The Project Title

A Study on Speech Emotion Recognition Based on Multi-Feature and Deep Models

Matthew Leeke

School of Computer Science

College of Engineering and Physical Sciences

University of Birmingham

2024-25

Abstract

Acknowledgements

Abbreviations

ACB	Apple Banana Carrot

Contents

Abstract **ii**

Acknowledgements **iii**

Abbreviations **iv**

List of Figures **vi**

List of Tables **vii**

1 Introduction **1**
1.0.1 Background 1
1.0.2 Main Contents of the Dissertation 2
1.0.3 Structure of the Dissertation 2

2 Literature Review **3**

3 Methodology **4**
3.1 Feature Extraction 4
3.1.1 MFCC 4
3.1.2 Mel-spectrogram 5
3.2 Model Developments 6
3.2.1 Multi-layer Perceptron 6
3.2.2 Shallow CNN 7
3.2.3 ResNet-18 7
3.2.4 Audio Spectrogram Transformer 8

4 Experiments **9**

5 Discussion **10**

6 Conclusion **11**

7 Chapter Title **12**

8 Chapter Title **13**

9 Chapter Title **14**

10 Chapter Title **15**

List of Figures

List of Tables

CHAPTER 1

Introduction

1.0.1 Background

Speech conveys rich affective information through prosody, articulation, and spectral characteristics, enabling listeners to infer speakers' intentions and internal states. Speech Emotion Recognition (SER) aims to computationally infer human emotion from acoustic signals, either as discrete categories (e.g., anger, happiness, sadness, fear, and neutral) or along continuous dimensions (e.g., valence, arousal, and dominance). Rooted in speech processing, affective computing, and psychology, early SER systems relied on handcrafted prosodic and spectral features combined with classical classifiers. Over the past decade, deep learning has transformed the field: convolutional networks on time–frequency representations, recurrent and attention-based models for temporal dynamics, and, more recently, transformer architectures and self-supervised pretraining have markedly improved generalization. The research focus has also shifted from acted laboratory corpora to spontaneous, in-the-wild data, emphasizing robustness to noise, channel variability, and cultural–linguistic diversity. In parallel, there is growing attention to transparency, privacy, and fairness, reflecting the societal implications of emotion inference technologies.

SER has broad practical value. In customer service and contact centers, emotion-aware analytics can monitor user satisfaction and support agent coaching in real time. In healthcare and mental health, non-invasive acoustic markers assist in screening and longitudinal monitoring of affective disorders, stress, and burnout. In human–computer interaction, conversational assistants and social robots can adapt responses and dialogue strategies based on users' emotional cues. In automotive and smart environments, continuous monitoring of driver frustration or fatigue enhances safety and personalization. Education technology, entertainment, and recommendation systems also benefit from affect-aware feedback and content curation. These applications motivate solutions that are accurate, low-latency, and resource-efficient for both cloud and on-device deployment.

The research significance of SER is twofold. Scientifically, SER advances

our understanding of how affect is encoded in speech and how to learn robust, interpretable representations under limited, imbalanced, and noisy supervision. Technically, it drives innovation in domain generalization, cross-lingual transfer, and multimodal fusion with text and vision, while promoting explainability to build user trust. Addressing ethical and legal considerations—including data privacy, informed consent, and bias mitigation—is essential to responsible deployment. Finally, rigorous benchmarking, reproducibility, and standardized evaluation protocols (e.g., unweighted/weighted accuracy for categorical tasks and concordance correlation for continuous estimation) are critical for cumulative progress. Altogether, SER research provides both theoretical insights and practical tools for empathetic, human-centered intelligent systems.

1.0.2 Main Contents of the Dissertation

This dissertation addresses the challenge of suboptimal recognition accuracy in Speech Emotion Recognition by developing a deep learning based framework that integrates speech signal preprocessing, feature engineering, attention mechanisms, and model optimization. We conduct a systematic study of SER methods from multiple perspectives, including audio preprocessing, data augmentation, feature selection, model architectures, and evaluation metrics. The central objective is to learn feature representations with stronger affective discriminability directly from speech, thereby improving accuracy, robustness, and generalization across speakers, recording conditions, and datasets. The proposed approach emphasizes practical deployability by balancing performance with computational efficiency for both cloud and on-device inference.

1.0.3 Structure of the Dissertation

The dissertation is organized as follows. Chapter 1 introduces the research background and motivation, identifies the key challenges in SER, and outlines the scope and structure of the thesis. Chapter 2 surveys related work on speech signal processing, feature extraction, the design of attention mechanisms, and emotion taxonomies and recognition paradigms, summarizing major international developments. Chapter 3 presents the methodology. We construct features from audio data using Mel-frequency cepstral coefficients (MFCCs) and Mel-spectrograms; we then investigate multiple models for SER, including a multilayer perceptron (MLP), a shallow convolutional neural network (CNN), ResNet-18 for 2D Mel-spectrogram inputs, and the Audio Spectrogram Transformer (AST), a ViT-based model for audio classification. Finally, we apply data augmentation by injecting noise during training to improve generalization and mitigate overfitting. Chapter 4 describes the experiments. We design comparative studies to evaluate the performance differences among features and models on SER tasks under consistent protocols and metrics. Chapter 5 provides discussion. We analyze the experimental results in depth, examine the relative strengths and weaknesses of different features and models, and articulate limitations and directions for future research. Chapter 6 concludes the thesis by summarizing the main findings and contributions, and highlighting potential avenues for subsequent work.

CHAPTER 2

Literature Review

CHAPTER 3

Methodology

3.1 Feature Extraction

This section details the feature extraction pipeline and unified notation used throughout the thesis. Let $x[n]$ denote the discrete-time speech signal sampled at rate f_s (Hz). Frames are indexed by m, frequency bins by k, and Mel bands by f. The frame length is N samples, hop size is H samples, the analysis window is $w[n]$ (Hann unless otherwise stated), the discrete Fourier transform (DFT) length is K, and the number of Mel filters is F. We use $X[k, m]$ for the short-time spectrum, $P[k, m] = |X[k, m]|^2$ for the power spectrum, $\mathbf{B} \in \mathbb{R}^{F \times K}$ for the Mel filterbank, and a small constant $\varepsilon > 0$ for numerical stability.

3.1.1 MFCC

Mel-frequency cepstral coefficients (MFCCs) are compact features that approximate human auditory perception by (i) warping frequency to the Mel scale, (ii) applying a band-pass filterbank to obtain perceptual energies, (iii) compressing the dynamic range, and (iv) decorrelating energies via a discrete cosine transform (DCT). The standard pipeline is as follows.

1) Pre-emphasis (optional) enhances high-frequency components:

$$y[n] = x[n] - \alpha\, x[n-1], \quad \alpha \in [0.95, 0.99]. \tag{3.1}$$

2) Framing and windowing create approximately stationary short-time segments:

$$X[k, m] = \sum_{n=0}^{N-1} x[n + mH]\, w[n]\, e^{-j2\pi kn/K}, \quad 0 \le k < K. \tag{3.2}$$

3) Power spectrum is computed as

$$P[k, m] = \frac{|X[k, m]|^2}{K}. \tag{3.3}$$

4) Mel filterbank energies are obtained by triangular filters spaced on the Mel scale:

$$S[f,m] = \sum_{k=0}^{K-1} B[f,k]\, P[k,m], \quad 1 \le f \le F. \tag{3.4}$$

The Mel frequency mapping for a linear frequency ν (in Hz) is

$$\mathrm{mel}(\nu) = 2595 \, \log_{10}\!\left(1 + \frac{\nu}{700}\right), \tag{3.5}$$

and the filterbank $\mathbf{B}$ is formed by piecewise-linear triangular filters uniformly spaced on the Mel axis and then mapped back to linear frequency bins.

5) Logarithmic compression reduces dynamic range and approximates loudness perception:

$$\widehat{S}[f,m] = \log\big(S[f,m] + \varepsilon\big). \tag{3.6}$$

6) DCT-II decorrelates the log-Mel energies, yielding MFCCs:

$$c_q[m] = \sum_{f=1}^{F} \widehat{S}[f,m] \cos\!\Big(\frac{\pi q}{F}\Big(f - \tfrac{1}{2}\Big)\Big), \quad 0 \le q < Q, \tag{3.7}$$

where Q is the number of cepstral coefficients retained (commonly $Q \in [12, 20]$). An optional sinusoidal lifter improves energy distribution across coefficients:

$$\widetilde{c}_q[m] = \Big(1 + \frac{L}{2} \sin\frac{\pi q}{L}\Big) c_q[m], \quad L > 0. \tag{3.8}$$

Temporal dynamics are often captured by regression-based derivatives using a symmetric window of radius R:

$$\Delta c_q[m] = \frac{\sum_{r=1}^{R} r\,\big(c_q[m+r] - c_q[m-r]\big)}{2\sum_{r=1}^{R} r^2}, \quad \Delta\Delta c_q[m] \text{ analogously.} \tag{3.9}$$

Advantages: MFCCs are compact, relatively robust to slowly varying convolutive effects, and widely validated in speech tasks. Limitations: information loss due to coarse spectral integration and DCT decorrelation, frame-level stationarity assumptions, and sensitivity to noise and channel mismatch without appropriate normalization or enhancement.

3.1.2 Mel-spectrogram

The Mel-spectrogram represents short-time spectral energy projected onto a perceptual frequency axis, typically used directly as a 2D input to convolutional or transformer models.

Given the short-time spectrum in (3.2) and power spectrum in (3.3), the Mel-spectrogram is

$$\mathrm{MelSpec}[f,m] = \sum_{k=0}^{K-1} B[f,k]\, P[k,m], \quad 1 \le f \le F. \tag{3.10}$$

A log or power-law compression is commonly applied to stabilize training and approximate loudness:

$$\text{LogMel}[f,m] = \log\big(\text{MelSpec}[f,m] + \varepsilon\big) \quad \text{or} \quad \text{MelSpec}[f,m]^{\gamma},\ \gamma \in (0,1]. \tag{3.11}$$

Design choices include the sample rate f_s, frame length N, hop size H, window type $w[n]$, number of Mel filters F, lower/upper frequency bounds of the filterbank, and whether to use magnitude or power spectra. Per-utterance or global mean-variance normalization is often applied to reduce channel variability.

Advantages: Mel-spectrograms preserve local time–frequency structure and are highly compatible with CNNs and attention-based encoders; they avoid the information loss introduced by the DCT in MFCC. Limitations: higher dimensionality increases memory and compute cost; the representation remains sensitive to additive noise and reverberation without enhancement or robust training; choices of F, bounds, and f_s affect resolution and cross-dataset transferability.

3.2 Model Developments

We consider a label set $\mathcal{C}$ with $|\mathcal{C}| = C$ emotion classes. For utterance-level classification, we denote the per-frame MFCC feature as $\mathbf{z}_m \in \mathbb{R}^Q$ from (3.7) (optionally concatenated with Δ and $\Delta\Delta$ features from (3.9)). A fixed-dimensional utterance representation is obtained by temporal mean pooling

$$\bar{\mathbf{z}} = \frac{1}{T}\sum_{m=1}^{T} \mathbf{z}_m \in \mathbb{R}^D, \quad (D = Q \text{ or } 3Q), \tag{3.12}$$

optionally augmented with standard deviation pooling. Let $\mathbf{y} \in \{0,1\}^C$ be the one-hot label vector.

3.2.1 Multi-layer Perceptron

We adopt a feed-forward multilayer perceptron (MLP) as a lightweight baseline operating on MFCC-based utterance vectors. Given input $\bar{\mathbf{z}}$ from (3.12), a depth-L MLP computes

$$\mathbf{h}^{(1)} = \sigma\big(\mathbf{W}^{(1)}\bar{\mathbf{z}} + \mathbf{b}^{(1)}\big), \tag{3.13}$$

$$\mathbf{h}^{(\ell)} = \sigma\big(\mathbf{W}^{(\ell)}\mathbf{h}^{(\ell-1)} + \mathbf{b}^{(\ell)}\big), \quad 2 \le \ell \le L-1, \tag{3.14}$$

$$\mathbf{o} = \mathbf{W}^{(L)}\mathbf{h}^{(L-1)} + \mathbf{b}^{(L)}, \tag{3.15}$$

where $\sigma(\cdot)$ is a pointwise nonlinearity (ReLU by default), $\mathbf{W}^{(\ell)}$ and $\mathbf{b}^{(\ell)}$ are trainable parameters. Class probabilities are

$$\mathbf{p} = \text{softmax}(\mathbf{o}), \quad p_c = \frac{\exp(o_c)}{\sum_{c'=1}^{C}\exp(o_{c'})}. \tag{3.16}$$

We train with cross-entropy and ℓ_2 regularization:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log p_c + \lambda\|\Theta\|_2^2, \tag{3.17}$$

where Θ collects all parameters and λ controls weight decay. Dropout between hidden layers improves generalization.

Role in this thesis: the MLP serves as a parameter-efficient baseline with MFCC inputs, quantifying the incremental benefits of time–frequency 2D models.

3.2.2 Shallow CNN

For time–frequency inputs we use the log-Mel representation $\text{LogMel} \in \mathbb{R}^{F\times T}$ from (3.11). A shallow 2D CNN treats LogMel as a single-channel image: $\mathbf{X} \in \mathbb{R}^{1\times F\times T}$. Let a 2D convolution with kernel $\mathbf{W} \in \mathbb{R}^{C_{\text{out}}\times C_{\text{in}}\times k_f\times k_t}$, stride (s_f, s_t), and padding (p_f, p_t) be

$$(\mathbf{W} * \mathbf{X})[c, f, t] = \sum_{c'=1}^{C_{\text{in}}} \sum_{i=0}^{k_f-1} \sum_{j=0}^{k_t-1} \mathbf{W}[c, c', i, j]\, \mathbf{X}[c', f+i-p_f,\, t+j-p_t]. \quad (3.18)$$

Each convolutional block applies convolution, batch normalization, ReLU, and pooling:

$$\mathbf{U}_1 = \text{ReLU}\big(\text{BN}(\mathbf{W}_1 * \mathbf{X})\big), \qquad \mathbf{V}_1 = \text{Pool}(\mathbf{U}_1), \quad (3.19)$$

$$\mathbf{U}_2 = \text{ReLU}\big(\text{BN}(\mathbf{W}_2 * \mathbf{V}_1)\big), \qquad \mathbf{V}_2 = \text{Pool}(\mathbf{U}_2). \quad (3.20)$$

We then apply global average pooling over time and frequency and a linear classifier to obtain logits $\mathbf{o}$ followed by (3.16).

Design: we use two convolutional blocks with moderate kernels (e.g., $k_f \in \{3, 5\}$, $k_t \in \{3, 5\}$), stride 1, and max-pooling with small windows to preserve resolution. Regularization includes dropout and additive noise during training to improve robustness.

Advantages: shallow CNNs are data-efficient and fast, capturing local spectro-temporal patterns. Limitations: weaker long-range modeling and limited receptive field without deeper stacks or dilations. In this thesis, the shallow CNN consumes log-Mel features to quantify the benefits of local 2D structure over MFCC-based MLPs.

3.2.3 ResNet-18

We adopt ResNet-18 as a deeper convolutional baseline to enhance receptive field and hierarchical feature learning on LogMel. A basic residual block computes

$$\mathbf{y} = \mathcal{F}(\mathbf{x};\, \{\mathbf{W}_i\}) + \mathcal{S}(\mathbf{x}), \quad (3.21)$$

where $\mathcal{F}$ is a stack of Conv→BN→ReLU→Conv→BN, and $\mathcal{S}$ is either identity or a 1×1 projection to match shape when stride/downsampling is used. The network stacks such blocks with increasing channels and occasional stride-2 to downsample along time and frequency. After global average pooling, a linear classifier produces logits $\mathbf{o}$ and probabilities via (3.16).

For spectrogram inputs, the first convolution is adapted to single-channel input ($C_{\text{in}} = 1$) with kernel size 7×7 and stride 2, followed by batch normalization and ReLU. Weight decay and label smoothing may be applied to improve generalization.

Advantages: ResNet-18 balances capacity and efficiency, modeling broader context and invariances. Limitations: increased compute relative to shallow CNNs and potential overfitting on small corpora without strong regularization. In this thesis, ResNet-18 processes LogMel to assess the gains of residual learning.

3.2.4 Audio Spectrogram Transformer

The Audio Spectrogram Transformer (AST) adapts the Vision Transformer (ViT) to audio by treating the log-Mel spectrogram as a 2D patch sequence. Given LogMel $\in \mathbb{R}^{F\times T}$, we partition it into non-overlapping patches of size $P_f \times P_t$, yielding $N = \lfloor F/P_f \rfloor \cdot \lfloor T/P_t \rfloor$ patches. Each patch is flattened and linearly projected to a token embedding:

$$\mathbf{z}_i = \mathrm{vec}(\mathrm{patch}_i)\,\mathbf{W}_e + \mathbf{b}_e \in \mathbb{R}^d, \quad i = 1, \ldots, N. \tag{3.22}$$

A learnable classification token $\mathbf{z}_{\mathrm{cls}}$ is prepended and positional embeddings $\mathbf{p}_i$ are added. The resulting sequence $\{\mathbf{z}_{\mathrm{cls}}, \mathbf{z}_1, \ldots, \mathbf{z}_N\}$ is processed by a stack of transformer encoder layers, each composed of multi-head self-attention (MHSA) and a positionwise feed-forward network (FFN) with residual connections and layer normalization. For a single head,

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3.23}$$

with queries $\mathbf{Q} = \mathbf{Z}\mathbf{W}^Q$, keys $\mathbf{K} = \mathbf{Z}\mathbf{W}^K$, values $\mathbf{V} = \mathbf{Z}\mathbf{W}^V$, and $\mathbf{Z}$ the input token matrix; MHSA concatenates multiple heads. The FFN uses two linear layers and a nonlinearity (e.g., GELU). The classification head reads the [CLS] token to produce logits $\mathbf{o}$ and probabilities via (3.16).

Design: patch sizes (P_f, P_t) balance spectral and temporal resolution; smaller patches increase sequence length and cost but improve detail. Regularization includes dropout and stochastic depth; data augmentation includes time/frequency masking and additive noise.

Advantages: AST captures global dependencies and long-range context beyond CNN receptive fields. Limitations: higher computational cost and sensitivity to data scale; careful regularization and pretraining can be beneficial. In this thesis, AST consumes log-Mel inputs to examine the benefits of attention-based modeling.

CHAPTER 4

Experiments

CHAPTER 5

Discussion

CHAPTER 6

Conclusion

CHAPTER 7

Chapter Title

CHAPTER 8

Chapter Title

CHAPTER 9

Chapter Title

CHAPTER 10

Chapter Title

Bibliography