian mixture model (GMM) and a support vector machine (SVM) [9, 10]. Recently, deep neural network (DNN)-based approaches have been intensively applied to emotion recognition, resulting in significantly higher performance [11–13]. One possible reason for this higher performance is that DNN can automatically learn the representations of emotions from the input data without the need to hand-craft the input features. To further improve the performance of SER, some studies combine different types of information. One study included feature fusion and ensemble learning on both speech and text from transcriptions [14, 15], including visual information and motion capture [16]. However, in many practical situations, such as in a call center conversation analysis system, the only usable information would be limited to speech. The emergence of automatic speech recognition (ASR) systems has enabled us to automatically obtain text information without requiring manual transcriptions. Combining this with acoustic information has improved SER performance. One main problem is that ASR is not robust to emotions. Some studies have proposed to solve this problem by fine-tuning ASR to be robust to emotions [17] and combining several different ASR and text encoding for SER [18]. These approaches are based on the idea presented in some studies [18, 19] that the lower word error rate in ASR correlates with higher SER performance.

To date, the use of SER in the world is still marginal due to the various challenges to realizing a high-performing SER. One of the challenges is that SER has yet to achieve satisfactory performance. Although there are many methods proposed to improve SER performance, such as by combining other types of information, incorporating ASR results, and lowering the error rates of ASR in recognizing emotional speeches, these methods require high computational resources. On the other hand, performance improvement is still not enough. For practical use, it is important to have an SER with sufficient performance that can be applied to situations with low computational resources.

Another challenge is the different specifications required to apply SER to different conditions in practical settings. For instance, SER designated to handle casual conversations in noisy places would require SER to be robust to noises, whereas SER designated to handle business conversations would require SER to handle more neutral speeches and be more sensitive in detecting emotions, as emotional speeches are uncommon. Therefore it is important to develop an SER method that specifically caters to certain practical condition settings.

The other challenge is that most conventional emotion recognition methods are classification-based methods, in which emotions are decided from the highest probability of a set of designated emotion classes. These methods mostly assume the balanced data population for each class. However, in practical conditions, the data population is highly imbalanced for each class, and the performance of some of the classes might be unequal. Therefore, classification-based methods might not be suitable to apply in practical situations. To make SER more applicable in practical situations, some methods, such as data augmentation, have been proposed [20]. However, the data augmentation may worsen the data imbalance problem if applied to the class containing most of the data population. Moreover, to date, there have not been many studies that address the data imbalance problem in emotion recognition, which is important to realize a high-performing SER in practical applications.

1.3 Research objectives

The main objective of this thesis is to investigate the problems of realizing SER for practical uses, ranging from improving the recognition performance and adapting SER to different specifications for practical use. First, we introduce the SER method to solve the performance degradation of ASR in emotional speeches. Second, we investigate the approach to solving SER problems in business conditions, addressing the low recognition performance of neutral speeches, which dominates business conversations. Finally, we explore a new training strategy for classification tasks that would solve real-life problems of data imbalance for emotional speeches.

1.4 Overview of thesis

This thesis consists of three parts. Chapter 2 provides a review of emotion recognition and speech emotion recognition along with the base method used in this study, as well as the evaluation metrics in this study. Chapter 3 introduces the speech emotion recognition method aiming to improve performance by solving the speech recognition error problem in emotional speeches. Chapter 4 introduces the speech emotion recognition method based on anomaly detection to solve the practicality problem in a business setting. Chapter 5 introduces a new approach to training classification-based methods and their particular application to speech emotion recognition. Finally, Chapter 6 concludes the entire contents and contributions in this dissertation.

Chapter 2

Emotion recognition

In this chapter, we will provide some preliminaries regarding emotion recognition and its components, especially in speech emotion recognition. We explain some of the most recent speech emotion recognition methods. We also provide some basic knowledge to understand our proposed approach throughout the thesis further. Finally, we review the evaluation criteria for speech emotion recognition.

2.1 Emotion recognition

Emotion recognition is an essential part of affective computing, in which machines attempt to identify the emotion in given inputs. Emotion recognition is recognizing the underlying emotions of the person conveying them. In practical use, emotion recognition is mainly formulated into the classifier-based task. Classifier-based emotion recognition involves identifying emotions based on discrete classes, such as happy and angry, represented using the emotion class probability. The method receives input from various types of information, such as speech, text, and images, extracts the feature, and classifies the emotion. Then, the method extracts the necessary features using a feature extractor and classifies the emotions with an emotion classifier. The flow of emotion recognition is illustrated in Figure 2.1.

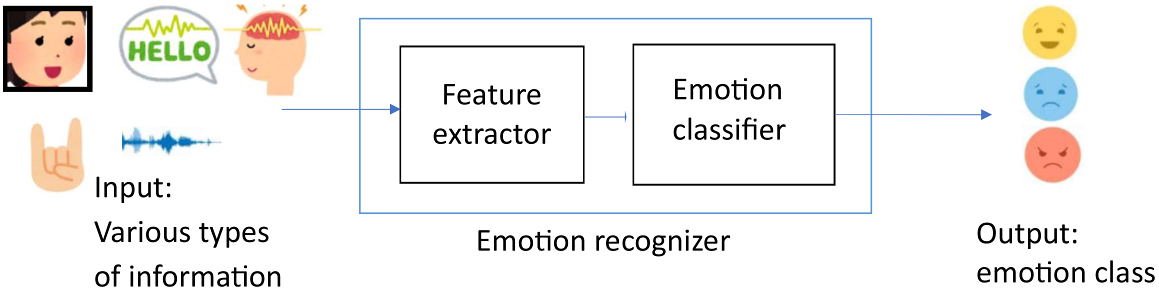


Figure 2.1: Classifier-based emotion recognition

2.2 Emotion modelling

One of the most important aspects of conducting effective emotion recognition is to choose the emotion model to adopt. As explained in one study [21], there are at least three models for emotions: categorical emotion, dimensional emotion, and appraisal emotion.

2.2.1 Categorical emotion

Categorical emotion, or known as a basic emotion, is emotion based on the available discrete labels. Categorical emotions seek to group emotions based on effective families. While experts have differing views, most emotion scientists agree that there are at least five core emotions. One proposal by Ekman [22] suggests six emotions: anger, fear, enjoyment, sadness, disgust, and surprise. On the other hand, Plutchik [23] further categorized the emotion into eight bipolar emotions: joy, sadness, trust, disgust, fear, anger, anticipation, and surprise. The emotion model proposed by Plutchik is illustrated in Figure 2.2. Another type is proposed by Cowen-Keltner [24]. Here, a total of 27 emotions were elicited from reviewing video samples. The study also reported that categorical labels such as amusement are more than capable of representing subjective experiences.

In emotion recognition, there are cases where it is not possible to classify the emotion into one of the defined emotion classes. One necessary step to alleviate this problem is to incorporate the neutral class as conducted throughout the thesis. The neutral class represents the situation where there is no emotion expressed, or the emotion is not expressed enough to be considered clearly as one of the emo-

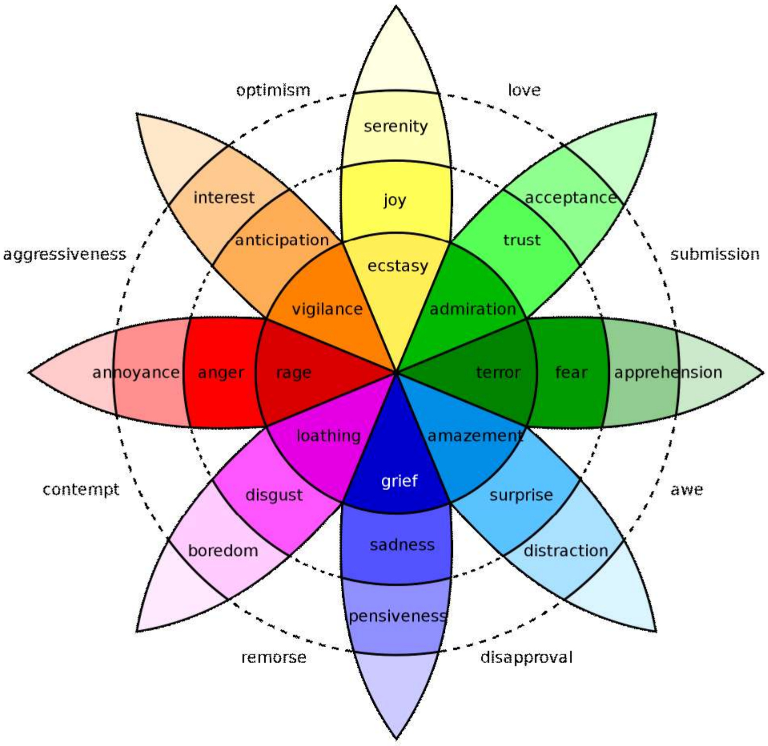


Figure 2.2: Plutchik wheel of emotions

tion classes. The neutral class does not include situations where there is a clear emotion that is other than the defined class, or there are multiple emotions from the defined classes. In such a case, it would be possible to generate the probabilities of the defined emotions and either present the multiple emotion classes with the highest probabilities or reject the data when the probabilities of being in a defined emotion are too low, which can be beneficial for practical situations.

In practical use, categorical emotion is the most commonly applied emotion model for emotion recognition. One of the main reasons is that although there are slight differences between the discrete labels in each model, several common emotions are found in almost all categorical emotion theories, such as happy, sad, and angry. These are the emotions that people find familiar; therefore, the results from using this model can be easily represented by many people.

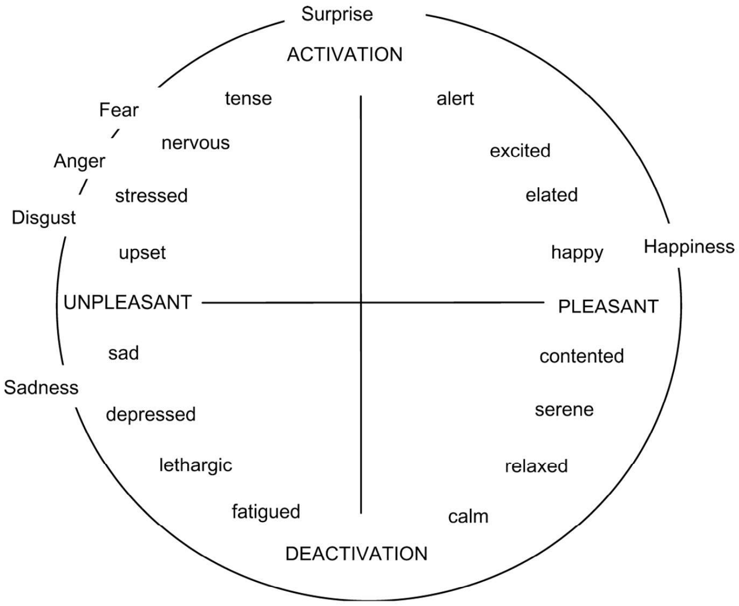


Figure 2.3: Dimensional emotion model

2.2.2 Dimensional emotion

Aside from categorical emotion, another way to model the emotion is through dimensional emotion. The dimensional emotion model views emotion as continuous values in two or three different attributes. The dimensional emotion was first developed by Russell [25], where emotion is modeled on a circumplex model based on subjective feelings. In the study, 28 emotion words are grouped based on perceived similarity, and the analysis in the study revealed two bipolar dimensions of valence and arousal. Valence measures how positive or negative an emotion is, while arousal measures the activation level of the emotion, namely sleepiness (low arousal) and awakedness (high arousal). The dimensional emotional model by Russell is illustrated in Figure 2.3. Another study by Fontaine [26] argues that emotion is not two but four dimensions (valence, arousal, dominance, and predictability).

As the dimensional emotion model measures emotion in terms of continuous values in two or more dimensions, it does not require any specific categorical labels and therefore eliminates the difference between emotional label names such as those used in categorical emotion. However, several notable issues exist in adopting the dimensional emotional model to practical use. First, the dimensional emotion model is less familiar and therefore needs to be represented back to categorical emotion. Second, the level of personal differences in the dimensional emotion model is higher than that of the categorical speech, which provides more challenges in determining the correct labels. Finally, the analysis from Cowen-Keltner revealed that dimensional attributes are less capable of capturing the subjective emotion evaluation used for practical situations than categorical labels. Therefore this approach is not used in this study.

2.2.3 Appraisal emotion

Other than categorical and dimensional emotion, there is appraisal emotion, another type of emotion model. Appraisal emotion is the hybrid of categorical and dimensional emotion, having the basic emotion labels and their intensity as represented in the dimensional emotion. One of example of appraisal emotion is the Emotion Geneva Wheel [27]. Since appraisal emotion is uncommon among researchers, not many datasets employ appraisal emotion as the label, and therefore not used in this study.

2.3 Speech emotion recognition (SER)

Speech is one of the most common methods of communication and one of the most readily available types of information. Speech contains linguistic information and other types of information, especially related to how it is conveyed. Due to these reasons, speech is mainly used to recognize emotions, leading to the study of SER. SER is the task of identifying the emotion expressed by the speaker in a given utterance. Like emotion recognition methods, SER comprises input feature extraction and classification steps.

The first known research paper on SER has existed since 1996 [28], with the idea possibly existing much earlier. In the first study on SER, statistical pattern recognition methods and the use of prosodic features were explored to classify emotional content in speeches. The method proposed in the study achieved performance comparable to humans using a limited amount of speech data.

The progress of studies in SER is further contributed by the increase in the datasets published. During the early 2000s, many datasets are published for emotion recognition, including speech datasets. Some notable datasets for SER include EmoDB [29], IEMOCAP [30], RAVDESS [31], and MSP-Podcast [32]. These datasets are commonly used in SER research and hasten the progress of studies in SER.

The advancement of SER studies, together with the increasing need for SER and the increasingly available amount of data available in real life, led to practical implementation in industries. SER has been implemented in various applications such as virtual assistants and call-center conversation analysis. The study of SER implemented to practical use was researched in this pilot study [33], where it was applied for call-center application. This early study showed tremendous potential for SER to be applied practically. Nowadays, the application of SER is widespread in the industry, while new methods on tackling real-life situations with SER are continuously being studied.

The availability of many data types in recent years has opened up the possibility of combining several different types of data to improve the performance of certain tasks, known as multimodal information processing. Emotion recognition, including SER, benefits from multimodal information processing. Multimodal information processing is inspired by the decision-making process in humans, where humans recognize situations based on different sensories. For instance, a person is convinced that someone is sad by looking into the tears, the speech uttered by the sad person, the body gestures, and many more. This is adapted to machine learning, which also benefits from multimodal information processing as machine learning improves with many inputs processed.

In the case of SER, which processes primarily from speech, it benefits from combining the information in speech: acoustic features and textual content. One method to obtain the textual content is through manual transcriptions. Transcriptions provide accurate speech content and are reliable for SER tasks. Studies showed that the methods using acoustic features and transcriptions are effective in recognizing emotions [14, 15]. However, obtaining transcriptions is impractical because it requires numerous human annotators and transcriptions are not available in real-time.

With the availability of ASR, it is possible to obtain speech content in the form of text by inputting speech to ASR. Therefore, it is possible to apply multimodal information processing in SER using speech and text, with speech as the main input source. One of the main issues that arise among the researchers in the SER community is whether multimodal information processing is necessary. However, recent studies [34] confirm the effectiveness of SER methods that applies multimodal information processing in comparison to the ones utilizing only speech or text. It is shown that the methods with the text feature from ASR results, in addition to other input features, still provide better performance for SER than those without, indicating that text information is still essential to improve SER performance.

Another issue is about the best way to apply multimodal information processing to SER, in other words, how to fuse speech and text in SER. Speech and text each represent different types of information and their importance. Due to this reason, speech and text are separated early in the processing and need to be fused back at a later processing, depending on which information to be considered more important. One of the known earliest works [35] combined speech and text information at the decision level. Several different approaches can accomplish the fusion of speech and text in SER.

2.3.1 Features

Input features are an essential part of developing many of the recent systems, including the emotion recognition system. In SER, the system mainly receives speech as the primary input and text as additional information to further enhance the performance of SER. However, speech and text are difficult to process without any further processing, and it would be better to have the system process only the important information in a way that is easily understandable to the system. An effective input feature extraction will yield high performance in SER and other systems. In this thesis, we divide the features into acoustic and textual features, representing the features from speech and text, respectively.

Acoustic features

Acoustic features contain much information about emotions and are correlated with emotion, with studies dating back to earlier years. In SER, there are two big divisions of acoustic features: hand-crafted features and deep learning-based features. Hand-crafted features is further divided into two categories, namely low-level descriptors (LLD) and high-level statistical features (HSF). LLD is mainly extracted per frame, while HSF is statistical features computed from LLDs and captures the changes among the frames.

LLD and HSF are further divided into several groups, namely signal energy, fundamental frequency (F0), voice quality, cepstral, time signal, and spectral, according to one study [36]. These LLDs, especially F0, intensity, and voice quality, are known to correlate strongly with emotion, as shown in this study [37].

LLD and HSF are most commonly used in SER and many speech-related tasks, such as ASR. In ASR, MFCC is widely used as it contains much information, such as the phoneme representation and the change of intonation. The use of MFCC has been proven in some studies to be better than other spectrogram-based features as MFCC provides the most informative acoustic features compared to other acoustic features.

The advancement of deep learning-based methods has made deep learning-based features available, which are obtained from extracting the acoustic representation in other deep learning-based tasks. Deep learning-based features are commonly used in recent years and have gained the attention of researchers in the field of SER.

Textual features

Textual features, known as text features, lexical features, linguistic features, and semantic features, represent important information from text inputs. As machines process inputs in numerical representation, text by itself is difficult to be processed and therefore needs to be converted to numerical values. In tasks that require text processing, including SER, there has been many ways presented to extract textual features. One of the simplest way, as presented in the early works [38], is to spot the keywords or phrases correlated to certain emotions. For example, the word “disappointed” can be represented as [(2, 0.2), (3, 0.6)] where 2 represents “angry” emotion and 3 represents “sadness” emotion. The values 0.2 and 0.6 represent the intensity of emotions.

The representation of textual features gradually becomes more systematic, as shown by TF-IDF. TF is defined as the frequency of a word in a particular document/utterance, whereas IDF is defined as a logarithm of the total number of documents ratio to the total number containing that word. TF-IDF is the multiplication of TF with IDF.

Another advancement is representing the text in a numerical feature vector, such as bag of words (BOW). First, a fixed integer is assigned to each word occurring in any document, i.e., building a dictionary from a corpus by assigning a word to integer indices. Second, count the number of occurrences of each word and store it as the value of feature j where j is the index of word w in the dictionary. Recently, BoW features have been expanded to their acoustic and visual counterparts (BoAW and BoVW).

Recent studies with deep learning-based methods have led to the study of deep learning-based word embeddings, such as word2vec, GloVe, BERT, and FastText. These models are trained from DNN models using a large linguistic corpus to generate word vectors and sometimes include relative positional embeddings or contextual information.

2.3.2 Classifier-based methods

Over the years, classifiers for SER has advanced from simple statistical-based method to support vector machine (SVM), and multilayer perception (MLP). Along with the rise of deep neural networks and the huge data availability in recent years, methods such as convolutional neural network (CNN), long short-term memory (LSTM), and attention mechanism have brought the benchmark performance of SER further. Here, we provide some preliminaries about the commonly-used classifier-based methods.

Support Vector Machine (SVM)

SVM is based on statistical learning theory and regression analysis. SVM [39] is a machine learning for classification problems, which conceptually implements these ideas: Find a hyperplane in a multidimensional space that separates the data points to their potential classes. SVM itself is applicable for many applications, such as EEG signal classification, cancer identification, seizure prediction, face recognition, speech disorder, and bioinformatics. In the early days, when the number of available data was still small, SVM was one of the best-performing methods for classification. Some of the early works [10,37] show promising results using SVM for SER.

Multilayer perception (MLP)

MLP is also known as the feedforward neural network. MLP is based on connectionist learning proposed in the early years [40]. MLP projects the input data into linearly separable space using non-linear transformation. MLP refers to the neural network containing several hidden layers, which is a layer between input and output containing many units or perceptrons. The increasing number of hidden layers in MLP creates deeper layers, which can be referred to as a simple deep neural network (DNN). One of the advantages of MLP is the ability to extract features from more complex, less-structured data. MLP can be formulated in Eqs. 2.1 Here, x denotes the input feature, y denotes the output, $\mathbf{W}$ denotes the hidden layer containing weights, and b denotes the bias.

$$y = g(\mathbf{W}^{\intercal}x + b_1), \tag{2.1}$$

MLP has been utilized for SER, and is proven to be effective against conventional statistical-based methods. Some of the studies [11, 41–43] reported the effectiveness of MLP in SER compared to the SVM-based methods. In another study [44], MLP is used to evaluate the effect of context information on categorical SER tasks and is shown to outperform most of the results using the baseline majority-class method.

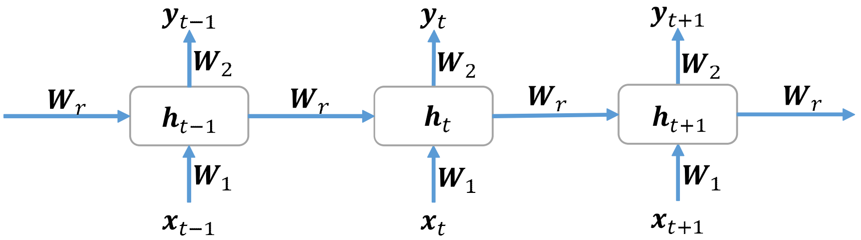


Figure 2.4: RNN architecture

CNN

Convolutional neural network, or CNN, is a type of neural network containing convolutional layers. In a convolutional layer, a convolution operation is conducted, in which the overlap of two functions are measured when one function assumed to be the input is shifted by another function assumed to be the kernel. The output of a convolution layer is a feature map. As CNN is deeply inspired by the mechanism of the animal visual cortex in processing the visual field, CNN is often applied to image-like data. However, processing time-series data using CNN is still possible by using 1-dimension CNN. In processing speech data, CNN usually receives spectrogram-based features as the input. In processing textual features, CNN works by computing n-gram vectors and grouping the vector afterward.

CNN has been widely used in SER and has shown promising results [45, 46] through efficient learning of salient features and spectrogram images. Using CNN for sequential data such as speech and language processing, has some drawbacks. One of the possible drawbacks is that CNN does not keep time-variant information, which can be alleviated by applying LSTM. Many studies in SER have taken advantage of combining CNN and LSTM, resulting in the improved benchmarks [12, 13, 47].

LSTM

Bidirectional Long Short-term Memory is a Recurrent Neural Network (RNN) variant. RNN is a neural network model that conducts sequential data processing, such as time-series data. RNN consists of a hidden layer that is accessed repeatedly, in which output from the previous sequence of the hidden layer is used as

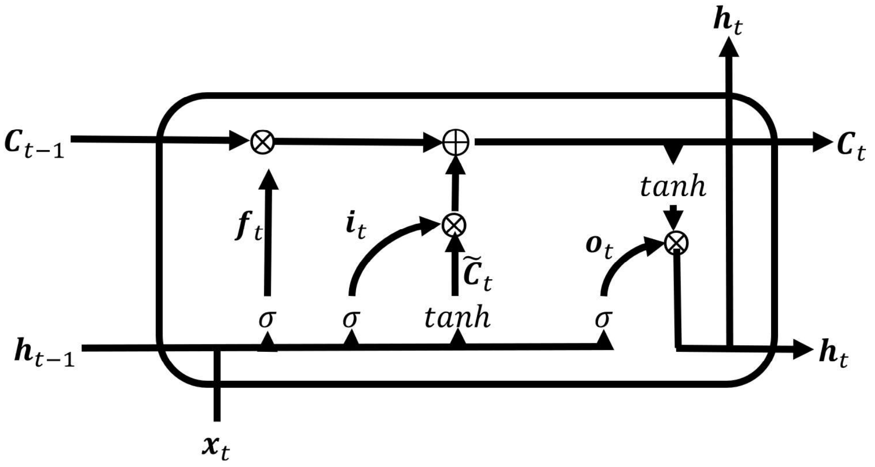


Figure 2.5: LSTM unit

input for the current sequence while maintaining the internal state. The architecture of RNN is shown in Figure 2.4. In the figure, $\mathbf{x}_t$, $\mathbf{h}_t$, and $\mathbf{y}_t$ represents the input sequence, hidden layer vector, and output vector from LSTM respectively.

RNN itself has several problems, including a vanishing gradient, which causes the inability to handle information on long-term dependencies. Although the information are still retained over the short term, they are lost over the long term, losing relevance to the current state. As a countermeasure, Long Short-term Memory (LSTM), a variant of RNN, is introduced. The improvement compared to RNN is that LSTM adds three gates: forget gate $\mathbf{f}_t$, input gate $\mathbf{i}_t$, and output gate $\mathbf{o}_t$ with a value between 0 to 1, where 0 indicating closed gate and 1 indicating open gate. These three gates enable information control, allowing the flow of long-term and short-term memory information. The single LSTM unit is shown in Figure 2.5. LSTM can be represented by

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f\right), \tag{2.2}$$

$$\mathbf{i}_t = \sigma\left(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i\right), \tag{2.3}$$

$$\widetilde{\mathbf{C}}_t = tanh\left(\mathbf{W}_C \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_C\right), \tag{2.4}$$

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \widetilde{\mathbf{C}}_t, \tag{2.5}$$

$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o\right), \tag{2.6}$$

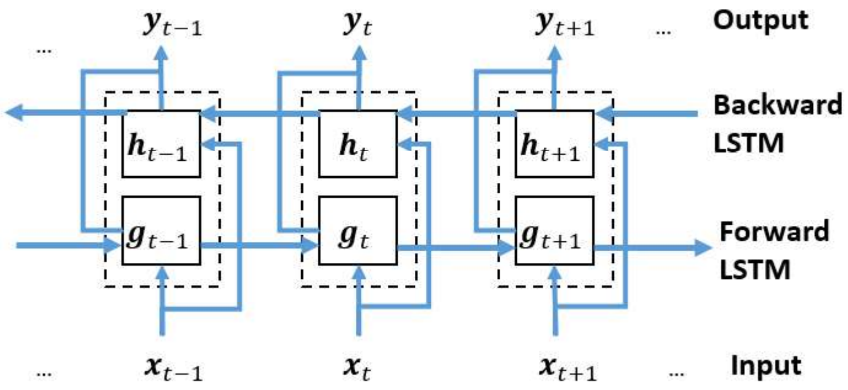


Figure 2.6: BLSTM architecture in general

$$\mathbf{h}_t = \mathbf{o}_t \cdot tanh\left(\mathbf{C}_t\right) \tag{2.7}$$

In the equations, $\mathbf{f}_t$, $\mathbf{i}_t$, and $\mathbf{o}_t$ represents value of forget gate, input gate and output gate at time t respectively. $\mathbf{C}_t$ and $\widetilde{\mathbf{c}}_t$ represents state value and new candidate state value. σ and $\odot$ represents local sigmoid function and point-wise multiplication.

BLSTM is a model that consists of two LSTMs, in which one LSTM runs to process the time sequence data in forward direction and the other runs the time sequence data in backward direction. Therefore, BLSTM can retain the information relevant to the previous state, performing better than LSTM. BLSTM is applied to utterance characteristics classification in this study. BLSTM architecture is shown on Figure 2.6. In this figure, $\mathbf{g}$ represents the forward LSTM states, $\mathbf{h}$ represents the backward LSTM states, and t represents the time. The time sequence input $\mathbf{x}_1$ to $\mathbf{x}_t$ is fed to $\mathbf{g}$ and $\mathbf{h}$. The output $\mathbf{y}_1$ to $\mathbf{y}_t$ is obtained by concatenating the resulting states from $\mathbf{g}$ and $\mathbf{h}$.

Attention mechanism

BLSTM-based networks have one demerit: information loss after going through long sequences. This issue causes some of the information in the earlier sequences to be considered unimportant. To solve this issue, recent studies focused on the important parts and used the weights to calculate another sequence. The attention mechanism [48], which is a neural-network-based mechanism to capture

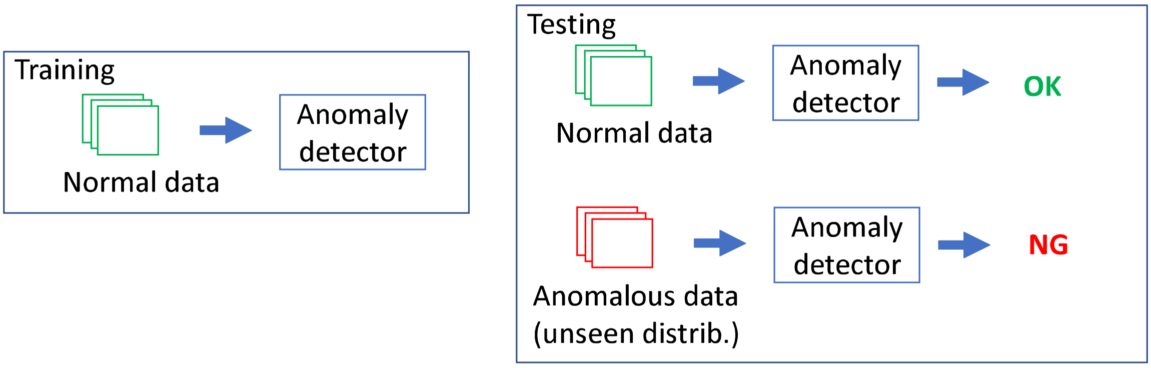


Figure 2.7: Anomaly detection

the contextual information from a sequence, has been introduced. This mechanism is based on the biological organism's ability to extract cues and identify which part to focus on next to obtain the key information.

Attention mechanism can be represented by

$$\mathbf{u}_i = tanh(\mathbf{W}\mathbf{e}_i + \mathbf{b}), \tag{2.8}$$

$$\boldsymbol{\alpha}_i = \frac{exp(\mathbf{u}_i^T \mathbf{u}_i)}{\sum_{i=1}^{t} exp(\mathbf{u}_i^T \mathbf{u}_i)}, \text{ and} \tag{2.9}$$

$$\mathbf{c} = \sum_{i=1}^{t} \boldsymbol{\alpha}_i \mathbf{e}_i. \tag{2.10}$$

The hidden state from the previous layer output e_i is fed to the attention mechanism to determine the attention weight $\boldsymbol{\alpha}_i$ of each frame, which is determined by Eqs. 2.8 and 2.9. The output of the attention mechanism is the weighted sum of e_i, represented by vector c, as shown in Eq. 2.10.

In particular, the self-attention mechanism [49] is widely used in classification tasks. The self-attention mechanism focuses on the important parts and applies weights in the same sequence. Together with LSTM-based networks, the use of the self-attention mechanism has improved the performance of many classification tasks, including SER [50–53].

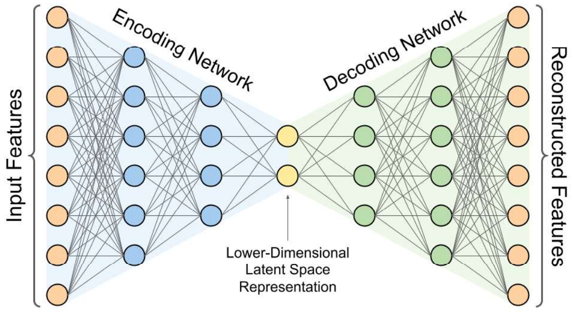


Figure 2.8: A typical structure of an autoencoder

2.3.3 Anomaly detection

Anomaly detection is the identification of data outside the distribution of the majority of the data. Anomaly detection is suitable for the condition where many data are of the same class, and many others are outside the class. The method of anomaly detection usually involves training the detector with normal data to learn their representation and reduce their error. Therefore, when the data that is out of the distribution, known as anomalous data, is inputted, it has a high error rate and can be considered an anomaly. The flow of anomaly detection is illustrated in Figure 2.7.

In general, there are many techniques that can be employed for anomaly detection, ranging from statistical-based methods, one-class SVM, to reconstruction methods. Due to the recent advancement in DNN-based methods, particularly autoencoder, many of the anomaly detection techniques have been able to handle larger dimensions of data.

Anomaly detection is a fundamental component in many applications, where spotting anomalies is vital such as in medical systems, critical infrastructures, security applications, and image defect detection. In the field of acoustics, there have been previous studies in detecting faulty machine sounds [54]. This study uses autoencoder to reconstruct spectrograms and ensures the reconstruction with minimal loss for normal machine sounds. The promising result shown in this study shows the potential of adopting anomaly detection to other acoustic-based tasks such as SER.

Autoencoder

Autoencoder [55] is a deep-learning architecture primarily used to represent higher-dimensional data, typically for dimensionality reduction. The autoencoder learns a representation for a data set, by training the network to ignore insignificant data. Autoencoder consists of an encoder and a decoder, which compress the input features to more compact bottleneck features and decompress the bottleneck features into a reconstructed output with the same shape as the input. An optimal autoencoder would perform as close to perfect reconstruction as possible, in other words, having a small reconstruction error. Typically, the autoencoder uses a fully connected neural network as the structure, with the encoder having units that are smaller than the previous layer and the decoder being the size of the encoder with reverse order. A typical autoencoder is illustrated in Figure 2.8. In recent years, many variants of autoencoder can be used, such as variational autoencoder (VAE) [56] used to augment the data by injecting a latent variable to autoencoder, and RNN-based autoencoder, which uses RNN as encoder and decoder instead of fully connected neural network. Autoencoder has been applied to many tasks such as data augmentation and anomaly detection through reconstruction.

In SER, the use of autoencoder has been explored for reconstruction-error-based learning [57]. The study used an autoencoder with RNN as the encoder and decoder, and RNN to predict emotion in continuous emotion recognition. The results show a high correlation between small reconstruction errors and performance improvement. Despite the great potential of using RNN as an autoencoder to reconstruct variable length data, it is still difficult to reconstruct using an RNN-based autoencoder, which prompts some studies to solve the problem related to the representation of autoencoder.

2.4 Base method

In this thesis, we will base our work on the basic SER method using speech and its ASR result as the input, as illustrated in Figure 2.9. Here, the speech features and their ASR result are extracted separately, representing different types of infor-

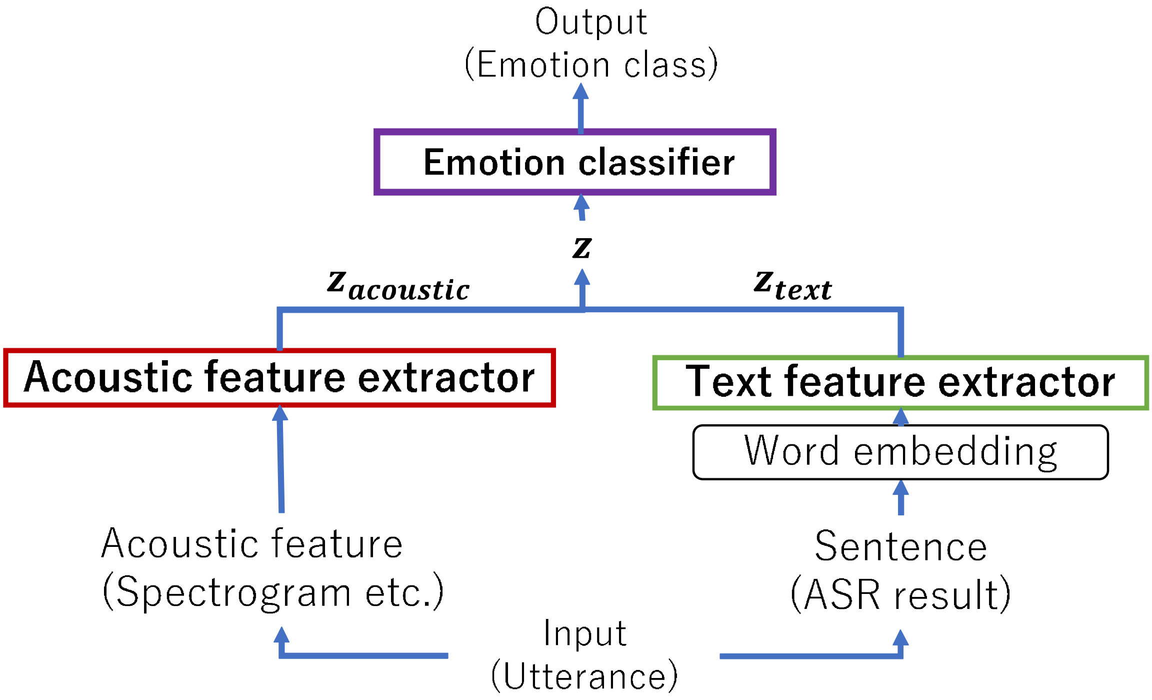


Figure 2.9: State-of-the-art SER method

mation.

Figure 2.10(a) illustrates the acoustic feature extractor in the basic SER method. The acoustic feature extraction in the state-of-the-art method uses bidirectional long short-term memories (BLSTM) and a self-attention mechanism, independently extracting speech features and their ASR result. For simplicity, we denote the acoustic features $x_1, ..., x_T$ where x represents the input acoustic feature and T represents the number of frames. These are then fed to the BLSTM to obtain $\mathbf{e}_i$, which is defined for each frame index i as

$$\mathbf{e}_i = \mathbf{g}_i \oplus \mathbf{h}_i, \tag{2.11}$$

where $\mathbf{g}$, $\mathbf{h}$, and $\oplus$ represent the forward hidden states of BLSTM, backward hidden states of BLSTM, and concatenation, respectively. Then, $\mathbf{e}_i$ is fed to the self-attention mechanism defined as

$$y_i = \mathbf{m}tanh(\mathbf{N}\mathbf{e}_i^T), \tag{2.12}$$

$$\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_T = softmax(y_1, ..., y_T). \tag{2.13}$$

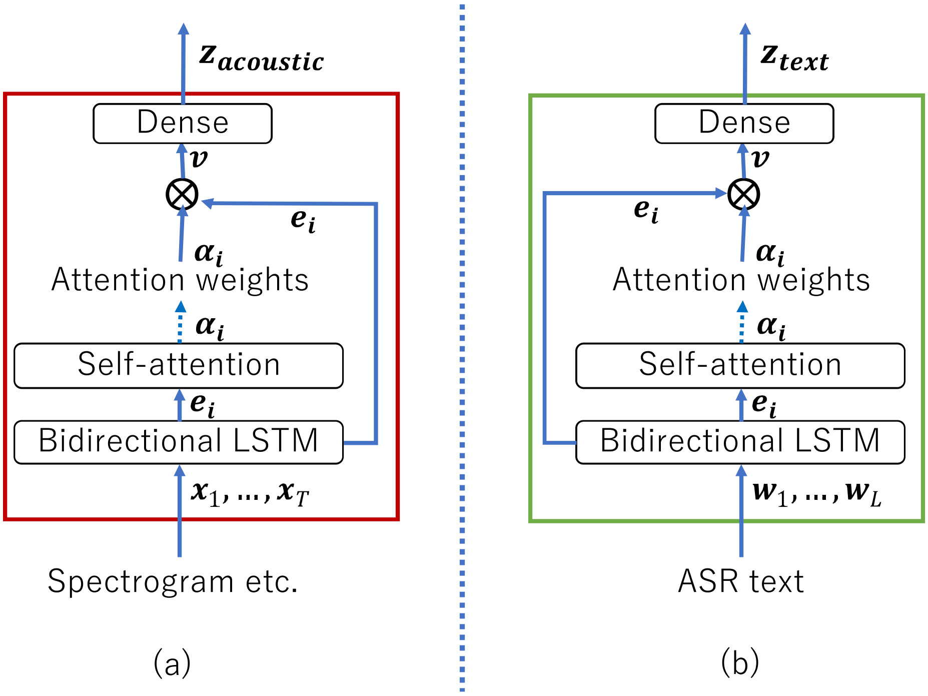


Figure 2.10: Feature extractors in the basic method: (a) acoustic feature extractor, (b) text feature extractor

α_i is the attention weight at frame i, and $\mathbf{m}$ and $\mathbf{N}$ are trainable parameters that can be represented as a layer of a dense neural network. The weighted sum **v** from BLSTM and attention weights are defined as

$$\mathbf{v} = \sum_{i=1}^{T} \boldsymbol{\alpha}_i \mathbf{e}_i. \tag{2.14}$$

After the weighted sum **v** is calculated, it is fed to a single fully connected layer to obtain a fixed-length intermediate layer representation, $\mathbf{z}_{acoustic}$, of acoustic features.

Figure 2.10(b) illustrates the text feature extractor in the basic SER method. The text feature extractor in the SER method uses the ASR text of the input utterance. ASR text is first encoded by bidirectional encoder representations from transformers (BERT) word embedding. The resulting embeddings with length L, defined as $\mathbf{w}_1, \ldots, \mathbf{w}_L$, are then fed to the text feature extractor by the same process as that of the acoustic feature extractor. Here, we obtain the fixed-length intermediate layer representation $\mathbf{z}_{text}$ of text features. The classification part then receives

$\mathbf{z} = \mathbf{z}_{acoustics} \oplus \mathbf{z}_{text}$ as the input and then outputs the emotion class probability. The final emotion class is taken from the highest emotion class probability.

2.5 Evaluation criteria

One necessary aspect for measuring the improvement of the emotion recognition methods in practical use is to evaluate the emotion recognition performance. Regarding the scope of this thesis, which assumes emotion recognition as a classification-based task, we use objective evaluation. To evaluate the overall emotion recognition performance, we use the unweighted accuracy (UA) and the weighted accuracy (WA) as the evaluation metrics. The UA is defined as the number of correctly classified samples divided by the total number of samples, whereas the WA is the average of the correctly classified sample in a class divided by the total number of data in each class. Aside from the UA and WA, we measure the performance of emotion recognition for each class using the F-score. The UA, WA, and F-score are respectively defined as

$$\text{UA} = \frac{\sum_{i=1}^{N} t_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N} t_{ij}}, \tag{2.15}$$

$$\text{WA} = \frac{1}{N}\sum_{i=1}^{N}\frac{t_{ii}}{\sum_{j=1}^{N} t_{ij}}, \tag{2.16}$$

$$\text{F-score} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}, \tag{2.17}$$

where N is the number of classes and t_{ij} is the number of data labeled as class i and predicted as class j. In practice, all of these criteria are calculated using Scikit learn. For all these metrics, the value lies between 0% to 100%, where higher percentage indicates better performance.

Chapter 3

SER based on self-attention weight correction for acoustic and text features

3.1 Introduction

SER using acoustic features and ASR results have suffered from performance degradation due to the ASR not being robust to emotions, resulting in many speech recognition errors in emotional speeches. Moreover, the state-of-the-art approaches of SER use a self-attention mechanism, which focuses on the important parts or segments in a sequence. They would end up focusing on the parts that contain speech recognition errors, contributing to the SER performance degradation. One possible way to mitigate the effects of speech recognition errors is to use the result of multiple ASR and text encoders, which has complex architecture and require access to several different ASR. Meanwhile, the speech segments containing speech recognition errors may contain cues beneficial to understand emotions. Therefore, the question is how to reduce the effects of speech recognition errors on SER by utilizing the information contained in those.

In this chapter, we introduce self-attention weight correction (SAWC) using confidence measures (CM) [58], a metric indicating the reliability of ASR results that corrects the weights of the attention mechanism. We investigate the applications of SAWC for acoustic and text features and examine the changes in the corrected attention mechanism.

3.2 Proposed method

3.2.1 Problems of the basic SER method

In this section, we discuss the problems of the basic SER method and then explain the details of our proposed method. One main issue with the basic SER method is that the SER performance deteriorates owing to ASR errors. One of the most prominent causes of these errors is the presence of emotion in speech because emotion changes the intonation and pronunciation of the intended speech content.

As a reference, we investigated the word error rate of ASR for neutral speeches and emotional speeches. In this study, we use a pretrained ASR based on the Kaldi speech recognition toolkit [59], using the speech data from Librispeech [60], which consists of English speech from audiobooks. The word error rate for this dataset is 3.8% under the clean condition. On the other hand, the word error rate for the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [61], the English emotional speech dataset used to evaluate our proposed method, using the same pretrained ASR is 43.5%. The high word error rate for the IEMOCAP dataset indicates that the presence of emotion considerably deteriorates ASR performance.

Since ASR is not robust to the presence of emotion in speech, the ASR text would contain many speech recognition errors upon recognizing emotional speech. As explained in the previous section, the acoustic and text feature extractors of the basic SER method use BLSTM and a self-attention mechanism. The self-attention mechanism focuses on the important words and speech segments to determine emotion. Focusing on the words incorrectly recognized by ASR results in many incorrectly recognized emotions, thus deteriorating the SER performance. One of the solutions to this problem is to improve ASR performance to be robust to emotions through retraining or fine-tuning as shown in Section II. However, this solution requires a high computational cost and might not effectively improve SER perfor-

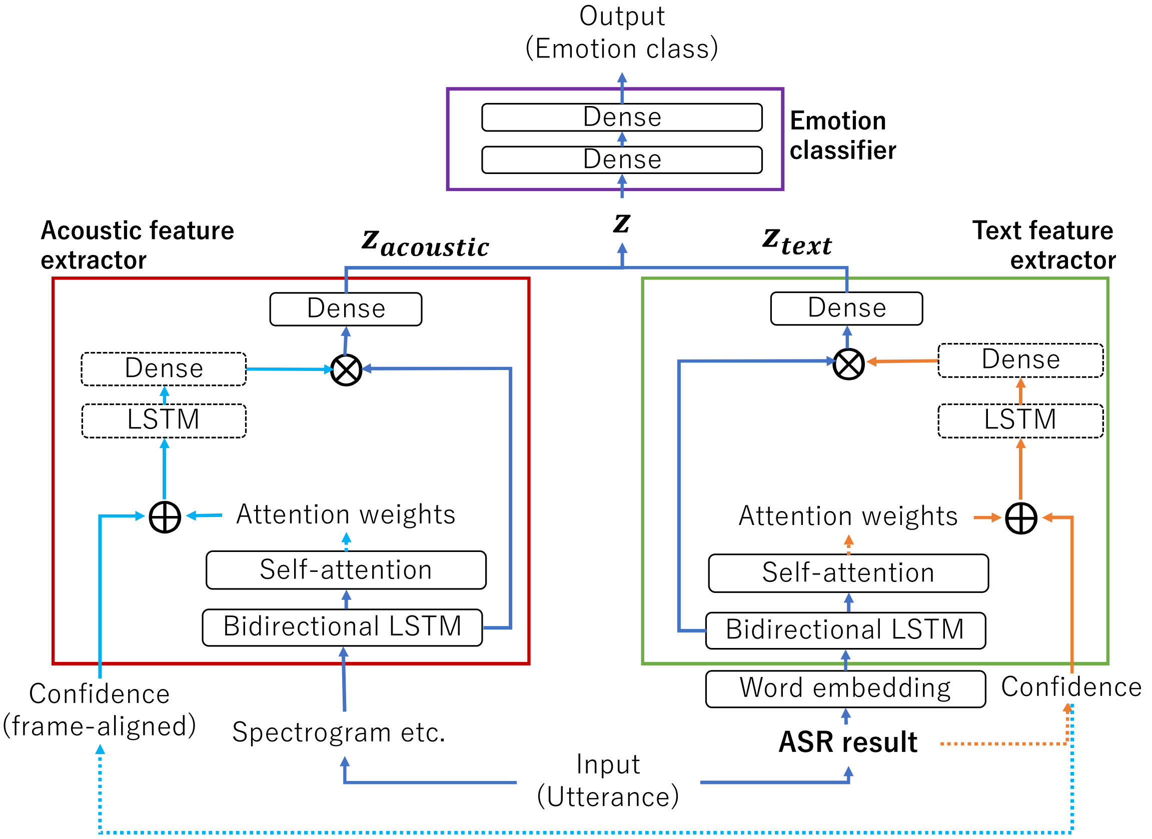


Figure 3.1: Proposed SER method

mance. The essential information regarding the presence of emotion in segments containing speech recognition errors, which can be focused on by a self-attention mechanism in acoustic feature extraction, might be lost owing to ASR being more robust to emotions.

We propose a method to improve the basic SER method by adjusting the self-attention weights using CM and named this method self-attention weight correction (SAWC). It is a critical component in acoustic and text feature extractors. SAWC resolves the issue without retraining or fine-tuning ASR to be robust to emotions. The proposed SER method is illustrated in Figure 3.1.

3.2.2 CM

In the field of speech recognition, one of the most prominently used metrics for ASR reliability is CM [58]. CM indicates how reliable ASR results is. CM falls in the range of 0 to 1; 0 indicates an unreliable result and 1 indicates a reliable result. CM

Table 3.1: Example of ASR results including the speech segments and their CMs

start (s)	duration (s)	result (1st candidate)	CM
0.70	0.65	CLEARLY	1.00
2.27	0.14	YOU	0.97
2.41	0.11	KNOW	0.86
2.52	0.42	D ‘ AVRIL	0.38
2.95	0.19	AGO	0.39
3.15	0.78	SUPERVISOR	1.00
3.93	0.13	OR	0.78
4.06	0.41	SOMETHING	1.00
4.47	0.33	YAH	0.44

has long been used in ASR to evaluate word-level and sentence-level recognition results, accurately discriminating parts that contain possible speech recognition errors. We employ CM in the Kaldi speech recognition toolkit, which is based on the lattice posterior estimation. An example of ASR results and CM aligned for each speech segment and its corresponding spectrogram are illustrated in Table 3.1 and Figure 3.2, respectively.

3.2.3 SAWC using CM

Text attention weight correction

In text features, SAWC aims to mitigate the effects of ASR error on SER performance. SAWC here uses CM to suppress incorrectly recognized words or emphasize the more correctly recognized words. Figure 3.3 illustrates the structures of the text feature extractors of the basic SER method and our proposed method with SAWC. In both text feature extraction structures, the flow begins with inputting the text feature consisting of BERT word embeddings. These features are fed to the text feature extractor using BLSTM and the self-attention mechanism. CM c_i is concatenated with self-attention weights α_i and will be fed to an LSTM network and dense network. The output from the network is then normalized using the

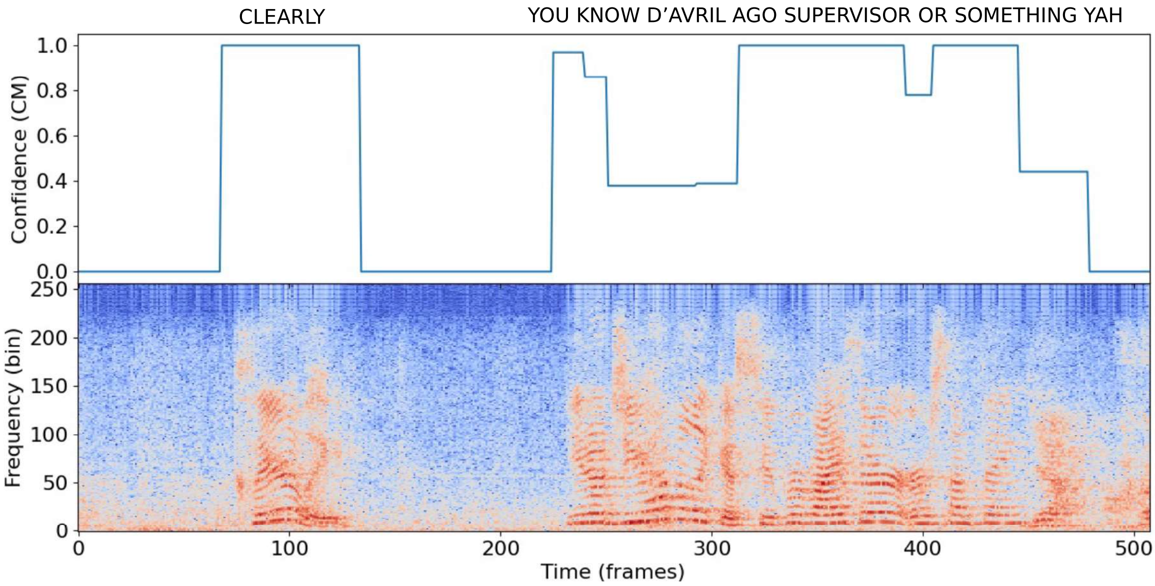


Figure 3.2: CM aligned for each speech segment and its corresponding spectrogram. For display purposes, the CM of the silent segment is set to 0.

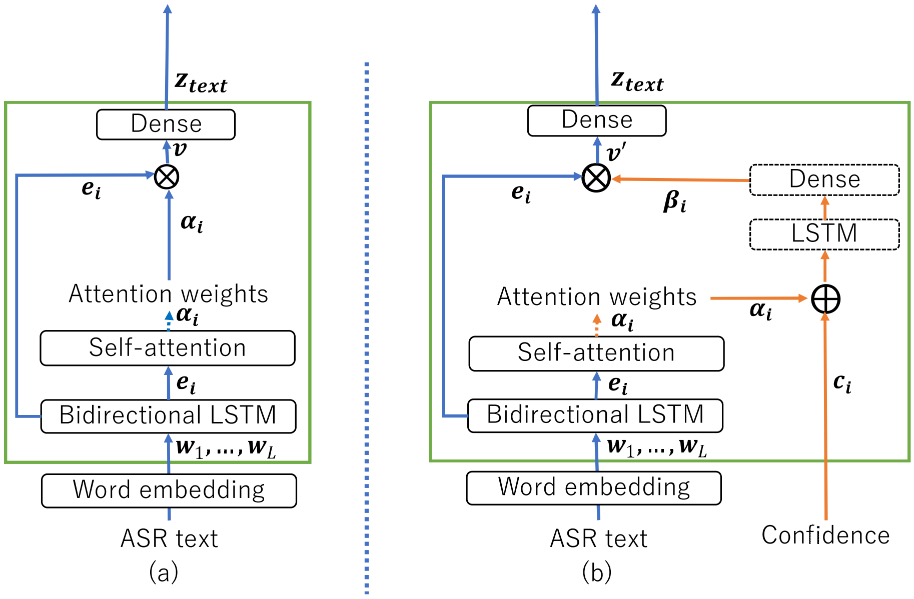


Figure 3.3: Structures of the text feature extractors of (a) the basic SER method and (b) our proposed method with SAWC

softmax function to obtain new attention weights. SAWC is defined as

$$s_i = Dense(LSTM(\boldsymbol{\alpha}_i \oplus c_i)), \tag{3.1}$$

$$\beta_1, ..., \beta_T = softmax(s_1, ..., s_T), \tag{3.2}$$

where $\boldsymbol{\beta}_1$, ..., $\boldsymbol{\beta}_T$ indicate the resulting self-attention weights. Here, the LSTM layer learns and adjusts the attention weights by also considering the CM sequence. $\boldsymbol{\beta}_1$, ..., $\boldsymbol{\beta}_T$ are then used to calculate the weighted sum of the BLSTM outputs defined as

$$\textbf{v'} = \sum_{i=1}^{T} \boldsymbol{\beta}_i \mathbf{e}_i, \tag{3.3}$$

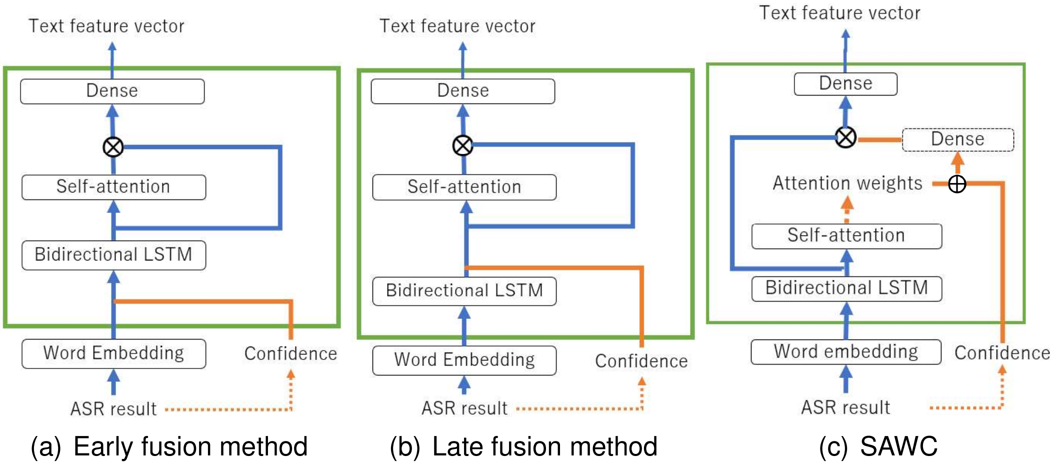
where **v'** represents the new weighted-sum feature, now used as the updated $\textbf{z}_{text}$.

We also investigated three applications of CM in text feature extraction: early fusion, late fusion, and our proposed SAWC. The illustration of the applications can be shown in Figure 3.4. The three applications are explained in the following:

Early fusion (Figure 3.4(a)) CM is treated directly as one of the textual embedding features as they are part of the ASR results and CM is represented by a sequence of weights, which might be suitable for extraction using BLSTM early on. In this method, CM is concatenated after the textual features have gone through BLSTM and before the self-attention mechanism.

Late fusion (Figure 3.4(b))) As CM is small in dimension compared with the textual features, extracting CM sequentially in the early stages might cause early information loss and CM would not have markedly reduce speech recognition errors. Therefore, it would be more effective when used as one feature to consider aside from the extracted sequential text features for the self-attention mechanism. In this method, CM is concatenated after the textual features have gone through BLSTM and before the self-attention mechanism.

SAWC (Figure 3.4(c))) The previous two mechanisms use CM both directly and indirectly as part of the textual features. These mechanisms would require many training data for the incorrectly recognized words and their weighting. To solve this, we concatenate CM directly to the self-attention mechanism weights and update the weights through a fully connected network. By this method, one



(a) Early fusion method (b) Late fusion method (c) SAWC

Figure 3.4: Architecture of the proposed method on the textual feature extraction part of SER

can decrease the CM dependence on the textual feature and train with fewer data. As CM indicates how reliable the ASR result is, CM values can act as another weight for ASR results, similarly to what the self-attention mechanism does for the textual feature extraction. The combination of two different weights provides more precise weights for textual information.

Acoustic attention weight correction

Figure 3.5 illustrates the structures of acoustic feature extractors of the basic SER method and our proposed method with SAWC. In both acoustic feature extractors, the flow begins with inputting the acoustic features MFCC, CQT, and F0. These features are fed to the acoustic feature extractor consisting of BLSTM and the self-attention mechanism.

On the basis of the application of CM in the text features, we apply SAWC to acoustic feature extraction. The aim of SAWC in the acoustic features is to utilize the information contained in the speech segments having a high probability of ASR errors and focus on these segments. The idea is that emotions cause pronunciation changes in specific speech segments resulting in ASR errors; therefore, the speech segments with a high probability of ASR errors contain information helpful in determining emotions. Here, we align the CM from each word to the corresponding speech segments in the acoustic features, as illustrated in Figure 3.2. Since

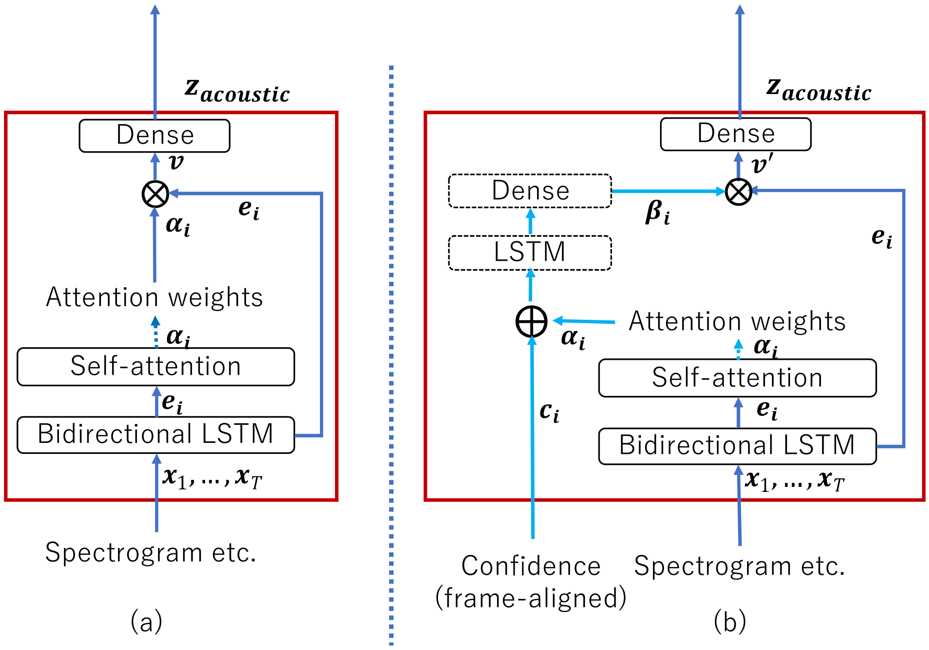


Figure 3.5: Structures of acoustic feature extractors of (a) the basic SER method and (b) our proposed method with SAWC

SAWC needs to have the same sequence length of CM as that of self-attention, for each of the words in the recognition result, the CMs on the start and end times are aligned to the corresponding speech segments. The silent segments are assumed to be correctly recognized, and CM on those segments is set to 1. The aligned CM, which has the same length as the acoustic feature and the calculated self-attention weights, is used for SAWC in the acoustic features. The self-attention weights of the acoustic features are concatenated with the aligned CM and updated, similarly to the calculations in Eqs. 3.1 and 3.2. The updated self-attention weights of the acoustic feature are then multiplied by the output of the BLSTM in the acoustic feature extractor similarly to the calculation in Eq. 3.3, and the updated $\mathbf{z}_{acoustic}$ is produced.

Table 3.2: Dataset specifications

Dataset	IEMOCAP	
Speakers	5 males and 5 females	
Utterance length	1−19 s	
# of utterances	Happy	1689
	Sad	1084
	Neutral	1708
	Angry	1103

Table 3.3: Number of speech data for each emotion class and recording session

Session	Happy	Sad	Neutral	Angry	Total
1	286	194	384	229	1093
2	335	197	362	137	1031
3	322	305	320	240	1093
4	303	143	258	327	1031
5	443	245	384	170	1242
Total	1689	1084	1708	1103	5584

3.3 Experiments and results

3.3.1 Datasets

We trained and evaluated our proposed method using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [61], which is one of the benchmark datasets for emotion recognition. The IEMOCAP dataset was developed when conventional machine learning methods, such as support vector machines, decision trees, logistic regression, and early neural networks, were the most commonly used methods to conduct SER. It is also used to evaluate the recent deep-learning-based methods. The IEMOCAP dataset is available upon request.

The IEMOCAP dataset is recorded from conversations spoken in English, which contain either scripted or improvised emotional speeches divided into five sessions, each containing one male and one female speaker. There are ten speakers (five males and five females) in the IEMOCAP dataset. The length of the IEMOCAP dataset is approximately 12 hours, comprising audiovisual data, including video, speech, motion capture of the face, and text transcriptions. For all experiments in this thesis, we only use the speech data, and the text transcriptions are used only for method comparison. The speech data in IEMOCAP dataset have a sampling rate of 16KHz and a format of 16-bit PCM. Each of the utterances in IEMOCAP datasets has class and dimensional emotion labels. The class labels in IEMOCAP follow the labels from Ekman models (neutral, happy, sad, angry, excited, frustrated, other). The class labeling process is conducted by three evaluators, and the label is decided through majority voting. The dimensional labels in IEMOCAP consist of valence, arousal, and dominance, each ranging from 1 to 5 as a result of Self-Assessment Manikin evaluation. The dimensional labels are decided by averaging the score given by two evaluators. Labeling by class is simpler to implement in practical situations, however, we use only the class labels. In addition, we used the data from four emotion classes (happy, sad, neutral, and angry). To make the dataset conditions similar to those in previous works, we grouped the utterances labeled as excited with the utterances labeled as happy. Throughout this thesis, we assume that each input data belongs to exactly one of the defined emotion classes. The experiments were performed with five-fold cross-validation, with each fold corresponding to each session. The training set consists of speech data from four sessions, and the test set consists of the speech data from the remaining one session, ensuring speaker independence. The details of the dataset used in this study are shown in Table 3.2, and the number of speech data for each emotion class and recording session is shown in Table 3.3.

3.3.2 Input features

Our proposed method receives two different types of input features: acoustic and text features. The acoustic features consist of 33 dimensions, consisting of 20-dimensional Mel-feature cepstrum coefficients (MFCCs), 12-dimensional constant Q-transform (CQT), and one-dimensional fundamental frequency (F0). All of the acoustic features are extracted using Librosa [62]. The text features is taken from the ASR result. First, we conducted ASR on the input speeches using a recognizer based on the Kaldi acoustic recognition toolkit pretrained with the Librispeech dataset. The Librispeech dataset consists of approximately 1000 h of read speeches in English, sampled at 16 kHz and in the format of 16-bit PCM. Next, we encoded ASR texts using BERT pretrained using lower-case English texts. The pretrained BERT consists of 12 layers and 110 M parameters, resulting in 768-dimensional text features. The pretrained BERT is named bert-based-uncased, which is publicly available.

3.3.3 Classifier specifications

The SER consists of an acoustic feature extractor, a text feature extractor, and an emotion classifier. The basic SER method comprises a two-layer BLSTM with 128 units and a self-attention mechanism. Each of the feature extractors consists of BLSTM with 128 units and a self-attention mechanism, resulting in a 128-dimensional vector representation for $\mathbf{z}_{acoustic}$ and $\mathbf{z}_{text}$ for the acoustic and text feature extractors, respectively. SAWC in acoustic and text feature extractors uses BLSTM that receives two-dimensional inputs, providing a two-dimensional output and a dense layer that outputs a one-dimensional output. The resulting intermediate layer representation $\mathbf{z}$ is a 256-dimensional vector consisting of a 128-dimensional vector from each of the acoustic and text features. The intermediate layer representation is fed to the emotion classifier, consisting of two dense layers with (256–64–4) units. In this experiment, we used Adam [63] as the optimizer with a learning rate of 0.0001 and a weight decay of 0.00001. The dropout was set to 0.3. The batch size was set to 40. The results were taken from the highest WA out of 100 epochs.

Figure 3.6 shows the graphical representation of loss and WA during the training on one of the folds in our proposed method using SAWC on both acoustic and text feature extractors. Here, the x-axis represents the number of model training epochs, and the y-axis respectively represents loss and WA in the left and right graphs. In the loss representation, the loss in the training phase decreases until the last epoch, whereas that in the testing phase decreases up to epoch 80 and starts to overfit afterward. On the other hand, in the WA representation, the accuracy in the training phase improves close to 100% until the last epoch, whereas

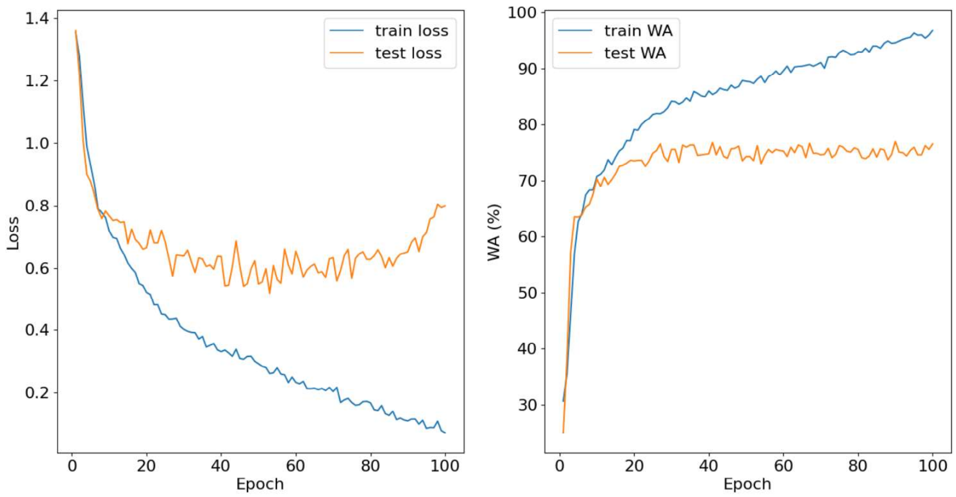


Figure 3.6: Graphical representation of loss and WA during the training of the proposed method

that in the testing phase increases to epoch 40 and tends to plateau afterward.

3.3.4 Computation environment

We conducted both the training and testing phases in the experiments using the GPU. Owing to the large amount of calculation needed during the training phase of our proposed method, it is recommended to use GPU for training. On the other hand, the testing phase can be conducted using either GPU or CPU. The computer used for the experiment has NVIDIA Quadro RTX 6000 GPU with 24 GB of RAM and Intel Core i9-10940X 3.3 GHz CPU with 32 GB of RAM. All the programs were run using the operating system Ubuntu 18.04. All the programs were written in the Python 3 programming language using PyTorch 1.4.0 [64] as the library. The evaluation metrics are calculated using the Scikit-learn [65] toolbox.

We measured the average computational time in our proposed method of SER using SAWC on both acoustic and text features. Our proposed method ran for 41.798 s on GPU for each epoch in the training phase. On the other hand, our proposed method ran for 0.011 s on GPU and 0.375 s on CPU for each utterance in the test phase.

Table 3.4: UA and WA performance comparison of the basic SER method with different input features and our proposed method with different SAWC combinations. Here, AC, TR, AT represents acoustic, transcriptions and ASR text, respectively.

Method	UA (%)	WA(%)
(Basic) AC	61.1	64.3
(Basic) TR	75.5	75.6
(Basic) AT	71.8	71.9
(Basic) AC + TR	78.6	78.4
(Basic) AC + AT	73.9	74.2
(Basic) AC w/0 SAWC + AT with CM early fusion	74.3	74.4
(Basic) AC w/0 SAWC + AT with CM late fusion	74.9	75.2
(Proposed) AC w/o SAWC + AT with SAWC	75.5	75.3
(Proposed) AC with SAWC + AT w/o SAWC	76.1	76.1
(Proposed) AC with SAWC + AT with SAWC	76.8	76.6

3.3.5 Experiment result

Comparison of the application of SAWC

Table 3.4 shows the performance of the basic SER method with different combinations of input features and SAWCs using CM. Here, we experimented with the basic SER method using only acoustic features, text features, and both features. The text features were divided into two types, one using human-based transcriptions provided in the dataset and the other using ASR text. The experiments were run with the same classifier specifications for each input feature.

From the comparison, the performance of the basic SER method using only acoustic features yields UA and WA of 61.1% and 64.3%, respectively. The SER method using only transcriptions yields the UA and WA of 75.5% and 75.6%, whereas that using ASR text as the input yields a lower performance of 71.8% and 71.9% in UA and WA, respectively. The method combining acoustic features and transcriptions achieved the UA and WA of 78.6% and 78.4%, respectively, which are significantly higher than those obtained by the method using only acoustic features or transcriptions. The same increase can also be observed by combining acoustic features and ASR text, achieving the UA and WA of 73.9% and 74.2%, respectively. The decline in the performance of the basic SER method using ASR text as the input text features compared with that using transcriptions is due to the performance deterioration of ASR caused by the presence of emotions in speeches, resulting in incorrect recognition results being used.

Now, we compare the SER performance of the proposed SER method using SAWC with CM on the acoustic and text features with the performance of the basic SER method using acoustic features and ASR text. First, applying SAWC with CM to the text feature only (Acoustic without SAWC + ASR text with SAWC) improved both the UA and WA compared with the basic SER method by 1.6% and 1.1% to 75.5% and 75.3%, respectively. One explanation is that SAWC considers the words with low CMs as speech recognition errors, thereby reducing the attention weights on these words and adjusting the attention weights to focus more on the correctly recognized words. These score is improved from the SER method using CM in early fusion or late fusion.

On the other hand, applying SAWC to the acoustic feature only (Acoustic with SAWC + ASR text without SAWC) also improved both the UA and WA compared with the basic method by 2.2% and 1.9% to 76.1% and 76.1%, respectively. The results show that SAWC can improve acoustic feature extraction by adjusting the importance weights of the speech segments in accordance with CM. One possible explanation is that the speech recognition errors in the speech segments contain information essential to determining the emotion; therefore, emphasizing these parts results in their being considered more in the decision of the emotional output label.

Furthermore, the proposed SER method combining SAWC with CM on acoustic and text features yields the UA and WA of 76.8% and 76.6%, respectively, which is a further performance improvement compared with the method applying SAWC on either acoustic or text feature extractors. The result implies that combining the two types of input enhanced with SAWC can improve the overall SER performance. The performance of our proposed method is close to that of the basic SER method using acoustic features and transcription.

We also evaluated the performance using F-score for each emotion class, as shown in Table 3.5. Overall, the trend of improvement of the F-score for each emotion on different input features is similar to those in UA and WA. The neutral

Table 3.5: F-score performance comparison of the basic SER method with different input features and our proposed method with different SAWC combinations. Here, AC, TR, AT represents acoustic, transcriptions and ASR text, respectively.

Method	F-score (%)			
	Happy	Sad	Neutral	Angry
(Basic) AC	47.6	70.5	56.4	69.7
(Basic) TR	82.8	77.8	65.3	75.4
(Basic) AT	80.4	74.6	58.6	70.9
(Basic) AC + TR	83.4	81.1	68.5	82.2
(Basic) AC + AT	79.7	71.8	64.7	81.1
(Basic) AC w/0 SAWC + AT with CM early fusion	79.0	73.8	64.5	78.9
(Basic) AC w/0 SAWC + AT with CM late fusion	79.5	76.5	64.7	81.1
(Proposed) AC w/o SAWC + AT with SAWC	81.4	76.3	65.8	79.7
(Proposed) AC with SAWC + AT w/o SAWC	82.5	77.2	64.4	80.9
(Proposed) AC with SAWC + AT with SAWC	83.0	77.6	66.2	80.6

class has the lowest F-score among the four emotion classes.

Comparison with state-of-the-art methods

Next, we compare the performance of our proposed method with those of state-of-the-art methods, as shown in Table 3.6. Most reports did not show the results in other metrics such as F-score for each emotion class. Therefore, we compare the performance of our proposed method with those of state-of-the-art methods only in terms of UA and WA. The state-of-the-art methods used for comparison are SER methods using acoustic and text features. The text feature is further separated into transcriptions and ASR text, where ASR text has a word error rate of 43.5%, indicating that many of the speech data contain incorrect text information, whereas the transcriptions can be assumed to have no such errors.

In terms of UA and WA, our proposed method outperforms the state-of-the-art SER methods using acoustic and ASR results as input information. Although our proposed method has yet to achieve the performance of the state-of-the-art SER methods using acoustic and transcriptions, the differences in UA and WA from those of the best state-of-the-art method are 1.6% and 0.9%, respectively, which

Table 3.6: Proposed and state-of-the-art methods

Method	Input	UA (%)	WA (%)
Yoon et al. [15]	Acoustic + Transcriptions	77.6	76.5
Wang et al. [66]	Acoustic + Transcriptions	77.1	76.8
Wu et al. [67]	Acoustic + Transcriptions	78.4	77.5
Kim and Shin [68]	Acoustic + ASR text	68.7	66.6
Xu et al. [69]	Acoustic + ASR text	69.5	70.4
Yoon et al. [15]	Acoustic + ASR text	73.9	73.0
Feng et al. [17]	Acoustic + ASR text	69.7	68.6
Heusser et al. [70]	Acoustic + ASR text	71.0	73.5
Wu et al. [67]	Acoustic + ASR text	75.6	74.7
Proposed method	Acoustic + ASR text + CM	76.8	76.6

means their accuracies are similar.

Confusion matrix of the SER classifier

Figure 3.7 shows the confusion matrices of the basic SER method and the proposed method, which combines both acoustic and text feature extraction with SAWC using CM. Compared with the basic SER method, most emotions except for neutral emotions have gains in the number of correctly classified speeches. The F-score for neutral emotions is only about 66.2%, whereas those for other classes are above 75%. Neutral speeches are mistaken for all classes, especially happy speeches. One possible explanation is that SAWC might have emphasized the speech segments that contain incorrect recognition results, indicating the possible presence of emotion despite the speech being neutral.

Visualization of SAWC

We discuss SAWC with CM in the proposed method through visualization. Here, we take an example of an utterance labeled as angry that had been incorrectly recognized as neutral by the basic SER method; it is the same utterance as the example shown in Table 3.1 and Figure 3.2.

Figure 3.8 shows SAWC with CM in the proposed method applied to the text

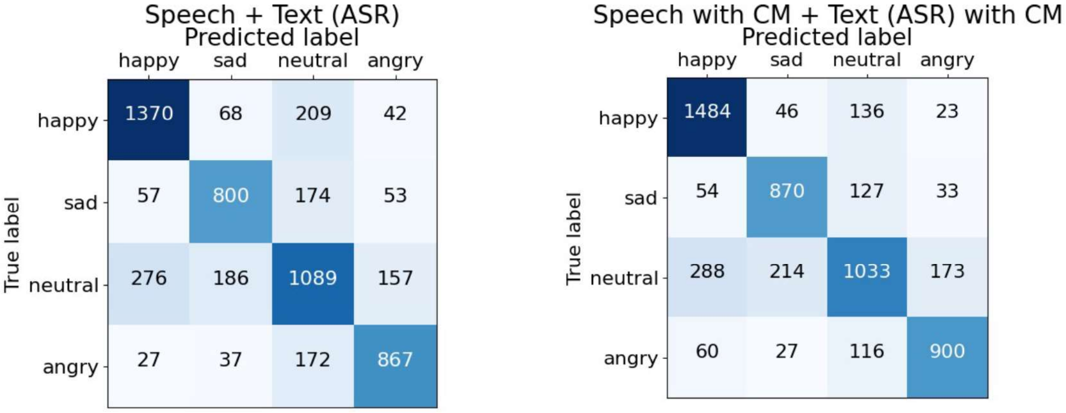


Figure 3.7: Confusion matrices of the basic SER method and our proposed method showing the best result

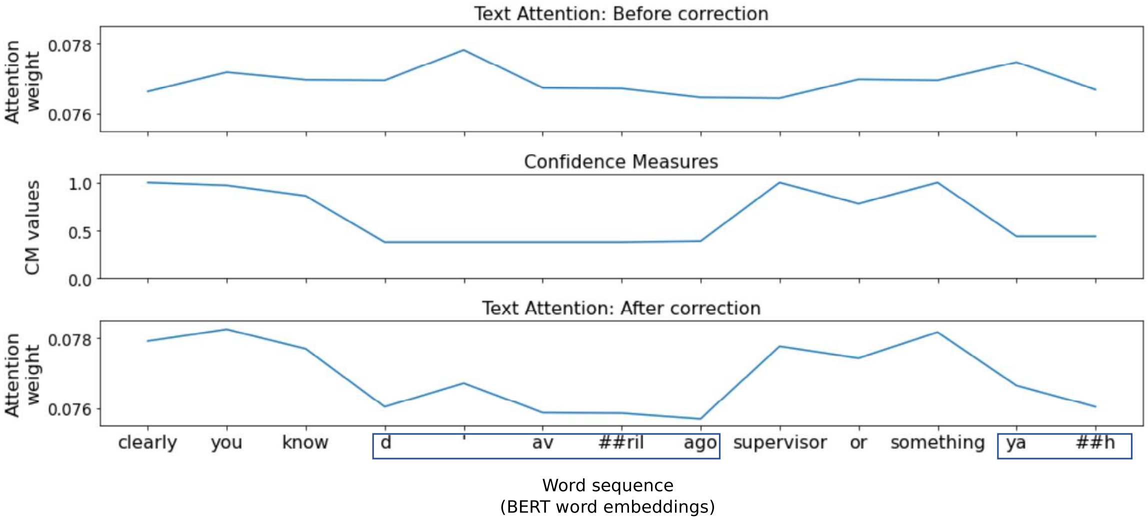


Figure 3.8: SAWC applied to the text feature extractor

feature extractor. The graphs from top to bottom respectively show the text attention weight before the update, CM aligned to each word, and the updated text attention weights. Here, some of the words with low CMs, which are more likely to be incorrectly recognized, were weighted more, whereas the words with high CM or the correctly recognized words were weighted less. By applying our proposed SAWC, we can reduce the text attention weights on the words with low CMs, whereas the words with high CMs are slightly emphasized. The visualization of the updated self-attention weights showed that applying the proposed method to the text features successfully improved the performance by suppressing the effects

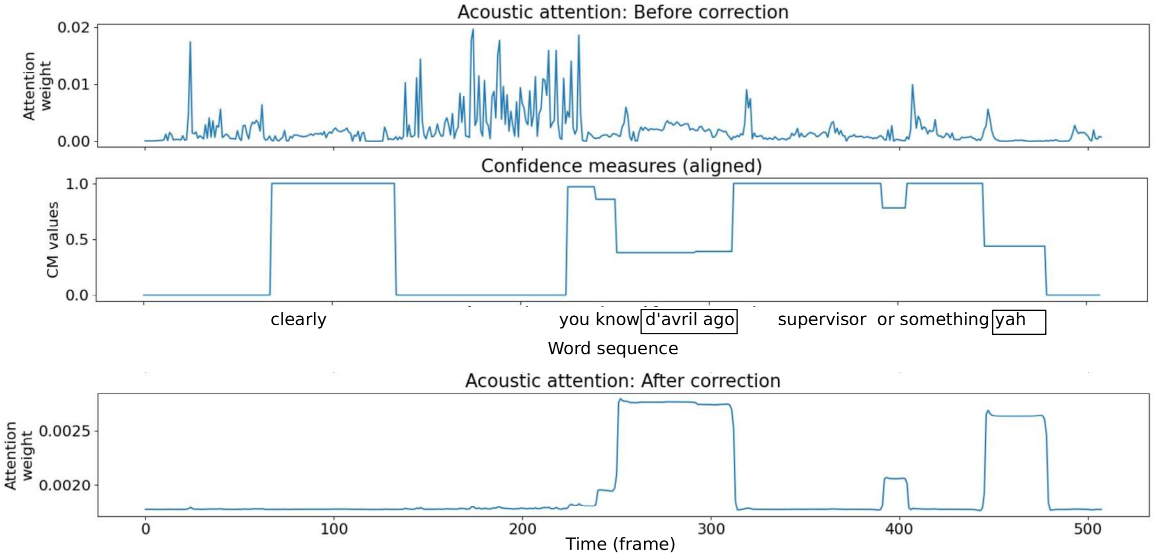


Figure 3.9: SAWC applied to the acoustic feature extractor

of ASR errors.

Figure 3.9 shows SAWC with CM in the proposed method applied to the acoustic feature extractor. The graphs from top to bottom respectively show the acoustic attention weight before the update, CM aligned to the acoustic frames and their corresponding ASR text, and the updated acoustic attention weights. Similarly to Figure 3.2, we set the silent speech segments to 0 for display purposes, in contrast to the experiment where the silent speech segments are set to 1. Here, the plot of updated attention weights resembles the inverted shape of the plot of CM aligned to the acoustic frame, where some parts contain the peak values from the attention weight before applying SAWC. SAWC works differently on the acoustic features from that on the text features. In the acoustic features, self-attention is adjusted to focus on the speech segments containing the speech recognition errors, which would likely have low CM. SAWC still considers the part previously focused on by self-attention, although not as much as CM. The visualization of SAWC confirms that the speech segment emphasized might be affected by emotion, thus containing information essential for SER.

Table 3.7: UA and WA of proposed method and basic SER method using CM as attention weight

Method	UA (%)	WA (%)
(Proposed) AC with SAWC + AT with SAWC	76.8	76.6
AC with CM as attention + AT w/o SAWC	76.3	76.3
AC with CM as attention + AT with SAWC	76.3	76.2

Comparison with the method using CM as attention weight

To extend our results and discussion, we also investigate whether replacing the self-attention mechanism in the acoustic feature extractor with the inversely aligned CM, which is obtained by replacing CM c_1, ..., c_T by $1 - c_1$, ..., $1 - c_T$, yields results similar to those obtained by SAWC with CM. The inversely aligned CM is considered to be due to the updated attention weights in the bottom part of Figure 3.9 showing a similar pattern. Here, we substitute the attention weights with the inversely aligned CM instead of applying the correction to the attention weights.

$$r_1, ..., r_T = softmax(1 - c_1, ..., 1 - c_T) \tag{3.4}$$

$$\textbf{c'} = \sum_{i=1}^{T} r_i \mathbf{e}_i \tag{3.5}$$

In this evaluation, we applied softmax to the inversely aligned CM weights and used them as the attention weights in the acoustic feature extractor.

Table 3.7 and Table 3.8 shows the UA and WA of the proposed method and the method with inversely aligned CM used as attention weights, and F-score of the comparison, respectively. The result shows that the inversely aligned CM used as attention weights yields a slightly lower UA, WA, and overall F-score for each emotion class. Despite the similarity of the updated attention weights to the inversely aligned CM, the attention weights in the proposed method before applying SAWC still hold some significance in the weights of the acoustic feature. It can be inferred that both the weights from the attention mechanism and the CM aligned to the acoustic frames are still essential in determining the important segment in the acoustic features.

Table 3.8: F-score of proposed method and basic SER method using CM as attention weight

Method	F-score (%)			
	Happy	Sad	Neutral	Angry
(Proposed) AC with SAWC + AT with SAWC	83.0	77.6	66.2	80.6
AC with CM as attention + AT w/o SAWC	81.7	77.2	64.0	80.0
AC with CM as attention + AT with SAWC	82.8	76.8	63.6	80.0

3.4 Summary of Chapter 3

In this chapter, we focus on the problem of SER, which is the ASR performance degradation in emotional speeches. Since ASR contains CM, which indicates how reliable the recognition result is and how likely the emotion recognition occurs, we propose using CM to mitigate the effect of speech recognition errors on SER. We investigated the use of CM in SER using acoustic features and ASR results in text, ranging from fusing CM as part of the input feature to using them to correct the attention mechanism, as presented in SAWC directly. The experimental results show that in the text feature, our proposed SAWC yields the highest SER performance compared to the other fusing approaches. We applied the proposed method to SER using acoustic features and ASR results as the text. Our results showed improvement in SER performance, where SAWC reduces the effects of speech recognition errors on the text features while emphasizing the segments containing speech recognition errors as cues for emotions. The method in this study provides a solution to the problem of ASR performance degradation in emotional speeches.

Chapter 4

SER improvement by neutral speech detection using autoencoder and intermediate representation

4.1 Introduction

In practical situations, such as business conversations, neutral speeches make up most of the speech data population. On the other hand, emotional speeches are uncommon and usually indicate some unexpected events or trouble. From the result in Chapter 3, the recognition result of neutral speeches is still low despite the number of data available. As a result, many neutral speeches would be incorrectly recognized as emotional speeches and outnumber the number of correctly recognized emotional speeches. In other words, low neutral speech recognition performance results in SER performance degradation. Therefore, maintaining the recognition performance for each emotion class, including neutral speech, is important to ensure a high-performing SER.

In several practical settings, such as business conversation analysis, most conversations do not contain emotions or are considered neutral. Emotional speeches might indicate potential trouble or unanticipated events in conversations. On the basis of this idea, we focus on the anomaly detection approach that uses only neutral speeches as training data. The anomaly detection model learns the representation of the normal data used in training and identifies anomalous data, that is the data deviating from normal behavior. Here, normal data can be regarded as neutral speeches and anomalous data as emotional speeches. The anomaly detection approach has been investigated in the acoustics domain, namely anomalous sound detection in machines [54]. In this approach, raw spectrograms are used as input and an autoencoder as a reconstructor. Here, the reconstructor is trained to minimize the reconstruction error of normal machine sounds. The normal machine sounds therefore will be successfully reconstructed, whereas the anomalous ones will not be reconstructed well. The anomalies in faulty machines were successfully detected from sounds using this approach.

There are several challenges to applying the anomaly detection approach to speech data in an SER domain. First, the reconstruction of speech is difficult because of the high dimensionality of a spectrogram and the variability of speech length. Second, it is difficult to deal with textual information, which is often utilized as input features in SER, to provide additional hints for possible anomalies in the emotion domain. These two problems can be solved by representing a raw spectrogram and textual information with fixed-length low-dimensional vectors. One way to realize it is to utilize the intermediate layer representation of the SER classifier as the input for anomaly detection, which has been studied in the field of image anomaly detection [71] and proven to be effective in improving the anomaly detection performance.

Considering the condition mentioned above, it would be possible to reformulate the problem as an anomaly detection problem, where neutral speeches are considered normal and emotional speeches are anomalous. This chapter introduces a method using neutral speech detection and applying them to screen neutral speeches. The neutral speech detection proposed uses an autoencoder to reconstruct the intermediate layer representation in SER for neutral speeches.

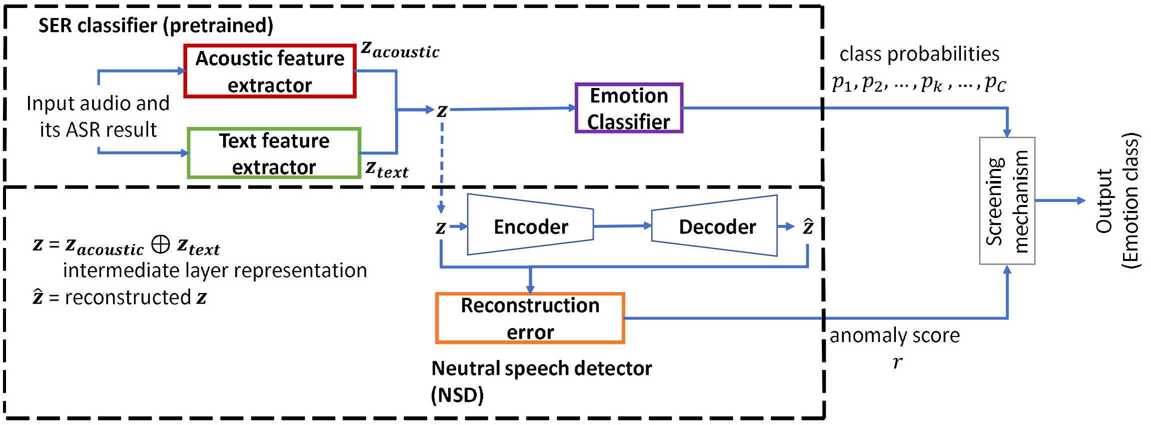

SER classifier (pretrained)
Input audio and its ASR result
Acoustic feature extractor
Text feature extractor
$z_{acoustic}$
z_{text}
z
Emotion Classifier
class probabilities
$p_1, p_2, \cdots, p_k, \cdots, p_C$
Screening mechanism
Output (Emotion class)
z
Encoder
Decoder
$\hat{z}$
$z = z_{acoustic} \oplus z_{text}$ intermediate layer representation
$\hat{z}$ = reconstructed z
Reconstruction error
anomaly score r
Neutral speech detector (NSD)

Figure 4.1: Proposed method using by detecting neutral speeches and using the result to correct the emotion class decision

4.2 Proposed method

4.2.1 Overview

The process flow of the proposed method by detecting neutral speeches and using the result to correct the emotion class decision is illustrated in Figure 4.1. The proposed method consists of the feature extractor, the NSD, and the screening mechanism part. The feature extractor is taken from the pretrained SER classifier explained in Chapter 3, following the steps until the output of the intermediate layer representation **z**, which is then fed to the autoencoder-based NSD. The NSD works by reconstructing **z**, resulting in the reconstructed feature vector $\hat{\mathbf{z}}$ and having the reconstruction error calculated as the anomaly score. When the anomaly score exceeds the decision threshold value, the input speech is classified as emotional (anomalous). Otherwise, it is classified as neutral (normal). Finally, the screening mechanism part decides the emotion class by correcting the neutral class probability based on the anomaly score.

4.2.2 NSD

The NSD of our proposed method consists of a deep autoencoder, which is a deep-learning architecture primarily used to represent higher-dimensional data, typically for efficient dimensionality reduction. In the proposed method, the autoencoder, which consists of the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ in the form of two neural networks, is used to learn the representation of neutral speech through the output of the intermediate layer **z** of the SER classifier. The most attractive feature of our NSD is that it can deal with not only acoustic information but also textual information as the target of reconstruction. **z** is transformed into a compact bottleneck representation **v** with the encoder $\mathcal{E}$, whereas the decoder $\mathcal{D}$ maps back the bottleneck representation into the reconstructed intermediate layer representation $\hat{\mathbf{z}}$. The process is defined as

$$\mathbf{v} = \mathcal{E}(\mathbf{z}|\theta_E), \tag{4.1}$$

$$\hat{\mathbf{z}} = \mathcal{D}(\mathbf{v}|\theta_D), \tag{4.2}$$

where θ_E and θ_D represent the parameter set of an encoder and a decoder respectively. The reconstruction error of the autoencoder, hereby defined as the anomaly score, is computed as the mean square error (MSE)

$$r = \sum_{i=1}^{dim} ||\mathbf{z}_i - \hat{\mathbf{z}}_i||^2, \tag{4.3}$$

where dim is the dimension of **z**. As the autoencoder is trained using only neutral speeches, **z** here represents the intermediate layer representation of neutral speeches.

We investigated the anomaly scores of the neutral speeches in the training data by the reconstruction experiment. As a result, it was found that the distribution of the anomaly scores is asymmetric. Therefore, the neutral/emotional decision is conducted using a decision threshold obtained from the value applied to the percentile point function of the Gamma distribution [72] of the anomaly scores in the training data. The distribution of the anomaly scores from the training data is illustrated in Figure 4.2.

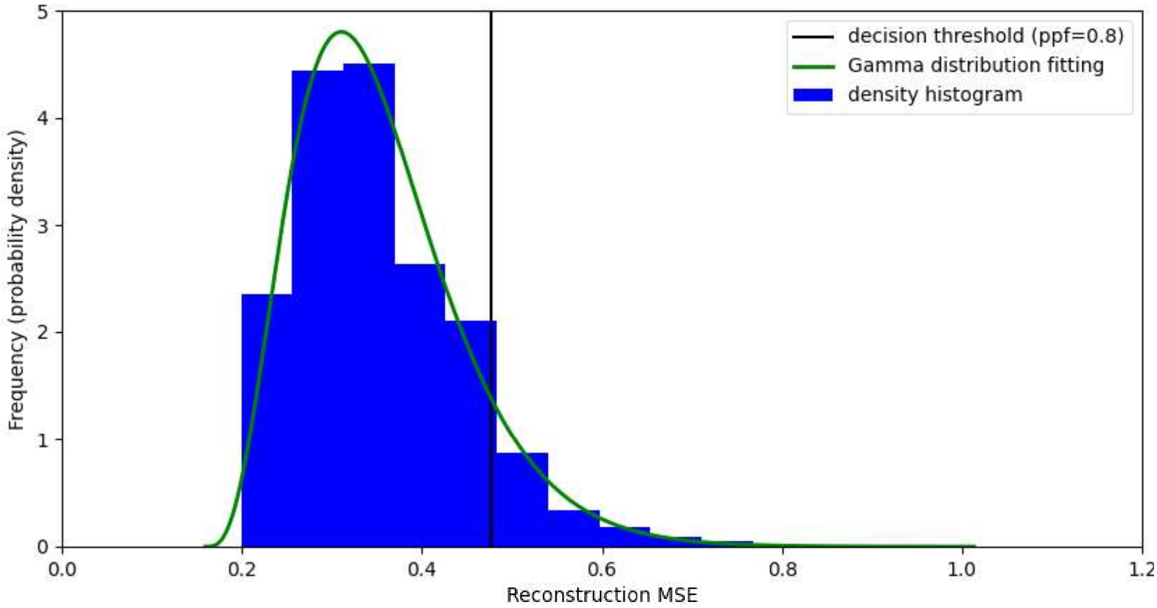


Figure 4.2: Distribution of the anomaly scores of from the training data

4.2.3 Screening mechanism

We introduce a screening mechanism to combine the results of SER and the NSD to improve the SER performance further. In the screening mechanism, the NSD is utilized as the main decider for the final class decision, where speeches detected as neutral are automatically regarded as neutral in the SER result. In the following equation, we will assume $p_1, p_2, ..., p_k, ..., p_C$ as the SER class probability, C as the number of emotion classes, p_k as the neutral probability, and r as the reconstruction error. We compare two screening mechanisms in this study.

Weak screening Speeches detected as neutral by the NSD are regarded as neutral in the final SER class decision. On the other hand, the class decision for speeches not detected as neutral will defer back to the initial SER class probability. This is described as

$$p_k = \begin{cases} 1, & r \leq T, \\ p_k, & r > T \end{cases} \tag{4.4}$$

where T is the decision threshold.

Strong screening This is similar to the weak screening in terms of speeches detected as neutral. However, speeches not detected as neutral are regarded as any class other than the neutral class. In this case, the SER class decision takes the neutral class probability out of the equation and takes the remaining class with the highest probability as the result. The mechanism can be described as

$$p_k = \begin{cases} 1, & r \leq T, \\ 0, & r > T. \end{cases} \tag{4.5}$$

4.3 Experiments and results

4.3.1 Dataset

In this study, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [61], one of the benchmark datasets for emotion recognition, to evaluate the effectiveness of the proposed method. The IEMOCAP dataset consists of scripted and improvised emotional speeches divided into five sessions, each containing one male and one female speaker. There are ten speakers (five males and five females) in the IEMOCAP dataset.

For the pretrained SER classifier, we used the data from four classes (happy, sad, neutral, and angry). To make it similar to previous works, we included the utterances labeled as excited to the utterances labeled as happy. The experiments were performed in five fold cross-validation. The training set comprises four sessions, and the test set comprises the remaining one session to ensure speaker independence. The F-score reported are based on the combined results from all five folds, not from averaging the F-score in each fold. The details of the dataset are shown in Table 4.1

For the NSD, we used the same five fold cross-validation setting with the pretrained SER classifier. However, because we aim to train the neutral data representation, the training set contains only the neutral speeches from each of the four sessions. On the other hand, the test set of the NSD uses the same dataset as that for the SER but with the labels being neutral and the rest of the classes being emotional.

Table 4.1: Dataset specifications

Dataset	IEMOCAP	
Speakers	5 males and 5 females	
Utterance length	1−19 s	
# of utterances	Happy	1689
	Sad	1084
	Neutral	1708
	Angry	1103

4.3.2 Input features

The input features used in this chapter is the same as those explained in Chapter 3, in which the features inputted to the pretrained SER classifier were divided into two parts for acoustic feature extraction and textual feature extraction. For the acoustic feature extraction, we extracted a 33-dimensional feature consisting of 20-dimensional Mel-frequency cepstral coefficients (MFCCs), 12-dimensional constant Q-transform (CQT), and one-dimensional fundamental frequency (F0). All of the acoustic features are extracted using Librosa [62]. For the textual features, first, we conducted ASR on the input speeches using a recognizer pretrained with the Librispeech [60] dataset and Kaldi speech recognition toolkit [59]. Librispeech consists of approximately 1000 hours of speech sampled at 16 kHz. Next, we encoded the ASR texts using pretrained BERT [73], which was trained from lowercase English texts. The pretrained BERT consists of 12-layer and 110M parameters, resulting in 768-dimensional textual features.

4.3.3 SER classifier and NSD specifications

The pretrained SER consists of a feature extractor (acoustic feature extractor and textual feature extractor) and the emotion classifier. The feature extractor used BLSTM with 128 units and a self-attention mechanism with 128 units for the acoustic feature extractor and the additional confidence measure-based correction mechanism for the text feature extractor. The resulting intermediate layer representation $\mathbf{z}$ from the SER is a 256-dimensional vector, consisting of a 128-dimensional vector from each of the acoustic and text features.

The NSD is an autoencoder consisting of nine layers with units (256–128–64–32–16–32–64–128–256). The optimizer is set to Adam [63] with a learning rate of 0.00001 and dropout to 0.2. For the anomaly score calculation, we use the Gamma distribution of the reconstruction error of neutral speeches. The decision threshold is taken from the distributions' percentile point function with a value of 0.8, which yields the best performance among the tested percentile values. We evaluate the results for the NSD using F-score for neutral and the results for the SER using the average unweighted accuracy (UA), average weighted accuracy (WA), and F-score of each emotion class. The pretrained SER model used as the feature extractor was taken from the model that yields the highest WA of the test data out of 100 epochs. Meanwhile, the results for the NSD were taken from the highest neutral F-score of the test data out of 100 epochs.

4.3.4 Experiment Results

Table 4.2 shows the F-score of our proposed method's neutral class in reconstructing the different features. Results of our experiment show that in all the different features reconstructed, the proposed method outperforms the SER method in terms of the neutral F-score. The base SER method (pretrained SER classifier) obtained a neutral F-score of 67.4%. The performance in reconstructing only the textual feature representation and the acoustic feature representation yields neutral F-scores of 61.1% and 76.0%, respectively. On the other hand, the reconstruction of both the acoustic and text features achieves a neutral F-score of 81.0%, which shows significant improvement from the base SER method and the reconstruction of a single feature. One possible explanation is that the intermediate layer output from the base SER method is produced by considering both the acoustic and text features in the training phase. Therefore, it is necessary for the NSD to use the representation from both acoustic and text features to achieve the best reconstruction. From the results, the NSD can be expected to have sufficient reliability as an input to the screening mechanism.

Table 4.3 shows UA and WA of our proposed method and the state-of-the-art SER classifiers with acoustic and text features as the input. Overall, our pro-

Table 4.2: Comparison of reconstructed feature

Reconstructed feature	Neutral F-score (%)
Base SER method	66.2
Text	61.1
Acoustics	76.0
Acoustics + Text	**80.3**

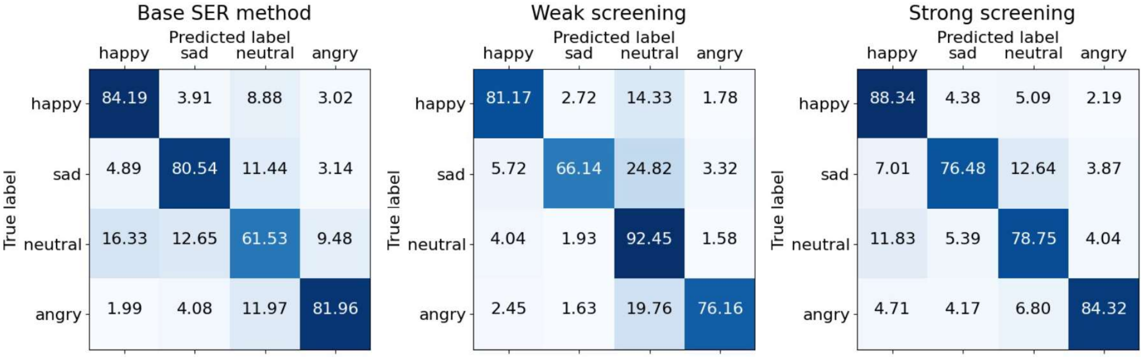
Table 4.3: SER performance comparison (UA, WA) with state-of-the-art methods. The symbol '–' means that the value is not described in the paper.

Method	UA (%)	WA (%)
Neumann and Vu [74]	–	56.1
Feng et al. [17]	69.7	68.6
Siriwardhana et al. [75]	75.5	–
Base SER method (Chapter 3)	76.8	76.6
Wang et al. [66]	77.1	76.8
Priyasad et al. [76]	79.2	80.5
(Proposed) Weak screening	81.0	**84.5**
(Proposed) Strong screening	**82.7**	83.2

posed method using the NSD for strong screening mechanism achieved UA and WA of 82.7% and 83.2% respectively. On the other hand, the use of NSD for the weak screening mechanism achieved UA and WA of 81.0% and 84.5% respectively. These results indicate the significant improvement of our method compared with the base SER method, achieving UA and WA of 75.9% and 76.1%, respectively. Table 4 shows the F-score of each emotion class of our proposed method and F-score reported in the state-of-the-art SER classifiers. The F-score of neutral is improved from 67.4% to 78.6% and 80.3% in the weak and strong screening mechanisms, respectively. The screening mechanism results indicate that the SER performance and the F-score of neutral speeches can be increased simply just by prioritizing the NSD screening result, where neutral speeches are automatically regarded as neutral. In results of the weak screening mechanism, most of the emotional classes show some performance increase because the speeches incorrectly recognized as emotional classes were corrected to neutral. However, the strong screening mechanism further improves the performance for neutral, angry,

Table 4.4: SER performance comparison (F-score) with state-of-the-art methods. The symbol '–' means that the value is not described in the paper.

Method	F-score (%)			
	Happy	Sad	Neutral	Angry
Neumann and Vu [74]	58.2	51.9	52.8	66.5
Feng et al. [17]	69.1	70.5	61.0	77.3
Siriwardhana et al. [75]	77.1	78.4	64.7	81.9
Base SER method (Chapter 3)	83.0	77.6	66.2	80.6
(Proposed) Weak screening	**85.2**	75.6	78.6	82.5
(Proposed) Strong screening	85.0	**78.0**	**80.3**	**85.3**



Base SER method

True label \ Predicted label	happy	sad	neutral	angry
happy	84.19	3.91	8.88	3.02
sad	4.89	80.54	11.44	3.14
neutral	16.33	12.65	61.53	9.48
angry	1.99	4.08	11.97	81.96

Weak screening

True label \ Predicted label	happy	sad	neutral	angry
happy	81.17	2.72	14.33	1.78
sad	5.72	66.14	24.82	3.32
neutral	4.04	1.93	92.45	1.58
angry	2.45	1.63	19.76	76.16

Strong screening

True label \ Predicted label	happy	sad	neutral	angry
happy	88.34	4.38	5.09	2.19
sad	7.01	76.48	12.64	3.87
neutral	11.83	5.39	78.75	4.04
angry	4.71	4.17	6.80	84.32

Figure 4.3: Confusion matrices (in %) for base SER method, weak screening mechanism and strong screening mechanism

and sad speeches by 1.7%–2.8% from those of the weak screening mechanism.

The confusion matrices from the base SER method, the proposed method with the weak screening mechanism, and the proposed method with the strong screening mechanism are shown in Figure 4.3. The strong screening mechanism improves the neutral classification performance from the base SER method by using only the neutral NSD detection result. As a result, it can be observed that the strong screening mechanism tends to improve the performance of both neutral and emotional classes in a well-balanced manner. On the other hand, the weak screening method drastically improves the neutral classification performance by using the NSD detection result and the SER class decision. However, in the weak screening mechanism, it can be observed that there is a tendency to incorrectly classify the emotional speeches as neutral.

4.4 Summary of Chapter 4

In this chapter, we introduced NSD, an SER method with anomaly detection approaches towards neutral speeches, which reconstructs the intermediate layer representation of SER. We propose to use the NSD to screen neutral speeches, and correct the class decision of SER. The results show that the reconstruction of neutral speeches achieved sufficient reliability as an input to the screening mechanism, and the screening mechanism achieved show significant improvement in the F-score of the neutral class and class-average weighted accuracy compared with the state-of-the-art SER classifiers.

Chapter 5

SER based on the reconstruction of acoustic and text features in latent space

5.1 Introduction

SER has been studied intensively, with most of the methods based on a classification approach, outputting the softmax probability of different emotion classes. One limitation is the need to balance the training data since, otherwise, it would result in a classifier being biased toward a certain class. The performance of the class with low performance can be improved by increasing the training data for that class. However, it would not be easy to maintain the balance of the training data. Another limitation is that in the case of additional emotional classes, it would be more difficult to add new emotion classes or to retrain the classifier from scratch.

In this chapter, we propose a novel training strategy for an imbalanced dataset based on reconstruction error. First, we extract the acoustic and text features in latent space by using a pretrained classifier. Second, the extracted features are fed into the reconstructor for each class. Finally, the emotion class is judged to have the lowest normalized reconstruction error. The main advantage of our proposed method is the possibility of training the autoencoder separately for each emotion class, therefore alleviating the need for data balancing. Furthermore, the

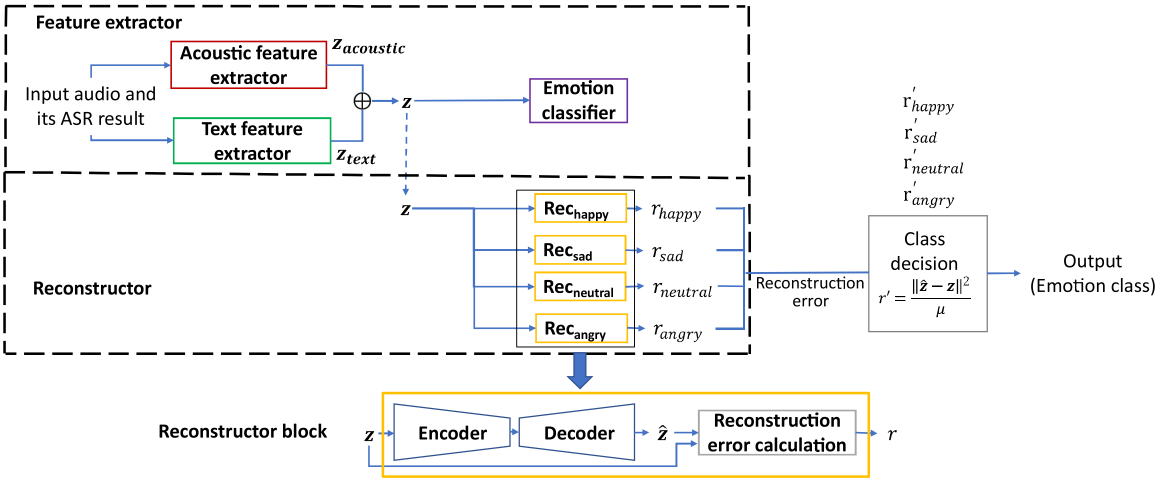


Figure 5.1: Proposed method flow of SER based on the reconstruction of acoustic and text features in latent representation

data augmentation of the latent space can be done specifically for each emotion class without being affected by the others. Finally, we compare the performance characteristics of our proposed method with those of the state-of-the-art SER classification methods.

5.2 Proposed method

Figure 5.1 illustrates our proposed method flow. Our proposed method consists of the feature extractor, the reconstructor for each emotion class, and a class decision. The feature extractor is taken from the acoustic and text feature extractor of the pretrained SER method explained in Chapter 2, resulting in the intermediate layer representation **z**. This is then fed to the autoencoder-based reconstructor for each target emotion class. Each reconstructor reconstructs **z**, resulting in the reconstructed feature vector $\hat{\mathbf{z}}$ and having the calculated reconstruction error. We select the emotion class with the lowest normalized reconstruction error.

5.2.1 Reconstructor

SER methods are mostly based on classification, which is trained to classify emotion classes. However, classification-based methods ideally require all classes to be trained with a balanced number of data to generalize well, which is difficult in practical situations with the data for each defined class being imbalanced. This can be handled by reducing the dependency for each class, which can be realized by reconstructing the features of each class separately and determining if the data is of the target class or not. The reconstruction-based method, on the other hand, has the potential to perform strongly in class-imbalanced situations compared to the classification method and therefore is proposed in this chapter.

Our proposed reconstructor has the architecture of a deep autoencoder and is made of an encoder and a decoder, both in the form of two neural networks. The autoencoder is mainly effective in dimensionality reduction and the representation of higher-dimensional data. The use of an autoencoder in our proposed method is similar to the NSD explained in Chapter 4, which was inspired by the anomaly detection-based approach applied in the anomalous sound detection task [54], which was previously handled using a classification-based approach. In anomalous sound detection, the autoencoder is used to reconstruct the spectrogram and detect the anomalous machine sound. The success of anomalous sound detection has made the autoencoder a commonly used solution for reconstruction and detection tasks. However, applying this approach to the SER domain has several challenges. First, reconstructing a spectrogram for speech is difficult owing to the high dimensions and variable lengths. Moreover, it is necessary to keep the textual information, providing cues to the target emotion.

One possible idea to solve these problems is to use latent representations as the reconstruction target for the autoencoder. The latent representation is usually taken from a pretrained method, has a fixed length, and contains a compact representation of the input features important to the task of the pretrained method. Using the autoencoder on the latent space has been proven effective in improving the anomaly detection performance in image anomaly detection [71]. Following the success of image anomaly detection, we introduced the use of the autoencoder and latent space to both our previous method and our proposed method.

In our reconstructor, the autoencoder learns the representation of speeches in a target emotion class from the latent representation $\mathbf{z}$ of the pretrained SER method. The main strength of our reconstructor is the ability to learn a more specific representation of the acoustic and textual information **z** as the reconstruction target. **z** is a compact representation of features prominent in SER, enabling easy reconstruction. **z** is transformed into a bottleneck representation **v** with the encoder $\mathcal{E}$, whereas the decoder $\mathcal{D}$ maps back the bottleneck representation into the reconstructed latent representation **ẑ**, where θ_E and θ_D represent the parameter set of an encoder and a decoder, respectively. The reconstruction error of the reconstructor is computed using the mean square error (MSE)

$$r = ||\mathbf{z} - \hat{\mathbf{z}}||^2, \tag{5.1}$$

where dim is the dimension of **z**.

5.2.2 Class decision

In reconstruction-based methods and in anomaly detection-based approaches, it is common to determine whether data is of the target class or not on the basis of the decision threshold. For example, the threshold used for the autoencoder in Chapter 4 is obtained from the Gamma distribution [72] percentile value of the training data. If the anomaly score exceeds the decision threshold, the data is considered anomalous, and vice versa. It is possible to use the decision threshold to decide whether the reconstructor result corresponds to the target emotion class or not. However, integrating reconstructor results from each emotion class obtained on the basis of the decision threshold would raise several issues, such as the difficulty in determining the optimum decision threshold value for each reconstructor and classifying a speech into exactly one emotion class. It can be said that the reconstruction-based method has difficulties with class decisions as it is not designed for classification tasks.

To solve these problems, we propose integrating the reconstructor results from each emotion class and determining the emotion class of a speech without using a decision threshold. First, we calculate r', which is the normalized reconstruction error, defined as

$$r' = \frac{||\mathbf{z} - \hat{\mathbf{z}}||^2}{\mu}, \tag{5.2}$$

$$\mu = \frac{1}{N}\sum_{n=1}^{N}||\mathbf{z}_n - \hat{\mathbf{z}_n}||^2, \tag{5.3}$$

where μ represents the average reconstruction error of the training data for the target emotion class, z and $\hat{z}$ represent the acoustic and text features in the latent space and its reconstruction version, and N represents the number of data in the training set. Then, we calculate r' for each target emotion class. Finally, we select the emotion class with the smallest r' as the final result.

5.3 Experiments and results

5.3.1 Datasets

The dataset used to conduct the experiment is IEMOCAP, which is the same dataset as Chapter 3 and 4. There are four classes (happy, sad, neutral, and angry) with utterances labeled as excited also included to the utterances labeled as happy. The experiments were performed in five fold cross-validation. The training set comprises four sessions, and the test set comprises the remaining one session to ensure speaker independence. The F-score reported are based on the combined results from all five folds, not from averaging the F-score in each fold.

The main difference is that the reconstructor from each emotions receive inputs from the specific classes. For the reconstructor, we used the same five fold cross-validation setting with the pretrained SER classifier. The reconstructor is trained separately for each class, therefore the training set contains only speeches on the first four training folds labeled as the designated emotion class for each reconstructor. On the other hand, the test set of each reconstructor uses the same dataset as that for the classifier-based SER but with the labels on the designated class remain the same and the rest of the class is labeled as not of the designated emotion class.

5.3.2 Input features

The input features for this method follow from Chapter 3. The input uses an extracted 33-dimensional acoustic features consisting of 20-dimensional Mel-frequency cepstral coefficients (MFCCs), 12-dimensional constant Q-transform (CQT), and one-dimensional fundamental frequency (F0). All of the acoustic features are extracted using Librosa [62]. In addition, the input also used text features, which were extracted from the ASR result on the input speeches using a recognizer pretrained with the Librispeech [60] dataset and Kaldi speech recognition toolkit [59]. Librispeech consists of approximately 1000 hours of speech sampled at 16 kHz. The text features were encoded from the ASR texts using pretrained BERT [73], which was trained from lower-case English texts. The pretrained BERT consists of 12-layer and 110M parameters, resulting in 768-dimensional textual features.

5.3.3 SER method and reconstructor specifications

The pretrained SER method consists of an acoustic feature extractor, a textual feature extractor, and an emotion classifier. The specifications for the pretrained SER method are the same as the one used in Chapter 3. The feature extractor for both the acoustic and text features used one layer of BLSTM with 128 units, a self-attention mechanism with 128 units for the acoustic feature extractor, and the additional CM-based correction mechanism for the text feature extractor. The resulting latent representation $\mathbf{z}$ from the base SER method is a 256-dimensional vector consisting of a 128-dimensional vector from each of the acoustic and text features. In the pretrained SER method, $\mathbf{z}$ is fed to an emotion classifier consisting of a fully connected network with (256–64–4) units. The output of the pretrained SER method is assigned to the softmax probability of the four emotion classes, where the highest probability identifies the final emotion class. The pretrained SER method uses softmax cross-entropy as the loss function. The optimizer is set to Adam [63] with a learning rate of 0.0001 and a dropout of 0.2.

The reconstructor for each emotion class is an autoencoder consisting of nine layers with (256–128–64–32–16–32–64–128–256) units. The autoencoder is trained with mean squared error as the loss function. The optimizer is set to Adam with a

Table 5.1: SER performance characteristics (UA, WA) of our proposed method and state-of-the-art methods. The symbol '–' means that the value is not described in the paper.

Method	UA (%)	WA (%)
Neumann and Vu [74]	–	56.1
Feng et al. [17]	69.7	68.6
Chen et al. [14]	75.3	74.3
Siriwardhana et al. [75]	75.5	–
Base SER method (Chapter 3)	76.8	76.6
Wang et al. [66]	77.1	76.8
Proposed method	**77.8**	**77.8**

Table 5.2: SER performance characteristics (F-score) of our proposed method and state-of-the-art methods. The symbol '–' means that the value is not described in the paper.

Method	F-score (%)			
	Happy	Sad	Neutral	Angry
Neumann and Vu [74]	58.2	51.9	52.8	66.5
Feng et al. [17]	69.1	70.5	61.0	77.3
Siriwardhana et al. [75]	77.1	78.4	64.7	**81.9**
Base SER method (Chapter 3)	83.0	77.6	66.2	80.6
Proposed method	**84.7**	**79.3**	**71.6**	74.3

learning rate of 0.00001 and a dropout of 0.2. We evaluate the SER performance using the average unweighted accuracy (UA), average weighted accuracy (WA), and F-score of each emotion class. The pretrained SER method used as the feature extractor was taken from the highest WA of the test data out of 100 epochs. In addition, the reconstructor for each emotion class, which is later integrated into our proposed method, was taken from the highest F-score for the emotion class of the test data out of 200 epochs.

5.3.4 Experiment results

Table 5.1 shows the UA and WA of our proposed method and state-of-the-art SER methods. Our proposed method achieved UA and WA of 77.8% and 77.8%, re-

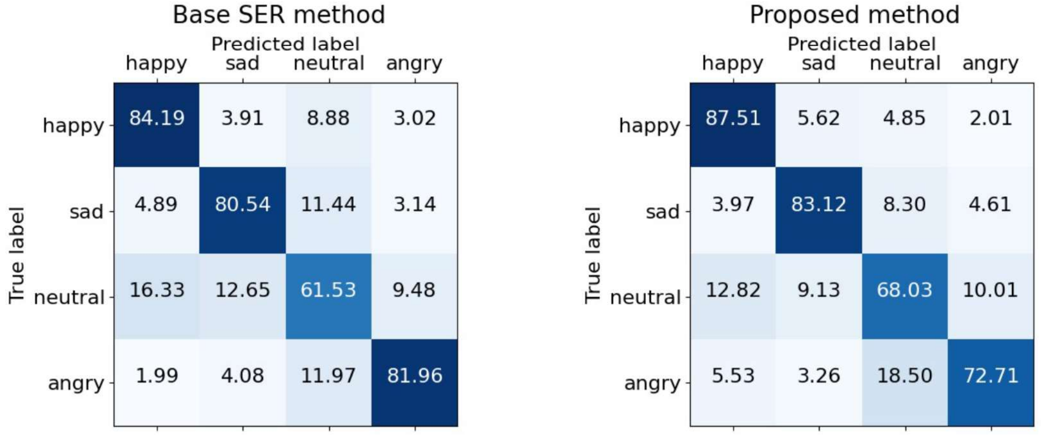


Figure 5.2: Confusion matrices (in %) for base SER method and our proposed method

spectively. These results indicate UA and WA improvement of 1.9% and 1.7% over the base SER method, which achieved UA and WA of 75.9% and 76.1%, respectively. Our proposed method outperformed the state-of-the-art SER methods in terms of UA and WA.

Table 5.2 shows the F-scores of our proposed method and state-of-the-art SER methods. Overall, the F-score shows the effectiveness of our proposed method of SER based on reconstruction error, which is superior to the F-score of happy, sad, and neutral classes of the base SER method by 3.1%–4.2%. Figure 5.2 shows the confusion matrices for the base SER method and our proposed method. Similar to the result of the F-score, our proposed method offers increased performance for the happy, sad, and neutral classes in terms of accuracy. On the other hand, there is a deterioration in the performance of the angry class, with many of the speeches being incorrectly recognized as a neutral class. One of the possible reasons is that the intermediate layer representation for the angry class is slightly harder to reconstruct compared to the other emotion classes. Another possibility is the effect of the resulting intermediate feature representation created from the feature extractor trained in Chapter 3, favoring the improvement of the happy and sad class over the angry class. Regardless, these results imply that the classifier-based SER methods and our proposed method have comparable overall performance but different strengths regarding performance for each emotion class. Therefore, integrating our proposed method with the classifier-based SER methods would potentially boost the SER performance.

5.4 Summary of Chapter 5

In this chapter, we introduced a novel training strategy to improve classification tasks on imbalanced data conditions that frequently occur in practical situations. We proposed a method by reconstructing the intermediate layer representations of SER independently for each class using autoencoder, and then combine the result using a simple class decision method that selects the emotion with the lowest normalized reconstruction error. We made a comparison between SER using classifier-based methods and our proposed method. The experimental results show that our proposed method shows the performance improvement in comparison to the classifier-based methods.

Chapter 6

Conclusion

6.1 Summary of this thesis

In this dissertation, we explained the importance of speech emotion recognition and presented the challenges related to its practical use in real life. The first challenge addressed is the effects of ASR performance degradation due to emotions and their influence on SER. The second challenge is training SER in practical conditions due to the imbalanced number of speeches in each emotion class, with many of the speeches being neutral. The third challenge is obtaining the balanced amount of emotional data used to train a robust SER model. We introduced SER methods based on classification and reconstruction to tackle these challenges.

In Chapter 3, we addressed the problem of ASR performance degradation in SER using acoustics and text information. The result from ASR is beneficial for practical use, especially for obtaining text information in real time. We proposed a BLSTM- and self-attention-based SER method using SAWC with CM. The idea is to mitigate the effects of ASR error on text feature extraction by reducing the weight of the words with low CM, which are likely to be a speech recognition error, and to emphasize the speech segments with low CM as segments with a higher probability of containing emotion in the acoustic feature. By utilizing the information from CM in ASR results and SAWC, our method can improve the SER performance. Our method does not require fine-tuning of ASR to be robust to emotion; this fine-tuning incurs a high computational cost and might lose the important emotional cues in the segments with speech recognition errors. The experimental results demonstrated that our proposed method using SAWC in acoustic and text feature extractors improved the classification performance parameters UA and WA by 2.9% and 2.4%, respectively, compared with those of the basic SER method. In addition, our proposed method outperformed the state-of-the-art SER methods.

In Chapter 4, we described the problem of SER for practical use, such as in business situations where most of the speech is neutral. Although classification-based SER methods have achieved high overall performance, these methods tend to have lower performance for neutral speeches. To solve the problem and improve the SER performance, we propose a neutral speech detector based on the anomaly detection approach, which uses an autoencoder, the intermediate layer output of a pretrained SER classifier, and only neutral data for training. The intermediate layer output of a pretrained SER classifier enables the reconstruction of both acoustic and text features, which are optimized for SER tasks. We then propose the combination of the SER classifier and the NSD used as a screening mechanism for correcting the class probability of the incorrectly recognized neutral speeches. Experimental results using the IEMOCAP dataset confirmed that the NSD has sufficient reliability as an input to the screening mechanism, and the screening mechanism achieved show significant improvement of 12.9% in the F-score of the neutral class to 80.3%, and 8.4% in the class-average weighted accuracy to 84.5% compared with the state-of-the-art SER classifiers.

In Chapter 5, we observed the problem related to the difficulty of obtaining emotional speeches, particularly in business conversations. Most of the existing SER methods are the classification-based method, which has some limitations, including maintaining the balance of the training data and the difficulty in handling additional emotional classes; it would be more difficult to add new emotion classes or to retrain the classifier from scratch. We proposed a novel training strategy for an imbalanced dataset based on the reconstruction error of acoustic and text features in latent space. The reconstructor for different emotion classes, including the neutral class, is used. The proposed method selects the emotion class with the lowest normalized reconstruction error as the SER result. Unlike the classifier approach, one reconstructor is dedicated to each emotion class and trained using only the data of the target emotion class. The main advantage of this method is the possibility of training the reconstructors of each emotion and augmenting the data for each emotion independently, reducing the dependency on the amount of data required for each emotion. Our experimental result obtained using the IEMOCAP datasets confirmed that our proposed method based on the reconstruction approach improves the overall SER performance by 1.9% on the UA and 1.7% on the WA for the IEMOCAP dataset, slightly outperforming most state-of-the-art methods based on the classification approach. In addition, the proposed method improved the SER performance for most emotion classes in terms of the F-score.

6.2 Future perspectives

We have proposed several methods for addressing problems related to SER and its improvement in fulfilling the conditions for practical situations, but there are still room for improvement.

- Reconstruction methods are, as the name implies, basically aimed to reconstruct the input feature representation with a minimum difference. Therefore, the reconstruction methods employed in this study can be used for data augmentation. Applying the proposed method to augment SER intermediate layer representation would be prospective to improve SER performance.

- The method of reconstructing the latent representation of acoustic and text features proposed in the thesis still relies on simple class decisions. This method is still unstable depending on the performance of the reconstructors of each emotion class. Other class decision methods, such as machine learning or neural network-based class decision methods, should be observed to alleviate this problem.

- Practical uses consider the computational time and cost, which should be as low as possible. However, adding the reconstruction step for the SER incurs extra computational time. Since the SER used to extract the feature in the proposed method is pretrained, and the reconstructors are independent for each emotion, the proposed method would benefit from parallel processing for each component, which we will investigate in the future.

- The proposed method has been verified using the IEMOCAP dataset, in which the data is slightly imbalanced. However, as the proposed methods based on reconstruction in Chapter 4 and Chapter 5 are to solve the imbalanced test dataset and imbalanced training dataset, respectively, the effectiveness of the proposed methods should be verified in the imbalanced dataset condition.

Bibliography

[1] L. Vidrascu and L. Devillers, “Real-life emotion representation and detection in call centers data,” *Proc. ACII 2005*, pp. 739–746, 2005.

[2] P.-S. Chiu, J.-W. Chang, M.-C. Lee, C.-H. Chen, and D.-S. Lee, “Enabling intelligent environment by the design of emotionally aware virtual assistant: A case of smart campus,” *IEEE Access*, vol. 8, pp. 62 032–62 041, 2020.

[3] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, “Survey on emotional body gesture recognition,” *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 505–523, 2018.

[4] H. Zhang, A. Jolfaei, and M. Alazab, “A face emotion recognition method using convolutional neural network and image edge computing,” *IEEE Access*, vol. 7, pp. 159 081–159 089, 2019.

[5] P. Zhong, D. Wang, and C. Miao, “Eeg-based emotion recognition using regularized graph neural networks,” *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1290–1301, 2022.

[6] P. Sarkar and A. Etemad, “Self-supervised learning for ecg-based emotion recognition,” *Proc. ICASSP 2020*, pp. 3217–3221, 2020.

[7] L. Canales and P. Martínez-Barco, “Emotion detection from text: A survey,” *Proc. JISIC 2014*, pp. 37–43, 2014.

[8] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, “A comprehensive review of speech emotion recognition systems,” *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.

[9] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, “Automatic emotion recognition using prosodic parameters,” *Proc. Interspeech 2005*, pp. 493–496, 2005.

[10] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, “Acoustic emotion recognition: A benchmark comparison of performances,” *Proc. ASRU 2009*, pp. 552–557, 2009.

[11] H. M. Fayek, M. Lech, and L. Cavedon, “Towards real-time speech emotion recognition using deep neural networks,” *Proc. ICSPCS 2015*, pp. 1–5, 2015.

[12] W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and recurrent neural networks,” *Proc. APSIPA 2016*, pp. 1–4, 2016.

[13] P. Tzirakis, J. Zhang, and B. W. Schuller, “End-to-end speech emotion recognition using deep neural networks,” *Proc. ICASSP 2018*, pp. 5089–5093, 2018.

[14] M. Chen and X. Zhao, “A Multi-Scale Fusion Framework for Bimodal Speech Emotion Recognition,” *Proc. Interspeech 2020*, pp. 374–378, 2020.

[15] S. Yoon, S. Byun, S. Dey, and K. Jung, “Speech emotion recognition using multi-hop attention mechanism,” *Proc. ICASSP 2019*, pp. 2822–2826, 2019.

[16] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, “Multimodal sentiment analysis: Addressing key issues and setting up the baselines,” *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[17] H. Feng, S. Ueno, and T. Kawahara, “End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model,” *Proc. Interspeech 2020*, pp. 501–505, 2020.

[18] S. Amiriparian, A. Sokolov, I. Aslan, L. Christ, M. Gerczuk, T. Hübner, D. Lamanov, M. Milling, S. Ottl, I. Poduremennykh *et al.*, “On the impact of word error rate on acoustic-linguistic speech emotion recognition: an update for the deep learning era,” *arXiv preprint arXiv:2104.10121*, 2021.

[19] S. Sahu, V. Mitra, N. Seneviratne, and C. Y. Espy-Wilson, “Multi-modal learning for speech emotion recognition: An analysis and comparison of asr out-

puts with ground truth transcription." *Proc. Interspeech 2019*, pp. 3302–3306, 2019.

[20] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using gans for speech emotion recognition." *Proc. Interspeech 2019*, pp. 171–175, 2019.

[21] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and cognition*, vol. 17, no. 2, pp. 484–495, 2008.

[22] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[23] R. E. Plutchik and H. R. Conte, *Circumplex models of personality and emotions*. American Psychological Association, 1997.

[24] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proc. National academy of sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[25] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[26] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[27] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[28] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," *Proc. ICSLP'96*, vol. 3, pp. 1970–1973, 1996.

[29] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." *Proc. Interspeech 2005*, pp. 1517–1520, 2005.

[30] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[31] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PLoS ONE*, vol. 13, no. 5, pp. 1–35, 2018.

[32] R. Lotfian and C. Busso, “Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings,” *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2019.

[33] V. Petrushin, “Emotion in speech: Recognition and application to call centers,” *Proc ANNIE 1999*, vol. 710, p. 22, 1999.

[34] Y. Li, P. Bell, and C. Lai, “Fusing asr outputs in joint training for speech emotion recognition,” *Proc. ICASSP 2022*, pp. 7362–7366, 2022.

[35] C. M. Lee, S. S. Narayanan, and R. Pieraccini, “Combining acoustic and language information for emotion recognition,” *Proc. ICSLP 2022*, pp. 873–876, 2002.

[36] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, “On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues,” *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010.

[37] B. Schuller, G. Rigoll, and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture,” *Proc. ICASSP 2004*, vol. 1, pp. I–577, 2004.

[38] Z.-J. Chuang and C.-H. Wu, “Multi-modal emotion recognition from speech and text,” *Proc. IJCLCLP 2004*, vol. 9, no. 2, pp. 45–62, 2004.

[39] C. Cortes and V. Vapnik, “Support vector machine,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[40] G. E. Hinton, “Connectionist learning procedures,” *Machine learning*, pp. 555–610, 1990.

[41] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” *Proc. Interspeech 2014*, 2014.

[42] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, “Deep neural networks for acoustic emotion recognition: Raising the benchmarks,” *Proc. ICASSP 2011*, pp. 5688–5691, 2011.

[43] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, A. N. Vembu, and R. Prasad, “Emotion recognition using acoustic and lexical features.” *Proc. Interspeech 2012*, vol. 2012, pp. 366–369, 2012.

[44] D. Griol, J. M. Molina, and Z. Callejas, “Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances,” *Neurocomputing*, vol. 326, pp. 132–140, 2019.

[45] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[46] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, “Speech emotion recognition from spectrograms with deep convolutional neural network,” *Proc. PlatCon 2017*, pp. 1–5, 2017.

[47] S. Zhang, X. Zhao, and Q. Tian, “Spontaneous speech emotion recognition using multiscale deep convolutional lstm,” *IEEE Transactions on Affective Computing*, 2019.

[48] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[49] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” *Proc. ICLR 2017*, 2017.

[50] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” *Proc. ICASSP 2017*, pp. 2227–2231, 2017.

[51] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, “Context-aware attention mechanism for speech emotion recognition,” *Proc. SLT 2018*, pp. 126–131, 2018.

[52] D. Luo, Y. Zou, and D. Huang, “Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition,” *Proc. Interspeech 2018*, pp. 152–156, 2018.

[53] Y. Li, T. Zhao, and T. Kawahara, “Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning,” *Proc. Interspeech 2019*, pp. 2803–2807, 2019.

[54] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, “Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212–224, 2018.

[55] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[56] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[57] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, “Reconstruction-error-based learning for continuous emotion recognition in speech,” *Proc. ICASSP 2017*, pp. 2367–2371, 2017.

[58] H. Jiang, “Confidence measures for speech recognition: A survey,” *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

[59] D. Povey, A. K. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The kaldi speech recognition toolkit," *Proc. ASRU 2011*, 2011.

[60] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *Proc. ICASSP 2015*, pp. 5206–5210, 2015.

[61] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[62] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," *Proc. 14th python in science conference*, vol. 8, pp. 18–25, 2015.

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[64] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[66] Y. Wang, G. Shen, Y. Xu, J. Li, and Z. Zhao, "Learning mutual correlation in multimodal transformer for speech emotion recognition," *Proc. Interspeech 2021*, pp. 4518–4522, 2021.

[67] W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," *Proc. ICASSP 2021*, pp. 6269–6273, 2021.

[68] E. Kim and J. W. Shin, “Dnn-based emotion recognition based on bottleneck acoustic features and lexical features,” *Proc. ICASSP 2019*, pp. 6720–6724, 2019.

[69] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, “Learning alignment for multimodal emotion recognition from speech,” *Proc. Interspeech 2019*, pp. 3569–3573, 2019.

[70] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, “Bimodal speech emotion recognition using pre-trained language models,” 2019.

[71] V. Abdelzad, K. Czarnecki, R. Salay, T. Denounden, S. Vernekar, and B. Phan, “Detecting out-of-distribution inputs in deep neural networks using an early-layer output,” *arXiv preprint arXiv:1910.10307*, 2019.

[72] K. O. Bowman and L. R. Shenton, “Gamma distribution,” *International Encyclopedia of Statistical Science*, pp. 573–575, 2011.

[73] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *Proc. NAACL 2019*, pp. 4171–4186, 2019.

[74] M. Neumann and N. T. Vu, “Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” *Proc. Interspeech 2017*, pp. 1263–1267, 2017.

[75] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, “Jointly fine-tuning” bert-like” self supervised models to improve multimodal speech emotion recognition,” *Proc. Interspeech 2020*, pp. 3755–3759, 2020.

[76] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, “Attention driven fusion for multi-modal emotion recognition,” *Proc. ICASSP 2020*, pp. 3227–3231, 2020.

Appendix A

List of Publications

Journal papers

[1] **Jennifer Santoso**, Takeshi Yamada, Kenkichi Ishizuka, Taiichi Hashimoto, and Shoji Makino, “Speech emotion recognition based on self-attention weight correction for acoustic and text features,” *IEEE Access*, vol. 10, pp. 115732–115743, 2022.

Peer-reviewed conference papers

[1] **Jennifer Santoso**, Takeshi Yamada, Shoji Makino, Kenkichi Ishizuka, Takekatsu Hiramura, “Speech emotion recognition based on attention weight correction using word-level confidence measure,” *Proc. Interspeech 2021*, pp. 1942–1947, 2021.

[2] **Jennifer Santoso**, Takeshi Yamada, Kenkichi Ishizuka, Taiichi Hashimoto, Shoji Makino, “Performance improvement of speech emotion recognition by neutral speech detection using autoencoder and intermediate representation,” *Proc. Interspeech 2022*, pp. 4700–4704, 2022.

[3] **Jennifer Santoso**, Rintaro Sekiguchi, Takeshi Yamada, Kenkichi Ishizuka, Taiichi Hashimoto, Shoji Makino, “Speech emotion recognition based on the reconstruction of acoustic and text features in latent space,” *Proc. APSIPA*

ASC 2022, pp. 1678–1683, 2022.

Non peer-reviewed conference papers

[1] **Jennifer Santoso**, Takeshi Yamada, Kenkichi Ishizuka, Taiichi Hashimoto, Shoji Makino, “Neutral/emotional speech classification using autoencoder and output of intermediate layer in emotion recognizer,” Spring Meeting of the Acoustical Society of Japan, 1-3P-3, pp. 1005–1008, 2022.