Research on Speech Emotion Recognition Based on Machine Learning Approaches

Liu Zheng

A Thesis submitted to Tokushima University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

September, 2023

Department of Information Science and Intelligent Systems
Graduate School of Advanced Technology and Science
Tokushima University, Japan

Contents

List of Tables

List of Figures

Abstract

With the continuous development of deep learning technology, research on human-machine emotional interaction has made significant progress. Speech is one of the critical means of information transmission, playing an indispensable role in human life by conveying not only semantic but also emotional information. In recent years, emotion recognition based on speech has received much attention due to the widespread application of deep learning technology and the rise of affective computing. Accurately extracting emotional information from speech signals is a vital issue in this field.

However, one of the significant challenges in developing high-performance speech emotion recognition systems is the lack of sufficient data. Constructing a high-quality emotional corpus requires a considerable investment of time and resources, including professional actors to perform voices with various emotions in specific scenes, and efficient data labeling to overcome the issue of imbalanced samples. Meanwhile, in order to improve the accuracy of speech emotion recognition, traditional methods only extract features from local datasets, which leads to overfitting of the models due to the weak robustness of the features. Therefore, constructing effective speech features is crucial to improving the accuracy of speech emotion recognition systems.

In this study, we propose an integrated active learning sampling strategy and an efficient framework to construct a speech emotional corpus effectively. Our method outperforms other active learning algorithms by improving sampling efficiency and selecting small category samples to be labeled with preference in imbalanced datasets. In actual corpus construction experiments, our method can prioritize selecting small class emotion samples, achieving an accuracy rate of 90% even with less than 50% of labeled data. This greatly enhances the efficiency of constructing the speech emotion corpus.

Additionally, we enhance the robustness of speech features using self-supervised learning and propose a feature fusion model (called Dual-TBNet) that consists of two 1D convolutional layers, two Transformer modules, and two bidirectional long short-term memory (BiLSTM) modules. Our model fuses five pre-trained features and acoustic features using the attention mechanism to capture the correspondence between the two features and enhance the contextual information of the fused features. In the comparison experiments, the Dual-TBNet model achieved a recognition accuracy and F1 score of 95.7% and 95.8% on the CASIA dataset, 66.7% and 65.6% on the eNTERFACE05 dataset, 64.8% and 64.9% on the IEMOCAP dataset, 84.1% and 84.3% on the EMO-DB dataset and 83.3% and 82.1% on the SAVEE dataset. The Dual-TBNet model effectively fuses acoustic features of different lengths and dimensions with pre-trained features, enhancing the robustness of the features, and achieved the best performance.

Keywords: Affective computing, Speech emotion recognition, Speech corpus construction, Active learning, Feature fusion

1 Introduction

1.1 Motivation

The continuous development of information technology has had a profound impact on human society, ranging from fulfilling basic survival needs to addressing spiritual needs. The advent of artificial intelligence (AI) technology has caused a significant stir in the era of information. Starting from the well-known achievement of Google's AlphaGo defeating the world Go champion in 2016 [102], which brought artificial intelligence into the limelight, to the recent introduction of ChatGPT, which revolutionized text productivity [13], artificial intelligence technology has once again garnered attention. However, the current state of artificial intelligence technology primarily revolves around perceptual intelligence, and a crucial step towards achieving cognitive intelligence lies in the exploration of emotions.

The notion of affective computing was initially proposed by Professor Picard in 1997, and for over two decades, it has emerged as one of the foremost research areas in human-computer interaction. The primary objective of affective computing is to equip intelligent systems with the capability to recognize, perceive, and generate human emotions [93, 94]. In recent years, with the advancement of foundational technologies such as 5G, blockchain, and cloud computing, coupled with the initiatives of internet giants like Facebook, the concept of the metaverse has gradually gained public attention [114, 59, 89]. The metaverse represents a digital realm constructed through technology that enables interaction with the real world, offering opportunities for learning, living, working, and more. For virtual characters within the metaverse to engage with humans more realistically, they require human-like thinking and emotions. Hence, research in the field of affective computing plays a crucial role in enhancing the interactive experience within the metaverse.

Speech emotion recognition is a crucial research area within the field of affective computing. As a significant mode of human communication, its objective is to identify and analyze the emotional content embedded within speech signals. By doing so, it empowers intelligent systems to enhance their service capabilities and adapt their attitudes towards humans more effectively [31, 103, 115, 1]. Speech emotion recognition has a wide range of applications and can have a profound impact on daily life and work. For instance, it can greatly contribute to enhancing the driving experience and promoting road safety by accurately detecting drivers' emotional states [123]. In the initial diagnosis of depression patients, speech emotion recognition can provide valuable reference information to assist doctors in making more accurate diagnoses [44]. In the field of customer service, recognizing customers' speech emotions can lead to better understanding of their needs and desires, enabling the provision of more personalized and professional services, and ultimately improving customer satisfaction [43]. Furthermore, within online learning and education systems, the utilization of speech emotion recognition technology to analyze students' states during classes can assist teachers in making informed decisions and enhancing the quality of instruction [50]. The development and application of speech emotion recognition technology will have a profound impact on human society. It not only facilitates intelligent systems in better understanding and serving humans, but also provides humans with more intelligent experiences in their daily lives and work, ultimately contributing to a better quality of life.

1.2 Significance of Research

Speech emotion recognition is an important research direction in the field of affective computing, which can analyze emotional information in human language expression, such as positive, negative, and neutral emotions. Deep learning models have shown excellent performance in speech emotion recognition tasks, but obtaining these high-performance models requires a large amount of training data, and building high-quality emotional corpora is a very expensive project that requires a lot of manpower and resources [113]. Currently, there are several shortcomings in the methods used to construct datasets. Firstly, most datasets are obtained from professional actors performing in specific scenarios, so emotional expressions are relatively exaggerated and less natural, and cannot cover emotional expressions in multiple languages and different application scenarios. Secondly, during the annotation process of speech data, due to the uneven distribution of different types of emotions, annotators need to listen to all speech data one by one for annotation, resulting in high annotation costs and the inability to prioritize rare samples. In addition, most datasets are collected in specific scenarios, which means that the accuracy of speech emotion recognition will be influenced by different languages, speaking styles, and application scenarios in practical application scenarios. Therefore, how to efficiently construct speech emotion datasets is currently a challenge.

Additionally, most research in this field is based on extracting features corresponding to the dataset for model training. These features only reflect the features of the current dataset, so they have poor robustness. To improve the robustness of features, pre-trained models can be used to extract high-dimensional feature spaces. For example, using speech representation learning methods to build feature spaces on a large amount of speech data, learning personalized speech features, and enhancing the robustness of speech features. For these features, exploring fusion solutions between different categories of speech features can effectively promote the improvement of speech emotion recognition accuracy.

1.3 Research Contents and Contributions

A speech emotion recognition system utilizes acoustic feature analysis and machine learning technologies to identify the emotional states conveyed in speech. As shown in Figure 1.1, the system begins by collecting speech sound waves through a microphone, with the sound card acquiring the analog signal from the microphone. The amplitude signal of the sound wave is then converted into a digital signal through processes like sampling, quantization, and encoding, which is stored in a computer. The fundamental steps of speech emotion recognition encompass feature extraction, feature processing, and emotion classification. Feature extraction involves techniques such as pre-emphasis, framing, windowing, and endpoint detection to break down the audio data and compute acoustic features, including spectral features. Feature processing aims to manipulate the

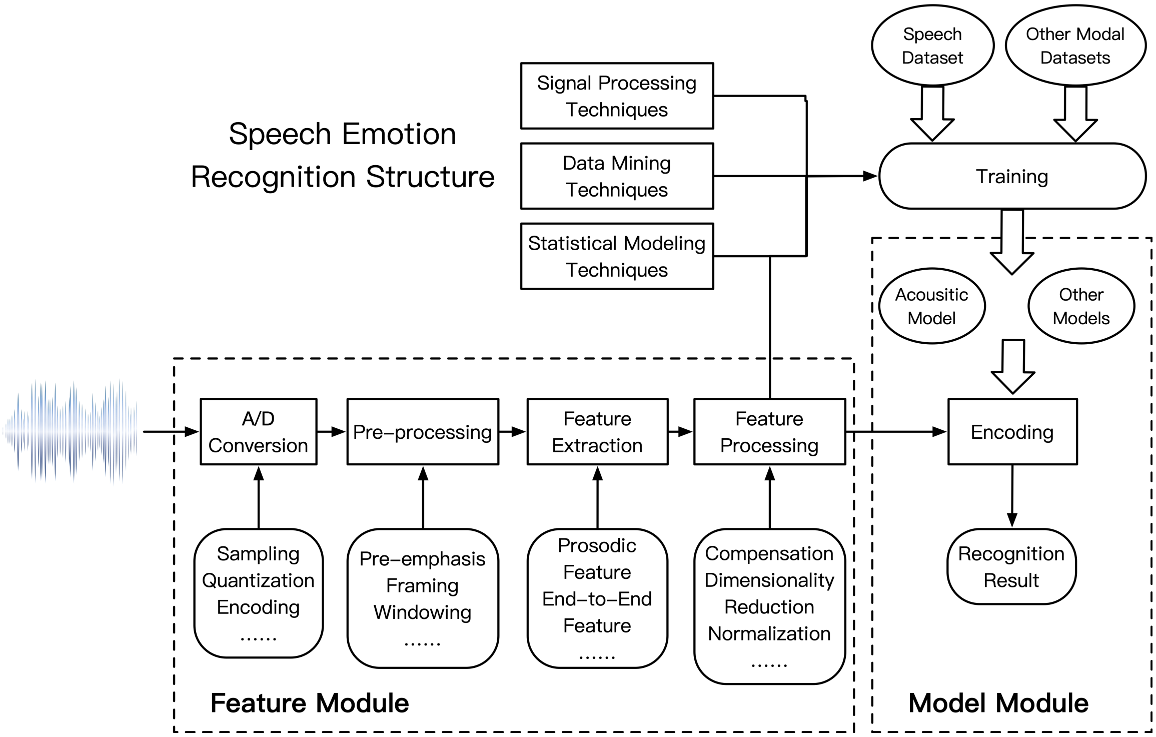

Speech Emotion Recognition Structure
Signal Processing Techniques
Data Mining Techniques
Statistical Modeling Techniques
Speech Dataset
Other Modal Datasets
Training
Acousitic Model
Other Models
Encoding
Recognition Result
A/D Conversion
Pre–processing
Feature Extraction
Feature Processing
Sampling Quantization Encoding ......
Pre–emphasis Framing Windowing ......
Prosodic Feature End–to–End Feature ......
Compensation Dimensionality Reduction Normalization ......
Feature Module
Model Module

Fig. 1.1. Overall architecture of speech emotion recognition system.

calculated features, incorporating tasks such as dimensionality reduction, compensation, elimination of redundant information, and inclusion of supplementary features that contribute to speech emotion expression. Emotion classification refers to the construction of a speech emotion recognition model, often accomplished by partitioning the speech dataset and employing a test set to evaluate the trained machine learning or deep learning recognition model. Through iterative adjustments, the final model is obtained. These steps enable the speech emotion recognition system to accurately identify and categorize emotional states within speech signals, ultimately enhancing people's speech interaction experience with greater intelligence.

To solve the problem of the lack of datasets in the field of SER, we designed a framework for speech emotion corpus construction. The framework integrates the newly proposed active learning strategy, and at the same time uses voice activity detection, feature extraction and other technologies. The raw voice of our framework is not limited to the performance of the actors, so it can be adapted to voice data collection in different scenarios.

In addition, to solve the problem of low speech emotion recognition rate caused by the weak robustness of acoustic features, we propose a novel feature fusion model Dual-TBNet, which contains two Transformers and two BiLSTM structures. In addition, to our knowledge, our work is the first to fuse self-supervised pre-trained features and acoustic features with different segment lengths and dimension sizes for speech emotion recognition.

The main contributions of our work are summarized as follows:

1. We propose a new active learning strategy using cross entropy distance for the sampling of imbalanced datasets. Based on the new strategy, we proposed an integrated active learning method, which can not only improve the efficiency of overall sampling, but also give priority to the small class samples of imbalanced dataset.

2. We design a framework for constructing emotional speech emotional corpus. The proposed framework can be applied to the audio collection in different scenarios.

Regarding the lack of corpus in the field of speech emotion recognition, our work provides an efficient method for constructing a speech emotion corpus.

3. In the exploration of acoustic features, we list and count the feature distribution under different segment lengths and prove that the features have a very diverse distribution under the length of 200ms, which is easy to fit the network model.

4. Traditional feature fusion schemes often fuse between acoustic features. Our study is the first in the field of speech emotion recognition to fuse pre-trained speech features and acoustic features to improve the robustness of the features.

5. For the feature fusion method, different from the traditional simple feature concatenation, we propose a novel feature fusion model Dual-TBNet, which is able to capture the correspondence between the two features with different segment lengths and dimension sizes at an early stage and fuse the two features more effectively.

1.4 Thesis Organizations

In this doctoral thesis, we reviewed and summarized various aspects of the field of speech emotion recognition, including speech emotion recognition models, speech emotion datasets, speech features, and speech recognition models. We also proposed our own solutions to challenges faced by this field, such as dataset construction and low accuracy in speech recognition. The organizational structure of this thesis is as follows:

Chapter 1: Introduction

In this chapter, we talk about the motivation, significance of this research and introduce the research contents and contributions of our work.

Chapter 2: Background

In this chapter, we conducted a comprehensive review of the different components within the domain of speech emotion recognition. This included emotion description models, speech emotion datasets, speech features, and speech emotion recognition models.

Chapter 3: Active Learning for Speech Emotion Corpus Construction

In this chapter, to address the issues of limited availability and low efficiency in constructing speech emotion datasets, we proposed a comprehensive active learning strategy. This strategy incorporated four data selection methods, namely uncertainty, representativeness, diversity, and complementarity, to identify more valuable data for annotation. Experimental results demonstrated the superiority of our approach.

Chapter 4: Dual-Transformer-BiLSTM for Speech Emotion Recognition

In this chapter, to enhance the accuracy of speech emotion recognition, we focused on improving the richness of speech features to enhance the robustness of the model. We proposed a novel feature fusion architecture that integrates two Transformer and BiLSTM modules, allowing for effective fusion of features with different dimensionalities and lengths. Experimental results demonstrated that our approach achieved state-of-the-art performance.

Chapter 5: Conclusion and Future work

In this chapter, we summarize the main contents of this thesis and give meaningful directions for future work.

2 Background

2.1 Emotion Description Models

Emotion recognition from speech is a crucial field within affective computing, aiming to identify human emotions by analyzing speech signals. Human emotions, responsive to intricate external stimuli, find expression through diverse channels such as facial expressions, voice tone, body language, and heart rate. These emotional states offer insights into individuals' internal psychological experiences, profoundly influencing their behavior [88]. For comprehensive research in affective computing, effective modeling and quantification of emotional states are crucial. Two primary methods stand out: categorical label-based representation and continuous dimensional representation of emotions.

Categorical models of emotional description assign distinct labels to individual emotions, such as happy, sad, or neutral. The categorical emotion model is more similar to the way people express themselves in daily life, it is simple and easy to understand. However, this approach isolates emotions from each other and doesn't capture intricate connections between them accurately. Its capacity to convey emotions with precision and continuity is limited. Furthermore, categorizing the entirety of human emotions using only a handful of labels proves inadequate. Moreover, discrepancies exist in how different researchers classify categorical emotions. In the classical categorical emotional models, the one introduced by Ekman in 1972 divides emotions into six basic categories: anger, disgust, fear, happiness, sadness, and surprise [34]. This categorization has gained substantial recognition. Another classical model is Plutchik's emotion wheel as shown in Fig.2.1, classifying emotions into eight primary types, with other emotions deriving from these. This model employs a color gradient to represent the intensity of emotions [90].

Dimensional emotional models, unlike categorical models, employ a continuous multi-dimensional emotional space to represent emotions, offering a more nuanced understanding. Each point in this space signifies an emotion, allowing for smoother transitions between emotions. Distances between emotions in this space symbolize their relationships. Dimensional models theoretically encompass the entirety of human emotions.

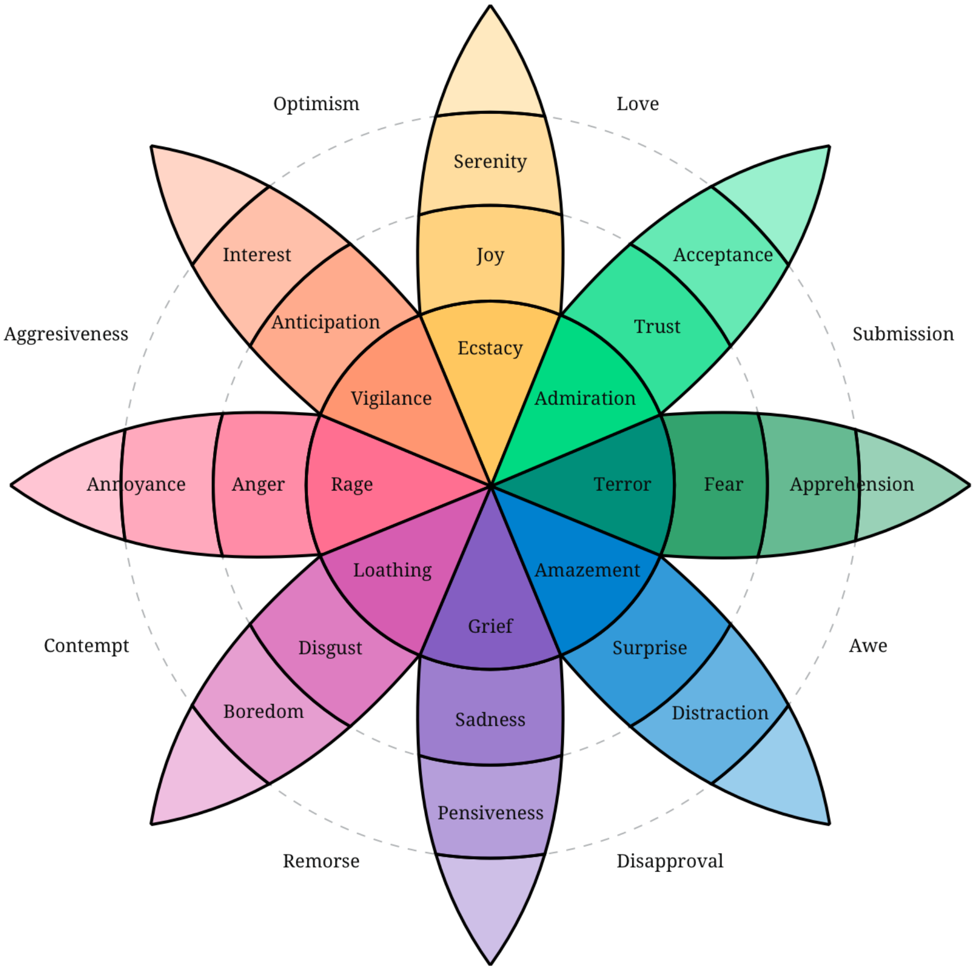


Fig. 2.1. Plutchik’s wheel of emotion.

Fontaine [38] proposed the PAD model as shown in Fig.2.2, which partitions emotional space into three dimensions: pleasure, arousal, and dominance. The pleasure dimension gauges pleasure levels, ranging from distress to ecstasy. The arousal dimension measures physiological and mental alertness, spanning from low arousal (sleepiness) to high arousal (tension). Dominance measures one's control over self and environment, varying from submission to control. The PAD model has gained widespread acceptance among researchers.

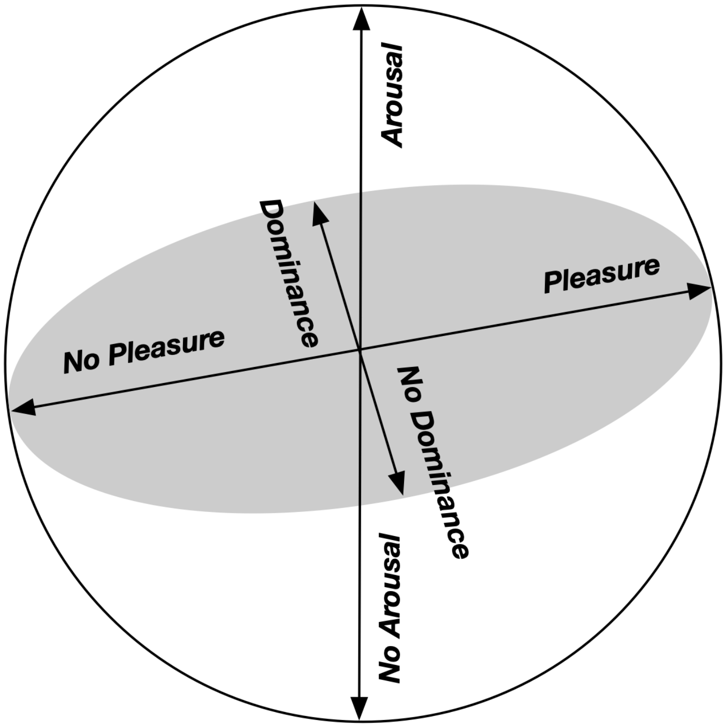


Fig. 2.2. PAD emotion model proposed by Fontaine.

2.2 Speech Emotional Datasets

Deep learning stands as the prevailing approach in contemporary artificial intelligence research, while datasets have emerged as an essential component for enabling intelligent systems to simulate real-world scenarios. Various datasets exhibit distinct performance levels in speech emotion recognition across diverse contexts, underscoring the critical significance of dataset construction and utilization. Within the realm of speech emotion recognition, datasets are primarily categorized into three types based on their collection and construction methods: acted datasets, elicited datasets, and natural datasets.

Acted datasets, often referred to as simulated datasets, are obtained by recording performances of actors or speakers trained to evoke specific emotions in predefined situations. The recording process of these datasets offers flexibility and control, free from constraints of time and location. Consequently, such datasets possess notable advantages in terms of both quantity and quality of data. Builders can infuse these datasets with rich emotional content using tailored emotional scripts, rendering them suitable for a wide array of emotional recognition and synthesis tasks. Furthermore, these datasets can effectively mitigate issues stemming from factors like microphone distance, encoder-decoder discrepancies, and noise, resulting in generally high speech quality. However, acted datasets lack the spontaneity and authenticity of emotions found in natural settings. Thus, the emotional speech generated tends to be more stylized, warranting careful consideration and processing during usage. Maintaining emotional authenticity, script diversity, and accounting for the actors' performance skills and accents are vital considerations.

For instance, the Emotional Speech Database (EmoDB) [14], curated by the Institute of Communication Science at the Berlin University of Technology and the Fraunhofer Institute for Open Communication Systems, is a German performance-based emotional speech database. It comprises 535 German phrases uttered by 10 actors, encompassing seven emotional categories: anger, disgust, fear, happiness, neutral, sadness, and surprise. Each sentence is meticulously annotated with its emotional category and intensity. This database has gained extensive traction in emotion recognition research and development, particularly for German emotion recognition tasks. The database's strengths encompass its inclusion of multiple emotional categories, diverse intensity levels of emotional expression, and reliable annotations. However, its limitations involve the exclusive focus on German speech data and a relatively modest data volume.

The elicited dataset, often referred to as the induced dataset, distinguishes itself from pure acted datasets. In its recording process, unexpected situations are introduced to align the actor's performance with a more natural state and authentic emotional speech expression. While the elicited dataset provides a closer simulation of emotional expression in a natural context compared to acted datasets, the presence of such settings during recording does influence the actors. Nevertheless, being artificially generated, the elicited dataset has a constrained variety of external scenarios, limiting the collection of emotional samples. Additionally, although the elicited dataset effectively addresses issues arising from factors like microphone distance, encoder-decoder effects, and noise, its recording process is somewhat less flexible than that of acted datasets. A prominent example of an elicited dataset is the Interactive Emotional Dyadic Motion Capture dataset (IEMOCAP) [15]. Created by the Speech Analysis and Interpretation Laboratory at the University of Southern California, this multimodal dataset encompasses audio, video, text, and action data. It features recordings of 10 actors in two sessions, each spanning around an hour. Conversations within IEMOCAP predominantly delve into emotional subjects such as daily struggles, personal experiences, and emotional events. This dataset is widely embraced within the emotional recognition field.

In the realm of the natural dataset, emotional speech expressions are genuine and reflect the primal and instinctual human emotions. Such datasets are rooted in real-world settings and typically entail recordings of conversations occurring in authentic situations. Sourcing from various contexts such as customer service phone calls, television broadcasts, dialogues between doctors and patients in rehabilitation centers, online interviews, and e-learning sessions, these datasets boast diverse origins. In contrast to datasets performed by trained actors, natural datasets offer convenience and authentic emotional expression. Moreover, they often encompass a broad spectrum of emotional categories suitable for various emotional recognition tasks. However, acquiring emotional samples from natural datasets is uncontrollable, and obtaining samples from smaller emotional categories can be challenging. Moreover, issues like noise and microphone distance in speech collection from natural environments require subsequent processing to enhance dataset quality and usability. An example of such a natural dataset is the FAU Aibo dataset [10], which contains emotional speech from robots. This dataset comprises 51 instances of interactive speech data with Sony Aibo robots, each exhibiting different emotional states. Compiled and organized by researchers at the Fraunhofer Institute for Applied Research (FAU) in Germany, the dataset encompasses over 600 distinct speech segments, spanning seven emotional states: happiness, sadness, anger, surprise, fear, boredom, and neutrality.

Through literature research and analysis, we have compiled a list of different types of datasets used in the field of speech emotion recognition. Tab.2.1 shows the datasets categorized by language, type, size, modality, and emotion categories. Based on the statistics, it can be observed that English datasets are the mainstream in emotion recognition research. In terms of dataset types, performance-based datasets account for over 60%, followed by naturalistic datasets. Regarding emotion labels, different datasets have different classifications, generally around 6 emotion categories.

2.3 Speech Features

Speech emotion recognition constitutes a significant area of research, wherein speech features serve as integral components of the speech system, profoundly influencing overall performance. To derive effective speech features, it is common practice to execute uncomplicated preprocessing procedures on the speech signal. These operations encompass conserving pivotal information while sieving out surplus data. At present, the prevailing acoustic features employed in speech emotion recognition can be categorized into two primary types: firstly, hand-crafted low-level features computed through temporal and spectral algorithms, and secondly, high-level features extracted either directly from raw signals or from low-level features through end-to-end neural network models. Both methodologies for feature extraction have found applications across diverse emotion recognition tasks. Hand-crafted low-level features have demonstrated efficacy in simpler assignments, whereas high-level features obtained through end-to-end model-driven extraction exhibit superior performance in more intricate undertakings.

Tab. 2.1. The commonly used datasets for speech emotion recognition, where ’an’ represents anger, ’ha’ represents happy, ’sa’ represents sad, ’ne’ represents neutral, ’fe’ represents fear, ’su’ represents surprise, and ’di’ represents disgust.

No	Dataset	Language	Type	Size	Modalities	Emotions
1	EMO-DB [14]	German	Acted	500	A	an, ha, sa, ne, di, boredom, and fe
2	SAVEE [45]	English	Acted	480	A,V	an, ha, ne, su, fe, di, and sa
3	eNTERFACE [79]	English	Elicited	1278	A,V	an, ha, su, fe, di, and sa
4	IEMOCAP [15]	English	Elicited	10039	A,V,T	an, ha, sa, ne
5	FAU-AIBO [10]	German	Natural	18216	A	an, emphatic, ne, joy, and rest
6	AESDD [111]	Greek	Acted	500	A	an, fe, di, sa, ha
7	MSP-PODCAST [75]	English	Natural	62140	A	an,sa, ha, su, fe, di, contempt and ne
8	RAVDESS [74]	English	Acted	7356	A	ha, an, ne, su, fe, di, and calm
9	RECOLA [95]	French	Natural	7 hours	A,V	arousal degree (1-5), valence degree (1-5)
10	AFEW[32]	English	Natural	1426	A,V	an, di, fe, sa,ha, ne, and su.
11	CHEAVD [63]	Chinese	Natural	2600	A,V	26 Non-prototypical emotions + 6 basic emotions and sa
12	CASIA [124]	Chinese	Acted	12000	A	ha, sa, an, su and ne
13	EHSD [8]	Hindi	Acted	6048	A	ha, an, sa, ne, su, and sarcastic
14	IIIT-H TEMD [92]	Telgu	Semi-Natural	2450	A	ha, an, sa, ne, su, fe, di, sarcastic, frustrated, relaxed, worried, shy, excited, and shout
15	IITKGP-SESC [56]	Telgu	Acted	12000	A	ha, an, ne, su, fe, di, sarcastic, and compassion
16	MELD [91]	English	Acted	13000	A,V	positive, negative, and ne
17	CREMA-D [17]	English	Acted	7442	A,V	an, ha, sa, fe, ne, and di
18	EMOVO [27]	Italian	Acted	588	A	di, joy, fe, su, sa, joy, and ne
19	JAVED [76]	Japanese	Acted	100 min	A,V	ha, an, sa, ne, and contentment
20	KVDERW [54]	Korean	Natural	1246	A,V	ha, an, sa,ne, su, di, and fe
21	GreThE [84]	Greek	Acted	500	A,V	valence and arousal
22	CaFE [42]	French	Acted	936	A	sa, ha, an, fe, di and su
23	EMOVIE [29]	Chinese	Acted	9724	A,V	positive, negative, and ne

2.3.1 Preprocessings

In a speech emotion recognition system, effective preprocessing of the collected dataset is essential to comprehend the speech data better and extract its features efficiently. The goal of speech data preprocessing is multifaceted: to safeguard and amplify crucial speech information, diminish or eliminate redundant data, and cater to various speech-related tasks. Standard speech preprocessing operations encompass framing, windowing, endpoint detection, normalization, noise reduction, among others. These operations prove instrumental in diminishing redundancy, preserving pivotal details, and augmenting feature extraction performance.

Framing emerges as a fundamental technique for managing speech tasks by segmenting continuous speech signals into fixed-length speech segments. Given the dynamic nature of speech signals, which encapsulate both emotional and textual nuances intrinsic to human communication, it's crucial to encapsulate these variations. Research has indicated that employing fixed segment lengths of 20 to 30 milliseconds maintains relatively consistent parameters across segmented waveforms. This approach ensures stability within each frame, establishing this time segment as the smallest analytical unit for audio assessment. Building on this foundation, common techniques like the discrete Fourier transform can further dissect sub-waves within each frame signal, thereby deriving distinctive frame characteristics of speech. These extracted features subsequently serve as valuable tools for addressing subsequent speech-related tasks.

The purpose of windowing is to make the amplitude of each framed speech signal fade to zero at both ends, which is convenient for subsequent Fourier transform processing, making the peaks on the spectrum finer, and reducing spectral leakage. Usually, a window function is used to transform the frame signal. Generally, the commonly used Hamming window for speech windowing is expressed as follows:

$$\omega(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \qquad 0 \le n \le M-1, \tag{2.1}$$

where $\omega(n)$ represents the speech signal within a frame, and M denotes the window size.

Endpoint detection entails the identification of the initiation and termination points of speech within a signal segment containing spoken content. The primary objective is to retain the voiced component of recorded speech, discard the unvoiced section, and effectively harness speech information. Speech signals can be categorized into three distinct types: voiced, unvoiced, and silence. Voiced segments encapsulate data about vocal fold vibration, unvoiced segments capture air turbulence stemming from vocal tract constriction, and silence signifies the absence of vocal tract engagement. A proficient endpoint detection algorithm should proficiently extract continuous voiced portions while excluding unvoiced and silent portions.

Regularization stands as a pivotal process in the optimization of speech algorithm models. Its role encompasses mitigating the influence of outlier samples on overall algorithm performance, curtailing the risk of speech feature overfitting, and enhancing the model's capacity for generalization. Among the commonly employed regularization techniques, the z-score normalization method holds prominence. It is expressed in the following formula, where "u" signifies the mean and " σ " denotes the standard deviation.

$$z = \left(\frac{x - \mu}{\sigma} \right) \tag{2.2}$$

In the speech signals of datasets, noise can arise due to environmental conditions or issues with microphone quality, potentially causing a significant decrease in the accuracy of conventional recognition models or even rendering them unable to identify speech. Therefore, noise reduction techniques are crucial for audio preprocessing. There are generally three approaches to noise reduction: one involves increasing the signal-to-noise ratio of the input speech signal to enhance the intelligibility of the speech. Another method involves purifying noise-contaminated speech features in the feature space of the speech recognition system, minimizing the mismatch between the training model and the recognition features. The third approach entails adjusting the parameters of the speech model to adapt to the testing speech environment, thereby improving recognition accuracy.

2.3.2 Handcrafted Features

Through time and frequency algorithms, manual computation and feature extraction can be performed, typically encompassing prosodic features, spectral features, and voice quality.

Prosodic features are unrelated to the semantic content of speech and are utilized to convey emotions by controlling speech rhythm, tempo, pauses, stress, and more. Different individuals expressing the same information can convey varying emotions due to diverse expressions, which can be effectively captured through prosodic features. Common physical parameters of prosodic features include duration-related parameters (speech rate, short-term zero crossing rate), fundamental frequency-related parameters (fundamental frequency, average fundamental frequency variation range), and energy intensity-related parameters (short-term average energy, energy variation rate, average amplitude). These parameters are used to describe perceptible speech characteristics for humans.

When producing sound, the vocal tract acts as a filter, with its shape governing the generated sound. Accurately modeling the shape of the vocal folds allows for precise depiction of sound transmission. Spectral features capture the relationship between vocal tract shape changes and phonation, providing insights into the characteristics of speech signals in the frequency domain. Different emotions are conveyed through different frequencies. Features related to the spectrum are primarily divided into linear spectral features and cepstral features. Common linear spectral features include Linear Predictive Coding (LPC) and Log-Frequency Power Coefficients (LFPC); common cepstral features include Linear Predictive Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC).

Figure 2.3 illustrates the process of extracting MFCC features. First, the raw waveform undergoes preprocessing steps such as frame segmentation, pre-emphasis, and windowing to reduce redundancy and enhance essential features. Next, the Fast Fourier Transform (FFT) is used to convert the time-domain signal of each frame into the frequency-domain signal. Due to the high computational complexity of the Discrete Fourier Transform (DFT), FFT is commonly used for spectrum calculation. Then, the absolute value or square operation can be applied to the complex spectrum to obtain the magnitude spectrum or power spectrum. As phase spectrum information is limited in speech signals, the magnitude spectrum is usually retained. Subsequently, a set of Mel triangular filters, relevant to human auditory perception, is applied to the spectrum for filtering. This step generates a feature widely used in the speech domain, known as Mel Spectral features (also referred to as FBANK features). Following this, the Mel Spectral features undergo a logarithmic operation to mimic the ear's perception of sound energy. This amplifies energy differences in low-energy regions, yielding the logarithmic Mel Spectral feature. Furthermore, the logarithmic FBANK feature undergoes Discrete Cosine Transform (D-CT), mapping the feature vector to a lower-dimensional space. This step primarily aims to eliminate inter-feature correlation and extract the most significant feature coefficients, known as MFCC features. To capture the dynamic variations of speech signals, first-order and second-order differentials can be computed based on static features, referred to as first-order and second-order dynamic features.

Voice quality is determined by the inherent physical characteristics of a sound and serves as a measure of its purity, clarity, and uniqueness. Feature parameters that capture sound quality include resonance peak frequency, bandwidth, frequency perturbation, amplitude perturbation, harmonic-to-noise ratio, vibrato, and glottal parameters, among others. There exists a robust correlation between speech quality and the emotional content conveyed by the speech.

Table 2.2 provides an overview of commonly utilized tools for extracting speech features. Among these, the openSMILE tool [37] encompasses multiple collections of speech emotion features, typically harnessed to derive low-level feature descriptors for speech segments. Librosa [81], a widely adopted tool within the Python environment, stands as a popular choice for extracting speech features and remains one of the predominant tools in numerous research endeavors.

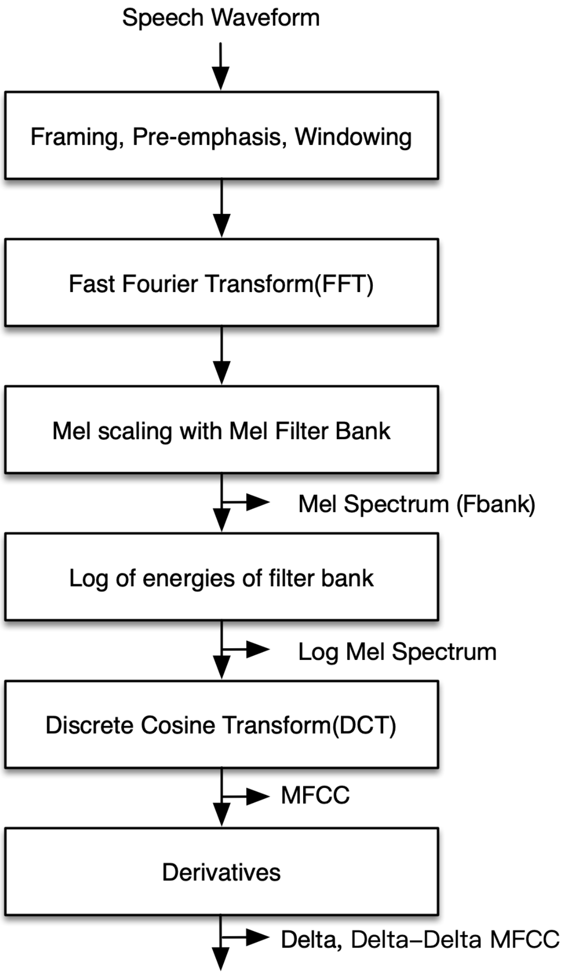

Speech Waveform
Framing, Pre-emphasis, Windowing
Fast Fourier Transform(FFT)
Mel scaling with Mel Filter Bank
Mel Spectrum (Fbank)
Log of energies of filter bank
Log Mel Spectrum
Discrete Cosine Transform(DCT)
MFCC
Derivatives
Delta, Delta–Delta MFCC

Fig. 2.3. MFCC feature extraction process.

Tab. 2.2. Speech feature extraction tools.

Tools	Platform	Description
Aubio	C/Python	Aubio is a tool designed for the extraction of annotations from audio signals. Its features include segmenting a sound file before each of its attacks, performing pitch detection, tapping the beat and producing midi streams from live audio
Essentia [12]	C++/Python	Essentia is an open-source library and tools for audio and music analysis, description and synthesis
Librosa [81]	Python	Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.
Madmom [11]	Python	Madmom is an audio signal processing library written in Python with a strong focus on music information retrieval (MIR) tasks.
pyAudioAnalysis [39]	Python	pyAudioAnalysis is a Python library for audio feature extraction, classification, segmentation and applications
Vamp-plugins	C++/Python	Vamp is an audio processing plugin system for plugins that extract descriptive information from audio data
Yaafe [80]	Python/Matlab	Yaafe is an audio features extraction toolbox.Features can be extracted in a batch mode, writing CSV or H5 files.
OpenSMILE [37]	C++/Python	openSMILE (open-source Speech and Music Interpretation by Large-space Extraction) is an open-source toolkit for audio feature extraction and classification of speech and music signals.
Praat	C++	Praat is a cross-platform multi-purpose speech learning software, mainly used for digital speech signal progress analysis, annotation, processing and synthesis, etc.
Voicebox	Matlab	Voicebox is a speech processing toolbox based on MATLAB, including speech framing, windowing, speech feature extraction, etc.

2.3.3 Features Extracted by End-to-End Models

In recent years, within the domain of speech emotion recognition, aside from manual extraction of low-level features, there has been a trend towards employing end-to-end neural network models to extract high-level features from raw signals or low-level features. This approach helps overcome potential subjectivity and limitations in manual feature extraction, thereby enhancing recognition accuracy and robustness.

A common approach to feature extraction involves extracting valuable information from raw speech signals, as demonstrated in [58], where a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) is employed to directly model raw speech signals. Results indicate that features captured from raw speech can achieve recognition performance similar to handcrafted features. However, directly modeling raw audio signals may not capture more intricate speech features due to information redundancy. Thus, some researchers utilize convolutional neural networks on speech spectrograms to extract higher-dimensional features. The most commonly used neural network architecture is CNN+RNN, where CNN captures local speech features, while RNN captures contextual relationships between speech features. In the study[67], using MFCC spectral features as a base, CNN+LSTM modules and triplet loss functions achieved state-of-the-art performance on the IEMOCAP dataset. To further capture relationships between speech sequences and identify crucial parts of the data, researchers introduced attention mechanisms, enabling models to automatically focus on distinct parts of speech sequences and allocate varying weights to each part [21, 130, 117]. In essence, neural network feature extraction methods bypass the intricate process of manual feature extraction, thereby enhancing the accuracy of speech emotion recognition.

Due to the labor-intensive process of constructing speech datasets, existing datasets often have limited sample sizes. When training neural networks on constrained datasets, there is a risk of underfitting, leading to inadequate accuracy in emotion recognition. To address this issue, recent research has focused on speech representation learning. Speech representation learning involves self-supervised methods that learn latent relationships between speech frames within large-scale datasets, aiming to capture higher-dimensional speech features.

Speech self-supervised learning primarily includes contrastive prediction and autoregressive predictive coding methods. Contrastive prediction methods involve using autoregressive models to predict future data, thereby learning speech feature representations. Initially, speech signals are divided into overlapping frames, and autoregressive models are then used to predict the features of the next frame. During training, contrastive prediction maximizes the similarity between the next frame and other future frames, commonly utilizing contrastive loss functions to learn high-quality, expressive, higher-dimensional speech feature representations [87]. Autoregressive predictive coding seeks to learn the structure and representation of speech signals through autoregressive models, often trained using cross-entropy loss functions. This method recursively predicts each sample point of the speech signal, progressively learning the structure and representation of the speech signal, resulting in high-quality speech representations [26]. Moreover, there are BERT-based masked reconstruction algorithms utilized for advanced speech feature learning [66, 65, 25].

Furthermore, from another perspective, speech features can also be classified. Table 2.3 summarizes the classification of speech features. Beyond traditionally manually extracted low-level features and currently popular deep learning techniques extracting high-level features, based on the length of the speech signal, speech features can also be classified into local features and global features. Local features focus on extracting localized information at the frame level, while global features encompass fusion and statistical analysis of features across the entire speech segment. Based on signal representation, they can be categorized as time-domain features, frequency-domain features corresponding to energy, and time-frequency features combining both.

Tab. 2.3. Summary of commonly used feature classification methods.

Classification Method	Type
Traditional Classification	Prosodic features, spectral features and sound quality
Feature Extraction	Low-level features and end-to-end features
Speech Signal Length	Local features and global features
Signal Representation	Time-domain features, frequency-domain features and time-frequency feature

2.4 Speech Emotion Recognition Model

2.4.1 Traditional Machine Learning

Traditional classifiers commonly used for speech emotion recognition encompass Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), and Hidden Markov Models (HMMs), among others. When employing SVMs for speech emotion recognition, it is customary to extract features such as MFCC and energy, representing prosodic aspects of the speech signal, and employ them as input for SVM classification. SVMs map these features to a higher-dimensional space to ascertain an optimal hyperplane that effectively separates speech data belonging to distinct emotional states. The training of SVM models necessitates a labeled dataset of speech data annotated with emotional labels, thereby segmenting the speech data into different emotion categories based on these labels. By adjusting SVM hyperparameters and utilizing appropriate kernel functions, the optimal classifier can be identified. For example, in [82], the SVM algorithm was harnessed for speech emotion recognition, achieving a 68% accuracy on the Berlin Emotional Speech Database utilizing 24 MFCC features. Recent studies frequently integrate SVM as an integral component of speech emotion recognition systems, enhancing overall system performance [35, 125, 46].

When exploiting HMMs for speech emotion recognition, a prevalent strategy involves segmenting the speech signal into discrete time frames and extracting features like MFCC and energy for each frame. These sequences of features serve as observation sequences for the HMM. In this framework, each emotional state is treated as a hidden state within the HMM, and classification entails leveraging transition probabilities, observation probabilities, and initial probabilities across the various emotional states [49, 109].

Gaussian Mixture Models (GMMs), well-known statistical models, find utility in tasks like probability density estimation and clustering. In the context of speech emotion recognition via GMMs, the acoustic features of speech are fed into the GMM as inputs. The GMM model separates speech data from different emotional states by learning the probability density distribution of the training dataset. During the training of the GMM model, speech data is categorized into different emotional classes based on emotion labels, and statistical methods such as Maximum Likelihood Estimation are employed to estimate the mean values and covariance matrices of Gaussian distributions corresponding to each emotional state. During prediction, the trained GMM model can be applied to classify new speech signals and discern their associated emotional classes [47, 86].

2.4.2 Deep Learning

In the field of speech emotion recognition, deep learning models have been widely applied and achieved remarkable performance across various datasets. Among them, Convolutional Neural Networks (CNNs) are commonly used models that extract frequency and temporal features from audio signals through convolutional layers, capturing local audio information. Hence, they are suitable for short-term speech emotion classification tasks. Conversely, Recurrent Neural Networks (RNNs) utilize recurrent layers to capture temporal dynamics within the signal, granting them stronger capabilities in handling temporal information, thus being suited for long-term speech emotion classification tasks. Long Short-Term Memory networks (LSTMs), an advancement of RNNs, introduce memory cells and gating mechanisms, effectively addressing the vanishing gradient problem in traditional RNN models and improving the accuracy of speech emotion recognition.

Self-attention mechanisms are deep learning components that have gained wide adoption in recent years to acquire richer context information from speech signals. By computing weighted sums of self-generated information for each input position, this mechanism obtains feature vector representations, intensifying the representation of relevant information while suppressing irrelevant information. This empowers the model to more accurately capture contextual information within speech signals. Additionally, Generative Adversarial Networks (GANs) have been applied to speech emotion recognition, employing adversarial frameworks to iteratively enhance both the generator and discriminator, thereby bolstering model robustness and generalization.

With the evolution of neural networks, various network variants have emerged to enhance speech emotion recognition performance. For example, Bidirectional LSTM networks consist of two RNN models—one processing input sequences in the forward direction and the other in reverse. Gated Recurrent Unit networks (GRUs), akin to LSTMs but with fewer parameters and faster computation speeds, perform well in short-sequence speech emotion recognition tasks. Emotion recognition models based on Variational Autoencoders (VAEs) compress and reconstruct data by learning latent distributions of data. Additionally, unsupervised transfer learning can lead to superior performance with limited annotated data, while simultaneously reducing training time and computational resources.

Deep learning offers several advantages over traditional machine learning. Traditional methods relying on feature extraction and classifiers necessitate manual feature design, whereas deep learning can automatically learn valuable features from raw speech signals, reducing reliance on domain expertise and manual processing. As speech signal quality might be influenced by factors like environmental noise and speaker variations, deep learning models can adaptively learn and address these variations, enhancing model robustness. Deep learning models demonstrate strong generalization abilities, performing well in new speech emotion recognition tasks, even when differences exist between training and testing data.

Table 2.4 presents the pros and cons of various algorithms, encompassing both traditional machine learning and deep learning algorithms, while also summarizing speech features utilized in previous literature. From the table, it's evident that traditional machine learning algorithms typically employ manually designed prosodic features as inputs. In [18], the authors proposed an approach that combines ranking and classification, utilizing openSMILE to extract acoustic and linguistic features for emotion recognition. By training SVM models, a 6.6% improvement in speech emotion recognition accuracy was achieved on the FAU AIBO dataset. Deep learning algorithms, on the other hand, further leverage spectral features for feature extraction. In [127], a Deep Convolutional LSTM (DC-LSTM) model utilizing multi-scale convolutional layers was introduced to classify and predict features of speech signals across multiple frequency components. This model achieved state-of-the-art results on the AFEW5.0 and BAUM-1s datasets. Moreover, an increasing number of studies are integrating deep learning and traditional machine learning methods to enhance the efficiency of speech emotion recognition systems [68, 35, 125].

Tab. 2.4. Advantages, disadvantages between different algorithms.

Algorithm	Advantages	Disadvantages
SVM	Able to address issues such as small sample size, high dimensionality, and non-linearity; Strong generalization ability; Avoidance of network structure selection	Sensitive to missing data; No universal solution for non-linear problems; Long training time.
HMM	Suitable for analyzing short-term stationary speech signals	Limited ability to process massive data
GMM	High fitting performance for speech emotion data	Needs to store parameters of each dimension and Gaussian component; Strong dependence on training data; Computationally complex
KNN	Simple and effective; Low cost of retraining; Suitable for automatic classification of class domains with large sample sizes	Class domains with small sample sizes are prone to misclassification; Weak output interpretability; High computational cost
ELM	Strong generalization ability; Fast learning speed, accurate prediction, and reduced computational burden	The calculation of the output layer decision values depends entirely on the labels
CNN	Strong ability to extract abstract features; Possesses characteristics of intra-class convergence and inter-class divergence	
DNN	Able to simulate any function and has strong emotion representation capability	
LSTM	Possesses memory characteristics and can effectively learn inter-frame correlations	Requires a large amount of resources and time to establish the model, needs a large amount of data to train the model, and parameters such as the number of hidden nodes and layers need to be repeatedly debugged
GAN	Strong perception ability for underlying probability distributions of raw data	
GCN	Capturing global features in a graph-based form	

3 Active Learning for Speech Emotion Corpus Construction

3.1 Introduction

Speech plays a crucial role in human society since it is one of the information conveyed by human communication. With the joint efforts of researchers in recent decades, speech emotion recognition has been rapidly developed [4, 85, 108], especially the emergence of machine learning in recent years has greatly promoted this boom [105, 53, 126]. Machine learning relies on large amounts of data for training, which makes the datasets in the current research on speech emotion recognition become precious and indispensable. The construction of corpus labeled data samples is a huge project with extremely high cost, which often requires countless manpower and material resources. It not only requires professional actors to record, but also the recorded voice fragments have to be manually screened and annotated.

The current methods of constructing datasets usually have the following shortcomings: 1. Most of the datasets are obtained by professional actors performing in a given scene, so the scene is relatively single, the emotional performance is relatively exaggerated, and the naturalness is not high. 2. In the process of labeling speech utterances, due to the uneven distribution of different types of emotions, it is impossible to prioritize the screening of rare samples. The annotator is required to listen to all the speech utterances one by one, and the labelling cost is high. 3. Datasets are mostly collected in specific scenarios. In actual application scenarios, the accuracy of speech emotion recognition will be low due to different languages, different speaking styles, and different application scenarios.

Lack of emotion corpus restricts the development of deep learning in the field of speech emotion recognition. To cope with the problem of insufficient datasets, researchers often use data enhancement in speech emotion experiments to increase the number of corpus samples [9, 20, 36], which expands the amount of data. Since the target voice is generated by the algorithm, it is impossible to ensure the quality of the voice. There are also some studies [104, 128, 77, 129] that make up for the shortcomings of the corpus through the form of cross-corpus, but it can't solve the problems fundamentally because of insufficient datasets. In the process of constructing the corpus, some tools for labelling emotion types were invented [28, 83, 95], which improved the efficiency of constructing the corpus to some extent but required the annotator to listen to the voice from the beginning to the end. Therefore, how to construct a corpus suitable for specific application scenarios quickly, efficiently and at low cost is the focus of this paper.

In classification tasks, supervised learning usually relies on many manually labeled training samples, and the process of labeling samples is very expensive [57]. The emergence of active learning has played a significant role in reducing the cost of marking. It is mainly composed of five core parts, including: unlabeled sample pool, screening strategy, labeling experts in related fields, labeled sample pool, and target classification model. Active learning combines the above five parts into the same process, and updates the performance of the classification model, unlabeled sample pool, and labeled data set in an iterative training method until the target model reaches the preset performance or no longer provides labeled data.

Regarding active learning research, the most important thing is how to choose the most valuable samples for labeling [99, 40, 98]. There are some classic sampling strategies developed so far, such as the random sampling strategy [97], which mainly randomly selects a certain proportion of samples from unlabeled samples and provides them to the model. Uncertain sampling strategy [60], mainly by combining the characteristics of the sample itself, calculating and selecting the least easily distinguishable sample, the sample with the best value. QBC query strategy [100], this algorithm will train multiple classifiers from different perspectives, and jointly screen samples for labeling experts. Active learning has attracted the attention of many researchers, and many methods have been expanded on this basis. For example, Density [33] uses a density map to find the most representative sample for labeling from all unlabeled samples. LAL [55] uses the pre-trained regression model to calculate the sample prediction error to determine the sample to be labelled. Query [48] finds the most informative and representative samples, and provides them to the annotators for labelling.

Over the years, active learning has been widely used in different fields. Goudjjl et al. [41] combined active learning with the SVM model, and selectively selected informative samples to train the SVM model, which achieved excellent results in text classification. Yan et al. [119] addressed the problem of difficult sample labeling in long text classification, combined with active learning to screen long text samples with high labeling value, and condenses the long text into words to facilitate labeling by the annotator. The papers[116, 120] combined active learning and deep learning to improve the efficiency of image segmentation and labeling in the biological field. Cao et al. [19] combined active learning and the CNN model to label complex spectral pictures. In the field of speech e-motion recognition, Mohammed et al. [2] proposed an iterative fast converging incremental adaptation algorithm that combines active learning and supervised domain adaptation to address the lack of generalization of speech emotion classifiers in real applications. Mohammed et al. [3] also used greedy sampling and DNN model to conduct speech e-motion recognition experiments, and the results show that active learning can improve the performance when the training set is limited. Vaaras et al. [110] combined CPC and various dimensionality reduction methods to explore the performance of clustering-based active learning under different feature conditions.

According to our knowledge, there is not much research and application for the improvement of the efficiency of labeling samples in the construction of speech emotion datasets. In this work, we present an active learning method to improve the construction efficiency of speech emotion data sets in the labelling process of constructing the emotion corpus. Our method is divided into the following steps. First, use a small number of labeled samples to train a logistic regression classifier, predict all unlabeled samples, and use the predicted probability to find the samples with the most unclear category, which we call uncertainty sampling. Then, use the feature relationship of all unlabeled samples to find the sample at the feature center in the feature space, and filter out the most a representative sample, which is representative sampling. Next calculate the feature distance between the unlabeled sample and the labeled sample one by one and select the samples that are different from the labelled samples, which is diversity sampling. Finally, by using the probability of the unlabeled data and labeled data, select the unlabeled samples which make all categories have the same distribution. Therefore, the samples selected by our method have the characteristics of strong uncertainty, strong class representativeness, strong feature diversity, and strong complementarity between classes. Compared with other active learning algorithms, the experimental results show that our algorithm is significantly better than other algorithms in screening small class samples.

To solve the problem of the lack of datasets in the field of SER, we designed a framework for speech emotion corpus construction. The framework integrates the newly proposed active learning strategy, and at the same time uses voice activity detection, feature extraction and other technologies. The raw voice of our framework is not limited to the performance of the actors, so it can be adapted to voice data collection in different scenarios. The framework first splits long raw data into segments by using endpoint detection. After simple filtering, uses openSMILE [37] feature extraction tool to extract the features of all filtered segments as unlabeled datasets. Finally, by using the proposed active learning method, screen out the samples for labelling. The framework can be applied to data collection in different voice scenarios, and has the characteristics of high naturalness, flexibility, and high efficiency. In this work, we take TV drama videos as the data source. After subjective listening and discriminating experiments, our scheme has achieved excellent results.

3.2 Methodology

In our work, we propose an integrated active learning method. The proposed method includes a commonly used logistic regression pre-classifier and four sampling strategies, uncertain sampling, representative sampling, diverse sampling and complementary sampling. Especially for complementary sampling, it is able to preferentially select rare samples and provide them to the annotator for labeling according to the distribution of the current samples.

For the logical classifier, suppose we have a total data set C with K categories. The dataset C contains the labeled data set L and the unlabeled data set U , namely $C = L + U$, where $L = \{(X_i, Y_i), ..., (X_m, Y_m)\}$. $X_i \in \mathbb{R}^N$ represents the features of the sample with N dimensions; $Y_i \in \mathbb{B}^k$represents the category of sample with K dimensions, and each dimension represents a different category value.

We train K binary logistic regression models for K different categories:

$$F_k : \mathbb{R}^N \to \mathbb{B} \tag{3.1}$$

Therefore, for each binary logistic regression, we can get:

$$Y_k = f_k(X) = \frac{1}{1 + exp(-W_k * X)} \tag{3.2}$$

Among them, W_k is the weight of each logistic regression classification model, Y_k is the probability of the logistic regression of the k-th category of each binary classification, and its value is between 0 and 1, and then P_k is used to represent Y_k in the following paper.

Uncertain sampling is mainly used for sampling ambiguous sample points at the decision boundary, which helps to clarify the decision boundary. Therefore, combined with the predicted probability of the sample, we use cross entropy to find the largest uncertain sample point. The following is the calculation formula for uncertain sampling. For sample X, the cross entropy H of each category is calculated separately, and the largest cross entropy is selected as the uncertainty value of the sample. We will sort all the samples and select the samples with the greatest uncertainty for subsequent sampling.

$$U(X) = max\{H(P_k)\,|k = 1, ..., K\} \tag{3.3}$$

$$H(P_k) = -P_k \log P_k - (1 - P_k) \log(1 - P_k) \tag{3.4}$$

Representative sampling is a kind of sampling that can best represent all unlabeled sample points. The sample point closest to the center of the unlabeled samples will be selected through the distance calculation of all sample features as the most representative sample point for labeling. In the following formula, for each unlabeled sample X, the feature distances of N dimensions between X and all other unmarked sample points will be calculated, and the average value is calculated as the feature similarity of the sample points.

$$R(X) = \frac{1}{|U|-1} \sum_{X' \in U-X} -Dis_{eu}(X,X') \tag{3.5}$$

$$Dis_{eu}(X,X') = \sqrt{\sum_{i=1}^{N} (X_i - X_i')^2} \tag{3.6}$$

Diversity sampling is mainly to find the unlabeled sample point closest to the feature center of all labeled samples. We assume that the feature center of the labeled sample is the boundary center of feature diversity, so that by looking for the closest to the feature center can effectively reduce the overfitting of the trained model. We use the following formula to calculate and find the point with the smallest feature distance between the unlabled sample X and all the labeled points X'.

$$D(X) = \min_{X' \in L} Dis_{eu}(X,X') \tag{3.7}$$

Complementary sampling can give preference to samples of small categories as much as possible, so that the overall category distribution of labeled data will become similar. Cross entropy is used to measure the similarity of the distribution using the category probability, the smaller the value, the more similar the distribution. In the following formula, P represents the sum of the category distribution with the probability of unlabeled points and the category of all labelled points, and Q is a reference, which represents the distribution of K same proportions, $(1/K, ..., 1/K)$. When the sum of the distribution of the unlabeled point we selected and all labelled points is close to the Q of the equal distribution, the cross-entropy value is the smallest, then the unlabeled points will be selected for labelling.

$$D(X) = \min_{X' \in L} Dis_{ce}(P(L(X)), Q) \tag{3.8}$$

$$L(X) = L \cup (X,Y) \tag{3.9}$$

$$Dis_{ce}(P,Q) = -\sum_{i=1}^{K} Q_i \log P_i \tag{3.10}$$

When speech emotion recognition is applied in a specific scene, it is usually necessary to construct an emotion corpus corresponding to the specific scene, which often

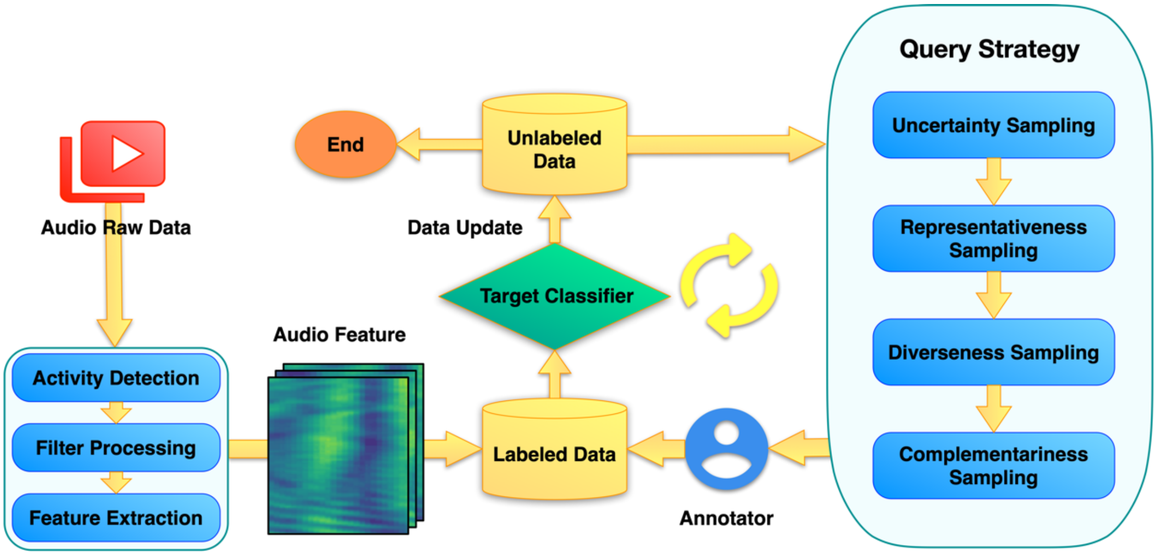


Fig. 3.1. Flow chart of the speech emotion corpus construction.

consumes large costs. To solve this problem, we propose a framework for constructing speech emotion corpus based on our active learning strategies.

The framework is shown in Fig.3.1. The audio processing part is mainly divided into three steps. First, the voice is split into segments through voice activity detection. Since the segments are split from the original data, data cleaning and manual filtering are required. After simple filtering, all the segments will be extracted for features.

For the labelling process, sample selection is an important part of active learning. It consists of five components, labeled data set, unlabeled data set, classifier, annotator, and selection strategy. All the features we get from the audio processing part will be divided into unlabeled data and labeled data. First, manually annotate a part of the samples. The classifier trained with this data will do a preliminary pre-classification of other unlabeled data and send all the feature values and probability values into the core selection strategy. Our method will repeatedly filter out a batch of high-quality data, let the annotator label, and the annotator will update the labelled data and unlabeled data respectively, then carry out the next round of screening. After several times of manual labeling, the finally trained target classifier replaces the manual labeling, so as to achieve the purpose of saving cost.

In the entire speech emotion corpus construction framework, the original audio source can be combined with the current application scenario, not only can be obtained from the network video, but also the audio from a specific scene can be recorded, which greatly improves the flexibility of the data source. For the voice activity detection part, we use the auditok tool to split the utterance into segments. In the feature extraction part, we use the openSMILE speech feature extraction tool to extract low-level descriptors of acoustic features as language features.

3.3 Active Learning Comparative Experiment

3.3.1 Datasets

To verify the effectiveness of our proposed active learning strategy in imbalanced class datasets, we selected 13 imbalanced class datasets which are commonly used in machine learning experiments and speech emotion recognition filed. Tab.3.1. shows the description of the used dataset in detail, which contains the name of the dataset, the number of instance, class and feature, the detail in different classes. Aggregation, Blood, Diabetes, qsar-biodeg, Vote, Vowel WBC and Thyroid are UCI data sets [52]. SAVEE [45], EMO-DB, CASIA [124],eNTERFACE05 [79] and IEMOCAP [15] are speech e-motional datasets. Since CASIA, eNTERFACE05 and IEMOCAP datasets are not so imbalanced datasets, we manually selected a part and processed them into CASIA_im, eNTERFACE05_im and IEMOCAP_im as imbalanced datasets.

3.3.2 Baseline

In addition, we also compared our proposed method with five baseline active learning strategies as follows:

1. QBC [100]: Active learning that trains multiple classifiers from different perspectives, and jointly select the samples for labeling.
2. Random: Random sampling, randomly selecting samples that need to be labeled from the sample pool to be labeled

Tab. 3.1. Detail description of the 16 datasets used in the experiment.

Dataset	Instance	Class	Feature	Detail
SAVEE	480	7	62	60;60;60;60;120;60;60
EMO-DB	535	7	62	127;81;46;69;71;79;62
CASIA	6000	5	62	1200;1200;1200;1200;1200
CASIA_im	1400	5	62	100;100;1000;100;100
eNTERFACE05	1257	6	62	210;210;210;209;209;209
eNTERFACE05_im	610	6	62	40;60;80;100;150;180
IEMOCAP	4983	4	62	1051;1707;1161;1064
IEMOCAP_im	1600	4	62	300;1000;200;100
Aggregation	788	7	3	45;170;102;273;34;130;34
Blood	748	2	5	570;178
Diabetes	768	2	9	500;268
qsar-biodeg	1055	2	42	356;699
Vote	435	2	17	267;168
Vowel	871	6	4	72;89;172;151;207;180
WBC	683	2	9	444;239
Thyroid	215	3	6	150;35;30

3. Unc [60]: Uncertainty sampling, select the most uncertain sample considered by the model.

4. Density [33]: Using a density map to find the most representative sample for labeling from all unlabeled samples

5. LAL [55]: LAL uses the pre-trained regression model to calculate the sample prediction error to determine the samples for labeling.

3.3.3 Evaluation Metrics

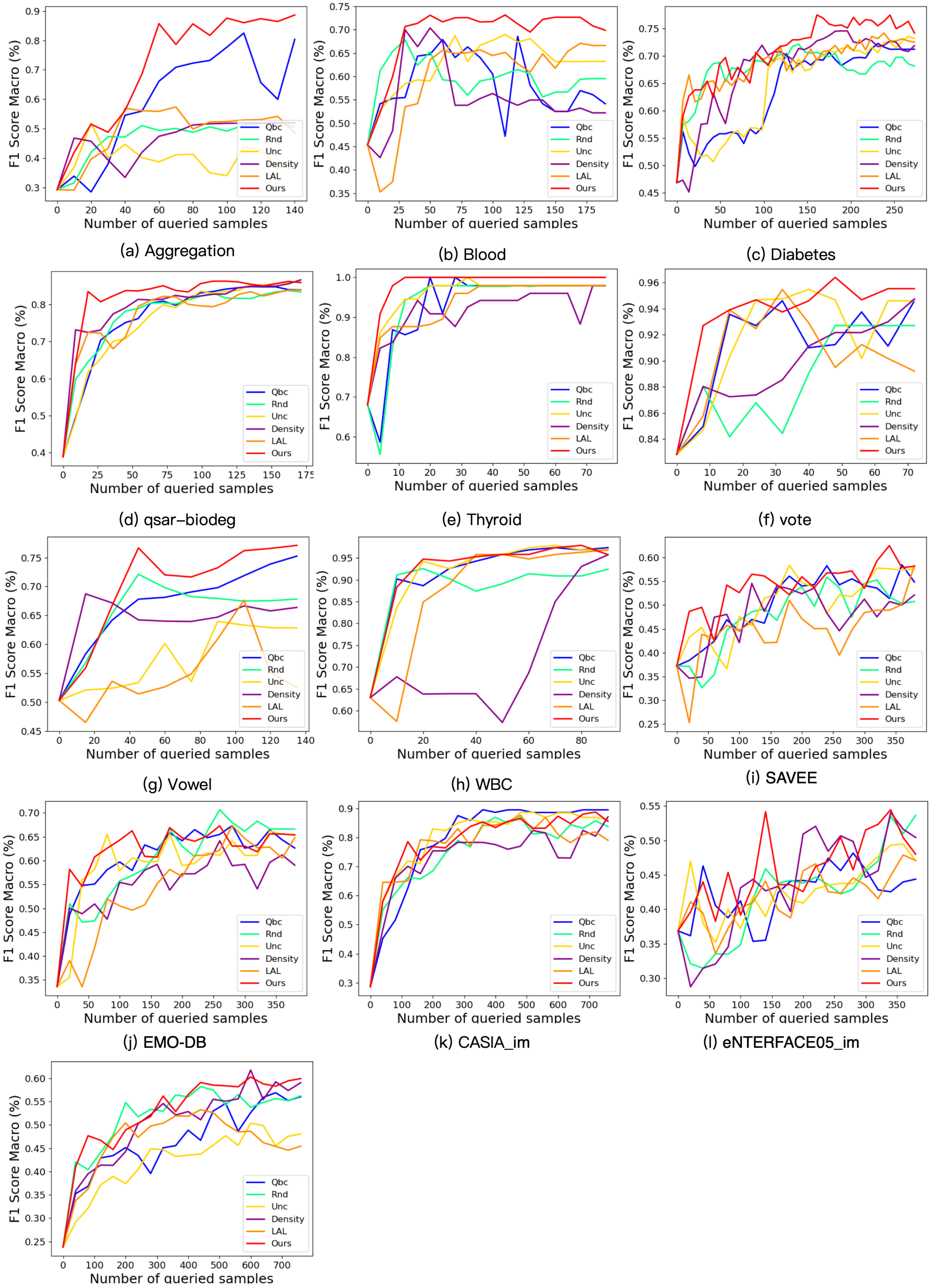
For the evaluation index of the experiment, we used the macro-average F1 value to measure the prediction effect of different methods. The F1 value is the harmonic average of the model's accuracy and recall, which indicates the two prediction performances of the model as follows:

$$F_1 = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{3.11}$$

3.3.4 Experimental Results

In the experiment, we used the Alipy [122] active learning tool developed by Huang Shengjun's team. A total of 5 active learning algorithms were used in our experiments on 13 imbalanced datasets. Fig.3.2 is the result of our experiment, where the horizontal axis represents the number of labelled samples, and the vertical axis represents the results of the macro average F1 value in the test set using the labelled samples, of which the test set accounts for 30%. The different colored lines in the figure represent the results of different algorithms. The red line is our proposed method and our method achieved excellent results, whether it is a two-class or multi-class imbalanced dataset.

To further illustrate the effectiveness of our method in imbalanced sample selection, we selected the results on the dataset qsar-biodeg and the multi-class Vowel dataset. Fig.3.3 shows the results on the qsar-biodeg dataset. The horizontal axis represents the total number of samples selected, and the vertical axis represents the number of samples in different classes. The labels 0 and 1 represent two different types of samples. The dataset



(a) Aggregation

(b) Blood

(c) Diabetes

(d) qsar–biodeg

(e) Thyroid

(f) vote

(g) Vowel

(h) WBC

(i) SAVEE

(j) EMO-DB

(k) CASIA_im

(l) eNTERFACE05_im

(m) IEMOCAP_im

Fig. 3.2. Comparison of different active learning methods on 13 datasets using the macro-average F1 value.

qsar-biodeg contains two classes with a ratio of 1:2. In the selection of random sampling (d), the selected sample has a uniform proportion, which best shows the proportion of the data set itself. (a)(b)(c)(e) methods do not calculate the class ratio of labelled samples, so they cannot achieve the balance of overall samples. Our algorithm (f) can give priority to the balance between different classes and select rare samples so that the proportions of different samples are close to each other.

Also in the multi-class emotional data set EMO-DB, as shown in Fig.3.4. Due to the limited number of labeled samples, the trained classification decision boundary cannot accurately determine the feature space of the category. Our method fully considers the category distribution of the labeled samples, so it can prioritize the selection of samples with rare classes to ensure that the selected sample classes are balanced.

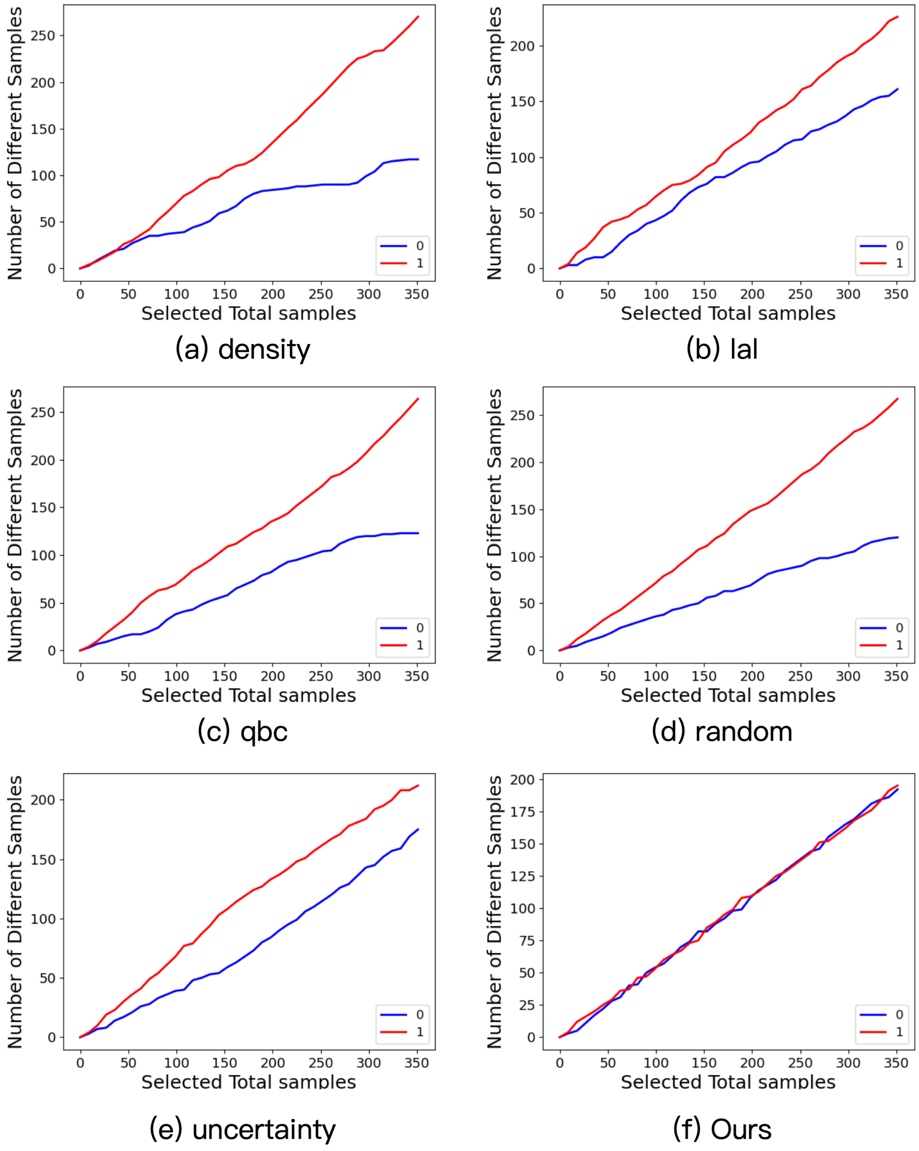
3.4 Speech Emotion Corpus Construction Experiment

3.4.1 Datasets

To evaluate the performance of our proposed method in the construction of speech emotion dataset, we perform the selection experiment on several speech emotional datasets. This experiment uses the following datasets commonly used in the field of SER. Each dataset represents a fixed speech annotation scene, so we can better test the annotation effect of our method on the speech emotion datasets.

1. CASIA [124]: A dataset constructed by the China Institute of Automation Science in 2005, which was recorded in a pure environment by 4 professional sound recorders, two men and two women. There are 5 emotions including happiness, sadness, anger, surprise, and neutrality.

2. EMO-DB [78]: A German emotional speech corpus recorded by the Technical University of Berlin, with 10 actors (5 males and 5 females) performing 7 emotions on 10 sentences (5 long and 5 short). The selection of corpus text follows the principle of semantic neutrality and no emotional tendency. Voice recording is done in a professional recording studio, requiring actors to reminisce their own real experience or



(a) density

(b) lal

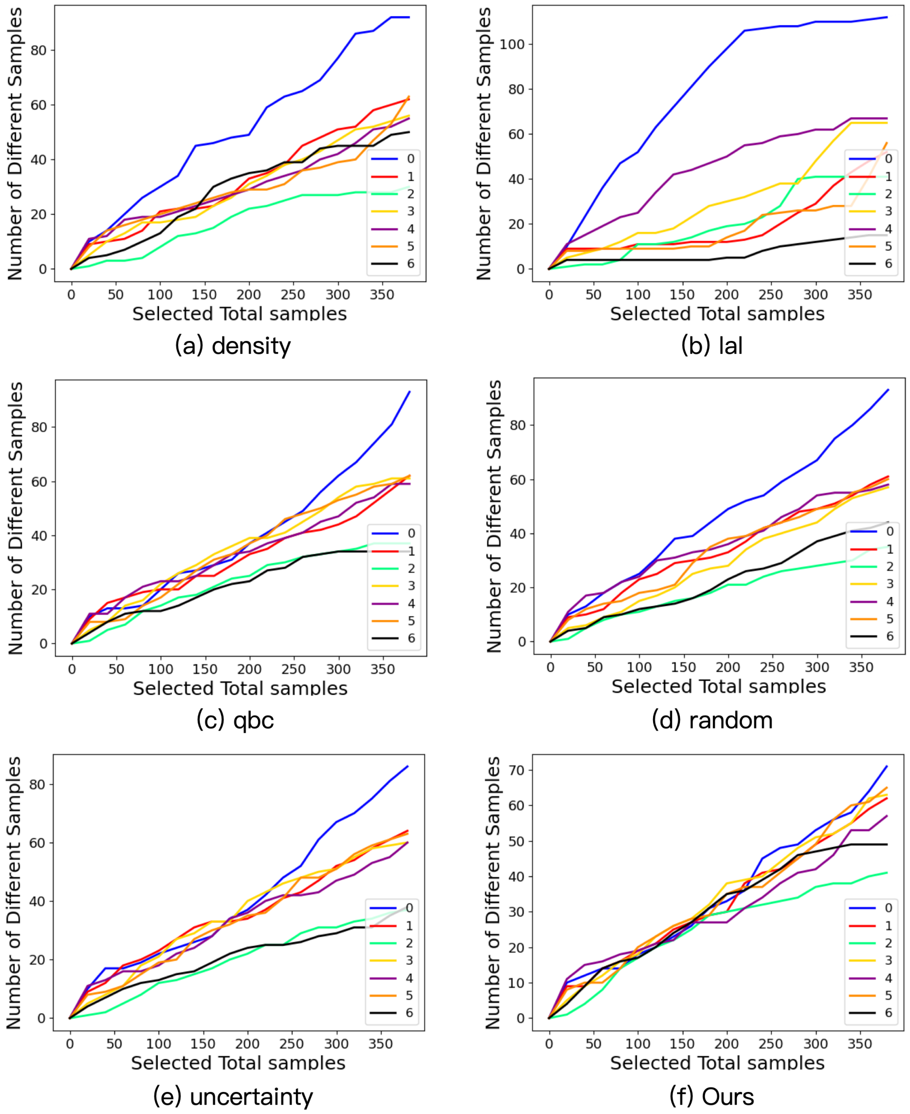
(c) qbc

(d) random

(e) uncertainty

(f) Ours

Fig. 3.3. Selection of samples of different categories in the binary class dataset qsar-biodeg.



(a) density

(b) lal

(c) qbc

(d) random

(e) uncertainty

(f) Ours

Fig. 3.4. Selection of samples of different categories in the multi-classification data set EMO-DB.

experience to brew emotions before interpreting a specific emotion to enhance the realism of emotions.

3. eNTERFACE05 [79]: It is an audition emotion dataset that contains six emotions such as anger, disgust, fear, happiness, sadness, and surprise. The dataset contains a total of 1166 video sequences. Of these 1166 video sequences, 264 female recordings (23%) and 902 male recordings (77%).

4. IEMOCAP [15]: Collected by the Sail Laboratory of the University of Southern California, it is a database of actions, multiple modes and multiple peaks. Completed by 10 actors and actresses, the dataset is about 12 hours of audiovisual data, which contains 10 emotions such as anger, neutrality, and excitement.

5. SAVEE [45]: The database contains a total of 480 British English audios from 4 male actors. These recordings have 7 different emotions: angry, disgusted, scared, happy, sad, surprised, neutral.

The detail description of emotion datasets used are shown in Tab.3.1.

3.4.2 Experiments and Results

We conducted two sets of experiments respectively. In the first group of experiments, the number of samples to be queried is fixed to 20 each query. In the second group of experiments, the samples to be selected are fixed to 5 percentage each query. We use macro average F1 valuation to measure our method. The train set and test set of each experiment were randomly divided five times, so the experimental results were averaged.

Tab.3.2 and Tab.3.3 show the selection efficiency results of our experiments respectively, the number 1 to 10 represent annotation times. The number 1 represents the initial labeling round. We observe that the trained classifier is positively correlated with the number of labeled samples, but the efficiency of proposed method is affected by the sample size of datasets. For example, the total size of CASIA and IEMOCAP data sets are 6000 and 4983, when the number of labelled samples is about 20%, the classifier will

Tab. 3.2. The active strategy selects 20 samples in each round and the performance of the target classifier accuracy.

Dataset	Test	Init	1	2	3	4	5	6	7	8	9	10	All
SAVEE	96	20	0.19±0.04	0.27±0.06	0.38±0.08	0.39±0.06	0.43±0.08	0.46±0.1	0.45±0.06	0.5±0.05	0.5±0.07	0.5±0.06	0.53±0.05
EMO-DB	107	20	0.34±0.06	0.38±0.05	0.43±0.04	0.49±0.03	0.52±0.02	0.55±0.03	0.55±0.05	0.57±0.05	0.61±0.06	0.61±0.05	0.69±0.05
CASIA	1200	20	0.24±0.02	0.31±0.03	0.39±0.04	0.42±0.02	0.43±0.02	0.46±0.02	0.48±0.02	0.5±0.02	0.52±0.02	0.53±0.01	0.62±0.01
CASIA_im	280	20	0.31±0.02	0.37±0.07	0.49±0.04	0.56±0.07	0.61±0.07	0.68±0.04	0.7±0.03	0.72±0.05	0.72±0.04	0.73±0.04	0.83±0.02
eNTERFACE05	251	20	0.16±0.03	0.19±0.02	0.23±0.04	0.24±0.04	0.25±0.04	0.26±0.03	0.28±0.03	0.3±0.03	0.32±0.04	0.33±0.06	0.42±0.04
eNTERFACE05_im	122	20	0.21±0.06	0.25±0.02	0.3±0.04	0.38±0.03	0.38±0.03	0.38±0.04	0.4±0.05	0.4±0.03	0.39±0.04	0.42±0.05	0.44±0.03
IEMOCAP	997	20	0.32±0.06	0.36±0.02	0.38±0.03	0.4±0.04	0.42±0.04	0.42±0.03	0.43±0.04	0.45±0.05	0.46±0.05	0.46±0.04	0.57±0.02
IEMOCAP_im	320	20	0.3±0.03	0.35±0.02	0.37±0.04	0.4±0.04	0.41±0.03	0.42±0.02	0.43±0.02	0.45±0.02	0.47±0.03	0.48±0.02	0.5±0.03

Tab. 3.3. The active learning strategy selects 5% in each round and the performance of the target classifier accuracy.

Dataset	Test	Init	1	2	3	4	5	6	7	8	9	10	ALL
SAVEE	96	19	0.19±0.04	0.29±0.03	0.38±0.06	0.39±0.05	0.44±0.04	0.47±0.08	0.47±0.09	0.49±0.07	0.49±0.05	0.49±0.05	0.54±0.06
EMO-DB	107	21	0.34±0.02	0.47±0.07	0.49±0.08	0.56±0.11	0.59±0.06	0.6±0.04	0.63±0.05	0.66±0.04	0.68±0.04	0.7±0.03	0.72±0.03
CASIA	1200	240	0.54±0.03	0.59±0.02	0.61±0.01	0.63±0.01	0.63±0.01	0.63±0.01	0.63±0.01	0.63±0.01	0.62±0.02	0.64±0.01	0.62±0.01
CASIA_im	280	56	0.41±0.03	0.64±0.07	0.72±0.05	0.73±0.02	0.75±0.04	0.76±0.02	0.78±0.03	0.8±0.03	0.79±0.03	0.78±0.05	0.83±0.03
eNTERFACE05	251	50	0.23±0.02	0.29±0.02	0.31±0.03	0.34±0.02	0.38±0.03	0.38±0.02	0.4±0.04	0.41±0.02	0.42±0.01	0.42±0.01	0.47±0.04
eNTERFACE05_im	122	24	0.21±0.02	0.25±0.03	0.36±0.01	0.31±0.02	0.36±0.08	0.38±0.03	0.4±0.05	0.4±0.06	0.39±0.06	0.41±0.08	0.43±0.06
IEMOCAP	997	199	0.48±0.02	0.54±0.02	0.55±0.01	0.56±0.01	0.56±0.01	0.57±0.02	0.57±0.02	0.57±0.02	0.57±0.01	0.58±0.02	0.57±0.02
IEMOCAP_im	320	64	0.39±0.04	0.44±0.02	0.48±0.03	0.48±0.02	0.51±0.03	0.52±0.02	0.52±0.04	0.51±0.03	0.52±0.04	0.52±0.04	0.5±0.04

have a good performance. For the datasets eNTERFACE05, EMO-DB and SAVEE, the total sample size is small, and the classifier can chive better results when the number of labelled samples is about 40

Based on the above experiments, we used our proposed framework to construct a lightweight speech emotional dataset. The speech audio was derived from Chinese TV dramas. After the processing of video-to-audio, voice activity detection, and simple filtering, 1151 speech segments were filtered out as the unlabeled pool. We divide emotion into four classes: angry, happy, neutral, and hurt. Using the proposed active learning method, we labelled a total of 404 samples on 20 rounds, and finally, get the dataset shown in Tab.3.4.

Tab. 3.4. The results of actual speech emotion dataset construction.

	Anger	Happy	Neutral	Sad	Total
Labeled Sample	50	32	298	24	404
Prediction Sample	16	16	712	3	747
Total Sample	66	48	1010	27	1151

Due to the selection of our proposed active learning strategies, many small-class samples are preferentially selected. The subjective audiometric test shows that when the amount of labelled data is less than 50%, the accuracy rate can reach 90%.

3.5 Discussion

3.5.1 Query Strategy Combination

To verify the efficiency of different strategy combinations, we randomly combined strategies into the following models, where C is a single complementary sampling, U+C is a combination of uncertainty sampling and complementary sampling, and R+C is a combination of representative sampling + Complementary sampling, D+C is combination of diversity sampling and complementary sampling, and U+R+D+C is a fusion of four sampling strategies.

We counted the average of the first 5 querying accuracy values. Tab.3.5 shows the comparison results of different strategy combinations. U+R+D+C model achieves best performance, where U+R+D is used to improve the accuracy of overall dataset, and C is used to give the priority of selecting the small-class samples.

Tab. 3.5. The average accuracy performance of different strategy combinations.

	C	U+C	R+C	D+C	U+R+D+C
SAVEE	0.305±0.03	**0.326±0.04**	0.295±0.05	0.296±0.03	0.291±0.03
EMO-DB	0.451±0.05	**0.476±0.05**	0.439±0.03	0.472±0.03	0.455±0.03
CASIA_im	**0.691±0.04**	0.684±0.02	0.68±0.03	0.68±0.03	0.666±0.04
eNTERFACE05_im	0.265±0.03	0.271±0.02	0.267±0.02	0.271±0.04	**0.286±0.03**
IEMOCAP_im	0.45±0.03	**0.465±0.05**	0.458±0.04	0.448±0.05	0.455±0.03
Aggregation	0.548±0.06	0.542±0.01	0.555±0.03	0.536±0.05	**0.564±0.02**
Blood	0.527±0.06	0.532±0.07	0.574±0.07	0.517±0.06	**0.584±0.06**
Diabetes	0.691±0.03	0.688±0.02	0.659±0.04	**0.694±0.02**	0.672±0.03
qsar-biodeg	0.79±0.03	0.796±0.02	0.791±0.03	0.802±0.03	**0.805±0.03**
Vote	0.923±0.02	0.923±0.01	**0.933±0.02**	0.924±0.02	0.932±0.02
Vowel	0.681±0.03	0.656±0.04	**0.687±0.03**	0.684±0.05	0.679±0.04
WBC	**0.941±0.02**	0.934±0.02	0.936±0.01	**0.941±0.02**	0.916±0.02
Thyroid	0.919±0.05	0.919±0.05	0.924±0.06	0.921±0.04	**0.928±0.04**

3.5.2 Ratio Value

The Ratio value represents the proportion of samples to be retained after querying by the four strategies. To further explore the efficiency of different ratio values, we perform the experiments with parameters 0.5, 0.6, 0.7, 0.8, and 0.9. As the Tab.3.6 shown, we observed that the ratio value around 0.8 can achieve better results.

3.5.3 Algorithm Running Speed

We reported the average CPU time of each query time for the datasets with different sample sizes. Fig.3.5 shows the running speed of our proposed method under the CentOS

Tab. 3.6. The accuracy performance of different Ratio values in the average of first 5 queries.

	R_0.5	R_0.6	R_0.7	R_0.8	R_0.9
SAVEE	0.317±0.06	0.313±0.04	0.299±0.06	0.311±0.04	**0.319±0.03**
EMO-DB	0.439±0.01	0.448±0.03	0.434±0.01	**0.46±0.05**	0.437±0.05
CASIA_im	0.589±0.03	0.653±0.02	0.68±0.03	**0.679±0.03**	0.679±0.04
eNTERFACE05_im	0.305±0.03	**0.306±0.06**	0.305±0.03	0.29±0.03	0.291±0.03
IEMOCAP_im	0.476±0.02	**0.479±0.02**	0.468±0.02	0.478±0.02	0.457±0.02
Aggregation	0.504±0.03	0.523±0.04	0.563±0.03	**0.571±0.03**	0.559±0.04
Blood	0.591±0.02	**0.599±0.03**	0.588±0.03	0.594±0.04	0.571±0.05
Diabetes	0.597±0.05	0.605±0.04	0.626±0.05	0.635±0.04	**0.641±0.05**
qsar-biodeg	0.808±0.02	0.807±0.02	**0.813±0.02**	0.799±0.02	0.8±0.01
Vote	0.902±0.04	0.907±0.04	0.91±0.03	0.916±0.02	**0.922±0.02**
Vowel	0.657±0.05	0.686±0.03	0.692±0.04	**0.706±0.06**	0.696±0.04
WBC	0.894±0.03	0.893±0.04	0.907±0.06	0.925±0.02	**0.929±0.02**
Thyroid	0.841±0.07	0.864±0.07	0.86±0.08	**0.884±0.04**	0.857±0.03

system with the machine configured as Intel(R) Core(TM) i7-6800K CPU @ 3.40GHz. The horizontal axis represents the number of queries, and the vertical axis represents the running time of each query. The CASIA and IEMOCAP datasets are too large and consume more time, and other data sets are kept within 100s.

3.5.4 Necessity of Constructing Speech Emotion Dataset

The extracted features of voice collected in different scenarios have different feature spaces. In the application of the specific scene, it is very important and necessary to construct a new dataset. To verify our thinking, we selected 4 datasets, each of which has 4 common emotions, Anger, Happy, Neutral, and Sad. We divide the training set and the test set with the ratio of 8:2 and train the classifier on each dataset. Then we used the four classifiers to view the performance of the four datasets.

Fig.3.6 shows the cross-prediction results for four datasets. We observe that the classifier trained on its own dataset has the best performance, but performs poorly on

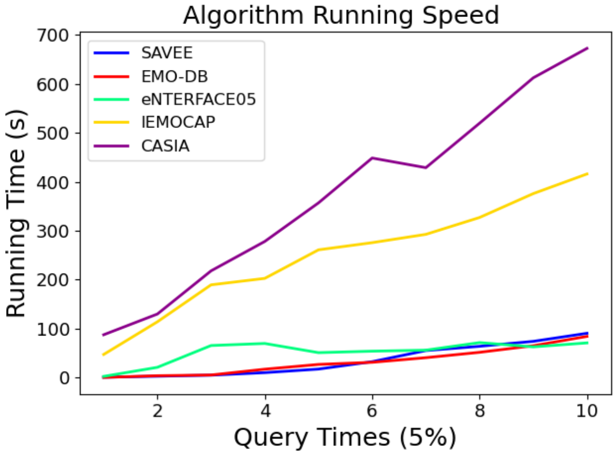


Fig. 3.5. The running speed of 10 queries on 5 speech emotional datasets.

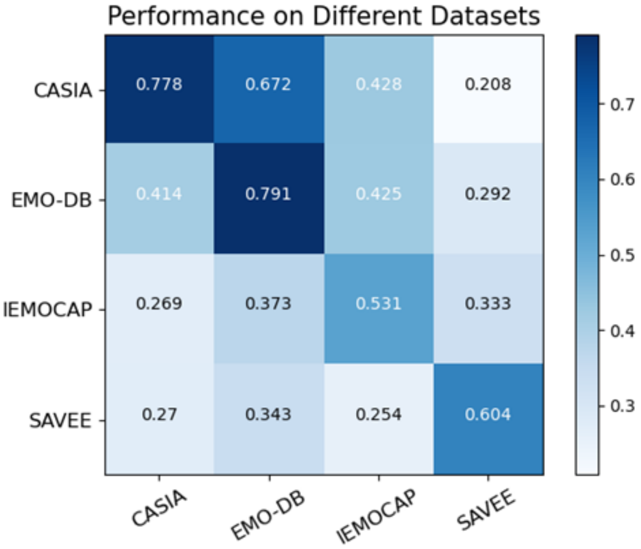


Fig. 3.6. The cross-prediction results on four datasets.

other datasets, which means that the contextual speaking styles in different scenarios will affect the recognition accuracy. Therefore, in the field of SER, it is very necessary to construct new corpus under different application scenarios.

3.6 Summary

To solve the high cost of the speech emotion corpus construction, especially for the problem of difficulty in sampling small-class samples, we propose an integrated active learning strategy and designed a framework for the construction of speech emotion dataset. In comparison experiments with other active learning strategies, our proposed active learning method achieved the best performance for sampling on different datasets, especially for the selection of small-class samples, which is significantly better than other methods. In another actual dataset construction experiment, our method in the process of constructing the dataset, when the total number of labeled samples is less than 50%, the recognition accuracy of emotion classes reaches more than 90%.

4 Dual-Transformer-BiLSTM for Speech Emotion Recognition

4.1 Introduction

In the feature construction process of speech emotion recognition, it is usually necessary to learn features from a large amount of speech data. However, in the case of a small number of samples in the speech emotion dataset, the extracted features will be over-fitted by the neural network model due to poor robustness, so the model cannot obtain a high emotion recognition accuracy. Speech self-supervised learning [70] is to find a higher-level information expression from massive speech, not limited to a single sample, and can be used for a variety of downstream tasks related to speech. The speech pre-trained features extracted by the speech pre-training model can represent the general features of speech, which can effectively improve the robustness of features. In recent years, in the field of speech signal processing, the speech representation obtained by unsupervised training on large-scale data sets through transfer learning technology has a greater performance in downstream tasks than traditional features that focus on a single sample.

Acoustic features are typically manually designed and based on signal processing. The advantages of acoustic features are their simplicity, effectiveness, and ease of calculation. The traditional manual extraction of acoustic features for modeling research is still the current mainstream. Some classic machine learning models, such as SVM, k-nearest neighbors(KNN), GMM, etc., often use acoustic features to classify emotions [106, 5]. Low-level descriptors (LLDs) are difficult to express contextual information in the temporal domain without incorporating a time-series model. [16] explored the use of different types of acoustic features and their influence on the LSTM time series network, and achieved excellent results. Issa [51] taked the acoustic features of audio MFCC and mel spectrum as input, and uses CNN neural network to achieve high emotion recognition rate on three data sets. Although CNN/RNN/LSTM models can capture higher-level features, they do not have the ability to pay attention between features. Therefore, Li [61] proposed the BiLSTM-DSA model, a BiLSTM with a self-attention mechanism, to improve the robustness of the acoustic features of the model.

The feature extraction process usually requires human experience and domain knowledge, which may introduce bias to the dataset and task and may not fully capture the dynamic characteristics of speech signals. Pre-trained features are automatically learned through neural networks, typically using large-scale speech datasets. Compared to acoustic features, pre-trained features can better capture the dynamic characteristics and contextual information of speech signals, with better representation and generalization abilities. Speech self-supervised learning has the following categories. One is self-supervised learning with the idea of contrast prediction [87], which uses an autoregressive model to predict the future of the latent space to learn advanced representations. Schneider [96] uses this idea to encode the upper part of the speech in combination with multi-layer CNN, and predicts the lower part of the speech by comparison .Another type is self-supervised learning using autoregressive predictive coding. Unlike the left-to-right prediction of contrast prediction, it can predict information at any location through context. Chung [26] used the autoregressive model to encode the context information of the past acoustic sequence of speech to predict the future frame information of the speech. There is also a way to encode speech information in a way of masking reconstruction, especially in recent years, the bert-based model learns the learning of high-dimensional information of speech [25, 65, 66].However, pre-trained features may face domain shift issues, where their performance may degrade on different datasets. Therefore, combining and fusing both types of features can increase their robustness and enhance the performance of speech recognition systems [72].

Among the current feature fusion methods in the field of speech emotion recognition, most methods fuse features after processing speech features through a multi-layer network [23, 131, 118]. These methods cannot capture the correspondence between the speech features in the early stage, so causing the loss of corresponding information of speech features. There are also some studies that try to fuse the features as early as possible [112, 131]. These methods only use concatenation methods and cannot effectively learn the correspondence between different types of features. In addition, due to the different segment lengths and dimensions of the features, part of the information contained will be forcibly converted into a uniform dimension for fusion, resulting in loss of information and making it difficult to achieve efficient feature fusion.

Therefore, in order to solve the problem of low speech emotion recognition rate caused by the weak robustness of acoustic features, we propose a novel feature fusion model Dual-TBNet, which contains two Transformers and two BiLSTM structures. In addition, to our knowledge, our work is the first to fuse self-supervised pre-trained features and acoustic features with different segment lengths and dimension sizes for speech emotion recognition. Experiments show that our model achieves higher speech emotion recognition rate on 5 speech emotion data sets.

4.2 Methodology

As shown in Fig.4.1, the entire framework mainly includes a feature construction part and a feature fusion part. The feature construction part includes the acoustic features and pre-trained features. The model part includes two 1D convolution modules, two Transformer-based feature fusion modules and two BiLSTM modules, as well as the final fully connected layer and Softmax classification module.

The 1D convolutional layer can effectively align features of different lengths and dimensions and convert them into the same dimension. Acoustic features and pre-trained features will first be processed by 1D convolutional layers, and the two features will be converted into vectors with the same dimensions. And then feed the output result into the Transformer attention fusion model for further fusion. The Transformer network structure is entirely composed of the Attention mechanism, which increases the training speed and can effectively capture the relationship between the input units. We use the self-attention mechanism in Transformer to capture the correspondence between acoustic features and pre-trained features. Then, the fused features are fed into BiLSTM to further learn the contextual relationship. Finally, the fully connected layers and softmax layers are used for emotional classification.

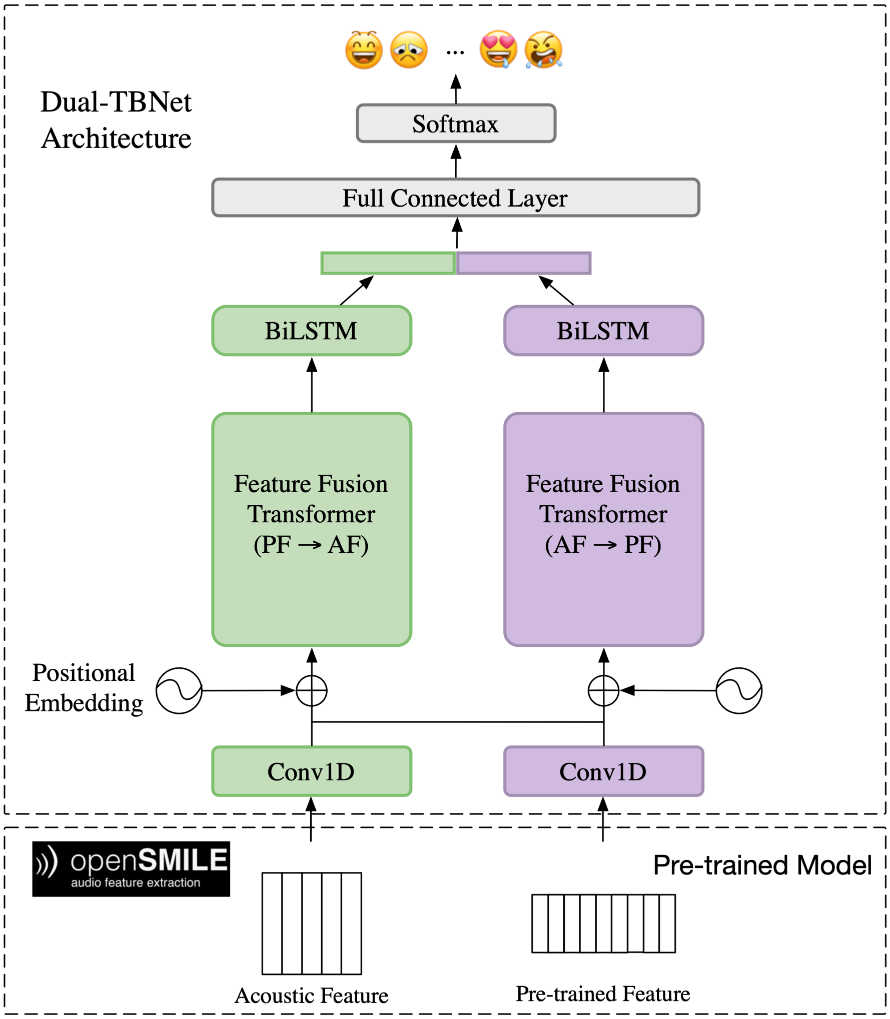


Fig. 4.1. Overall architecture for fusing pre-trained features and acoustic features.

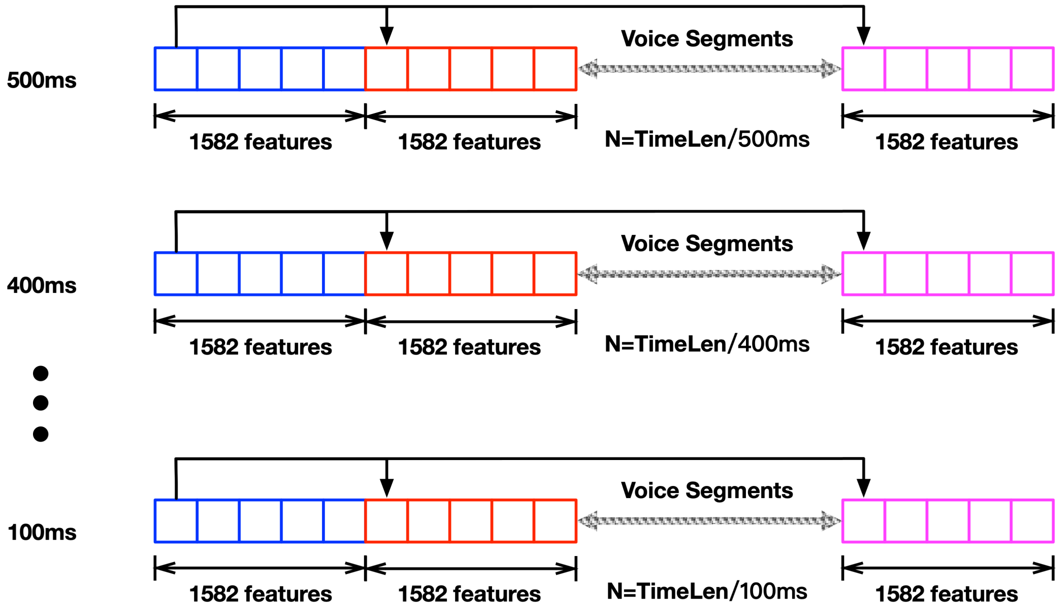


Fig. 4.2. Feature distribution calculation process for exploring the diversity of acoustic features.

4.2.1 Feature Extraction

In the research, we use five speech pre-trained models, Tera [65], Audio Albert [25], NPC [64], Wav2Vec [96] and Vq-wav2vec [7] combined with acoustic features to improve the accuracy of speech emotion recognition. TERA is a self-supervised speech pre-training method, which is used for pre-training on a large number of unlabeled speech by masking the speech spectrum along three orthogonal axes to obtain Transformer encoding model. AUDIO ALBERT uses the ALBERT self-supervised learning model, which is trained on large-scale speech data sets, and can be used for feature extraction of downstream tasks such as speech-related tasks, or as a fine-tuning participation model training. NPC is also a self-supervised learning method. It only relies on the partial information of the speech to represent the speech in a non-autoregressive manner. Wav2vec uses contrastive loss as the loss function for training, and the final representation can replace the acoustic feature. Vq-wav2vec adds a quantization module to wav2vec to improve the performance of the model.

For the acoustic features, each utterance is segmented into several sub-segments with a fixed time length. We used the openSMILE tool to extract the acoustic features of each sub-segment to the dimension of 1582, constructing a total of $n * 1582$ feature blocks, where n is the number of sub-segments after different speech segments. Acoustic features can represent the static characteristics of a certain segment of speech, but cannot represent the context information of entire utterance.

4.2.2 Acoustic Feature Statistics

As shown in Fig.4.2, to explore the effect of segmentation time on the model, we divided each utterance into several speech segments by taking 100ms, 200ms, 300ms, 400ms and 500ms as time units and use the standard deviation to calculate the distribution of the 1582 features of a speech file in different length segments.

Eq.4.1 is the standard deviation calculation formula, n represents the number of segments of the speech divided by the specified time length, $\bar{x}$ represents the average value of one of 1582 features, and x_i represents the features of different segments. σ represents the distribution of the value of the feature over the time length. The larger the value, the more diverse the feature distribution. By using the statistic method, to calculate the distribution of the feature value in different time lengths. The more diverse the distribution of feature values, the better the recognition results of the model [132]. We found that the features of the same dimension are very diversely distributed in the 200ms segments. So we use the 200ms features as our acoustic features.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \tag{4.1}$$

4.2.3 Model Structure

In Fig.4.3, the attention mechanism in Transformer can focus on the correspondence between speech segments. Therefore, with the help of this mechanism, it can fully promote the full fusion of acoustic features and pre-trained features. In our experiment, we use a 6-layer Transformer encoder to fuse the features. The attention mechanism in the

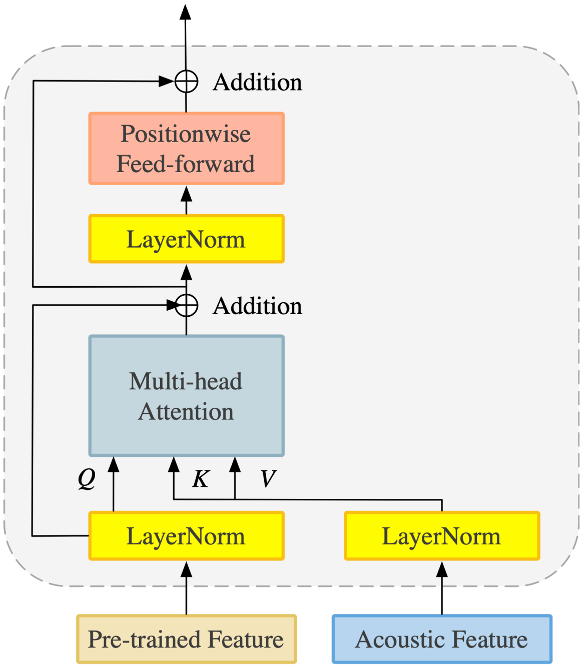


Fig. 4.3. Feature fusion Transformer for fusing acoustic features and pre-trained features.

traditional Transformer is represented by Eq.4.2, Q(Query) represents the query vector, and K(Key)V(Value) represents the vector being queried. QKV are derived from the model's input vector through three different matrix multiplication linear transformations.

$$\text{Attention}\ (Q,K,V) = \text{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{4.2}$$

In our method, since pre-trained features and acoustic features have different maximum lengths and dimensions, we first use 1D convolutional network to convert the acoustic features and pre-trained features to the same dimensional features. Then use the one feature as query Q, and the another feature as K and V, and fuse them by using the Eq.4.3. X_α and X_β respectively represent the acoustic features and pre-trained features. We define the Querys as $Q_\alpha = X_\alpha W_{Q_\alpha}$, Keys as $K_\beta = X_\beta W_{K_\beta}$ and Values as $V_\beta = X_\beta W_{V_\beta}$. The adaptation from pre-trained features(PF) to acoustic features(AF) is presented as $AF_{\beta\rightarrow\alpha}\left(X_\alpha, X_\beta\right)$.

$$\begin{aligned} AF_{\beta\rightarrow\alpha}\left(X_{\alpha},X_{\beta}\right) &= \text{Softmax}\left(\frac{Q_{\alpha}K_{\beta}^{\mathrm{T}}}{\sqrt{d_k}}\right)V_{\beta} \\ &= \text{Softmax}\left(\frac{X_{\alpha}W_{Q_{\alpha}}W_{K_{\beta}}^{\mathrm{T}}X_{\beta}^{\mathrm{T}}}{\sqrt{d_k}}\right)X_{\beta}W_{V_{\beta}} \end{aligned} \tag{4.3}$$

In contrast, the adaptation from acoustic features(AF) to pre-trained features(PF) is presented as $PF_{\alpha\rightarrow\beta}\left(X_{\alpha},X_{\beta}\right)$ in Eq.4.4. Through the two Transformer fusion structures, the features can be fully fused and then fed to the BiLSTM model.

$$\begin{aligned} PF_{\alpha\rightarrow\beta}\left(X_{\alpha},X_{\beta}\right) &= \text{Softmax}\left(\frac{Q_{\beta}K_{\alpha}^{\mathrm{T}}}{\sqrt{d_k}}\right)V_{\alpha} \\ &= \text{Softmax}\left(\frac{X_{\beta}W_{Q_{\beta}}W_{K_{\alpha}}^{\mathrm{T}}X_{\alpha}^{\mathrm{T}}}{\sqrt{d_k}}\right)X_{\alpha}W_{V_{\alpha}} \end{aligned} \tag{4.4}$$

In Fig.4.4, we use the BiLSTM structure to further learn the front and back spatial information fused by Transformer to enhance the robustness of features. BiLSTM is composed of two reverse LSTM networks. LSTM is a long and short-term memory network, a structure further optimized on the basis of RNN. LSTM can learn what information to remember and what information to forget through the training process, and solves the problem of gradient explosion and gradient disappearance caused by the RNN as the sentence length is too long. LSTM is often used in tasks with time series context information such as text data and speech data. Since it is unable to encode the information from the back to the front of the sentence, BiLSTM uses two LSTM structures in the opposite direction [101].

4.3 Experimental Setup

4.3.1 Datasets

In the experiment, we used 5 kinds of data sets, CASIA, eNTERFACE05, IEMOCAP, EMO-DB, and SAVEE. The specific division of the datasets used in the experiment is shown in Tab.4.1. The total number of utterances in the five datasets, CASIA, eNTERFACE05, IEMOCAP, EMO-DB, and SAVEE, are 6000, 1257, 4660, 535, and 480,

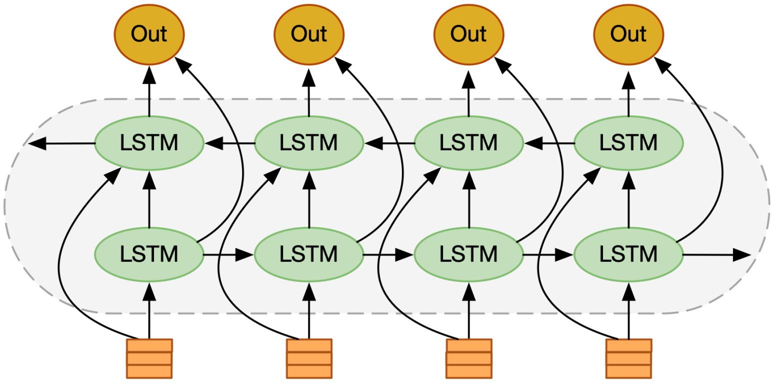


Fig. 4.4. BiLSTM module for capturing contextual information.

respectively. We divided the datasets into training, validation, and test sets in a ratio of approximately 8:1:1. However, due to the presence of overly long utterances in the IEMOCAP dataset, we removed these samples from the comparison experiments.

Tab. 4.1. The division of training set, validation set and test set of 5 kinds of data sets.

	Total	train	valid	test
CASIA	6000	5440	280	280
eNTERFACE05	1257	1017	120	120
IEMOCAP	4660	3705	464	491
EMO-DB	535	426	40	69
SAVEE	480	408	36	36

4.3.2 Baseline Models

In order to compare the performance of different feature fusion models, we designed a total of four feature fusion models. The details of the four special fusion models are shown in Fig.4.5.

1. TB_af: The TB_af model contains two 1D convolutional layers, a Transformer feature fusion module and BiLSTM module. Acoustic features are used for Q query vectors, and pre-trained features are used for KV vectors.

2. TB_pf: The TB_pf model is similar to the TB_af model. The only difference is that TB_pf uses pre-trained features for the Q query vector, and acoustic feature features for the KV vectors.

3. DT: The DT model contains two 1D convolutional layers, two Transformer feature fusion modules and two common Transformer modules. Acoustic features and pre-trained features are respectively used as Q query vectors, and finally output to two Transformers for further fusion.

4. Dual-TBNet: DBual-TBNet is our proposed model, which is similar to the DT network. The only difference is that after the features are fused with the output of the two common Transformer, the context information is further learned through two BiLSTM modules.

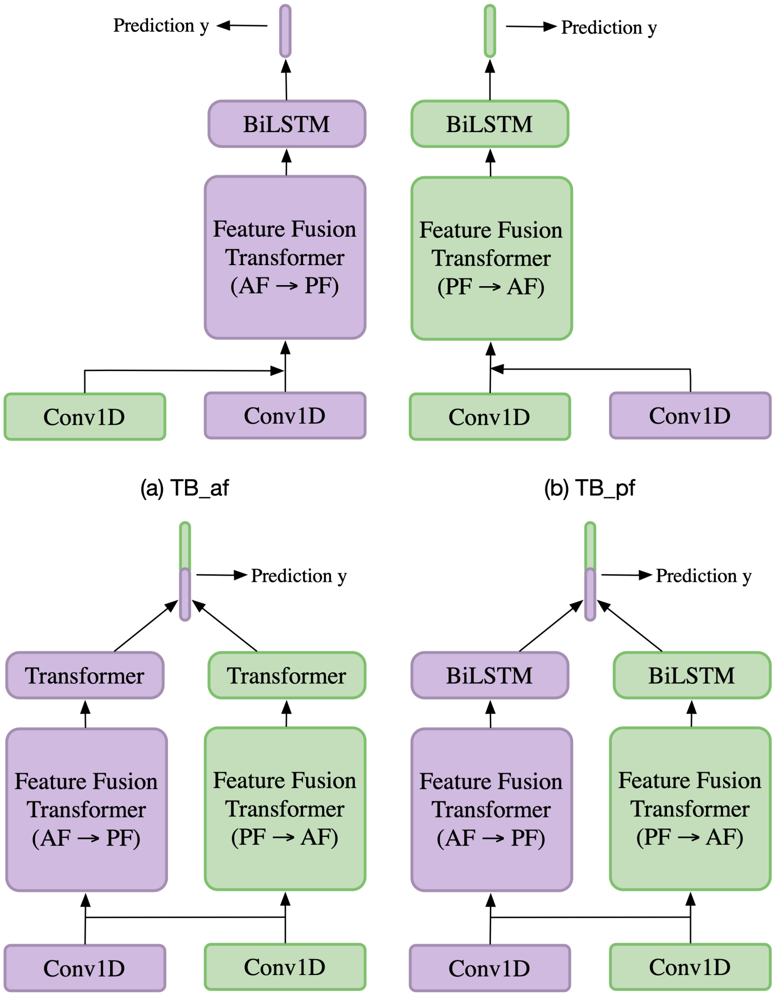
For the feature input of all models, we use acoustic features to fuse with Tera, Audio albert, NPC, Wave2vec and Vqwav2vec pre-trained features. Each model will have 5 feature fusion combinations, and finally 20 sets of experiments to compare the final speech recognition performance.

4.3.3 Evaluation Metrics

In this experiment, for the results of speech emotion recognition, we used two evaluation metrics, namely F1 and Ac. The F1 is the weighted average of precision and recall, and the F1 formula is expressed as Eq.4.5. Among them, TP (True Positive) indicates that the model prediction result is positive, and the sample is also positive, TN (True Negative) indicates that the model prediction result is positive, but the sample is negative, FP (False Positive) indicates that the prediction is negative, and the sample is positive, FN (False Negative) indicates that the prediction is negative, and the sample is also negative.

$$F_1 = \frac{2*TP}{2*TP+FP+FN} \tag{4.5}$$

Ac is the classification accuracy score, which refers to the percentage of all classifications that are correct. The Eq.4.6 is as follows:


Prediction y
BiLSTM
Feature Fusion Transformer (AF → PF)
Conv1D
Conv1D
Prediction y
BiLSTM
Feature Fusion Transformer (PF → AF)
Conv1D
Conv1D
Prediction y
Transformer
Transformer
Feature Fusion Transformer (AF → PF)
Feature Fusion Transformer (PF → AF)
Conv1D
Conv1D
Prediction y
BiLSTM
BiLSTM
Feature Fusion Transformer (AF → PF)
Feature Fusion Transformer (PF → AF)
Conv1D
Conv1D

(a) TB_af

(b) TB_pf

(c) DT

(d) Dual-TBNet

Fig. 4.5. Four baseline models for feature fusion experimental comparison.

$$Ac = \frac{TP+TF}{TP+TF+FP+FN} \tag{4.6}$$

4.4 Experimental Results

4.4.1 Acoustic Feature Statistical Results

Tab.4.2 shows the statistical results of feature diversity under different time segments using the standard deviation as the measurement standard. For all data sets, nearly half of the samples have more diverse distributions of features in the case of the 200ms time length and the length is the best for the recognition accuracy of the time series model [73]. Therefore, we use 200ms as the time segmentation length to extract static acoustic features.

Tab. 4.2. The statistical results of the feature diversity of 5 data sets under different time lengths.

	100ms	200ms	300ms	400ms	500ms
CASIA	660	**3298**	562	621	859
eNTERFACE05	347	**525**	95	129	161
IEMOCAP	1448	**2065**	348	365	757
EMO-DB	58	**383**	38	14	42
SAVEE	71	**205**	47	48	109

4.4.2 Emotion Recognition Results and Discussion

Tab.4.3 to Tab.4.6 respectively show the recognition results of the four models using five kinds of feature combinations. The 4 tables correspond to four fusion models respectively. Each table shows the Ac and F1 values with five pre-trained features and acoustic features on 5 data sets.

From the table, we observe that for all feature combination schemes, the recognition results of Tera and acoustic feature fusion schemes are better than Audio Albert, NPC,

Tab. 4.3. The results of Dual-TBNet model with different pre-trained features and acoustic features, where A presents acoustic features.

	A+NPC		A+AAlBert		A+Wav2vec		A+Vqwav2vec		A+Tera	
	Ac	F1	Ac	F1	Ac	F1	Ac	F1	Ac	F1
CASIA	0.882	0.881	0.918	0.918	0.886	0.886	0.846	0.846	**0.957**	**0.958**
eNTERFACE05	0.533	0.513	0.575	0.567	0.558	0.552	0.517	0.495	**0.667**	**0.656**
IEMOCAP	0.554	0.547	0.629	0.627	0.623	0.617	0.556	0.555	**0.648**	**0.649**
EMO-DB	0.565	0.504	0.725	0.723	0.754	0.738	0.812	0.811	**0.841**	**0.843**
SAVEE	0.722	0.677	0.722	0.71	0.778	0.769	0.639	0.64	**0.833**	**0.821**

Tab. 4.4. The results of TB_af model with different pre-trained features and acoustic features.

	A+NPC		A+AAlBert		A+Wav2vec		A+Vqwav2vec		A+Tera	
	Ac	F1	Ac	F1	Ac	F1	Ac	F1	Ac	F1
CASIA	0.796	0.799	0.821	0.821	0.779	0.781	0.796	0.796	**0.825**	**0.826**
eNTERFACE05	0.525	0.501	0.525	0.511	0.533	0.511	**0.533**	**0.523**	0.467	0.44
IEMOCAP	0.55	0.544	**0.578**	**0.574**	**0.578**	**0.574**	0.568	0.567	0.556	0.553
EMO-DB	**0.536**	**0.495**	0.507	0.455	0.551	0.481	0.565	0.502	0.522	0.463
SAVEE	0.639	0.63	**0.722**	**0.695**	0.583	0.547	0.639	0.618	0.611	0.607

Tab. 4.5. The results of TB_pf model with different pre-trained features and acoustic features.

	A+NPC		A+AAlBert		A+Wav2vec		A+Vqwav2vec		A+Tera	
	Ac	F1	Ac	F1	Ac	F1	Ac	F1	Ac	F1
CASIA	0.861	0.861	0.914	0.915	0.857	0.859	0.843	0.843	**0.943**	**0.942**
eNTERFACE05	0.45	0.422	**0.592**	**0.579**	0.525	0.516	0.525	0.516	0.558	0.532
IEMOCAP	0.57	0.571	0.576	0.566	0.635	0.634	0.599	0.595	**0.642**	**0.64**
EMO-DB	0.522	0.469	0.725	0.717	**0.797**	**0.798**	0.696	0.697	0.754	0.757
SAVEE	0.556	0.48	0.611	0.589	0.611	0.52	0.639	0.594	**0.722**	**0.726**

Tab. 4.6. The results of DT model with different pre-trained features and acoustic features.

	A+NPC		A+AAlBert		A+Wav2vec		A+Vqwav2vec		A+Tera	
	Ac	F1	Ac	F1	Ac	F1	Ac	F1	Ac	F1
CASIA	0.879	0.878	0.893	0.893	0.882	0.883	0.839	0.84	**0.911**	**0.91**
eNTERFACE05	0.5	0.475	0.508	0.497	0.492	0.486	0.467	0.442	**0.5**	**0.491**
IEMOCAP	0.568	0.562	0.56	0.557	**0.65**	**0.649**	0.566	0.566	0.589	0.585
EMO-DB	0.507	0.505	**0.667**	**0.667**	0.623	0.584	0.623	0.591	0.58	0.569
SAVEE	0.472	0.404	0.639	0.634	0.611	0.605	**0.639**	**0.635**	0.611	0.605

Wav2vec and Vq-wav2vec. Tera is a new generation of pre-trained model based on BERT, which has a stronger advantage for capturing the information between speech data frames.

In the case of the same feature fusion scheme, the result of TB_pf is better than the result of TB_af. Different from TB_af, TB_pf uses pre-trained features as Q query vector, which has better performance for model recognition. Comparing TB_pf and DT, TB_pf is also better than DT model, because BiLSTM further extracts the fusion context information in the output layer. The Dual-TBNet model performs best among all models, not only has the attention mechanism but also uses BiLSTM to capture contextual information, fully fusion of pre-trained features and acoustic features. Tera's pre-trained features are more suitable for model recognition. In general, our proposed model achieved the accuracy of 95.7%, 66.7%, 64.8%, 84.1%, and 83.3% on the 5 data sets of CASIA, eENTERFACE05, IEMOCAP, EMO-DB, and SAVEE, respectively.

To analyze the performance of our Dual-TBNet model with Tera and acoustic features on different types of emotion across five datasets, as shown in Tab.4.7, we conducted a statistical analysis on all categories using precision(P), recall(R), and F1 score(F1). We also plotted a confusion matrix in Fig.4.6, which provides a more visual representation of the performance of emotion recognition. In the CASIA data set, happy, neutral, and sad all achieved 98% recognition rate. Due to similar emotional polarity, some samples of anger and surprise were incorrectly identified as happy. Our model performs well on the Chinese act-based dataset. In the data set of eENTERFACE05, the recognition re-

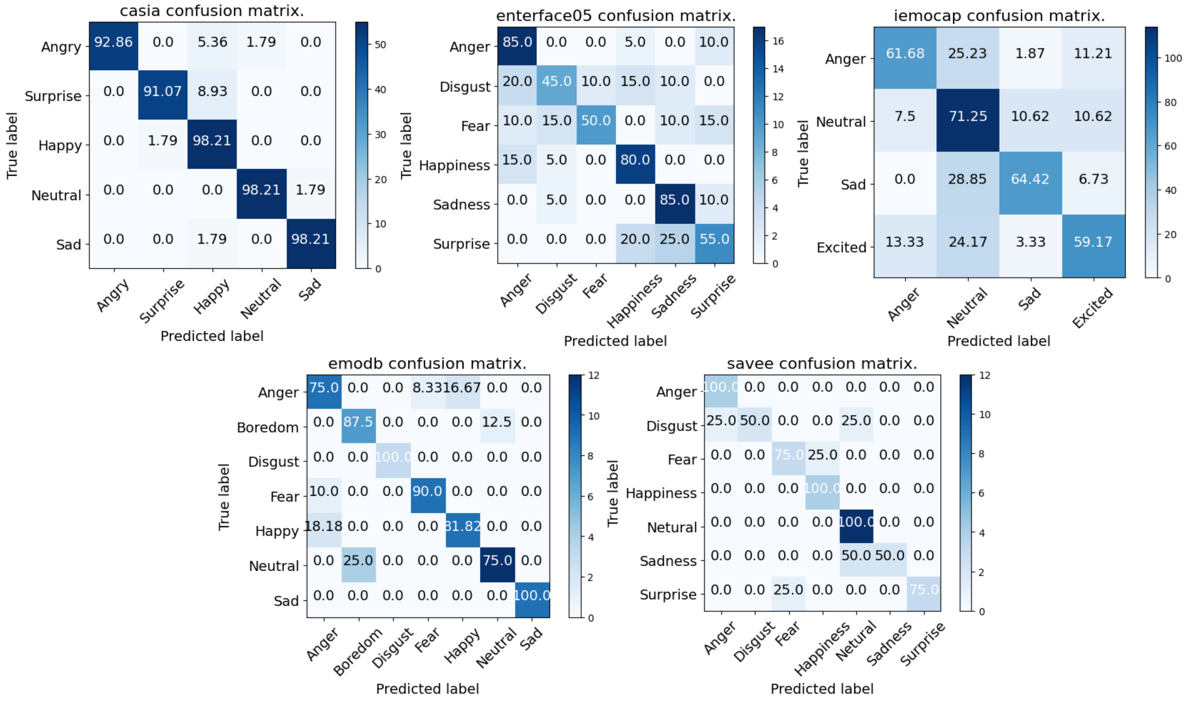


Fig. 4.6. Confusion matrix of Dual-TBNet emotion recognition results on CASIA, eNTERFACE05, IEMOCAP, EMO-DB, SAVEE datasets.

sults of Disgust and Fear are poor due to the imbalance of the categories of the data set. The neutral emotion prediction results in the IEMOCAP data set are the best. Although the EMO-DB and SAVEE data sets have a small sample size, they also achieve a high accuracy rate.

To compare with previous studies, we collected the data results of researches in the field of speech emotion recognition in the past three years. As shown in Tab.4.8, we list the emotion recognition results of different studies according to the different datasets. From the table, we can find that the models with attention mechanism have achieved good results. Among them, our proposed model achieves state-of-the-art results on the CASIA and SAVEE datasets, and achieves mid-to-upper results on eNTERFACE05, IEMOCAP, and EMO-DB. CASIA is a Chinese speech emotion dataset, and SAVEE is an English dataset, both recorded in a pure environment and composed of many short sentences. For such data, our model has the best results. For speech datasets collected in natural environ-

Tab. 4.7. Performance of the Dual-TBNet on different emotions in CASIA, eNTERFACE05, IEMOCAP, EMO-DB, and SAVEE datasets.

	IEMOCAP			CASIA			eNTERFACE05			EMO-DB			SAVEE		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Angry	0.7	0.62	0.66	1	0.93	0.96	0.65	0.85	0.74	0.75	0.75	0.75	0.8	1	0.89
Sad	0.74	0.64	0.69	0.98	0.98	0.98	0.65	0.85	0.74	1	1	1	1	0.5	0.67
Neutral	0.57	0.71	0.63	0.98	0.98	0.98	-	-	-	0.92	0.75	0.83	0.8	1	0.89
Excited	0.66	0.59	0.63	-	-	-	-	-	-	-	-	-	-	-	-
Happy	-	-	-	0.86	0.98	0.92	0.67	0.8	0.73	0.82	0.82	0.82	0.8	1	0.89
Disgust	-	-	-	-	-	-	0.64	0.45	0.53	1	1	1	1	0.5	0.67
Fear	-	-	-	-	-	-	0.83	0.5	0.62	0.9	0.9	0.9	0.75	0.75	0.75
Surprise	-	-	-	0.98	0.91	0.94	0.61	0.55	0.58	-	-	-	1	0.75	0.86
Bored	-	-	-	-	-	-	-	-	-	0.64	0.88	0.74	-	-	-

Tab. 4.8. Performance comparison between the proposed model with other models on the CASIA, eNTERFACE05, IEMOCAP, EMO-DB and SAVEE emotion corpus.

Models	CASIA	eNTERFACE05	IEMOCAP	EMO-DB	SAVEE
CNN+LSTM [30]	-	-	0.775	0.785	0.781
RNN+Attenttion [61]	-	-	0.643	**0.861**	-
3DRNN+Attention [22]	-	-	0.647	0.828	-
CNN(Spectrogram and Phoneme Features) [121]	-	-	0.64	-	-
CNN(Phonological Features) [112]	-	-	0.6002	-	-
SVM+SISMOTE [71]	0.902	-	-	0.8582	0.75
DenseNet-GRU [24]	0.8	-	-	-	-
FaceNet [69]	0.9	-	0.689	-	-
SVM+Decision Tree [107]	0.853	-	-	0.858	-
SVM [6]	-	0.564	-	0.815	0.7563
CNN+Frequential Attention [62]	-	**0.758**	**0.804**	0.833	0.565
Dual-TBNet	**0.957**	0.667	0.648	0.841	**0.833**

ment such as eNTERFACE05, IEMOCAP, our model also achieves good performance.

4.5 Summary

To improve the accuracy of speech emotion recognition on limited datasets, we use the pre-trained features learned on large-scale datasets by self-supervised learning to enhance the robustness of acoustic features. Furthermore, we propose a new feature fusion model called Dual-TBNet, which mainly consists of dual Transformer and BiLSTM modules. With the help of attention mechanism and bidirectional time series module, our model can fully learn the corresponce information betwwen acoustic features and pre-trained features with different segment lengths and dimension sizes.

A total of four fusion models are designed to fuse five pre-trained features and acoustic features. Among them, the Dual-TBNet model achieve 95.7%, 66.7%, 64.8%, 84.1% and 83.3% accuracy in the comparative experiments of CASIA, eNTERFACE05, IEMOCAP, EMO-DB and SAVEE datasets respectively.

5 Conclusion and Future Work

5.1 Conclusion

Speech emotion recognition can help us better understand and interpret the emotional states expressed by people in verbal communication. Emotion is an essential component of human interaction, as it can influence our behavior, attitudes, and decisions. By accurately recognizing and understanding emotions in speech, we can gain deeper insights into others' inner feelings, thus establishing stronger interpersonal relationships.

Currently, there are several challenges in constructing such datasets. Firstly, collecting emotional data requires the participation of many subjects, which requires a lot of time and manpower resources. Additionally, since emotion is a subjective experience, the quality and quantity of the data collected depend on the participants' emotional expression. Secondly, to train machine learning models, emotional data needs to be labeled with emotional categories. Annotators need to listen to all speech materials one by one, which is a costly process and cannot prioritize rare samples. Thirdly, emotional data is often imbalanced, as some emotions may be more common than others, leading to a shortage of samples for certain categories in the dataset. This may result in poor model performance when predicting less common emotions. Fourthly, the data collected is often specific to certain scenarios, which affects the accuracy of speech emotion recognition in different languages, speaking styles, and application scenarios. Therefore, the dataset needs to include participants with different languages, cultural backgrounds, ages, genders, and personalities. Finally, the accuracy of speech emotion recognition is affected by many factors, such as audio quality, speaker pronunciation and language habits, environmental noise, etc.

Most current research in this field is based on extracting speaker-specific features for model training. These features only reflect the characteristics of the current dataset, and therefore have poor robustness. In order to improve feature robustness, future research can combine the use of pre-training models to extract high-dimensional feature spaces. For example, using speech representation learning methods to construct feature spaces on massive speech data, to learn personalized speech features and enhance the robustness of speech features. In addition, with the continuous deepening of research on speech features, various emotion-related features have been proposed. For these features, exploring fusion schemes between different categories of speech features can effectively promote the improvement of speech emotion recognition accuracy.

To address the two challenges in the field of speech emotion recognition mentioned above, we conducted research on the following aspects to promote its development.

1. We propose an effective active learning strategy to overcome challenges related to limited dataset availability and construction efficiency. By incorporating uncertainty, representativeness, diversity, and complementarity data selection methods, the proposed strategy identifies valuable data for annotation, resulting in superior outcomes.

2. We present a novel feature fusion architecture that combines two Transformer and BiLSTM modules. This architecture enhances the accuracy of speech emotion recognition by enriching the richness and robustness of speech features. Experimental evaluations demonstrate the state-of-the-art performance achieved by the proposed approach.

5.2 Future Work

The architecture of the active learning method we proposed is serial, which cannot achieve parallel operations. In the future, we will try a parallel sampling architecture to further improve the overall running speed of the method. Additionally, the sampling algorithm we propose combines the logistic regression classifier for sample selection. In the future, we will explore other classification models to improve the sampling efficiency. Moreover, in the actual application of active learning, we will further explore the application of active learning methods in the construction of multimodal emotional corpus to adapt to the increasingly complex mass data.

For the speech emotion recognition model, our Dual-TBNet model is in the form of a frozen pre-trained model to extract the speech feature information. In the future, we will try to add the pre-trained model to our framework to fine-tune the model. In addition, we will further explore the accuracy of speech emotion recognition under natural conditions.

Acknowledgement

I am deeply grateful for the people who have supported me throughout my academic journey and personal life. Firstly, I want to express my gratitude to Prof. Ren Fuji for providing me with a conducive learning environment, access to various resources and invaluable research guidance. Secondly, I would like to thank Assistant Prof. Kang Xin for providing me with valuable suggestions on research details, which greatly enhanced my academic ability. Additionally, I would like to extend my heartfelt appreciation to Prof. Fukeda, Prof. Shishibori, and Prof. Terada for reviewing my doctoral thesis and all the members of the A1 laboratory cohort, especially Wang Linhuang, for their advice, support and assistance.

As an international student, I faced numerous challenges when I first enrolled at Tokushima University. I didn't know how to speak Japanese, and I struggled with research. However, over the course of five years, I have learned Japanese and adjusted to life in Japan. In terms of academics, I have also made significant progress, having gained mastery in research methods, learned how to conduct research, how to write papers, and how to effectively present my research findings to others. Overall, both my personal and academic growth have been significant during my time at Tokushima University.

In the past few years, many unfortunate events have also occurred. My mother was diagnosed with late-stage ovarian cancer, and we faced financial difficulties during that time. These experiences caused me a lot of pain and sadness. However, with our collective efforts, we have successfully overcome challenging times, and now everything has become better. I would like to express my gratitude to my family for their unwavering support in my academic pursuits, as well as to my wife Mrs. Li for her invaluable companionship and support during significant phases of my journey.

My student days are coming to an end, while my academic career is just beginning. In the future, I will maintain a positive attitude and continue to pursue my dreams and goals. I also hope that my life and career will continue to improve. Once again, I want to thank all the professors and friends who have helped me during these times.

References

[1] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249, 2021.

[2] Mohammed Abdelwahab and Carlos Busso. Incremental adaptation using active learning for acoustic emotion recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5160–5164. IEEE, 2017.

[3] Mohammed Abdelwahab and Carlos Busso. Active learning for speech emotion recognition using deep neural network. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

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

[5] Mohammed Jawad Al Dujaili, Abbas Ebrahimi-Moghadam, and Ahmed Fatlawi. Speech emotion recognition based on svm and knn classifications fusion. *International Journal of Electrical and Computer Engineering*, 11(2):1259, 2021.

[6] J Ancilin and A Milton. Improved speech emotion recognition with mel frequency magnitude coefficient. *Applied Acoustics*, 179:108046, 2021.

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

[8] Sweeta Bansal and Amita Dev. Emotional hindi speech database. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE, 2013.

[9] Fang Bao, Michael Neumann, and Ngoc Thang Vu. Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In *INTERSPEECH*, pages 2828–2832, 2019.

[10] Anton Batliner, Stefan Steidl, and Elmar Nöth. Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. In Laurence Devillers, Jean-Claude Martin, Roddy Cowie, Ellen Douglas-Cowie, and Anton Batliner, editors, *Proc. of a Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect*, pages 28–31, Marrakesh, 2008.

[11] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. Madmom: A new python audio and music signal processing library. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1174–1178, 2016.

[12] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Herrera Boyer, Oscar Mayor, Gerard Roma Trepat, Justin Salamon, José Ricardo Zapata González, Xavier Serra, et al. Essentia: An audio analysis library for music information retrieval. In *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR), 2013.

[13] Tom B Brown, Benjamin Mann, Nick Ryder, Meghana Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Girish Shyam, Gaurav Sastry, Alexander Askell, et al. Language models are few-shot learners. In *International Conference on Learning Representations*, 2020.

[14] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

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

[16] Sung-Woo Byun and Seok-Pil Lee. A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms. *Applied Sciences*, 11(4):1890, 2021.

[17] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[18] Houwei Cao, Ragini Verma, and Ani Nenkova. Combining ranking and classification to improve emotion recognition in spontaneous speech. In *Interspeech 2012*, pages 358–361. ISCA, September 2012.

[19] Xiangyong Cao, Jing Yao, Zongben Xu, and Deyu Meng. Hyperspectral image classification with convolutional neural network and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7):4604–4616, 2020.

[20] Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, and Shrikanth Narayanan. Data augmentation using gans for speech emotion recognition. In *Interspeech*, pages 171–175, 2019.

[21] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, October 2018.

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

[23] Qiupu Chen and Guimin Huang. A novel dual attention-based blstm with hybrid features in speech emotion recognition. *Engineering Applications of Artificial Intelligence*, 102:104277, 2021.

[24] Siyuan Cheng, Dongya Zhang, and Didi Yin. A densenet-gru technology for chinese speech emotion recognition. In *International Conference on Frontiers of Electronics, Information and Computation Technologies*, pages 1–7, 2021.

[25] Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350. IEEE, 2021.

[26] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240*, 2019.

[27] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. Emovo corpus: an italian emotional speech database. In *International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3501–3504. European Language Resources Association (ELRA), 2014.

[28] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[29] Chenye Cui, Yi Ren, Jinglin Liu, Feiyang Chen, Rongjie Huang, Ming Lei, and Zhou Zhao. Emovie: A mandarin emotion speech dataset with a simple emotional text-to-speech model. *arXiv preprint arXiv:2106.09317*, 2021.

[30] Ranjana Dangol, Abeer Alsadoon, PWC Prasad, Indra Seher, and Omar Hisham Alsadoon. Speech emotion recognition using convolutional neural network and

long-short termmemory. *Multimedia Tools and Applications*, 79(43):32917–32934, 2020.

[31] Javier de Lope and Manuel Graña. An ongoing review of speech emotion recognition. *Neurocomputing*, 2023.

[32] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(03):34–41, 2012.

[33] Sandra Ebert, Mario Fritz, and Bernt Schiele. Ralf: A reinforced active learning formulation for object class recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633. IEEE, 2012.

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

[35] Mehmet Bilal Er. A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features. *IEEE Access*, 8:221640–221653, 2020.

[36] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch. Cnn+ lstm architecture for speech emotion recognition with data augmentation. *arXiv preprint arXiv:1802.05630*, 2018.

[37] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[38] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

[39] Theodoros Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12):e0144610, 2015.

[40] RA Gilyazev and D Yu Turdakov. Active learning and crowdsourcing: A survey of optimization methods for data labeling. *Programming and Computer Software*, 44(6):476–491, 2018.

[41] Mohamed Goudjil, Mouloud Koudil, Mouldi Bedda, and Noureddine Ghoggali. A novel active learning method using svm for text classification. *International Journal of Automation and Computing*, 15(3):290–298, 2018.

[42] Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. A canadian french emotional speech dataset. In *Proceedings of the 9th ACM multimedia systems conference*, pages 399–402, 2018.

[43] Wenjing Han, Tao Jiang, Yan Li, Björn Schuller, and Huabin Ruan. Ordinal learning for emotion recognition in customer service calls. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6494–6498. IEEE, 2020.

[44] Lasse Hansen, Yan-Ping Zhang, Detlef Wolf, Konstantinos Sechidis, Nicolai Ladegaard, and Riccardo Fusaroli. A generalizable speech emotion recognition model reveals depression and remission. *Acta Psychiatrica Scandinavica*, 145(2):186–199, 2022.

[45] Sanaul Haq, Philip JB Jackson, and J Edge. Speaker-dependent audio-visual emotion recognition. In *AVSP*, volume 2009, pages 53–58, 2009.

[46] Jia-Hao Hsu, Ming-Hsiang Su, Chung-Hsien Wu, and Yi-Hsuan Chen. Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1675–1686, 2021.

[47] Hao Hu, Ming-Xing Xu, and Wei Wu. GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, pages IV–413–IV–416, Honolulu, HI, April 2007. IEEE.

[48] Shengjun Huang, Rong Jin, and Zhihua Zhou. Active learning by querying informative and representative examples. *Advances in neural information processing systems*, 23, 2010.

[49] Yusuke Ijima, Makoto Tachibana, Takashi Nose, and Takao Kobayashi. Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4157–4160, Taipei, Taiwan, April 2009. IEEE.

[50] Maryam Imani and Gholam Ali Montazer. A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications*, 147:102423, 2019.

[51] Dias Issa, M Fatih Demirci, and Adnan Yazici. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894, 2020.

[52] Bache K and Lichman M. UCI machine learning repository, 2013.

[53] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019.

[54] Trinh Le Ba Khanh, Soo-Hyung Kim, Gueesang Lee, Hyung-Jeong Yang, and Eu-Tteum Baek. Korean video dataset for emotion recognition in the wild. *Multimedia Tools and Applications*, 80(6):9479–9492, 2021.

[55] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. *Advances in neural information processing systems*, 30, 2017.

[56] Shashidhar G Koolagudi, Sudhamay Maity, Vuppala Anil Kumar, Saswat Chakrabarti, and K Sreenivasa Rao. Iitkgp-sesc: speech database for emotion analysis. In *International conference on contemporary computing*, pages 485–492. Springer, 2009.

[57] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[58] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Julien Epps. Direct Modelling of Speech Emotion from Raw Speech. In *Interspeech 2019*, pages 3920–3924. ISCA, September 2019.

[59] Lik-Hang Lee, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *arXiv preprint arXiv:2110.05352*, 2021.

[60] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994.

[61] Dongdong Li, Jinlin Liu, Zhuo Yang, Linyu Sun, and Zhe Wang. Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Systems with Applications*, 173:114683, 2021.

[62] Shuzhen Li, Xiaofen Xing, Weiquan Fan, Bolun Cai, Perry Fordson, and Xiangmin Xu. Spatiotemporal and frequential cascaded attention networks for speech emotion recognition. *Neurocomputing*, 448:238–248, 2021.

[63] Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. Cheavd: a chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, 8(6):913–924, 2017.

[64] Alexander H Liu, Yu-An Chung, and James Glass. Non-autoregressive predictive coding for learning speech representations from local dependencies. *arXiv preprint arXiv:2011.00406*, 2020.

[65] Andy T Liu, Shang-Wen Li, and Hung-yi Lee. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366, 2021.

[66] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE, 2020.

[67] Jiawang Liu and Haoxiang Wang. A Speech Emotion Recognition Framework for Better Discrimination of Confusions. In *Interspeech 2021*, pages 4483–4487. ISCA, August 2021.

[68] 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, Barcelona, Spain, May 2020. IEEE.

[69] Shuhua Liu, Mengyu Zhang, Ming Fang, Jianwei Zhao, Kun Hou, and Chih-Cheng Hung. Speech emotion recognition based on transfer learning from the facenet framework. *The Journal of the Acoustical Society of America*, 149(2):1338–1345, 2021.

[70] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[71] Zhen-Tao Liu, Bao-Han Wu, Dan-Yun Li, Peng Xiao, and Jun-Wei Mao. Speech emotion recognition based on selective interpolation synthetic minority over-sampling technique in small sample environment. *Sensors*, 20(8):2297, 2020.

[72] Zheng Liu, Xin Kang, and Fuji Ren. Improving speech emotion recognition by fusing pre-trained and acoustic features using transformer and bilstm. In *Intelligent Information Processing XI: 12th IFIP TC 12 International Conference, IIP 2022, Qingdao, China, May 27–30, 2022, Proceedings*, pages 348–357. Springer, 2022.

[73] Zheng Liu, Fuji Ren, and Xin Kang. Research on the effect of different speech segment lengths on speech emotion recognition based on lstm. In *Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering*, pages 491–499. WCSE, 2019.

[74] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[75] Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2017.

[76] Nurul Lubis, Randy Gomez, Sakriani Sakti, Keisuke Nakamura, Koichiro Yoshino, Satoshi Nakamura, and Kazuhiro Nakadai. Construction of japanese audio-visual emotion database and its application in emotion recognition. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (L-REC'16)*, pages 2180–2184, 2016.

[77] Hui Luo and Jiqing Han. Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2047–2060, 2020.

[78] Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman. A novel adaptive fractional deep belief networks for speaker emotion recognition. *Alexandria Engineering Journal*, 56(4):485–497, 2017.

[79] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8. IEEE, 2006.

[80] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. Yaafe, an easy to use and efficient audio feature extraction software. In *ISMIR*, volume 2010, pages 441–446. Citeseer, 2010.

[81] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, 2015.

[82] A. Milton, S. Sharmy Roy, and S. Tamil Selvi. SVM Scheme for Speech Emotion Recognition using MFCC Feature. *International Journal of Computer Applications*, 69(9):34–39, May 2013.

[83] Jon D Morris. Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research*, 35(6):63–68, 1995.

[84] Maria Moutti, Sofia Eleftheriou, Panagiotis Koromilas, and Theodoros Giannakopoulos. A dataset for speech emotion recognition in greek theatrical plays. *arXiv preprint arXiv:2203.15568*, 2022.

[85] Mumtaz Begum Mustafa, Mansoor AM Yusoof, Zuraidah M Don, and Mehdi Malekzadeh. Speech emotion recognition research: an analysis of research focus. *International Journal of Speech Technology*, 21(1):137–156, 2018.

[86] Daniel Neiberg and Kjell Elenius. Automatic recognition of anger in spontaneous speech. In *Interspeech 2008*, pages 2755–2758. ISCA, September 2008.

[87] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[88] Andrew Ortony and Terence J Turner. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.

[89] Sang-Min Park and Young-Gab Kim. A metaverse: taxonomy, components, applications, and open challenges. *IEEE access*, 10:4209–4251, 2022.

[90] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[91] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[92] Banothu Rambabu, Kishore Kumar Botsa, Gangamohan Paidi, and Suryakanth V Gangashetty. Iiit-h temd semi-natural emotional speech database from professional actors and non-actors. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1538–1545, 2020.

[93] Fuji Ren. Affective information processing and recognizing human emotion. *Electronic notes in theoretical computer science*, 225:39–50, 2009.

[94] Fuji Ren and Yanwei Bao. A review on human-computer interaction and intelligent robots. *International Journal of Information Technology & Decision Making*, 19(01):5–47, 2020.

[95] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013.

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

[97] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, volume 2, page 6. Citeseer, 2000.

[98] Christopher Schröder and Andreas Niekler. A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv:2008.07267*, 2020.

[99] Burr Settles. Active learning literature survey. 2009.

[100] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992.

[101] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3285–3292. IEEE, 2019.

[102] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[103] Youddha Beer Singh and Shivani Goel. A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, 2022.

[104] Peng Song. Transfer linear subspace learning for cross-corpus speech emotion recognition. *IEEE Trans. Affect. Comput.*, 10(2):265–275, 2019.

[105] Peng Song and Wenming Zheng. Feature selection based transfer subspace learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, 11(3):373–382, 2018.

[106] Linhui Sun, Sheng Fu, and Fu Wang. Decision tree svm model with fisher feature selection for speech emotion recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2019(1):1–14, 2019.

[107] Linhui Sun, Qiu Li, Sheng Fu, and Pingan Li. Speech emotion recognition based on genetic algorithm–decision tree fusion of deep and acoustic features. *ETRI Journal*, 2022.

[108] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

[109] Monorama Swain, Subhasmita Sahoo, Aurobinda Routray, P. Kabisatpathy, and Jogendra N. Kundu. Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition. *International Journal of Speech Technology*, 18(3):387–393, September 2015.

[110] Einari Vaaras, Manu Airaksinen, and Okko Räsänen. Analysis of self-supervised learning and dimensionality reduction methods in clustering-based active learning for speech emotion recognition. *arXiv e-prints*, pages arXiv–2206, 2022.

[111] Nikolaos Vryzas, Rigas Kotsakis, Aikaterini Liatsou, Charalampos A Dimoulas, and George Kalliris. Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society*, 66(6):457–467, 2018.

[112] Wei Wang, Paul A Watters, Xinyi Cao, Lingjie Shen, and Bo Li. Significance of phonological features in speech emotion recognition. *International Journal of Speech Technology*, 23(3):633–642, 2020.

[113] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022.

[114] Yuntao Wang, Zhou Su, Ning Zhang, Rui Xing, Dongxiao Liu, Tom H Luan, and Xuemin Shen. A survey on metaverse: Fundamentals, security, and privacy. *IEEE Communications Surveys & Tutorials*, 2022.

[115] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814, 2021.

[116] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[117] Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Bjorn Schuller. Speech Emotion Classification Using Attention-Based LSTM. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1675–1685, November 2019.

[118] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. Learning alignment for multimodal emotion recognition from speech. *arXiv preprint arXiv:1909.05645*, 2019.

[119] Yifan Yan, Shengjun Huang, Shaoyi Chen, Meng Liao, and Jin Xu. Active learning with query generation for cost-effective text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6583–6590, 2020.

[120] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 399–407. Springer, 2017.

[121] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. Speech emotion recognition using spectrogram & phoneme embedding. In *Interspeech*, volume 2018, pages 3688–3692, 2018.

[122] Tang Yingpeng, Li Guoxiang, and Huang Shengjun. ALiPy: Active learning in python. Technical report, Nanjing University of Aeronautics and Astronautics, January 2019. available as arXiv preprint `https://arxiv.org/abs/1901.03802`.

[123] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W Picard. Driver emotion recognition for intelligent vehicles: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–30, 2020.

[124] JTFLM Zhang and Huibin Jia. Design of speech corpus for mandarin text to speech. In *The Blizzard Challenge 2008 workshop*, 2008.

[125] Shiqing Zhang, Aihua Chen, Wenping Guo, Yueli Cui, Xiaoming Zhao, and Limei Liu. Learning Deep Binaural Representations With Deep Convolutional Neural Networks for Spontaneous Speech Emotion Recognition. *IEEE Access*, 8:23496–23505, 2020.

[126] 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, 2017.

[127] 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, April 2022.

[128] Weijian Zhang and Peng Song. Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:307–318, 2019.

[129] Weijian Zhang, Peng Song, Dongliang Chen, Chao Sheng, and Wenjing Zhang. Cross-corpus speech emotion recognition based on joint transfer subspace learning and regression. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[130] Ziping Zhao, Yu Zheng, Zixing Zhang, Haishuai Wang, Yiqin Zhao, and Chao Li. Exploring Spatio-Temporal Representations by Integrating Attention-based Bidirectional-LSTM-RNNs and FCNs for Speech Emotion Recognition. In *Interspeech 2018*, pages 272–276. ISCA, September 2018.

[131] Chunjun Zheng, Chunli Wang, and Ning Jia. An ensemble model for multi-level speech emotion recognition. *Applied Sciences*, 10(1):205, 2020.

[132] Fuji Ren Zheng Liu and Xin Kang. Research on the effect of different speech segment lengths on speech emotion recognition based on lstm. In *9th International Workshop on Computer Science and Engineering*, pages 491–499, 2019.