university of groningen

campus fryslân

# SERCT: End-to-End Speech Emotion Recognition based on CNN-Transformer

Siqi Zheng

university of groningen

campus fryslân

## University of Groningen - Campus Fryslân

# SERCT: End-to-End Speech Emotion Recognition based on CNN-Transformer

**Master's Thesis**

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
**Dr. Shekhar Nayak** (Voice Technology, University of Groningen)
with the second reader being
**Dr. XXX** (Voice Technology, University of Groningen)

**Siqi Zheng (S5407958)**

June 11, 2024

# Acknowledgements

First and foremost, I want to express my heartfelt gratitude to my family, particularly my parents, for their unwavering support during my journey to the University of Groningen. Without their assistance, I would not have had the opportunity to pursue my education at this esteemed institution and attain my current master's degree.

I am profoundly grateful to my supervisor, Dr. Shekhar Nayak, whose invaluable guidance, support, and expertise have been instrumental throughout the research process. His insightful feedback and unwavering encouragement have significantly shaped this thesis. My sincere appreciation goes to both Dr. Shekhar Nayak for his steadfast commitment and dedication to my academic journey. His mentorship has been a cornerstone of my development as a researcher.

Additionally, I am grateful to the faculty and staff of Voice Technology for their support and the resources they provided, which were crucial for the successful completion of this study.

Finally, I want to thank my family and friends for their endless encouragement and understanding throughout this journey. Their unwavering support has been a constant source of motivation for me.

# Abstract

Speech Emotion Recognition (SER) plays a crucial role in various applications such as human-computer interaction, emotion-driven systems, and sentiment analysis. Traditional SER approaches involve complex feature extraction and analysis processes, which often require domain knowledge and manual intervention. In recent years, the development of end-to-end systems has emerged as a promising approach to address these challenges by eliminating the need for explicit feature engineering.

In this thesis, we propose a architecture called CNN-Transformer (SERCT) for end-to-end Speech Emotion Recognition. The CNN-Transformer architecture combines the strengths of Convolutional Neural Networks (CNNs) and Transformers, enabling a more convenient and efficient framework for building SER applications. CNNs are known for their ability to capture local patterns and relationships in speech signals, while Transformers excel at modeling long-range dependencies and capturing global contextual information.

The proposed CNN-Transformer architecture consists of two main components: a CNN module and a Transformer module. The CNN module performs initial feature extraction and captures local acoustic patterns, while the Transformer module captures high-level contextual information and long-range dependencies. The two modules are integrated in a sequential manner, allowing the network to learn discriminative representations directly from raw speech signals without the need for handcrafted features.

To evaluate the effectiveness of the proposed CNN-Transformer architecture on SER task, we conducted experiments on two widely used speech emotion datasets named TESS and ESD. The experimental results demonstrate that the approach we proposed almost outperforms all baselines, while significantly reducing the complexity of feature engineering and analysis. The end-to-end nature of CNN-Transformer architecture enables more efficient training and deployment of SER systems, making it a promising solution for real-world applications. The experimental code and the code for the application we built can both be available at https://github.com/0105zhengsiqi/SERCT.

***Index Terms*** - Speech Emotion Recognition, End-to-End, CNN, Transformer

# Contents

**1 Introduction 8**

**2 Literature Review 11**

2.1 End-to-End SER . . . . . 11

2.2 CNNs & RNNs . . . . . 12

2.3 Attention Mechanism . . . . . 13

2.4 Transformer . . . . . 13

**3 Research Question and Hypothesis 16**

**4 Methodology 19**

4.1 Overview of the Model . . . . . 19

4.2 Pre-processing . . . . . 21

4.3 CNN Module . . . . . 22

4.3.1 Convolutional Layers . . . . . 23

4.3.2 Activation Functions and Normalization . . . . . 24

4.4 Transformer Module . . . . . 25

4.4.1 Transformer Architecture Overview . . . . . 25

4.4.2 Stacked Transformer Layers . . . . . 27

4.4.3 Deformable Speech Transformer . . . . . 27

**5 Experiments Setup 30**

5.1 Dataset . . . . . 30

5.2 Baselines . . . . . 31

5.3 Implementation Details . . . . . 32

5.4 Metrics . . . . . 33

**6 Results 36**

6.1 Main Results . . . . . 36

6.2 Language Generalization . . . . . 37

6.3 Effect of Transformer Layer Depth . . . . . 37

6.3.1 Performance Analysis . . . . . 38

6.3.2 Loss Analysis . . . . . 39

6.4 Ablation Study . . . . . 39

**7 Discussion 42**

7.1 Validation of All Hypotheses . . . . . 42

7.2 Ethical Considerations . . . . . 43

7.3 Limitations . . . . . 44

**8 Conclusion 47**

8.1 Main Contributions . . . . . 47

8.2 Limitations and Future Work . . . . . 47

**References** **49**

**Appendices** **55**
A Demonstrator . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

# 1 Introduction

Emotion is a crucial aspect of human communication, conveying valuable meta-information beyond the literal meaning of spoken words. Speech Emotion Recognition (SER) aims to automatically identify the emotional state of a speaker from their vocal cues, with wide-ranging applications in areas such as human-computer interaction, mental health assessment, and customer service. The ability to accurately recognize and understand emotions expressed in speech can greatly enhance communication between humans and machines, leading to improved user experiences and more personalized services. Despite the significant progress made in this field, accurately recognizing subtle and variable emotional expressions remains a challenging task.

Most SER systems rely on complex and time-consuming feature extraction processes, which involve extracting a wide range of acoustic features such as pitch, intensity, and spectral characteristics. These features are then fed into machine learning models for emotion classification. However, this approach not only requires expert knowledge and domain-specific feature engineering but also hampers the efficiency and scalability of the system.

The advent of deep neural networks (DNNs) has revolutionized the field of SER, enabling the transition from hand-crafted acoustic features, such as low-level descriptors (LLDs), to data-driven deep emotion embeddings. Various DNN architectures, including convolutional neural networks (CNNs), long short-term memory (LSTMs), time-delay neural networks (TDNNs), residual networks (ResNets), and dilated residual networks (DRNs), have been extensively explored in recent SER research (Zhang et al., 2018; Etienne et al., 2018; Meng et al., 2016; Li et al., 2019a).

A common approach in modern SER systems is to adopt a two-module pipeline: a feature extraction module that generates temporally-relevant acoustic features, and an aggregation module that pools these features into a compact, global representation (i.e., the emotional embedding) at the utterance level. This approach aims to effectively capture both the local and global contextual information in the speech signal, which is crucial for accurately recognizing the underlying emotional state.

Recent SER studies have focused on developing novel module architectures to yield more effective emotional embeddings. For example, the mixed CNN-LSTM architecture proposed in Etienne et al. (2018) leverages CNNs to extract feature sequences from raw spectrograms and LSTMs to aggregate long-term feature dependencies. Similarly, both 1D and 2D CNN-LSTM-based SER systems have been explored, with the 2D variant showing superior performance by capturing both local correlations and global contextual information from spectrograms (Zhao et al., 2019). Motivated by the attention mechanism (Bahdanau et al., 2016), RNN-attention-based SER systems (Mirsamadi et al., 2017) have also been proposed, where RNNs extract temporal features and attention mechanisms focus on emotionally salient features. Temporal modeling approaches that learn deep emotion features directly from raw waveforms have also been investigated (Sarma et al., 2018). Furthermore, the incorporation of multi-head attention (MHA) in architectures like CNN-BLSTM-attention (Li et al., 2019b) and DRN-attention (Li et al., 2019a) has demonstrated the benefits of exploring different representation subspaces at different positions. Contextual information has also been explored, with methods like contextual LSTM attention (Xie et al., 2019) considering the dependencies among surrounding utterances to enhance the emotional recognition performance.

Despite the progress made in previous studies, there is still room for improvement in SER performance. Interestingly, in other research domains, it has been frequently reported that replacing recurrent architectures with stacked transformer layers (STLs) can yield significant performance improvements. For example, STLs-based acoustic models (Wang et al., 2020) have achieved the best results on the Librispeech benchmark, and STLs-based architectures have also shown substantial performance improvements in the domains of question answering (Tran and Niedereée, 2018) and image captioning (Lee et al., 2018; Zhu et al., 2018).

In light of these observations, this thesis aims to investigate the potential of STLs in the context of SER and proposes an innovative approach to SER by introducing an end-to-end architecture based on a combination of Convolutional Neural Networks (CNNs) and Transformers. The primary motivation behind this research is to overcome the limitations of feature-engineering-based SER systems and provide a streamlined and efficient solution for speech emotion recognition.

The main contributions of this thesis are summarized as follows:

- We proposed an end-to-end CNN-Transformer architecture for SER, eliminating the need for manual feature extraction. This allows the model to learn the most relevant features directly from the raw speech data, simplifying the overall system.
- We conducted extensive experiments to thoroughly validate the effectiveness of the proposed model. Through rigorous testing, we demonstrate the strong performance of the CNN-Transformer approach for SER tasks.
- We explore the indispensable roles of the CNN module and Transformer module in the overall architecture. Furthermore, we investigate the tradeoffs between the number of Transformer layers, performance and duration of training or inference. This analysis provides valuable insights for balancing model complexity and efficiency.
- Additional experiments were designed to verify the outstanding zero-shot generalization capabilities of the CNN-Transformer model across different languages. This indicates the robustness and broad applicability of the proposed model, making it suitable for deployment in multilingual scenarios.
- An application based on the CNN-Transformer model has been successfully developed.

The rest of the paper is organized as follows. Section 2 provides a comprehensive literature review, summarizing previous research and existing knowledge in the field. Section 3 defines the research question and hypothesis, setting the stage for the investigation. Section 4 details the methodology employed to conduct the research, explaining the approaches and techniques used. Section 5 describes the experimental setup, outlining the configurations and conditions under which the research was carried out. Section 6 presents the results obtained from the experiments. Section 7 discusses these results, interpreting their implications and integrating them with existing literature. Finally, Section 8 concludes the paper, summarizing the findings and suggesting potential areas for future research.

# 2 Literature Review

In the section of Literature Review, we delve into key computational methods and frameworks that have significantly shaped the field of speech emotion recognition (SER). This section is crucial for setting the context of our research within the broader landscape of machine learning technologies applied to audio signal processing and emotion analysis.

The exploration begins with a comprehensive overview of subsection 2.1 End-to-End SER. These systems represent a pivotal shift in how emotional cues are extracted and interpreted directly from raw speech data without the need for manual feature selection. This approach leverages the intrinsic power of deep learning architectures to learn complex patterns and nuances in emotional expressions, providing a more holistic and potentially more accurate emotion recognition process.

Following this, the discussion transitions to the integration of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) in SER applications. CNNs are adept at handling spatial hierarchies in data, making them suitable for processing the spectrogram representations of speech signals, whereas RNNs excel in capturing the temporal dynamics inherent in speech. The synergy of these networks in hybrid architectures allows for robust feature extraction and sequential data processing, crucial for understanding the temporal progression of emotional states.

The subsection 2.3 introduces the Attention Mechanism, a transformative concept that has improved the performance of neural networks by enabling them to focus on salient parts of the speech signal that are more indicative of emotional content. This mechanism addresses the limitations of earlier models by providing a means to weigh the importance of different temporal segments, enhancing the model's ability to discern subtle emotional cues within a conversation.

Finally, the subsection 2.4 on Transformers discusses how this architecture has revolutionized SER by providing a framework that handles long-range dependencies more effectively than traditional RNNs. With its self-attention layers, the Transformer model can analyze entire sequences of data in parallel, dramatically improving the speed and efficiency of training models on large datasets. This capability is particularly beneficial in capturing the complex and varied emotional nuances present in human speech.

## 2.1 End-to-End SER

Although many SER systems using conventional machine learning techniques rely on handcrafted acoustic features (e.g., Mel-frequency cepstral coefficients (MFCC) features) as input, the selection of these features can introduce bias and require additional time. Recently, there has been increasing interest in DNNs that directly process raw audio signals, eliminating the need for manual feature extraction and potentially providing more comprehensive representations for SER tasks.

**Train-from-scratch Models** Tzirakis et al. (2018) (denoted as Tzirakis18) introduced a neural network structure that integrates convolution and recurrent layers, specifically for ongoing recognition of emotional states such as arousal and valence. This model incorporates three convolutional layers to analyze raw audio data, followed by a two-layer Long Short-Term Memory (LSTM) system to handle context-based dependencies. Similarly, Zhao et al. (2019) (denoted as Zhao19) proposed a network with four convolutional layers and a duo of LSTM layers targeting discrete emotion recognition. Both Tzirakis18 and Zhao19 were examined and contrasted in Tzirakis et al. (2021a) focusing on the auditory domain. Meanwhile, Zhang et al. (2019b) applied an attention mechanism alongside multi-task learning techniques to boost the accuracy of emotion recognition through enhanced audio features. Sun (2020) added a gender-specific module to a residual CNN framework to refine the accuracy of their emotion recognition model. Furthermore, Tzirakis et al. (2021b) merged the high-level semantic inputs from the Word2Vec and Speech2Vec with low-level paralinguistic characteristics obtained through CNN layers to elevate the model's effectiveness.

**Foundation Models** Extensive pre-training on a substantial volume of data allows foundation models to acquire robust representations that simultaneously encapsulate the acoustic and linguistic characteristics of speech. This enables the models to effectively transfer their acquired knowledge to a diverse range of speech processing applications.

The Wav2vec 2.0 (Baevski et al., 2020) framework comprises two key components - a Convolutional Neural Network (CNN) module that serves as a feature encoder to derive latent representations of speech, and a Transformer module that captures the global contextual dependencies within the speech data. Wav2vec 2.0 employs a self-supervised learning approach, where the model is trained on a vast corpus of speech data using a contrastive objective to learn highly discriminative representations. Due to its exceptional capability in extracting robust speech representations, the Wav2vec 2.0 model has been widely adopted and leveraged in a variety of SER research efforts (Chang et al., 2023; Wagner et al., 2023; Chen and Rudnicky, 2023; Pepino et al., 2021; Cai et al., 2021).

The Hidden-Unit BERT (HuBERT) framework (Hsu et al., 2021) adopts a similar architectural approach to that of Wav2vec 2.0. HuBERT also employs a self-supervised learning strategy, but with the addition of auxiliary tasks, such as frame-level feature prediction, which further enhances the model's ability to learn integrated acoustic and linguistic representations from raw speech data. Morais et al. (2022) fine-tuned the pre-trained HuBERT model and leveraged the generated utterance embeddings as an upstream input for their emotion classification task.

Building upon the promising results of self-supervised pre-training in speech processing, the WavLM framework (Chen et al., 2022) aims to tackle a comprehensive range of speech processing tasks by leveraging large-scale unlabeled data. To better capture the sequential information inherent in audio data, WavLM further extends the HuBERT approach by incorporating a gated relative position bias within its Transformer structure, and also augmenting the training data through an utterance mixing strategy. Feng et al. (2024) utilized the WavLM model for extracting speech embeddings and also explored the trustworthiness of these learned representations.

## 2.2 CNNs & RNNs

Speech is inherently a continuous time-series signal, and both CNN and RNN have been prominently utilized in Speech Emotion Recognition (SER). Drawing inspiration from CNN applications in computer vision, networks such as AlexNet (Krizhevsky et al., 2017) and ResNet (He et al., 2016) have demonstrated significant success in image classification tasks, prompting their exploration in SER. A notable advancement is the Global Aware Multi-scale (GLAM) neural network proposed by Zhu and Li (2022), which employs a global perception fusion module to extract multi-scale feature representations, emphasizing emotional content. Additionally, the multi-time-scale (MTS) method introduced in Guizzo et al. (2020) enhances CNNs by adjusting and resampling the convolutional kernel along the temporal axis, thereby improving temporal flexibility. Liu et al. (2020) introduced a local-global-aware deep representation learning system combining CNNs and Capsule Networks to capture both local and comprehensive global information.

RNNs excel at modeling the temporal aspects of speech and capturing long-term dependencies within the speech signal. The HNSD network proposed by (Cao et al., 2021) efficiently integrates static and dynamic features of SER, utilizing LSTM to encode these features and a gated multi-features unit (GMU) for frame-level feature fusion to represent emotional intermediates. Furthermore, Xu et al. (2020) proposed the hierarchical grained and feature model (HGFM), leveraging recurrent neural networks to process both discourse-level and frame-level speech information. Recognizing that CNNs are adept at capturing local feature details while RNNs excel at modeling temporal information, numerous studies have integrated these approaches to achieve remarkable results. For instance, Li (2021) utilized a BiLSTM neural network to extract location information from MFCC and VGGish features, subsequently fusing these features for emotion prediction. Similarly, Liu and Wang (2021) combined triplet loss with CNN-LSTM models to enhance the discriminative power of sentiment information, achieving outstanding experimental results. Additionally, Zou et al. (2022) employed CNN, BiLSTM, and wav2vec2 to extract various levels of speech information, including MFCC, spectrogram, and acoustic data, fusing these features through an attention mechanism. Lastly, Zhang et al. (2019a) utilized CNNs to learn segment-level features in spectrograms, while a deep LSTM model captured the temporal dependencies between speech segments.

## 2.3 Attention Mechanism

In recent years, attention mechanisms have garnered significant attention across various fields, enhancing task processing by concentrating on the most pertinent information among numerous inputs. Squeeze-and-Excitation (SE), a channel attention mechanism introduced by Hu et al. (2018), assigns weights to individual channels and adaptively recalibrates their feature responses. Woo et al. (2018) developed a convolutional block attention module that integrates spatial and channel dimensions to achieve superior attention results. Moreover, researchers have leveraged deep learning methods for speech feature extraction and enhancement of feature maps through attention mechanisms. An innovative approach based on attention pooling was introduced by Li et al. (2018), effectively combining class-agnostic bottom-up attention maps and class-specific top-down attention maps, surpassing traditional average and maximum pooling techniques. Kwon et al. (2021) proposed a self-attentive module (SAM) for SER systems, utilizing a multilayer perceptron (MLP) to capture global channel information and a specialized dilated CNN to generate an attentional map for both channels, significantly reducing computational and parametric overhead. Furthermore, a spectrotemporal-channel (STC) attention mechanism was introduced by Guo et al. (2021), which creates attention feature maps across three dimensions: time, frequency, and channel. Xi et al. (2022) employed an attention mechanism based on time and frequency domains to incorporate long-distance contextual information.

## 2.4 Transformer

Transformers have seen rapid advancement in the realm of natural language processing (NLP) in recent years, achieving remarkable success. Their powerful ability to capture global information has led to their extension into speech processing fields. An end-to-end speech emotion recognition model was introduced to enhance global feature representation by employing stacked transformer blocks at the model's end (Wang et al., 2021). Another study leveraged multiple models, incorporating residual BLSTM to boost learning capabilities and proposed a convolutional neural network coupled with an E-transformer module for learning both local and global information Hu et al. (2022).

Recently, transformer-based self-supervised methods have been applied to speech, yielding significant successes in automatic speech recognition (ASR) with models such as wav2vec (Schneider et al., 2019), VQ-wav2vec (Baevski et al., 2019), and wav2vec 2.0 (Baevski et al., 2020). These models are also being employed in speech emotion recognition through transfer learning techniques. For instance, a pre-trained wav2vec 2.0 model is used as the network input, combining outputs from multiple layers to generate a richer speech feature representation (Pepino et al., 2021). Another approach introduced a multi-task learning (MTL) framework that utilizes the wav2vec 2.0 model for feature extraction while simultaneously training for speech emotion classification and text recognition Cai et al. (2021).

In the domain of computer vision, the Vision Transformer (ViT) (Dosovitskiy et al., 2020) pioneered the direct application of transformers to image patch sequences, marking a groundbreaking shift. ViT's superior structure and reduced computational resource consumption offer advantages over convolutional neural networks. Similar methods have been adopted in speech processing. For example, ViT was adapted to speech and enhanced by focusing on spectrogram properties, leading to the development of a separable transformer (SepTr) that processes tokens concurrently within the same frequency interval Ristea et al. (2022). Additionally, an approach that integrates ViT with original log-Mel spectrograms and first-order time-frame and frequency bin differential log-Mels 3D features was applied to improve infant cry recognition Xu et al. (2022).

# 3 Research Question and Hypothesis

In light of the preceding discussion, the research question at the core of this study can be formulated as follows:

> **Does the SERCT demonstrate effective performance and surpass all baseline approaches?**

From which the following subquestions are derived:

- Using the English ESD and TESS data respectively to train the baseline from scratch, what would the performance of the baseline be across various evaluation metrics?
- Does SERCT outperform the baseline across all the evaluation metrics?
- Are both the CNN module and Transformer module in SERCT necessary? What would happen if one of them was removed?
- What number of Transformer layers in SERCT achieves the best balance between performance and duration of training or inference?
- How is SERCT's generalization capability on other languages? Is it the best?

The baselines we will use are CNN, CNN-BiLSTM, LSTM, MLP and SVM (Hearst et al., 1998), which have been widely employed in speech emotion recognition tasks and have generally demonstrated robust capabilities in capturing relevant acoustic and linguistic features (Jain et al., 2020; Mou et al., 2021; Abdul Qayyum et al., 2019). When trained on the well-curated ESD (Zhou et al., 2021) and TESS (Pichora-Fuller and Dupuis, 2020) datasets, which provide a diverse and high-quality collection of emotional speech samples, these baseline models are expected to leverage the richness of the data to learn effective representations for emotion classification. Thus, baseline models trained from scratch using the English ESD and TESS datasets would exhibit promising performance across various evaluation metrics, such as Weighted Accuracy (WA), Unweighted Accuracy (UA), and Weighted F1-score (WF1). However, it is important to note that the baseline models, while exhibiting good performance, may not necessarily outperform the more advanced SERCT which combines the strengths of convolutional neural networks and transformer architectures, could potentially offer superior performance by effectively capturing both local and global dependencies in the speech data, thereby yielding more accurate and robust emotion recognition results. This is our hypothesis regarding the main question and the first two sub-questions.

As for the third sub-question, both the CNN module and Transformer module in SERCT are necessary components, and the removal of either one would likely lead to a decline in the model's performance. The CNN module is responsible for extracting low-level features from the input speech data, which are then passed to the Transformer module for higher-level processing and emotion classification (Wang et al., 2021). Without the CNN module, the Transformer would lack the necessary input features to effectively learn the complex patterns and relationships underlying speech emotion recognition. Conversely, the Transformer module is crucial for modeling the long-range dependencies and contextual information in the speech data, which the CNN module alone may struggle to capture. Therefore, the synergistic interaction between the two modules is essential for SERCT to achieve optimal speech emotion recognition accuracy. Removing either component would result in a suboptimal model that is unable to fully leverage the strengths of both architectures, leading to a degradation in the overall system performance.

Regarding the fourth sub-question, the number of Transformer layers in SERCT should strike a balance between performance and the duration of training or inference. As the number of Transformer layers increases, the model's performance is likely to improve due to the enhanced capability of the Transformer architecture to capture long-range dependencies and complex patterns in the speech data. However, adding more Transformer layers also increases the model complexity, which can lead to longer training and inference times. Therefore, the ideal number of Transformer layers would be the one that provides the best trade-off between achieving high performance while maintaining reasonable training and inference durations. This hypothesis suggests that there exists an optimal point where the performance gains from additional Transformer layers are outweighed by the corresponding increase in computational cost, and the goal would be to identify this sweet spot through empirical evaluation.

For the last sub-question, given that SERCT is an end-to-end speech emotion recognition model that leverages a combination of Convolutional Neural Networks (CNNs) and Transformers, it is reasonable to hypothesize that SERCT's generalization capability on other languages is likely to be superior compared to other speech emotion recognition models. The use of Transformers, which have demonstrated strong cross-lingual transfer learning abilities (Lu et al., 2024), coupled with the robustness of CNNs in feature extraction, suggests that SERCT should be able to effectively generalize to speech emotion recognition tasks in languages beyond the ones it was trained on. Additionally, the end-to-end nature of SERCT may further contribute to its ability to capture language-agnostic emotional cues, leading to the best generalization performance among comparable models. However, this hypothesis would need to be empirically verified through rigorous experimentation and evaluation on diverse language datasets to confirm SERCT's superior cross-lingual generalization capabilities.

# 4 Methodology

In this section, we delineate the methodology underpinning our research, structured into four distinct subsections to provide clarity and depth to our approach. Initially, in subsection 4.1, we introduce the theoretical framework and the structural blueprint of the proposed model (i.e., SERCT), outlining its objectives, scope, and the key features. This provides a foundation for the subsequent, more technical discussions. Following this, subsection 4.2 details the specific techniques and procedures applied to prepare the data before it is fed into the neural network. This includes steps like determining audio frequency, padding and the handling of missing data, which are crucial for the model's performance, accuracy and training efficiency. The subsection 4.3, CNN Module, focuses on the implementation of Convolutional Neural Network layers within the proposed model. Here, we discuss the architecture of these layers, their role in feature extraction from the input data, and how they are optimized to enhance model efficiency. Finally, subsection 4.4 describes the integration of Transformer architecture, emphasizing its advantages in modeling long-range dependencies within data. We elaborate on the configuration of the attention mechanisms and how they contribute to the model's ability to discern complex patterns and relationships.

## 4.1 Overview of the Model

As shown in Figure 1, the model is designed to process raw audio inputs and classify them into various emotional categories. The model architecture follows a systematic approach starting from raw audio preprocessing to final emotion classification, leveraging several advanced deep learning techniques to ensure accuracy and efficiency. Initially, the raw audio signals undergo a preprocessing stage, which includes determining audio frequency and padding to convert them into a standardized time series format. This step is crucial for ensuring that all audio inputs are of uniform length and amplitude, facilitating consistent processing in subsequent stages. The preprocessed audio data is then passed through a convolutional layer. This layer performs convolution operations to extract local features from the audio signal, capturing essential patterns that are relevant for emotion recognition. Convolutional layers are adept at identifying spatial hierarchies in data, making them suitable for this initial feature extraction phase. Following the convolutional layer, the data undergoes a projection step. This projection maps the low-dimensional features extracted by the convolutional layer into a higher-dimensional space, increasing the model's expressive power to enable it to learn more complex patterns and features and enhancing the model's ability to generalize across different audio samples. The projected data is then fed into a series of transformer layers, which include a specialized transformer layer with a deformable speech attention (DSA) mechanism (Chen et al., 2023). Transformers are powerful for capturing long-range dependencies and contextual information in sequential data. The DSA mechanism further refines this by enabling the model to focus on different parts of the sequence simultaneously, thus improving the recognition of nuanced emotional cues in the audio signal. Subsequently, the output from the transformer layers is directed through dense layers that act as an emotion classifier. These dense layers apply non-linear transformations to the extracted features, gradually narrowing down the possible emotion categories based on learned patterns and relationships. To finalize the classification process, average pooling is applied. This pooling layer aggregates the features, reducing their dimensionality while preserving essential information, which enhances the robustness of the model's predictions. The aggregated features are then passed through a softmax function, which computes the probabilities of the audio sample belonging

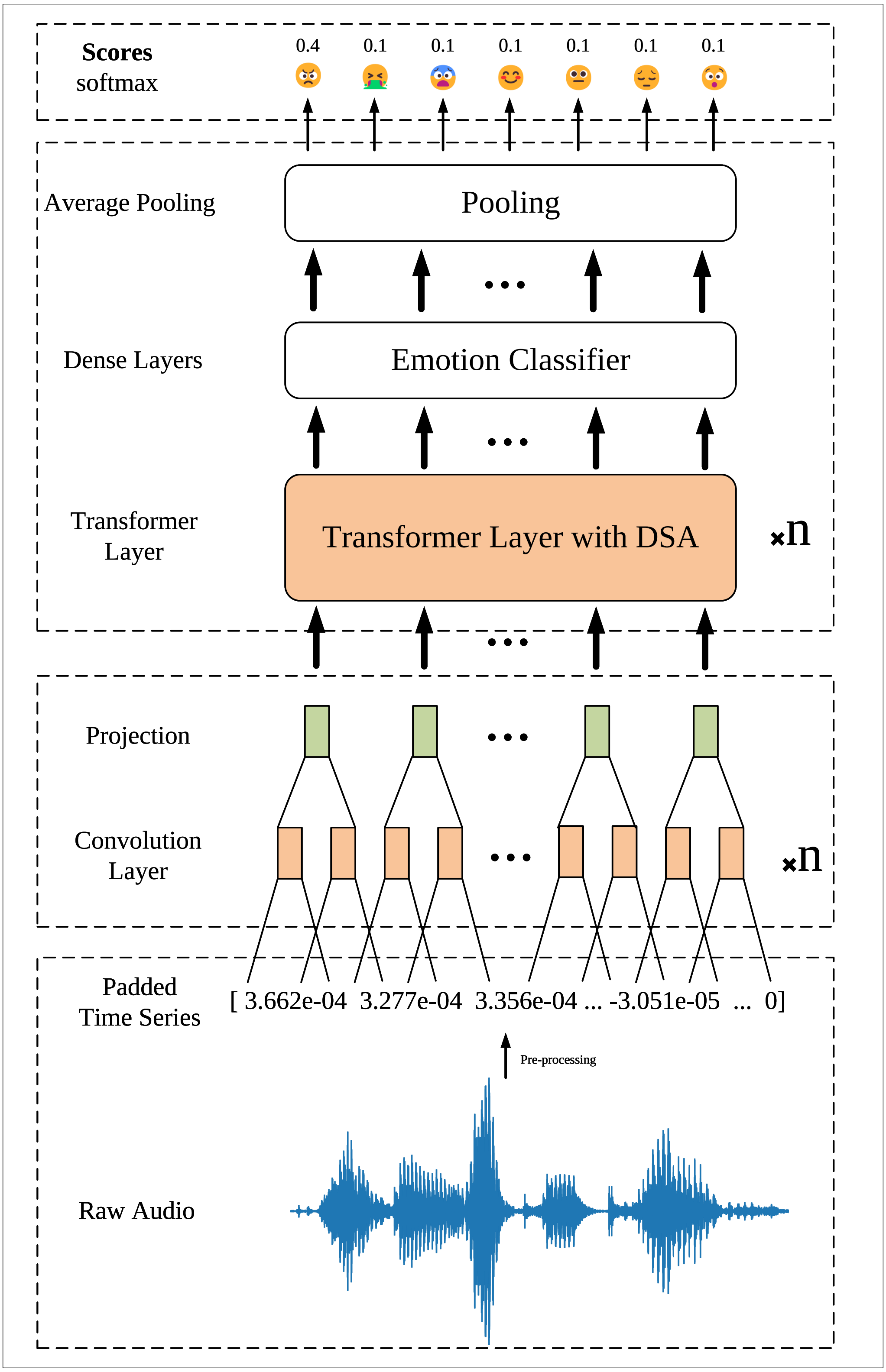


Figure 1: An illustration of the proposed speech emotion recognition framework, SERCT, consisting of two modules: i) CNN Module, which efficiently captures the local information of the features through lightweight CNN layers; ii) Transformer Module, which captures the long-range dependencies in the features and enhance the time-frequency domain features of speech through Transformer Layers.

to each emotional category. The softmax function ensures that the output scores sum up to one, providing a clear probabilistic interpretation of the classification results.

In summary, SERCT's architecture efficiently combines convolutional and transformer layers with dense neural networks and pooling mechanisms. This integrated approach allows for accurate and reliable emotion classification from raw audio inputs, leveraging advanced deep learning techniques to capture both local and global patterns in the data. The following subsections describe main stages in detail, providing an overview of their roles and contributions to the overall model.

## 4.2 Pre-processing

**Dataset Preparation** The dataset consists of speech samples, each ranging in length from 1 to 3 seconds. These samples are stored in audio files, and the dataset includes corresponding labels for each speech sample. Additionally, the dataset is organized into separate train and test directories. Within each of these directories, the speech samples are further categorized into different emotion classes. Each emotion class has its own dedicated directory, and the audio files within these directories are named in an incrementing sequence. This hierarchical organization allows for efficient management and easy access to the data during the testing and evaluation step. By dividing the dataset into train and test subsets, we can evaluate the performance of the neural network on unseen data. The emotion-based categorization within each subset enables the network to learn and recognize specific emotional patterns in the speech samples. During the pre-processing step, one of the primary objectives is to convert the audio files into a format that is suitable for input into a neural network. This typically involves transforming the audio data into a numeric representation, such as a spectrogram or a Mel-frequency cepstral coefficient (MFCC) matrix (but we didn't choose these numerical representation methods; instead, we opted for a simpler processing approach). These representations capture the underlying acoustic features of the speech samples, which are essential for the neural network to analyze and learn from the data effectively. Furthermore, the naming convention of the audio files in each emotion directory, following an incremental sequence, ensures that the order and integrity of the data are maintained. This systematic approach facilitates the organization and retrieval of specific speech samples during the training and evaluation stages.

**Data Loading** The audio files are loaded using the librosa library (McFee et al., 2024), a powerful Python package specifically designed for music and audio analysis. It provides a wide range of functionalities for audio processing, including loading audio files, extracting audio features, and performing various transformations. When loading the audio files with librosa, we made use of the librosa.load() function. To ensure that we preserve the original sampling rate of the audio, we set the sr parameter of librosa.load() to None. By doing so, the function will automatically detect and use the original sampling rate of the audio file. Once the audio files are loaded, they are converted into one-dimensional NumPy arrays. This conversion allows us to represent the audio data in a numerical format that can be easily manipulated. The NumPy array represents the amplitude values of the audio signal over time. By setting the sr parameter to None in librosa.load(), we ensure that the audio data is sampled consistently at its original sampling rate. This is crucial for maintaining the integrity and fidelity of the audio samples. Without consistent sampling, the audio data may be distorted or lose important details, making it difficult to perform accurate analysis and processing.

**Padding and Truncation** To ensure consistency in batch processing, it is crucial to standardize the duration of the original speech samples, which may vary in length. In this study, one-dimensional NumPy arrays of a fixed length of 144,000 are used. This length corresponds to a duration of 3 seconds when sampled at a rate of 48 kHz, which is a commonly used sampling rate for high-quality audio recordings (Pras and Guastavino, 2010). To achieve this standardized length, each speech sample undergoes processing. If the length of one-dimensional NumPy arrays of a speech sample is shorter than 144,000, it is padded with zeros at the end. This padding ensures that the sample reaches the required length without altering its content. By appending zeros, the speech sample effectively extends to meet the fixed length, preserving the temporal structure of the original recording. Conversely, if a speech sample exceeds the 144,000-length limit, it is truncated or shortened to fit the required length. Truncation involves discarding the excess samples from the end of the speech sample, ensuring that it aligns with the fixed length. This step prevents any distortion or loss of information caused by retaining unnecessary data beyond the defined duration. Adopting this approach guarantees that all input data shares a uniform shape, with each speech sample precisely fitting the fixed length of 144,000. The uniformity in shape facilitates efficient batch processing within the neural network. By ensuring that the input data adheres to a consistent format, the network can process multiple speech samples simultaneously, enabling parallelization and optimizing computational efficiency.

**Batch Processing** To enhance the training process, the dataset is partitioned into batches. Batch processing plays a crucial role in boosting the efficiency of training by enabling the model to handle multiple samples simultaneously. This approach significantly reduces the computational overhead and allows for parallelized computations, ultimately expediting the learning process. To ensure the effectiveness of batch processing, the data is shuffled prior to partitioning. Shuffling the data guarantees that each batch comprises a diverse and representative subset of the entire dataset. By incorporating varied samples within each batch, the model is exposed to a broader range of patterns and instances during training. This exposure enhances the model's ability to generalize and perform well on unseen data, improving its overall performance. When working with PyTorch (Paszke et al., 2019), the DataLoader module from the PyTorch library becomes a valuable tool for managing batch processing. The DataLoader is configured with a custom collate function, which serves as a preprocessing step for the speech samples. This collate function handles tasks such as padding and truncation, ensuring that all samples within a batch possess consistent shapes and formats. By aligning the lengths of speech samples through padding or truncation, the collate function guarantees that the input data is properly formatted and ready for ingestion by the neural network. In short, the utilization of batch processing in training offers significant advantages, such as improved computational efficiency and enhanced model generalization. The combination of shuffling the data and employing the PyTorch DataLoader with a custom collate function facilitates streamlined batch processing, ensuring consistent batch shapes and formats for effective neural network training.

The pre-processing stage is a critical part of preparing the speech dataset for training the proposed SERCT. By standardizing the length of the audio samples and organizing them into batches, the pre-processing pipeline ensures that the input data is in a suitable format for efficient and effective training. This approach leverages powerful Python libraries to handle audio loading, normalization, padding, and batch processing, providing a robust framework for speech data processing.

## 4.3 CNN Module

In this section, we delve into the details of the Convolutional Neural Network (CNN) module utilized in this thesis. The CNN architecture is carefully designed to extract meaningful features from the input data, which is crucial for the subsequent tasks. The following description outlines the

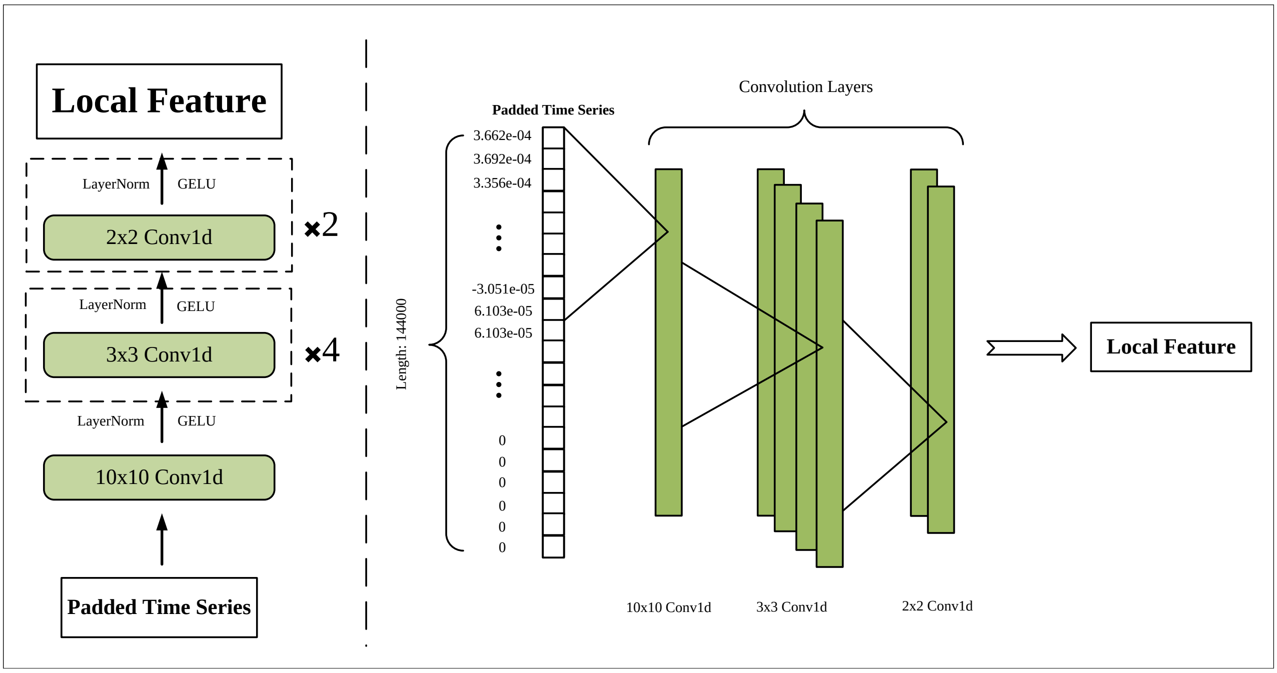


Figure 2: Overview structure of the CNN module.

structure and functionality of each layer in the CNN module as depicted in the Figure 2.

### 4.3.1 Convolutional Layers

The core of the CNN module is constituted by a series of one-dimensional convolutional layers. These layers are pivotal for capturing local patterns and features within the time series data. In the simplest case for a 1D convolution, the output value of the layer with input size $(N, C_{in}, L)$ and output $(N, C_{out}, L_{out})$ can be precisely described as

$$\operatorname{out}\left(N_i, C_{\text{out}_j}\right) = \operatorname{bias}\left(C_{\text{out}_j}\right) + \sum_{k=0}^{C_{\text{in}}-1} \operatorname{weight}\left(C_{\text{out}_j}, k\right) \star \operatorname{input}\left(N_i, k\right) \tag{1}$$

where $\star$ is the valid cross-correlation operator, $N$ is a batch size, $C$ denotes a number of channels, $L$ is a length of signal sequence. In the CNN module of this thesis, the sequence of convolutional layers is as follows:

**10x10 Conv1d Layer** In the given architecture, the first convolutional layer plays a crucial role in extracting significant information from the input time series. It achieves this by applying a set of 512 filters, each with a size of 5x5, to the input data. The purpose of this layer is to capture low-level features present in the data, which can include localized trends and short-term dependencies. By convolving the filters over the input time series, this layer performs a series of operations that preserve the temporal structure while simultaneously enhancing the representation of features. The convolution operation involves sliding the filters across the input, computing dot products at each position, and producing an output feature map. This process helps in identifying patterns and relationships within the data that are essential for subsequent layers to analyze. The application of 512 filters provides the model with a rich set of feature maps, allowing it to learn a diverse range of low-level patterns. These learned features serve as building blocks for higher-level representations in deeper layers of the network, enabling the model to capture more complex and abstract temporal patterns as the network deepens.

**3x3 Conv1d Layer** The second convolutional layer employs 512 filters, each of size 3x3. This layer builds upon the features extracted by the previous layer, which also utilized 512 filters of the same size. By using an equal number of filters, the second convolutional layer can maintain the richness and diversity of the features learned from the initial layer, ensuring that no important information is lost as the data progresses through the network. The 3x3 filter size is a common choice in convolutional neural networks (CNNs) because it strikes a balance between capturing detailed spatial information and computational efficiency. In the second convolutional layer, these filters focus on slightly larger patterns compared to the first layer, allowing the network to capture more complex and abstract relationships within the data. This progression is crucial for tasks such as image recognition and classification, where understanding higher-level features is essential for accurate predictions. Furthermore, by stacking multiple convolutional layers with the same filter size, the network can effectively increase its receptive field. This means that each filter in the second layer can see a larger portion of the input, integrating information over a broader area. As a result, the network can detect more intricate patterns and dependencies.

**2x2 Conv1d Layer** The final convolutional layer employs 512 filters of size 2x2. This layer aims to distill the most pertinent features, facilitating the capture of nuanced patterns that might be crucial for the task at hand. Positioned at the pinnacle of the convolutional hierarchy, this layer is preceded by a sequence of convolutional and normalization operations designed to progressively refine the feature map. The use of a 2x2 filter in the final convolutional layer is strategic, as it enables the network to focus on smaller, yet highly significant local patterns within the data. This fine-grained analysis is particularly beneficial in tasks requiring high precision and attention to detail. Additionally, the integration of 512 filters allows the network to capture a diverse set of features, thereby enhancing its ability to generalize across different instances of the input data. The layered architecture culminating in a 512-filter 2x2 convolutional layer exemplifies a robust design aimed at extracting and refining the most relevant features from the input data. This configuration not only optimizes the network's capacity to recognize complex patterns but also ensures that the final feature representation is rich and discriminative, facilitating superior performance in the task.

### 4.3.2 Activation Functions and Normalization

Following each convolutional layer, the network incorporates activation functions and normalization techniques to enhance learning and stability. Activation functions, such as GELU (Hendrycks and Gimpel, 2016), introduce non-linearity, enabling the network to learn complex patterns. Normalization techniques, like layer normalization (Ba et al., 2016), standardize the activations within a layer across the features for each data point, which accelerates training and improves convergence. Together, these components ensure that the neural network can effectively learn from data while maintaining stability and efficiency during the training process.

**GELU Activation** The Gaussian Error Linear Unit (GELU) activation function is applied after each convolutional layer in the network architecture. The choice of GELU is particularly strategic due to its unique properties that enhance the performance of deep learning models. Unlike traditional activation functions such as ReLU or Sigmoid, GELU introduces non-linearity in a more refined manner by considering the input's distribution. This probabilistic approach ensures a smoother transition and avoids sharp discontinuities, which is highly advantageous for gradient-based optimization methods. GELU's ability to maintain a smooth gradient is crucial during the backpropagation process. It prevents the common issue of vanishing or exploding gradients, thereby facilitating more stable and efficient training of deep neural networks. This smooth gradient property allows the network to learn more effectively, even in deeper architectures, by providing consistent gradient information throughout the training process. Moreover, GELU enhances the network's ability to capture complex patterns within the data. Its non-linear characteristics enable the model to represent intricate relationships that linear functions cannot. By activating neurons based on the input's Gaussian distribution, GELU introduces a level of sophistication in the activation process that can lead to improved generalization and performance on unseen data. According to Hendrycks and Gimpel (2016), GELU is defined as

$$GELU(x) = xP(X \leq x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + erf(x/\sqrt{2})\right] \tag{2}$$

**Layer Normalization** Layer normalization is applied to stabilize and accelerate the training process in the CNN module. This technique normalizes the outputs of each layer, ensuring that the distribution of features remains consistent across different layers, which helps maintain a stable learning environment. By addressing the issue of internal covariate shift, layer normalization mitigates fluctuations in the input distribution as it moves through the layers, leading to more efficient and stable training. This consistency not only speeds up the convergence process but also enhances the model's performance and generalization capabilities. Additionally, layer normalization is particularly beneficial in deep learning models, where the depth and complexity of the network can make training more challenging. By providing a more stable learning landscape, it allows for the development of more accurate and reliable models. Ba et al. (2016) compute the layer normalization statistics over all the hidden units in the same layer as follows

$$\mu^l = \frac{1}{H}\sum_{i=1}^{H} a_i^l \qquad \sigma^l = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(a_i^l - \mu^l)^2} \tag{3}$$

where $H$ denotes the number of hidden units in a layer.

## 4.4 Transformer Module

Inspired by the foundational work on transformer architectures, particularly the seminal "Attention is All You Need" by Vaswani et al. (2017), our research initially aimed to employ a model with stacked transformer layers for enhanced performance in speech emotion recognition tasks. However, upon further exploration and evaluation, we integrated the Deformable Speech Transformer (DST) as presented in the recent study by Chen et al. (2023), due to its advanced capabilities and superior performance in handling speech data.

### 4.4.1 Transformer Architecture Overview

Transformers have indeed revolutionized the field of natural language processing (NLP) and have been adapted for various other domains due to their unparalleled ability to model long-range

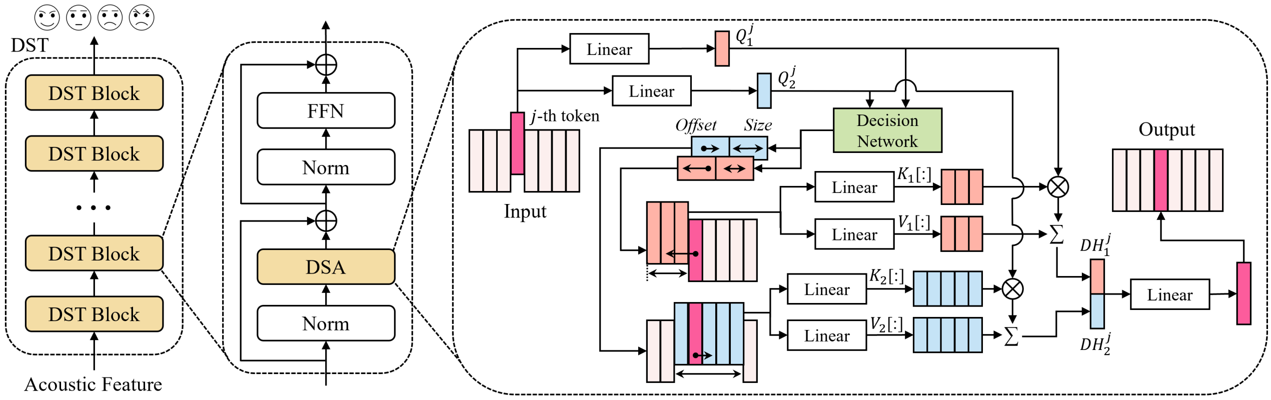


Figure 3: Overview structure of the Transformer module. The only difference between DST block and the vanilla Transformer is the replacement of MSA with DSA. In DSA, here shows only $h$ = 2 attention heads and omit the softmax operation for a clear presentation. $\otimes$ and $\oplus$ represent the matrix multiplication and addition, respectively. $\sum$ represents the weighted summation. Reprinted from Chen et al. (2023).

dependencies through self-attention mechanisms. The significance of transformers lies in their architecture, which primarily consists of an encoder and a decoder, both made up of multiple layers of self-attention and feedforward neural networks. This dual-component structure enables transformers to process and generate complex data sequences effectively. The self-attention mechanism at the heart of transformers allows the model to dynamically weigh the importance of different parts of the input data. This is achieved through the computation of attention scores, which quantify the relevance of each input element relative to others within the sequence. Consequently, the model can capture intricate relationships and dependencies, regardless of the distance between elements in the sequence. This capability is particularly advantageous for tasks involving sequential data, such as language translation, text summarization, and speech recognition. In the encoder, self-attention layers process the input data to produce a contextual representation that captures the relationships between all tokens. Each token's representation is adjusted based on its interactions with all other tokens, leading to a rich and holistic understanding of the input. The encoder's output is then fed into the decoder, which also incorporates self-attention layers to process the previously generated tokens along with the encoder's output. This dual attention mechanism in the decoder enables it to generate coherent and contextually appropriate outputs. Specifically, the MSA mechanism can be formulated as

$$Q_i = QW_i^Q, \quad K_i = KW_i^K, \quad V_i = VW_i^V \tag{4}$$

$$H_i = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_Q}}\right) V_i \tag{5}$$

$$MSA(Q,K,V) = \text{concat}(H_1, \ldots, H_h) W^O \tag{6}$$

where $\boldsymbol{Q,K,V}$ are query, key and value matrices, respectively; $d_Q$ is a scaling factor and $h$ denotes the number of attention heads; $W_i^Q, W_i^K, W_i^V$ and $W^O$ are to be learned parameters.

### 4.4.2 Stacked Transformer Layers

Stacked transformer layers are a fundamental architectural component in many state-of-the-art natural language processing (NLP) models, revolutionizing the field with their remarkable ability to capture intricate patterns in data. By stacking multiple transformer modules atop one another, these models iteratively refine and abstract information, enabling them to comprehend and generate text with unprecedented accuracy and fluency. The core principle underlying stacked transformer layers lies in their hierarchical processing of input data. At each layer, the transformer module receives the output from the preceding layer and applies a series of self-attention mechanisms and feedforward neural networks to extract and distill relevant features. This iterative process allows the model to gradually build complex representations of the input, capturing both local and global dependencies within the data. One key advantage of employing stacked transformer layers is their capacity to learn rich, hierarchical representations of language. By leveraging multiple layers of abstraction, the model can encode a wide range of linguistic phenomena, from basic syntactic structures to subtle semantic nuances. This hierarchical approach not only enhances the model's ability to understand and generate coherent text but also facilitates the transfer of knowledge across different tasks and domains. Furthermore, stacked transformer layers have demonstrated significant improvements in various NLP applications, particularly those involving the interpretation of nuanced information, such as emotion recognition in speech (Wang et al., 2021). By leveraging the hierarchical nature of stacked transformers, these models can effectively capture the subtle cues and contextual clues essential for discerning emotions conveyed through speech signals. This capability holds immense promise for applications in sentiment analysis, dialogue systems, and affective computing, where accurately understanding and responding to human emotions are paramount.

### 4.4.3 Deformable Speech Transformer

The Deformable Speech Transformer (DST), shown in Figure 3, extends the traditional transformer architecture by incorporating a deformable attention mechanism, significantly enhancing its capability to process speech emotion recognition tasks. Traditional transformers use fixed attention weights, where each token in the input sequence uniformly attends to all other tokens through a weighted sum. This approach, while powerful, does not account for the varying relevance of different parts of the input sequence, particularly in speech data where important emotional cues can be scattered non-uniformly.

According to Chen et al. (2023), the core innovation of DST is the Deformable Speech Attention (DSA), which adapts the attention mechanism to be more flexible and context-sensitive. Instead of using fixed-size attention windows, DSA allows the model to dynamically adjust both the size and the position of the attention windows based on the input data. This dynamic adjustment is facilitated by a decision network that predicts the optimal window size and position for each token in the sequence. This deformable attention mechanism enables DST to focus more precisely on the most relevant parts of the input sequence, which is crucial for tasks like speech emotion recognition where emotional cues are not uniformly distributed. Emotions in speech are often conveyed through subtle variations in tone, pitch, and timing, which may occur sporadically rather than uniformly throughout the speech signal. The deformable attention mechanism allows the model to dynamically adjust its focus to these salient parts, capturing the emotional content more effectively. By adapting the window sizes and positions, DST can handle the varying nature of speech data better than traditional transformers. This leads to a more accurate and nuanced understanding of the emotional states conveyed in the speech, as demonstrated by superior performance on benchmark datasets like IEMOCAP and MELD. Chen et al. (2023) formulates DSA as follows

$$DH_i^j = \text{softmax}\left(\frac{(Q_i^j K_i[L_{ij} : R_{ij}]^\top)}{\sqrt{d_Q}}\right) V_i[L_{ij} : R_{ij}] \tag{7}$$

$$DSA(Q,K,V) = \text{concat}(DH_1,\ldots,DH_h)W^o \tag{8}$$

where $K_i[L_{ij} : R_{ij}]$ and $V_i[L_{ij} : R_{ij}]$ consist of the $L_{ij}$-th to the $R_{ij}$-th tokens of $K_i$ and $V_i$ matrices, respectively; $DH_i^j$ denotes the $j$-th output token of the $i$-th attention head.

# 5 Experiments Setup

In this section, we outline the experimental setup used to evaluate the performance of various models on SER task. The experiments are designed to provide a comprehensive analysis of different approaches, leveraging established datasets, a range of baseline models, detailed implementation practices, and robust evaluation metrics. In subsection 5.1, we introduce the datasets employed in our experiments. The Toronto Emotional Speech Set (TESS) (Pichora-Fuller and Dupuis, 2020) and the Emotional Speech Dataset (ESD) (Zhou et al., 2021) serve as the primary sources of data. These datasets are selected for their diverse emotional content and high-quality recordings, which are essential for training and evaluating emotion recognition models effectively. Next, in subsection 5.2, we describe the baseline models utilized in our study. The chosen models include convolutional neural networks (CNN), hybrid CNN-BiLSTM networks, long short-term memory networks (LSTM), multilayer perceptrons (MLP), and support vector machines (SVM). These models are selected to cover a broad spectrum of machine learning approaches, from deep learning architectures to traditional machine learning techniques. In the implementation details (i.e., subsection 5.3), we mainly introduce the hyperparameter settings and training protocols used for each model in our study. This includes specific configurations such as learning rates, batch sizes and the number of epochs. Finally, in subsection 5.4, we outline the metrics used to evaluate model performance. Weighted Accuracy (WA), Unweighted Accuracy (UA), and Weighted F1 Score (WF1) are the primary metrics.

## 5.1 Dataset

**TESS** [1] The Toronto Emotional Speech Set (TESS) is a valuable resource for researchers studying speech and emotion recognition. This dataset includes 2800 audio files in the format of wav, all recorded in English with a sampling frequency of 24kHz, ensuring high-quality sound. Each audio file is designed to represent one of seven emotional categories: angry, disgusted, fearful, happy, neutral, sad, and surprised. These categories provide a broad range of emotional expressions, making TESS particularly useful for training and testing machine learning models in emotion detection and speech analysis. TESS was developed with the goal of enhancing the understanding of how emotions are conveyed through speech. The dataset features recordings from two female speakers, aged 26 and 64, who were asked to speak a set of target phrases in each of the seven emotional tones. This setup ensures that the data captures variations not only in emotional expression but also in vocal characteristics due to age differences. We can access the TESS dataset via the University of Toronto's TSpace repository, making it easily available for academic and commercial use. The structured and well-documented nature of the dataset allows for straightforward integration into our experimental setups.

**ESD** [2] The Emotional Speech Dataset (ESD) is a comprehensive and versatile resource designed to facilitate research in the field of speech emotion recognition. The dataset encompasses a total of 35,000 audio files in the format of wav, each sampled at a frequency of 16kHz, ensuring high-quality sound representation suitable for various analytical techniques. ESD categorizes the emotional content of its audio samples into five distinct classes: angry, happy, neutral, sad, and surprised. These recordings are available in both English and Chinese, offering a bilingual dataset that

[1] TESS is available at https://tspace.library.utoronto.ca/handle/1807/24487

[2] ESD is available at https://github.com/HLTSingapore/Emotional-Speech-Data

Table 1: Statistical analysis of the sample size in datasets. Here, ESD-en refers to the English audio samples in the ESD dataset, while ESD-zh denotes the Chinese audio samples in the same dataset.

| **Emotions** | **TESS** | **ESD-en** | **ESD-zh** | **EMO-DB** |
|---|---|---|---|---|
| Angry | 400 | 3500 | 3500 | 127 |
| Disgusted | 400 | - | - | 46 |
| Fearful | 400 | - | - | 69 |
| Happy | 400 | 3500 | 3500 | 71 |
| Neutral | 400 | 3500 | 3500 | 79 |
| Sad | 400 | 3500 | 3500 | 62 |
| Surprised | 400 | 3500 | 3500 | - |
| Total | 2800 | 17500 | 17500 | 454 |

enhances the robustness and applicability of research findings across different linguistic contexts. The inclusion of both English and Chinese language data broadens the dataset's utility, making it a valuable asset for studies that aim to investigate cross-linguistic emotional expression. We can leverage this diversity to compare and contrast emotional speech patterns across these two widely spoken languages, potentially uncovering universal and language-specific emotional markers.

**EMO-DB** [3] The Berlin Database of Emotional Speech (EMO-DB), created by the Institute of Communication Science, Technical University, Berlin, Germany, is a widely recognized and extensively utilized dataset in the field of emotion recognition and affective computing. Ten professional speakers (five males and five females) participated in data recording. It comprises a collection of 535 high-quality audio recordings in the format of wav, each recorded at a 48-kHz sampling rate and then down-sampled to 16-kHz. The recordings feature German-language utterances, specifically designed to represent a diverse range of emotional expressions. The dataset categorizes these emotional states into six distinct classes: angry, happy, neutral, sad, disgusted, and fearful.

The statistical information on emotion categories for TESS, ESD and EMO-DB is shown in Table 1. From the results in the table, it is evident that TESS and ESD-en have a very balanced number of samples across different emotion categories. Therefore, in the subsequent experiments, we did not use upsampling or downsampling methods to achieve data balance.

## 5.2 Baselines

In our experiments, we employed several baseline models, including Convolutional Neural Networks (CNN), CNN-BiLSTM, LSTM, Multi-Layer Perceptron (MLP) and Support Vector Machines (SVM). Below is a detailed description of each model setup.

**CNN** CNNs have proven effective in SER by leveraging their ability to analyze spectral and temporal representations of speech. These networks typically employ layers of convolutions that can extract and learn hierarchical features from raw speech data, which is crucial for recognizing emotions embedded in varying speech patterns (Anvarjon et al., 2020). The architecture generally includes convolutional layers followed by pooling layers, which reduce the spatial size of the repre-

[3] EMO-DB is available at http://www.emodb.bilderbar.info/download/

sentations, thus decreasing the number of parameters and computations in the network. This setup helps in capturing the essential characteristics of emotional speech without the overhead of processing raw audio at its original resolution.

**CNN-BiLSTM** The CNN-BiLSTM combines the spatial feature extraction capabilities of CNNs with the sequential data processing strength of bidirectional long short-term memory networks (BiLSTMs). This model is particularly useful in SER for capturing both the local features through CNN and the contextual dependencies in speech via the forward and backward LSTM layers. By processing data in both time directions, the BiLSTM layers help in understanding the emotion context that might be lost in models only considering past context, thereby providing a robust mechanism for emotion detection across varied speech segments.

**LSTM** LSTMs are critical in modeling long-range dependencies in time-series data, such as speech, which is essential for identifying emotions that may unfold over several time frames. Unlike standard recurrent neural networks, LSTMs can learn and remember over long intervals with their gated mechanism, which regulates the flow of information. This ability prevents the vanishing gradient problem common in traditional RNNs, making LSTMs particularly suitable for tasks where understanding the temporal dynamics of speech is crucial for recognizing emotional states.

**MLP** MLPs are a type of feedforward neural network that consists of at least three layers: an input layer, hidden layers, and an output layer. In the context of SER, MLPs can be utilized to classify emotional states from a set of features extracted from the speech signal, such as pitch, energy, and MFCCs (for simplicity and to allow for fair comparison, we use Padded Time Series as the raw input for the MLP). By learning a non-linear combination of these input features, MLPs can effectively distinguish between different emotional states.

**SVM** SVMs are well-suited for classification tasks like SER, where the goal is to find a hyperplane that best divides a dataset into classes in a high-dimensional space. In the context of emotion recognition, SVMs can classify emotional states based on features extracted from speech. The effectiveness of SVMs in this domain often depends on the choice of the kernel, which allows the data to be transformed and analyzed in a manner that the linear separation is possible even when the original feature space might not allow it. SVMs are particularly valued for their ability to handle large feature spaces and maintain effectiveness even with a limited number of training samples, which is often the case in SER tasks.

These baseline models provide a comprehensive toolkit for tackling the complex problem of speech emotion recognition. Each model offers unique strengths, making them collectively beneficial for comparative studies in SER. This diversity allows us to explore various aspects of emotional expression in speech, enhancing the robustness and accuracy of emotion recognition systems.

## 5.3 Implementation Details

Table 2: Comparison of different models based on time taken (in hours).

| | SERCT | CNN | CNN-BiLSTM | LSTM | MLP | SVM |
|---|---|---|---|---|---|---|
| Time (h) | 2.67 | 0.60 | 0.72 | 0.65 | 0.12 | 0.17 |

In section 4, we detail the implementation of our proposed Speech Emotion Recognition model, SERCT, which integrates a CNN-Transformer architecture. The input to our SERCT model is a custom-designed Padded Time Series, optimized to handle varying lengths of audio signals efficiently. To seamlessly connect the CNN and Transformer modules, we introduce a randomly initialized linear projection layer, facilitating smooth transitions between these two architectures.

For the training of the SERCT model, we set the learning rate to 0.001, which balances the speed of convergence with the stability of training. The training process spans 20 epochs, providing sufficient time for the model to learn intricate patterns without overfitting. We also apply a weight decay of 0.005 to regularize the model and prevent overfitting. The batch size is set to 16, which is a compromise between computational efficiency and the gradient estimate's stability. The optimization of our model was performed using the Adam optimizer, which is well-suited for handling sparse gradients in our CNN-Transformer setup. The loss function used was cross entropy criterion, appropriate for our classification task. Additionally, we employed the learning rate scheduler StepLR with a step size of 10 and a gamma of 0.1 to adjust the learning rate dynamically during training. To ensure the reproducibility of our experiments, we set the random seeds for PyTorch (Paszke et al., 2019), NumPy (Harris et al., 2020) and the random module to 3.

Focusing on the CNN part of the SERCT, we configure it with six convolutional layers. Each layer does not use dropout, emphasizing the retention of all features during training. We set bias to False, meaning no bias term is included in the convolutional layers. The GELU (Hendrycks and Gimpel, 2016) activation function is employed, known for its smooth and non-linear properties, which help in capturing complex patterns. Layer normalization is applied to stabilize and accelerate the training process.

As for the Transformer part, we configure it with an embedding dimension of 1024, providing a rich feature space for the audio signals. The feedforward network within the Transformer layers has a dimension of 512, and the model consists of four Transformer layers. Each layer utilizes eight attention heads to capture different aspects of the input sequence. A dropout rate of 0.1 is applied to prevent overfitting, and we exclude bias terms in the Transformer layers. The ReLU (Agarap, 2019) activation function is used, promoting sparsity and computational efficiency.

The dataset is split into training and testing sets in a 9:1 ratio, ensuring a robust evaluation of the model's performance. For baseline models, the hyperparameters for learning rate, epoch, and weight decay are kept consistent with those of the SERCT to ensure a fair comparison. Other hyperparameters for the baseline models are kept at their default values.

Concerning the hardware used in experiments, the experiments were conducted on the Hábrók high-performance computing cluster of the University of Groningen. The GPU used is an Nvidia A100 GPU accelerator card with 40 GB of VRAM available. As for the training time of all models, please refer to Table 2.

## 5.4 Metrics

In this section, we will describe the evaluation metrics used in our experiments to assess the performance of the proposed model, SERCT. The metrics employed include Weighted Accuracy (WA), Unweighted Accuracy (UA), and Weighted F1 Score (WF1). To thoroughly understand these metrics, we first need to clarify some fundamental concepts: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). True Positive (TP) refers to the instances where the model correctly identifies positive cases. For example, in a binary classification task where the model is designed to detect a specific condition, TP represents the number of correctly identified cases with that condition. False Positive (FP) occurs when the model incorrectly identifies a positive case, i.e., the model predicts the presence of the condition when it is absent. True Negative (TN) denotes the correct identification of negative cases, where the model accurately predicts the absence of the condition. Lastly, False Negative (FN) happens when the model fails to identify a positive case, predicting the absence of the condition when it is actually present. With these definitions in mind, Accuracy, Precision, Recall and F1-score can be expressed as

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{11}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

Then, we can obtain the calculation methods for WA, UA and WF1 as follows

$$\text{WA} = \sum_{i=1}^{n} w_i \cdot \text{Accuracy}_i \tag{13}$$

$$\text{UA} = \frac{1}{n} \sum_{i=1}^{n} \text{Accuracy}_i \tag{14}$$

$$\text{WF1} = \sum_{i=1}^{n} w_i \cdot \text{F1}_i \tag{15}$$

where $w_i$ represents the weight of the $i$-th class and $\text{Accuracy}_i$ denotes the accuracy for the $i$-th class. Additionally, $n$ is the number of classes and $\text{F1}_i$ is the F1-score for the $i$-th class.

# 6 Results

In this section, we present a comprehensive evaluation of our proposed model. Subsection 6.1 details the main results, demonstrating the effectiveness of SERCT in accurately identifying emotions from speech. Subsection 6.2 discusses the model's capability for language generalization, showcasing its performance across various languages. Subsection 6.3 explores the impact of the number of transformer layers on the model, aiming to achieve a balance between performance and time overhead. Finally, subsection 6.4 covers the ablation study, highlighting the contributions of different components within the SERCT model to its overall performance. These comprehensive evaluations affirm the robustness and versatility of our proposed approach.

## 6.1 Main Results

Table 3: Performance for different models on TESS and ESD-en datasets.

| **Model** | **TESS** | | | **ESD-en** | | |
|---|---|---|---|---|---|---|
| | **WA** | **UA** | **WF1** | **WA** | **UA** | **WF1** |
| MLP | 0.3857 | 0.3857 | 0.3736 | 0.2306 | 0.2306 | 0.2290 |
| LSTM | 0.1311 | 0.1311 | 0.0301 | 0.1083 | 0.1887 | 0.0499 |
| CNN-BiLSTM | 0.1429 | 0.1429 | 0.0357 | 0.2003 | 0.2000 | 0.0669 |
| SVM | 0.1607 | 0.1607 | 0.0757 | 0.1998 | 0.2000 | 0.0667 |
| CNN | 0.8214 | 0.8214 | 0.8190 | 0.4732 | 0.4731 | **0.4705** |
| **SERCT** | **0.9893** | **0.9893** | **0.9894** | **0.5200** | **0.5200** | 0.4249 |

In this study, we propose the Speech Emotion Recognition model named SERCT and compare its performance with several baseline models, including MLP, LSTM, CNN-BiLSTM (Deschamps-Berger et al., 2021), SVM and CNN, on two datasets: TESS and ESD-en. The results are shown in Table 3. The performance metrics used for evaluation are Weighted Accuracy (WA), Unweighted Accuracy (UA) and Weighted F1-score (WF1).

For the TESS dataset, the proposed SERCT model significantly outperforms all baseline models across all evaluation metrics. The SERCT model achieves a WA of 0.9893, a UA of 0.9893 and a WF1 of 0.9894. In comparison, the closest performing model, CNN, achieves a WA of 0.8214, a UA of 0.8214 and a WF1 of 0.8190. Other models, such as MLP, LSTM, CNN-BiLSTM and SVM, show considerably lower performance, with WA, UA and WF1 scores below 0.39, demonstrating the superior accuracy and robustness of the SERCT model.

On the ESD-en dataset, the SERCT model also demonstrates superior performance, achieving a WA and UA of 0.5200 and a WF1 of 0.4249. While the CNN model shows competitive performance with a WA of 0.4732, a UA of 0.4731 and a WF1 of 0.4705, it still falls short compared to the SERCT model. Other baseline models, including MLP, LSTM, CNN-BiLSTM and SVM, perform poorly with WA, UA and WF1 scores not exceeding 0.23.

The results clearly indicate that the SERCT model is highly effective for Speech Emotion Recognition tasks, significantly outperforming other established models on both TESS and ESD-en datasets.

This demonstrates the robustness and accuracy of the proposed SERCT model in recognizing emotions from speech, highlighting its potential for real-world applications in emotion detection systems.

## 6.2 Language Generalization

Table 4: Performance comparison of different models on EMO-DB and ESD-zh datasets.

| **Model** | **EMO-DB** | | | **ESD-zh** | | |
|---|---|---|---|---|---|---|
| | **WA** | **UA** | **WF1** | **WA** | **UA** | **WF1** |
| MLP | 0.2320 | 0.2320 | 0.2307 | 0.2048 | 0.1863 | 0.1956 |
| LSTM | 0.2677 | 0.1366 | 0.1023 | 0.1987 | 0.1987 | 0.0497 |
| CNN-BiLSTM | 0.2797 | 0.1667 | 0.1223 | 0.2000 | 0.2000 | 0.0667 |
| SVM | 0.1740 | 0.1704 | 0.1090 | 0.2000 | 0.2000 | 0.0698 |
| CNN | 0.2203 | 0.1699 | 0.1910 | 0.4800 | 0.4800 | 0.4672 |
| **SERCT** | **0.3877** | **0.3712** | **0.3892** | **0.5480** | **0.5480** | **0.5224** |

In this subsection, we test the performance of our proposed Speech Emotion Recognition model, SERCT, on datasets in languages other than English, despite being trained exclusively on English data. This evaluation aims to determine if SERCT can effectively generalize to new languages without additional training. Specifically, we tested the model on two non-English datasets: EMO-DB (German) and ESD-zh (Chinese).

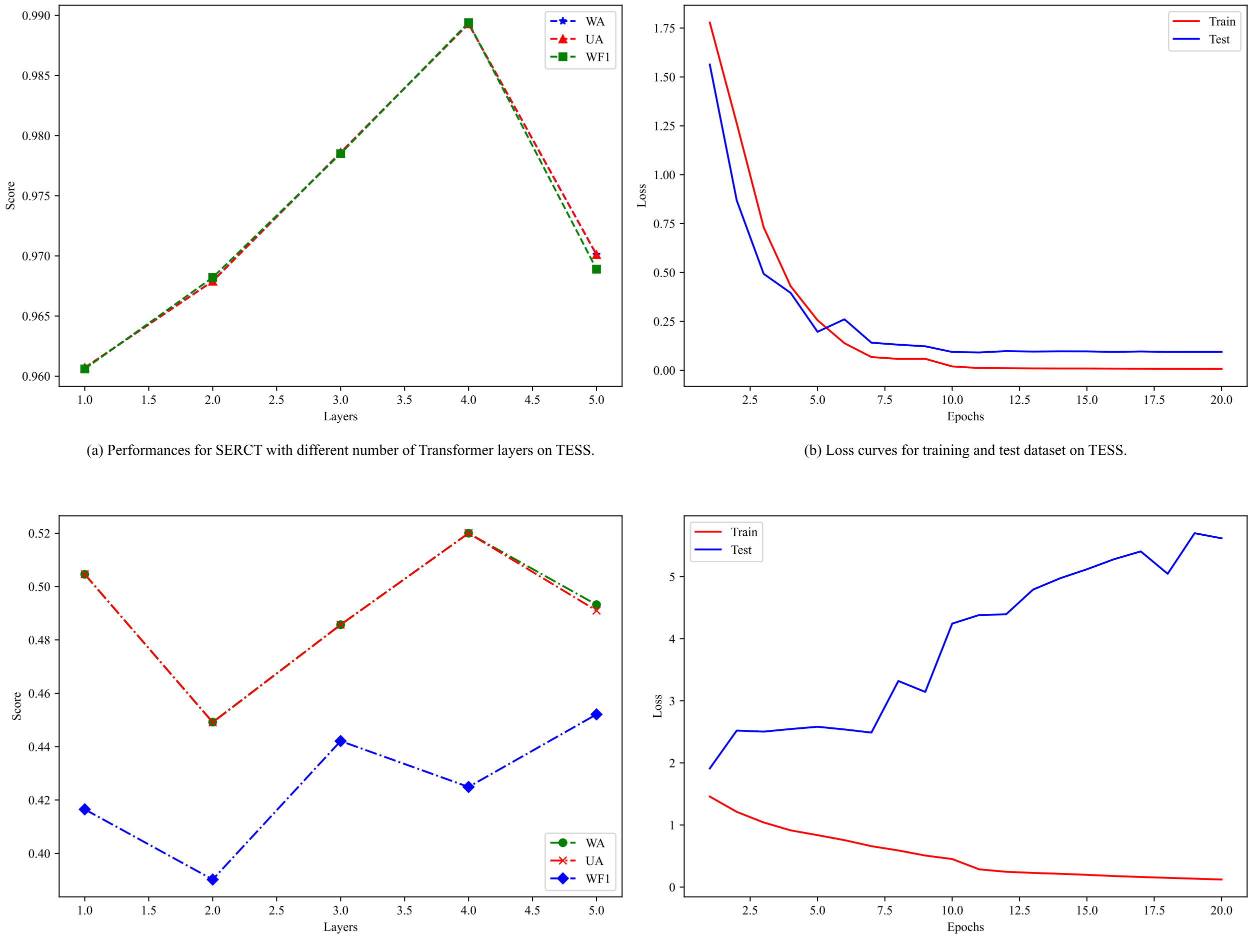
The performance metrics we used to evaluate the models include Weighted Accuracy (WA), Unweighted Accuracy (UA) and Weighted F1 score (WF1). The results are presented in the table 4, where we compare SERCT with other established models such as MLP, LSTM, CNN-BiLSTM, SVM and CNN.

As shown in the results, SERCT significantly outperforms the other models across all three metrics on both EMO-DB and ESD-zh datasets. For the German EMO-DB dataset, SERCT achieved a WA of 0.3877, UA of 0.3712 and WF1 of 0.3892. In comparison, the next best-performing model, CNN-BiLSTM, achieved considerably lower scores, with a WA of 0.2797, UA of 0.1667 and WF1 of 0.1223. Similarly, on the Chinese ESD-zh dataset, SERCT achieved a WA of 0.5480, UA of 0.5480 and WF1 of 0.5224, whereas the CNN model, the second-best performer, achieved a WA of 0.4800, UA of 0.4800 and WF1 of 0.4672.

These results highlight the robust generalization capabilities of SERCT, demonstrating its effectiveness in recognizing speech emotions across different languages without the need for additional language-specific training. The superior performance of SERCT in comparison to other models indicates its potential as a highly adaptable and reliable solution for speech emotion recognition in a multilingual context.

## 6.3 Effect of Transformer Layer Depth

In this section, we study the impact of the number of Transformer layers on the performance of our proposed Speech Emotion Recognition model, SERCT. The objective of this analysis is to



(a) Performances for SERCT with different number of Transformer layers on TESS.

(b) Loss curves for training and test dataset on TESS.

(c) Performances for SERCT with different number of Transformer layers on ESD-en.

(d) Loss curves for training and test dataset on ESD-en.

Figure 4: Further investigation results on the number of Transformer layers.

understand how varying the number of Transformer layers affects the model's performance and to find an optimal balance between accuracy and computational efficiency. From the analysis below, it is evident that the number of Transformer layers has a significant impact on the performance of SERCT. The results indicate that a Transformer layer number of 4 layers might strikes a balance between achieving high performance and maintaining computational efficiency. Adding more layers beyond this offers minimal performance improvements and increases the time complexity, making it less desirable for practical applications.

### 6.3.1 Performance Analysis

To evaluate the effect of the number of Transformer layers, we conducted experiments using two distinct datasets: TESS (Toronto Emotional Speech Set) and ESD-en (English Emotional Speech Dataset). For each dataset, we trained the SERCT model with varying numbers of Transformer layers and recorded the performance metrics, namely Weighted Accuracy (WA), Unweighted Accuracy (UA) and Weighted F1 Score (WF1).

Figure 4(a) illustrates the performance metrics of SERCT with different number of Transformer layers on the TESS dataset. As observed, the performance in terms of WA, UA and WF1 improves significantly with an increase in the number of layers up to a certain point. Specifically, the model achieves optimal performance at around 3 to 4 layers, beyond which the gains plateau and even slightly diminish. Similarly, Figure 4(c) presents the performance metrics on the ESD-en dataset. The trend observed is consistent with that of the TESS dataset, with the model performance peaking around 3 to 4 layers. The diminishing returns beyond this point suggest that additional layers may not contribute significantly to performance improvements and might introduce unnecessary computational overhead.

### 6.3.2 Loss Analysis

To further understand the training dynamics, we examined the loss curves for both the training and test datasets at the optimal number of Transformer layers identified in the performance analysis. In this thesis, we have chosen 5 as the optimal number of Transformer layers for our proposed Transformer module.

Figure 4(b) shows the loss curves for the training and test datasets on the TESS dataset. The training loss decreases steadily, while the test loss also reduces but at a slower rate, indicating good generalization ability of the model. The convergence of the loss curves suggests that the model is effectively learning the underlying patterns without overfitting. Figure 4(d) depicts the loss curves for the ESD-en dataset. Unlike the TESS dataset, the training loss declines progressively, while the test loss exhibits an upward trend. This behavior indicates potential overfitting, where the model performs well on the training data but struggles to generalize to unseen data, resulting in increased test loss.

## 6.4 Ablation Study

Table 5: Performance with different experimental settdings.

| **Model** | **TESS** | | | **ESD-en** | | |
|---|---|---|---|---|---|---|
| | **WA** | **UA** | **WF1** | **WA** | **UA** | **WF1** |
| W/o transformer (DST) module | 0.8214 | 0.8214 | 0.8190 | 0.4732 | 0.4731 | **0.4705** |
| W/o CNN module | 0.9500 | 0.9500 | 0.9503 | 0.3779 | 0.3780 | 0.3127 |
| W/ transformer (Vanilla) module | 0.9612 | 0.9612 | 0.9613 | 0.4099 | 0.4099 | 0.3576 |
| **SERCT** | **0.9893** | **0.9893** | **0.9894** | **0.5200** | **0.5200** | 0.4249 |

In this section, we investigate the contributions of different components of our proposed SERCT, which integrates a CNN module and a Transformer module with DST. We conducted ablation experiments to determine the significance of each component by systematically removing or altering them. The results of these experiments are summarized in the table 5. The following are the components for comparison:

- **W/o transformer (DST) module**: In this configuration, we removed the Transformer module with DST, leaving only the CNN module. This setup allows us to assess the performance of the CNN module in isolation.

- **W/o CNN module**: we excluded the CNN module, retaining only the Transformer module with DST. This helps in understanding the role of the Transformer module with DST independently.

- **W/ transformer (Vanilla) module**: In this variation, we replaced the Transformer (DST ) module with a Vanilla Transformer module to compare their effectiveness.

First, we examined the performance of the model without the Transformer (DST) module, retaining only the CNN module. The results show that this configuration yields a WA, UA and WF1 of 0.8214, 0.8214 and 0.8190 respectively on the TESS dataset, and 0.4732, 0.4731 and 0.4705 on the ESD-en dataset. This indicates a notable drop in performance compared to the full SERCT model, especially on the ESD-en dataset, highlighting the significant role of the Transformer (DST) module in handling more complex and diverse emotional speech data.

Next, we evaluated the model without the CNN module, leaving only the Transformer (DST) module. The results demonstrated an even more pronounced decrease in performance, with WA, UA and WF1 values of 0.9500, 0.9500 and 0.9503 on the TESS dataset, and 0.3779, 0.3780 and 0.3127 on the ESD-en dataset. This suggests that while the Transformer (DST) module is crucial, the CNN module also plays a vital role in feature extraction, particularly for datasets with less complexity like TESS.

Furthermore, we tested a variant of the model where the Transformer (DST) module was replaced with a Vanilla Transformer module. The results showed an improvement over the configurations lacking either module but still fell short of the full SERCT model's performance. Specifically, the WA, UA and WF1 scores were 0.9612, 0.9612 and 0.9613 on the TESS dataset, and 0.4099, 0.4099 and 0.3576 on the ESD-en dataset. This underscores the efficacy of the DST variant of the Transformer module in capturing the temporal dependencies and contextual information necessary for accurate emotion recognition.

Finally, the complete SERCT model achieved the highest performance across all metrics and datasets, with WA, UA and WF1 scores of 0.9893, 0.9893 and 0.9894 on the TESS dataset, and 0.5200, 0.5200 and 0.4249 on the ESD-en dataset. These results clearly demonstrate the complementary strengths of both the CNN and Transformer (DST) modules in our SERCT model, affirming that their integration is essential for achieving excellent performance in Speech Emotion Recognition tasks.

# 7 Discussion

In this section, subsection 7.1 will evaluate the results in relation to our initial questions and hypotheses. Following this, subsection 7.2 will address ethical considerations, ensuring that all aspects of the research were conducted with the standards of ethical integrity. Finally, subsection 7.3 will discuss the limitations of the study, acknowledging the constraints and potential areas for improvement in the research.

## 7.1 Validation of All Hypotheses

In section 3, we presented the questions and hypotheses of this thesis, which include one main question and five derived sub-questions, corresponding to which are six hypotheses. This subsection discusses the validation of the six hypotheses proposed in the thesis. Each hypothesis will be evaluated based on the results and discussions presented throughout the thesis.

**Validation of the First Hypothesis** The first hypothesis is that SERCT will demonstrate effective performance and surpass all baseline approaches. Based on the data from Table 3 and Table 4, as well as the results of the ablation experiments, we can easily conclude that SERCT is indeed an excellent model. However, it does not outperform baseline models in all metrics. Therefore, it can be tentatively considered that the current hypothesis is only half correct.

**Validation of the Second Hypothesis** The second hypothesis is that baseline models trained from scratch using the English ESD and TESS datasets will exhibit promising performance across various evaluation metrics, but not necessarily outperform the more advanced SERCT. The performance of baseline models was evaluated on the ESD and TESS datasets. The results indicated that the baseline models showed good performance but were outperformed by the proposed SERCT model. Specifically, the CNN model, the closest baseline, achieved a WA of 0.8214, UA of 0.8214 and WF1 of 0.8190 on the TESS dataset, while the SERCT model achieved significantly higher scores of 0.9893, 0.9893 and 0.9894 respectively. On the ESD dataset, SERCT also significantly outperforms the second-best CNN in two metrics. Thus, we believe the current hypothesis holds true.

**Validation of the Third Hypothesis** The third hypothesis is that SERCT will outperform baseline models across all the evaluation metrics. The current hypothesis is very similar to the first hypothesis. However, due to the overfitting of SERCT on the ESD dataset, this has resulted in CNN's WF1 score being significantly higher than that of SERCT on the ESD dataset. Therefore, we conclude that the current hypothesis is not valid.

**Validation of the Fourth Hypothesis** The fourth hypothesis is that both the CNN module and Transformer module in SERCT are necessary components, and the removal of either one would likely lead to a decline in the model's performance. Ablation studies demonstrated that removing either the CNN or the Transformer module from the SERCT significantly degraded its performance. Without the Transformer (DST) module, the model's performance on the ESD dataset dropped to a WA, UA and WF1 of approximately 0.4732, 0.4731 and 0.4705, respectively . Conversely, removing the CNN module resulted in even lower scores. These results indicate that both modules are crucial for optimal performance, with the CNN module extracting low-level features and the Transformer module capturing long-range dependencies and contextual information. Based on this, the current hypothesis holds true.

**Validation of the Fifth Hypothesis** The fifth hypothesis is that the number of Transformer layers in SERCT should strike a balance between performance and the duration of training or inference. We evaluated the impact of varying the number of Transformer layers on the model's performance. The results showed that performance improved with an increasing number of layers up to a point, beyond which the gains plateaued and even slightly diminished. The optimal configuration was found to be around 3-4 layers, balancing performance with computational efficiency. This proves that the current hypothesis is valid.

**Validation of the Sixth Hypothesis** The sixth hypothesis is that SERCT's generalization capability on other languages is likely to be superior compared to other speech emotion recognition models. The SERCT model was tested for its generalization capabilities on non-English datasets (EMO-DB and ESD-zh). The model demonstrated robust performance, significantly outperforming other models. On the German EMO-DB dataset, SERCT achieved a WA of 0.3877, UA of 0.3712 and WF1 of 0.3892, while on the Chinese ESD-zh dataset, it achieved 0.5480, 0.5480 and 0.5224, respectively . These results validate that SERCT can generalize effectively across different languages without additional language-specific training.

## 7.2 Ethical Considerations

**Use of Public Datasets** The research conducted for this thesis on Speech Emotion Recognition (SER) strictly utilized publicly available datasets. These datasets are explicitly marked for academic research use, ensuring compliance with legal and ethical standards regarding data usage. The transparency of dataset availability and their intended purpose supports the ethical foundation of this research by adhering to permissions and avoiding any unauthorized use of proprietary data.

**Data Collection and Privacy** Throughout the research process, no new data collection involving human subjects was undertaken. This absence of personal data collection mitigates potential ethical concerns related to privacy, consent and the handling of sensitive information. Utilizing only existing, open-source data ensures that the study does not infringe on individuals' privacy rights or ethical research practices.

**Open Source Development** The development of the application based on this research was carried out with an emphasis on openness and transparency. All code and tools used are open source, and the complete codebase is accessible on GitHub with comprehensive documentation. This approach promotes transparency, allowing others to validate and extend the work without any restrictions. The open-source nature of the tools and data ensures that the research community can benefit and build upon the findings without legal or ethical hindrances.

**Contribution to the Field** The application developed from this research has the potential to advance SER technology across various industries, providing practical benefits while adhering to ethical guidelines. Additionally, the open-source nature of the application fosters innovation and ethical use in further developments.

This thesis on Speech Emotion Recognition has been conducted with a strong commitment to ethical standards, ensuring that data usage, development processes and application usage are all aligned with best practices in research ethics. The emphasis on open-source tools and transparent methodologies further strengthens the ethical integrity of the thesis.

## 7.3 Limitations

In conducting the research and developing the Speech Emotion Recognition based on CNN-Transformer (SERCT), several limitations were identified that could impact the generalizability and robustness of the findings. These limitations include the lack of data augmentation, insufficient exploration of the impact of CNN layer quantity on performance, absence of hyperparameter tuning, reliance on limited datasets and the simplicity of the developed SER application. Addressing these limitations in future work could enhance the model's performance and applicability in real-world scenarios.

**Lack of Data Augmentation** One significant limitation of this study is the absence of data augmentation techniques. Data augmentation is a crucial process in machine learning that involves creating new training samples through various modifications of the existing data, such as noise addition, pitch alteration, time stretching and other manipulations. This process helps improve the model's robustness and ability to generalize to unseen data by exposing it to a wider variety of training examples. In our study, the raw audio data was used as-is without any augmentation. As a result, the model may not perform optimally on real-world data that exhibits variations not present in the training set. Future studies should incorporate data augmentation to create a more diverse and representative dataset, which can help the model learn more robust features and improve its generalization capabilities.

**Insufficient Exploration of CNN Layer Quantity** Another limitation is the lack of investigation into the effect of varying the number of CNN layers on the performance of the SERCT model. Convolutional Neural Networks (CNNs) play a critical role in feature extraction from raw audio signals, and the depth of the CNN (i.e., the number of layers) can significantly influence the model's ability to capture intricate patterns and features. In this thesis, a fixed CNN architecture was used without experimenting with different configurations. As a result, it is unclear whether the chosen architecture is optimal or if performance improvements could be achieved by adjusting the number of CNN layers. Future research should conduct a systematic study on the impact of CNN layer depth, experimenting with both shallow and deep architectures to identify the optimal configuration for speech emotion recognition tasks.

**Absence of Hyperparameter Tuning** We did not conduct an extensive hyperparameter tuning process for the CNN and Transformer components of the SERCT model. Hyperparameters, such as learning rate, batch size, dropout rate and the number of attention heads in the Transformer, are critical to the model's performance. Proper tuning of these hyperparameters can lead to significant improvements in accuracy and efficiency. In this research, default or manually selected values were used without a comprehensive search for optimal settings. This limitation could result in suboptimal model performance. Future work should employ techniques such as grid search, random search or more advanced optimization methods like Bayesian optimization to systematically tune the hyperparameters and achieve the best possible performance.

**Limited Dataset Diversity** The experiments were conducted using only two datasets: the Toronto Emotional Speech Set (TESS) (Pichora-Fuller and Dupuis, 2020) and the Emotional Speech Dataset (ESD) (Zhou et al., 2021). While these datasets are well-established in the field, they may not fully represent the diversity and variability of speech and emotions in real-world scenarios. The limited number of datasets could lead to biased or overfitted models that perform well on the specific datasets used in training but fail to generalize to other data sources. Additionally, both datasets include only a small number of speakers and limited emotional categories, which may not capture the full spectrum of human emotional expression. Future studies should incorporate a broader range of datasets, including those with more speakers, different languages and diverse emotional expressions, to ensure the model's robustness and applicability across different contexts.

**Simplistic SER Application Development** The SER application developed based on the research findings is a simplified version and does not fully consider the complexities of real-world application environments. Real-world deployment of speech emotion recognition systems often involves handling noisy data, real-time processing, integration with other systems and ensuring privacy and security. The current application serves as a proof of concept but lacks the advanced features and optimizations required for practical use. For instance, it does not address issues such as latency, computational efficiency and the ability to process continuous speech streams in real-time. Future work should focus on developing a more sophisticated and robust application that can operate effectively in real-world conditions, including the implementation of noise reduction techniques, real-time processing capabilities and integration with broader systems.

The research on Speech Emotion Recognition based on CNN-Transformer (SERCT) has shown promising results, but several limitations need to be addressed to enhance the model's performance and applicability. The lack of data augmentation, insufficient exploration of CNN layer quantity, absence of comprehensive hyperparameter tuning, reliance on limited datasets and the simplistic nature of the developed application all present areas for improvement. Addressing these limitations in future research will be crucial for developing a more robust and generalizable speech emotion recognition system that can be effectively deployed in diverse real-world scenarios.

# 8 Conclusion

In this thesis, we proposed an end-to-end Speech Emotion Recognition (SER) model based on the integration of Convolutional Neural Networks (CNNs) and Transformer architecture, termed SERCT. Our research aimed to address the limitations inherent in traditional SER approaches that rely heavily on handcrafted features and domain-specific knowledge, which can be time-consuming and less adaptable to diverse datasets.

## 8.1 Main Contributions

The primary contributions of this thesis are multifaceted:

1. **End-to-End Architecture:** We introduced the CNN-Transformer architecture for SER, which eliminates the need for manual feature extraction by learning features directly from raw speech signals. This end-to-end approach streamlines the process, making it more efficient and less reliant on domain-specific expertise.
2. **Integration of CNNs and Transformers:** The proposed model leverages the strengths of both CNNs and Transformers. CNNs are adept at capturing local acoustic patterns, while Transformers excel at modeling long-range dependencies and global contextual information. This combination enables the model to effectively capture both fine-grained and high-level features necessary for accurate emotion recognition.
3. **Experimental Validation:** Extensive experiments were conducted on two widely used speech emotion datasets, TESS and ESD. The results demonstrated that the SERCT model significantly outperforms several baseline models, including MLP, LSTM, CNN-BiLSTM and SVM, across various performance metrics such as Weighted Accuracy (WA), Unweighted Accuracy (UA) and Weighted F1-score (WF1).
4. **Language Generalization:** We tested the model's ability to generalize across different languages without additional language-specific training. The SERCT model showed robust performance on non-English datasets, EMO-DB (German) and ESD-zh (Chinese), outperforming other models and highlighting its potential for multilingual applications.
5. **Ablation Studies:** Through ablation studies, we confirmed the essential roles of both the CNN and Transformer modules in achieving optimal performance. The removal of either module led to a significant drop in accuracy, underscoring the complementary strengths of these components.
6. **Application Development:** An application based on the proposed model was developed, demonstrating its practical utility and potential for real-world deployment. This application serves as a proof of concept for integrating SER capabilities into various systems such as customer service platforms and mental health assessment tools.

## 8.2 Limitations and Future Work

Despite the promising results, this study has several limitations that warrant further investigation:

1. **Data Augmentation:** We did not explore extensive data augmentation techniques, which could enhance the model's robustness and generalization capabilities, especially in low-resource settings.

2. **Hyperparameter Tuning:** The study could benefit from a more comprehensive exploration of hyperparameter tuning to optimize the model further.
3. **Real-World Application Complexity:** The developed application is a simplified version that does not fully account for the complexities of real-world environments, such as handling noisy data, real-time processing and system integration. Future work should focus on addressing these challenges to enhance the practicality of the SERCT model.
4. **Computational Efficiency:** While the model demonstrates high accuracy, the computational efficiency and latency issues associated with real-time deployment need to be thoroughly evaluated and optimized.
5. **Broader Dataset Diversity:** Future studies should incorporate a broader range of datasets, including those with more speakers, different languages and diverse emotional expressions, to ensure the model's robustness and applicability across various contexts.

In conclusion, the SERCT model represents an advancement in the field of Speech Emotion Recognition by providing a more streamlined, efficient and robust framework for emotion detection from speech. The integration of CNNs and Transformers leverages the best of both architectures, resulting in superior performance across multiple datasets and languages. Addressing the identified limitations in future work will further enhance the model's applicability and effectiveness in real-world scenarios. This research lays a basic foundation for the continued development and refinement of end-to-end SER systems, with potential applications spanning human-computer interaction, mental health assessment and beyond.

# Bibliography

Alif Bin Abdul Qayyum, Asiful Arefeen, and Celia Shahnaz. Convolutional neural network (cnn) based speech-emotion recognition. In *2019 IEEE International Conference on Signal Processing, Information, Communication Systems (SPICSCON)*, pages 122–125, 2019. doi: 10.1109/SPICSCON48833.2019.9065172.

Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019.

Tursunov Anvarjon, Mustaqeem, and Soonil Kwon. Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors*, 20(18), 2020. ISSN 1424-8220. doi: 10.3390/s20185212. URL https://www.mdpi.com/1424-8220/20/18/5212.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church. Speech emotion recognition with multi-task learning. In *Interspeech*, volume 2021, pages 4508–4512. Brno, 2021.

Qi Cao, Mixiao Hou, Bingzhi Chen, Zheng Zhang, and Guangming Lu. Hierarchical network based on the fusion of static and dynamic features for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6334–6338. IEEE, 2021.

Yi Chang, Zhao Ren, Thanh Tam Nguyen, Kun Qian, and Björn W Schuller. Knowledge transfer for on-device speech emotion recognition with neural structured learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Li-Wei Chen and Alexander Rudnicky. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. Dst: Deformable speech transformer for emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Théo Deschamps-Berger, Lori Lamel, and Laurence Devillers. End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings. In *2021 9th International*

*Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch. Cnn+lstm architecture for speech emotion recognition with data augmentation. In *Workshop on Speech, Music and Mind (SMM 2018)*. ISCA, September 2018. doi: 10.21437/smm.2018-5. URL http://dx.doi.org/10.21437/SMM.2018-5.

Tiantian Feng, Rajat Hebbar, and Shrikanth Narayanan. Trust-ser: On the trustworthiness of fine-tuning pre-trained speech embeddings for speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11201–11205. IEEE, 2024.

Eric Guizzo, Tillman Weyde, and Jack Barnett Leveson. Multi-time-scale convolution for emotion recognition from speech audio signals. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6489–6493. IEEE, 2020.

Lili Guo, Longbiao Wang, Chenglin Xu, Jianwu Dang, Eng Siong Chng, and Haizhou Li. Representation learning with spectro-temporal-channel attention for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6304–6308. IEEE, 2021.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

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

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998. doi: 10.1109/5254.708428.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Desheng Hu, Xinhui Hu, and Xinkang Xu. Multiple enhancements to lstm for learning emotion-salient features in speech emotion recognition. In *INTERSPEECH*, pages 4720–4724, 2022.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE* *conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R, and Rajesh Kumar Muthu. Speech emotion recognition using support vector machine, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Soonil Kwon et al. Att-net: Enhanced emotion recognition system using lightweight self-attention module. *Applied Soft Computing*, 102:107101, 2021.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching, 2018.

Chang Li. Robotic emotion recognition using two-level features fusion in audio signals of speech. *IEEE Sensors Journal*, 22(18):17447–17454, 2021.

Pengcheng Li, Yan Song, Ian Vince McLoughlin, Wu Guo, and Li-Rong Dai. An attention pooling based representation learning method for speech emotion recognition. 2018.

Runnan Li, Zhiyong Wu, Jia Jia, Sheng Zhao, and Helen Meng. Dilated residual network with multi-head self-attention for speech emotion recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6675–6679, 2019a. doi: 10.1109/ICASSP.2019.8682154.

Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *Proc. Interspeech 2019*, pages 2803–2807, 2019b. doi: 10.21437/Interspeech.2019-2594.

Jiawang Liu and Haoxiang Wang. A speech emotion recognition framework for better discrimination of confusions. In *Interspeech*, pages 4483–4487, 2021.

Jiaxing Liu, Zhilei Liu, Longbiao Wang, Lili Guo, and Jianwu Dang. Speech emotion recognition with local-global aware deep representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7174–7178. IEEE, 2020.

Xin Lu, Yanyan Zhao, and Bing Qin. Vanilla transformers are transfer capability teachers, 2024.

Brian McFee, Matt McVicar, Daniel Faronbi, Iran Roman, Matan Gover, Stefan Balke, Scott Seyfarth, Ayoub Malek, Colin Raffel, Vincent Lostanlen, Benjamin van Niekirk, Dana Lee, Frank Cwitkowitz, Frank Zalkow, Oriol Nieto, Dan Ellis, Jack Mason, Kyungyun Lee, Bea Steers, and Waldir Pimenta. librosa/librosa: 0.10.2.post1, June 2024. URL `https://doi.org/10.5281/zenodo.11192913`.

Hongying Meng, Nadia Bianchi-Berthouze, Yangdong Deng, Jinkuang Cheng, and John P. Cosmas. Time-delay neural network for continuous emotional dimension prediction from facial expression sequences. *IEEE Transactions on Cybernetics*, 46(4):916–929, 2016. doi: 10.1109/TCYB.2015.2418092.

Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231, 2017. doi: 10.1109/ICASSP.2017.7952552.

Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. Speech emotion recognition using self-supervised features. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6922–6926. IEEE, 2022.

Wei Mou, Pei-Hsuan Shen, Chu-Yun Chu, Yu-Cheng Chiu, Tsung-Hsien Yang, and Ming-Hsiang Su. Speech emotion recognition based on CNN+LSTM model. pages 43–47, Taoyuan, Taiwan, oct 2021. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). URL https://aclanthology.org/2021.rocling-1.6.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*, 2021.

M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (TESS), 2020. URL https://doi.org/10.5683/SP2/E8H2MF.

Amandine Pras and Catherine Guastavino. Sampling rate discrimination: 44.1 khz vs. 88.2 khz. In *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

Nicolae-Catalin Ristea, Radu Tudor Ionescu, and Fahad Shahbaz Khan. Septr: separable transformer for audio spectrogram processing. *arXiv preprint arXiv:2203.09581*, 2022.

Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak. Emotion Identification from Raw Speech Signals Using DNNs. In *Proc. Interspeech 2018*, pages 3097–3101, 2018. doi: 10.21437/Interspeech.2018-1353.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Ting-Wei Sun. End-to-end speech emotion recognition with gender information. *IEEE Access*, 8: 152423–152438, 2020.

Nam Khanh Tran and Claudia Niedereée. Multihop attention networks for question answer matching. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 325–334, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210009. URL https://doi.org/10.1145/3209978.3210009.

Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5089–5093. IEEE, 2018.

Panagiotis Tzirakis, Jiaxin Chen, Stefanos Zafeiriou, and Björn Schuller. End-to-end multimodal affect recognition in real-world environments. *Information Fusion*, 68:46–53, 2021a.

Panagiotis Tzirakis, Anh Nguyen, Stefanos Zafeiriou, and Björn W Schuller. Speech emotion recognition using semantic information. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6279–6283. IEEE, 2021b.

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

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Xianfeng Wang, Min Wang, Wenbo Qi, Wanqi Su, Xiangqian Wang, and Huan Zhou. A novel end-to-end speech emotion recognition network with stacked transformer layers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6289–6293. IEEE, 2021.

Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, Christian Fuegen, Geoffrey Zweig, and Michael L. Seltzer. Transformer-based acoustic modeling for hybrid speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2020. doi: 10.1109/icassp40776.2020.9054345. URL http://dx.doi.org/10.1109/ICASSP40776.2020.9054345.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Yu-Xuan Xi, Yan Song, Li-Rong Dai, Ian McLoughlin, and Lin Liu. Frontend attributes disentanglement for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7712–7716. IEEE, 2022.

Yue Xie, Ruiyu Liang, Zhenlin Liang, and Li Zhao. Attention-based dense lstm for speech emotion recognition. *IEICE Transactions on Information and Systems*, 102(7):1426–1429, 2019. doi: 10.1587/TRANSINF.2019EDL8019.

Hai-tao Xu, Jie Zhang, and Li-rong Dai. Differential time-frequency log-mel spectrogram features for vision transformer based infant cry recognition. *Proc. Interspeech 2022*, pages 1963–1967, 2022.

Yunfeng Xu, Hua Xu, and Jiyun Zou. Hgfm: A hierarchical grained and feature model for acoustic emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6499–6503. IEEE, 2020.

Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590, 2018. doi: 10.1109/TMM.2017.2766843.

Shiqing Zhang, Xiaoming Zhao, and Qi Tian. Spontaneous speech emotion recognition using multiscale deep convolutional lstm. *IEEE Transactions on Affective Computing*, 13(2):680–688, 2019a.

Zixing Zhang, Bingwen Wu, and Björn Schuller. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6705–6709. IEEE, 2019b.

Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical signal processing and control*, 47:312–323, 2019.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE, 2021.

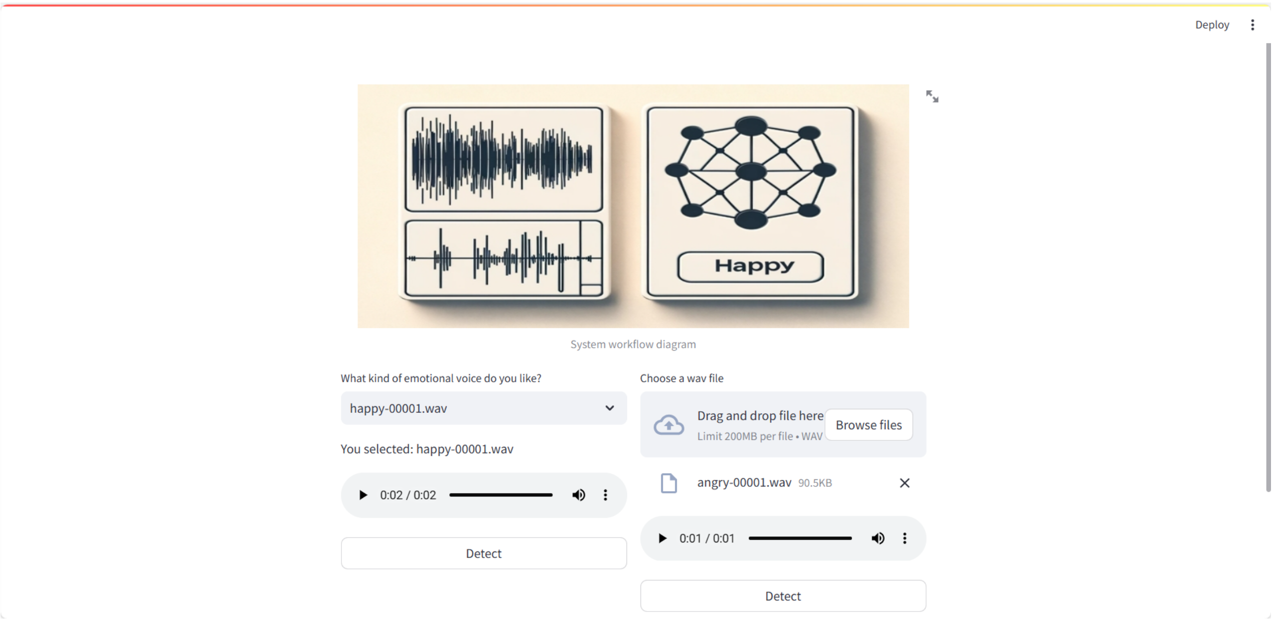
Wenjing Zhu and Xiang Li. Speech emotion recognition with global-aware fusion on multi-scale feature representation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6437–6441. IEEE, 2022.

Xinxin Zhu, Lixiang Li, Jing Liu, Haipeng Peng, and Xinxin Niu. Captioning transformer with stacked attention modules. *Applied Sciences*, 8(5), 2018. ISSN 2076-3417. doi: 10.3390/app8050739. URL https://www.mdpi.com/2076-3417/8/5/739.

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

# Appendices

## A Demonstrator



System workflow diagram

What kind of emotional voice do you like?

happy-00001.wav

You selected: happy-00001.wav

0:02 / 0:02

Choose a wav file

Drag and drop file here

Limit 200MB per file • WAV

angry-00001.wav 90.5KB

0:01 / 0:01

Figure 5: The user interface of our developed app.

We have developed a simple Speech Emotion Recognition (SER) application using Streamlit [4], an open-source app framework in Python. Streamlit allows the creation of web applications with minimal effort and is particularly useful for data scientists and machine learning practitioners who want to showcase their models and data in an interactive way.

The application interface, as shown in Figure 5, is straightforward and user-friendly. At the top of the interface, there is a system workflow diagram that visually represents the process of emotion detection from speech. The diagram includes representations of audio waveforms and a neural network, emphasizing the underlying technology used to analyze the emotional content of the voice.

To use the application, the user can select an emotional voice file from a dropdown menu on the left side of the interface. For example, the user might choose a file named ”happy-00001.wav”. Once selected, the filename is displayed below the dropdown menu and the user can play the audio by clicking the play button. This feature allows users to listen to the audio file to verify their selection before running the emotion detection.

On the right side of the interface, there is an option to upload a custom WAV file. The user can either drag and drop their file into the designated area or browse their local files to select one. This functionality ensures that users can test the application with their audio files, making the application versatile and applicable to various use cases. Once the file is uploaded, its name and size are displayed, and the user can also play this audio file using the provided play button. After selecting or uploading an audio file, the user can click the ”Detect” button to initiate the emotion detection

[4] Streamlit is available at https://streamlit.io/

process. The application processes the audio file and identifies the emotional content using a pre-trained model. The emotional state detected from the audio is displayed on the interface, typically showing emotions such as happy, sad, angry or neutral.