Siddique Latif
Queensland University of Technology, Brisbane City, QLD, 4000, AUSTRALIA

Muhammad Usama
National University of Computer and Emerging Science, Faisalabad, 38000, PAKISTAN

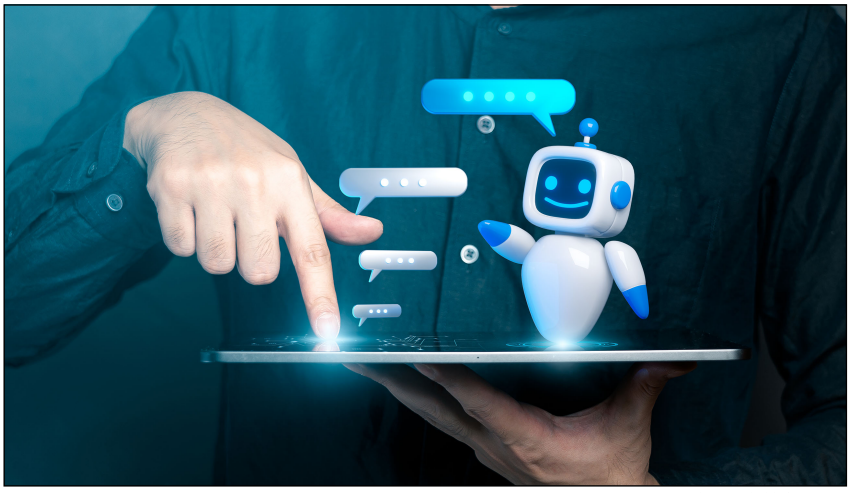
Muhammad Ibrahim Malik
EmulationAI, Karachi, 75530, PAKISTAN

Björn W. Schuller
Imperial College London, SW7 2AZ, London, U.K. and also University of Augsburg, 86159, Augsburg, GERMANY

Can Large Language Models Aid in Annotating Speech Emotional Data? Uncovering New Frontiers

Abstract

Despite recent advancements in speech emotion recognition (SER) models, state-of-the-art deep learning (DL) approaches face the challenge of the limited availability of annotated data. The advent of large language models (LLMs) has revolutionised our understanding of natural language, introducing emergent properties that broaden comprehension in language, speech, and vision. This paper explores the potential of LLMs, such as ChatGPT, to annotate abundant speech data with the goal of advancing the state-of-the-art in SER. Specifically, it proposes a method that integrates audio representations and gender information with textual prompts to enhance the annotation process using LLMs. Our evaluation encompasses single-shot and few-shots scenarios, revealing performance variability in SER. Notably, this work achieves improved results through data augmentation by incorporating ChatGPT-annotated samples into the existing datasets. Our work also uncovers new frontiers in speech emotion classification, highlighting the increasing significance of LLMs in this field moving forward.



©SHUTTERSTOCK/PINGINGZ

I. Introduction

The rapid growth in Natural Language Processing (NLP) has led to the development of advanced conversational tools, often called large language models (LLM) [1]. These tools are capable of assisting users with various language-related tasks, such as question answering, semantic parsing, proverbs and grammar correction, arithmetic, code completion, general knowledge, reading comprehensions, summarisation, logical inference, common sense reasoning, pattern recognition, translation, dialogues, joke explanation, educational content, and language understanding [1]. LLMs are trained on an enormous amount of general-purpose data and human-feedback-enabled reinforcement learning. A new

Digital Object Identifier 10.1109/MCI.2024.3504833
Date of current version: 9 January 2025

This article was recommended for publication by Associate Editor Jen-Wei Huang. Corresponding author: Siddique Latif (e-mail: siddique.latif@qut.edu.au).

1556-603X

field of study called "Foundational Models" has emerged from these LLMs, highlighting the interest of the academic community and computing industry [2]. The foundational models have demonstrated the ability to perform tasks for which they were not explicitly trained. This ability, known as emergence, is considered an early spark of artificial general intelligence (AGI) [3]. The emergence properties of the foundational models have sparked a wide range of testing of these models for various tasks, such as sentiment analysis, critical thinking skills, low-resource language learning and translation, sarcasm and joke understanding, classification, and other affective computing challenges.

Speech emotion recognition (SER) is a fundamental problem in affective computing. The need for SER has evolved rapidly with the rapid integration of modern technologies in every aspect of our lives. SER systems are designed to understand the wide range of human emotions from the given input data (audio, video, text, or physiological signal) using traditional and modern machine learning (ML) techniques. However, the limited availability of larger annotated data remains a challenging aspect for SER systems, thereby prompting the need for further investigation and exploration of new methods.

The use of crowd-sourced and expert intelligence for data annotation is a common practice. The annotated data serves as the ground truth for ML models to learn and generate predictions. This annotation policy is mostly opted in computational social science (sentiment analysis, bot detection, stance detection, emotion classification, etc.), human emotion understanding, and image classification [4]. However, these strategies are prone to a variety of biases, ranging from human biases to situational biases. These annotation techniques also necessitate a big pool of human annotators, clear and straightforward annotator instructions, and a verification rationale that is not always available or dependable [5]. Although there are a few unsupervised techniques for data annotations, these techniques necessitate a high sample size of the data; unfortunately, the generated annotations do not embed the context [6].

Annotating speech emotion data is a doubly challenging process. The annotators listen to a speech recording and assign an annotation to a data sample using the pre-defined criteria. Human emotions are highly context-dependent, and annotating emotions based on a brief recording in a specific controlled situation may limit the accuracy of the annotations. While the state-of-the-art human-annotated emotion classification is robust, the generalizability of the learning to unseen data with slightly different circumstances can hinder the effectiveness of the SER system. The recent availability of several LLMs (ChatGPT, Google Bard, etc.) has unearthed the possibility of replacing or assisting human annotators in natural language processing tasks. LLMs are trained on enormous text corpora, allowing them to learn and grasp complicated language patterns. Their emergence property [7] makes them well-suited for data annotations and various studies (e. g., [8], [9]) explored LLMs for annotations of various NLP tasks. However, none of the studies explores them to annotate speech emotion data based on the transcripts.

To address this limitation, this paper introduces a composite approach that enriches the annotation process by combining audio features, gender information, and transcripts. The paper makes the following contributions:

1) It evaluates the effectiveness of LLMs like ChatGPT in annotating emotional speech data for SER, highlighting through experiments that annotations based solely on text content do not generalise well to emotional speech data due to the absence of audio context.
2) To address this challenge, our study introduces a novel approach for speech representation using a VQ-VAE. This method effectively transforms speech into a discrete feature representation of fixed length, allowing for a more comprehensive capture of audio context, which transcends the limitations of using textual transcripts alone. This innovative encoding captures the nuances of emotional speech and improves the generalisation of annotations.
3) Building on these advancements, the research pioneers the application of LLMs for annotating speech emotion data, with a focus on improving the classification of speech emotions with data augmentation. It also includes a comparative analysis of LLM-based data annotations using publicly available datasets such as IEMOCAP, MSP-IMPROV, MELD, and RECOLA, demonstrating the practical implications and effectiveness of these contributions.

The subsequent section presents a concise literature review on the use of LLMs for data annotation, emphasising the disparity between conventional annotations and those made with LLMs. The methodology employed in this study is described in Section III, while Section IV outlines the experimental setup. The experiments and their corresponding results are presented in Section V. Finally, Section VI concludes the paper, highlighting the potential for extending this work.

II. Related Work

Speech Emotion Recognition (SER) is a rapidly evolving field, as evidenced by the diverse range of techniques being employed, from data augmentation [10], [11] to sophisticated transformer architectures [12], [13], [14], [15], [16]. The development of SER systems critically hinges on the availability of accurately labelled data, essential for training and evaluating these advanced models. The quality and correctness of speech emotion data annotations are crucial, highlighting the imperative need for rigorous and precise data annotation methods. A diverse array of annotation approaches is available, comprising expert human annotations, bulk annotations, semi-supervised techniques, and crowdsourced methods, each possessing

its unique strengths and weaknesses [17]. Expert human annotators can provide high-quality data but face issues like bias, costliness, and scalability. While bulk annotation methods offer the advantages of speed and cost-effectiveness, they may come at the expense of annotation quality. Semi-supervised methods merge the benefits of expert and bulk annotations but are complex and face robustness issues. Crowdsourcing annotation methods, while efficient for handling large datasets, encounter challenges in maintaining consistent quality. These varying strategies, with their inherent trade-offs, highlight the complexity of obtaining reliable and accurate data for training SER systems.

Recently, a few studies have investigated the efficacy of LLMs (i. e., ChatGPT) for data annotations. The goal of these experiments is to explore the potential of ChatGPT for data annotation and to find out whether ChatGPT can achieve full emergence in downstream tasks such as classification. Zhu et al. [8] explored the ability of ChatGPT to reproduce the human-generated annotations for five seminal computational social science datasets. The datasets include stance detection (two datasets), hate speech detection, sentiment analysis, and bot detection. Their results indicate that ChatGPT is capable of annotating the data, but its performance varies depending on the nature of the tasks, the version of ChatGPT, and the prompts. The average re-annotation performance is 60.9% across all five datasets. For the sentiment analysis task, the accuracy of ChatGPT re-annotating the tweets is reported at 64.9%, and for the hate speech task, the ChatGPT performance has gone down to 57.1%.

Fact-checking is a well-known way to deal with the misinformation epidemic in computational social science. Hose et al. [18] evaluated the ability of LLMs, specifically ChatGPT, to assist fact-checkers in expediting misinformation detection. The authors used ChatGPT as a zero-shot classifier to re-annotate 12,784 human-annotated ("true claim", "false claim") fact-checked statements. ChatGPT successfully re-annotates 72.0% of the statements. The study further suggests that ChatGPT performs well on recent fact-checked statements with "true claim" annotations. Despite the reasonable performance of ChatGPT on fact-checking, it is hard to suggest that it will replace human fact-checkers anytime soon. Yang et al. [19] explored the rating of news outlet credibility by using ChatGPT to re-annotate news domains as a binary task. ChatGPT achieves reasonable performance, re-annotating 7,523 domains with a Spearman correlation coefficient of $\rho = 0.54$. Tornberg [20] used ChatGPT-4 as a zero-shot classifier to re-annotate 500 political tweets, finding that ChatGPT-4 surpasses experts and crowd annotators in accuracy, reliability, and bias. Similarly, Gilardi et al. [21] reported that ChatGPT as a zero-shot classifier, outperforms crowd-sourced text annotations across five content moderation tasks. The use of LLMs, particularly ChatGPT, spans various computational social science tasks, including election opinion mining [22], intent classification [23], genre identification [24], stance detection [25], and sentiment analysis [26], with additional studies exploring its application in diverse domains [27], [28], [29], [30].

Amin et al. [31] assessed ChatGPT's performance in three key NLP classification tasks within affective computing: personality recognition, suicide tendency prediction, and sentiment analysis. They found that ChatGPT significantly outperforms Word2Vec models [32] in noisy data environments, matches the performance of Bag-of-Words (BoW) and Word2Vec models in clean data, and surpasses a specific-affective-task-trained RoBERTa model [33]. Specifically, ChatGPT achieves an 85.5% unweighted average recall in sentiment analysis, nearly 20% higher than BoW and Word2Vec models. In suicide tendency prediction, ChatGPT equals the performance of Word2Vec and BoW models, each with around 91.0% recall, while RoBERTa leads with 97.4%. For personality recognition, RoBERTa tops the charts with a 62.3% recall, whereas ChatGPT lags at 54.0%, slightly behind Word2Vec and BoW models. Wang et al. [34] advocated GPT-3 as a cost-effective tool for annotating data in natural language tasks, showing it can cut annotation costs by 50% to 96%. They highlighted GPT-3âs lower reliability compared to humans in sensitive scenarios. Further analysis of ChatGPT's performance in NLP tasks is in [35]. Huang et al. [9] found that laypeople prefer ChatGPTâs explanations to plain annotations, with an 80% agreement rate with human judgements.

In contrast to the aforementioned studies, our research explores the untapped potential of LLMs in annotating emotions in speech data. This paper presents a novel approach that incorporates audio context into LLMs to improve the precision of SER annotations. To our knowledge, no prior research has investigated the utilisation of LLMs for annotating speech emotion data.

III. Methodology

In our exploration of emotional data annotation, a series of experiments is conducted. Initially, samples are annotated using only textual data. The approach is then expanded to include audio features, such as the average energy and pitch of each utterance, along with gender information, to enhance the annotations. These audio contexts along with textual data are passed to ChatGPT. Additionally, the use of VQ-VAE is proposed to generate a 64-dimensional discrete representation of the audio data, which is also provided to ChatGPT as the audio context. For speech-emotion classification, a convolutional neural network (CNN) and a bi-directional Long-Short Term Memory (BLSTM)-based classifier are trained. The subsequent section details the proposed methodology. The entire process is illustrated in Figure 1.

A. VQ-VAE for Speech Code Generation

In this work, a VQ-VAE [36] is used to learn a discrete representation from the speech data. In contrast to traditional

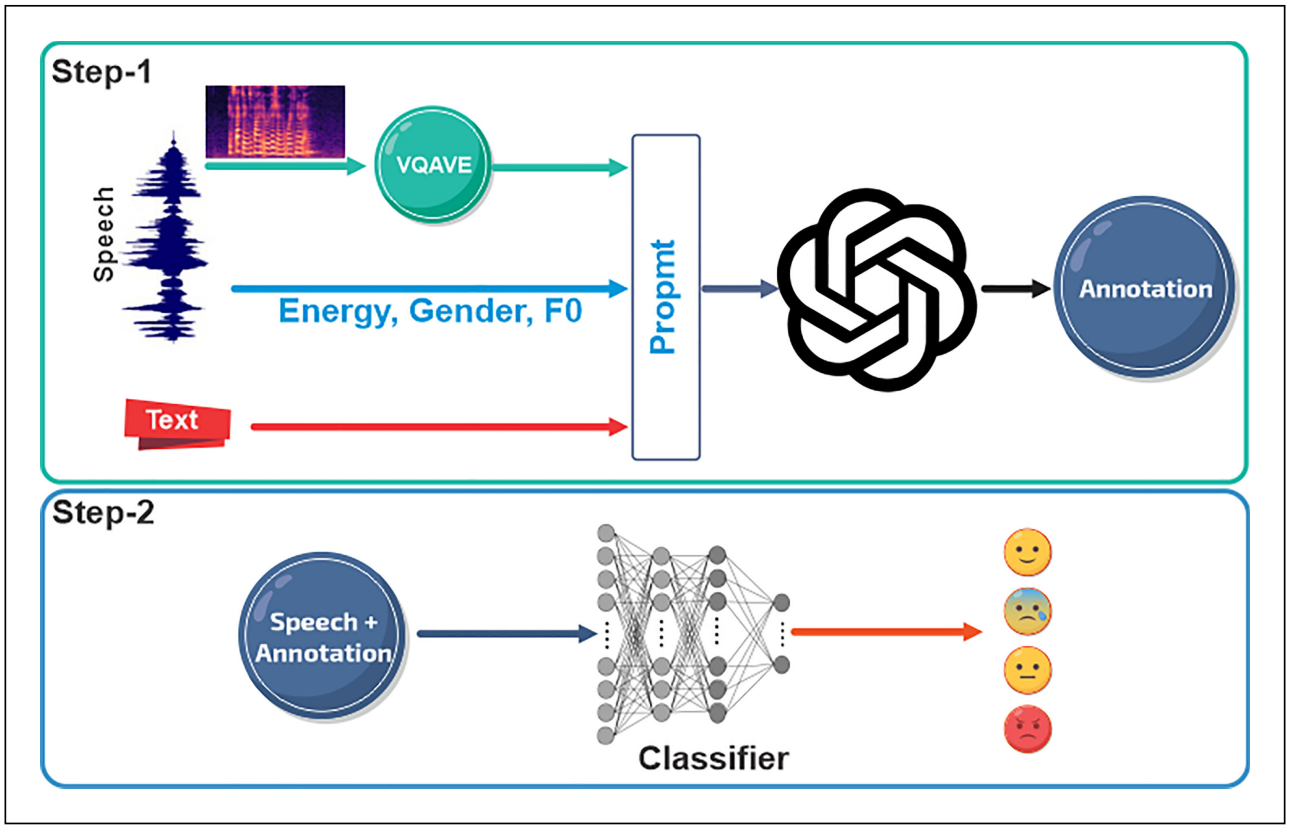


FIGURE 1 The process begins with encoding the speech using a VQ-VAE (as shown in Figure 2). Speech features such as energy, gender, and F0 are then extracted. These encoded and extracted features are sent to ChatGPT along with the speech text, prompting it to generate annotations for the speech text. The audio with the ChatGPT-based annotation is then used to train a speech emotion classifier. Additionally, the annotated text is also used for data augmentation, which involves generating synthetic speech samples with varying emotional expressions.

VAEs where a continuous discrete space is present, in VQ-VAE, the latent space is expressed as a set of discrete latent codes, and the prior is learned instead of being fixed. As depicted in Figure 2, the model consists of three main components: the encoder, the vector quantiser, and the decoder.

The encoder takes the input in the form of Mel-spectrograms and passes it through a series of convolutional layers having a shape of (n, h, w, d) where n is

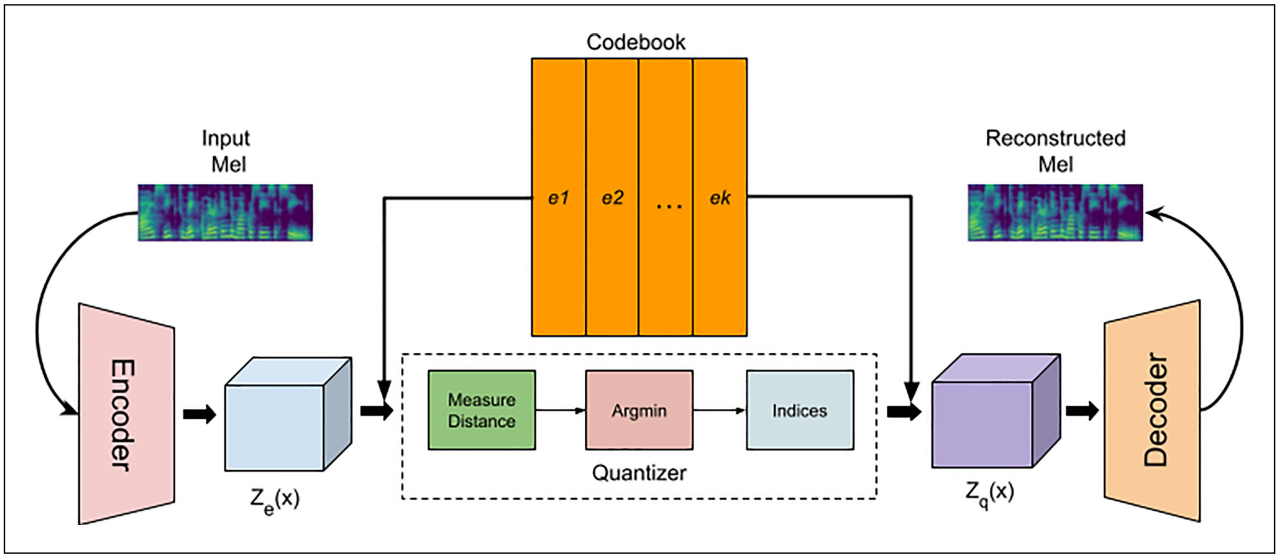


FIGURE 2 Model Diagram of the VQ-VAE. The input Mel-spectrogram is converted into the latent space as a series of vectors, each with a unique index of its own. These vectors are then processed to recreate the original Mel-spectrogram as output.

the batch size, h is the height, w is the width, and d represents the total number of filters after convolutions. The encoded output is denoted as z_e. The vector quantiser component consists of an embedding space that contains a total of k vectors, each has d dimensions. The main goal of this component is to output a series of embedding vectors denoted as z_q. To accomplish this, z_e is first reshaped into the form of $(n * h * w, d)$, and the distance is calculated for each of these vectors with the vectors in the embedding dictionary. For each of the $n * h * w$ vectors, the closest vector from the embedding space is found among the k vectors, and the closest vector from the embedding space is indexed for each $n * h * w$ vector. The discrete indices of each vector in the embedding space are referred to as codes, and a unique series of codes is obtained for each input to the model. The selected vectors are then reshaped back to match the shape of z_e. Finally, the reshaped vector embeddings are passed through a series of transpose convolutions to reconstruct the original input Mel-spectrogram. One problem with this approach is that the process of selecting vectors is not differentiable. To tackle this problem, the authors of [36] simply copied the gradients from z_q to z_e.

The total loss is composed of three loss elements: the reconstruction loss, the code book loss, and the commitment loss. The reconstruction loss is responsible for optimising the encoder and decoder and is represented by:

$$\text{Reconstruction Loss} = -log(p(x|z_q)). \tag{1}$$

In this work, a code book loss is employed, which causes the vector embeddings to be moved closer to the encoder output z_e.

$$\text{Code Book Loss} = ||sg[z_e(x)] - e||^2, \tag{2}$$

where sg is the stop gradient operator, this essentially freezes all gradient flows. e are the vector embeddings and x is the input to the encoder. Finally, a commitment loss is added to ensure that the encoder is committed to an embedding.

$$\text{Commitment Loss} = \beta||z_e(x) - sg[e]||^2, \tag{3}$$

where β is a hyperparameter that determines the weight assigned to the commitment loss. Overall, this work trains the VQ-VAE model to represent the audio in the form of a discrete list of integers or "codes." These audio representations, along with the transcriptions, are fed to ChatGPT for annotation. The subsequent section will delve into the details of the annotation procedure.

B. Emotion Label Annotation Using LLMs

The experiments began by annotating the training data of IEMOCAP, passing textual transcripts to ChatGPT and annotating the data in both zero-shot and few-shot settings. In the few-shot scenario, 10 samples are randomly selected from the training data and provided to ChatGPT as context. The classifier is then trained using the training samples annotated by ChatGPT, and unweighted average recall (UAR) is computed. The procedure is repeated by passing audio features along with textual information. Initially, average pitch and energy for a given utterance are used for re-annotation in both zero-shot and few-shots settings, and classification UAR is measured using a CNN-BLSTM-based classifier. Since the female voice typically has a higher pitch and energy, gender information is also provided for data annotation. Subsequently, an audio representation by VQ-VAE (Section III-A) is introduced and passed to ChatGPT as an audio context. The OpenAI API with the "ChatGPT pro" version is utilised to annotate the data. In this approach, multiple prompts are meticulously designed and curated for annotating the data, leveraging ChatGPT for the annotation process. The classifier is trained on the text-based annotated dataset, and UAR is computed, serving as a benchmark for evaluating classification performance. To enhance this benchmark, additional experiments are conducted, exploring various prompts to improve the classification results beyond the established performance level.

C. Speech Emotion Classifier

In this research, CNN-BLSTM-based classifiers are implemented due to their popularity in SER research. It has been observed that the performance of BLSTM can be enhanced by providing it with a robust emotional representation. Therefore, a CNN is utilised as an emotional feature extractor from the given input data [37]. A CNN layer functions as a data-driven filter bank and is capable of modelling emotionally salient features. These emotional features are then passed to the BLSTM layer to learn contextual information. Since emotions in speech are primarily contained in the temporal dimension, the BLSTM layer is crucial for modelling these temporal relationships [37]. The outputs from the BLSTM are subsequently passed to an attention layer to aggregate the emotionally salient attributes distributed over the given utterance. For a given output sequence h_i, utterance level salient attributes are aggregated as follows:

$$R_{\text{attentive}} = \sum_i \alpha_i h_i, \tag{4}$$

where α_i represents the attention weights that can be computed as follows:

$$\alpha_i = \frac{\exp W^T h_i}{\sum_j \exp W^T h_j}, \tag{5}$$

where W is a trainable parameter. The attentive representation $R_{\text{attentive}}$ computed by the attention layer is passed to the fully connected layer for emotion classification. Overall, the classifier is jointly empowered by the CNN layers to capture an abstract representation, the BLSTM layer for context capturing, the attention layer for aggregating emotionally salient attributes, and the fully connected layer for emotion classification.

TABLE I **The mapping rules for IEMOCAP and RECOLA are as follows: The elements listed inside the brackets are included, while the elements listed inside the parentheses are not included.**

CORPUS	LOW/NEGATIVE	HIGH/POSITIVE
IEMOCAP	[1, 2.5]	(2.5, 5]
RECOLA	[-1, 0]	(0, 1]

IV. Experimental Setup

A. Datasets

To assess the effectiveness of the proposed approach, the study employs four datasets including IEMOCAP, MSP-IMPROV, MELD, and RECOLA, which are commonly used in speech emotion classification research [38], [39]. The details of these datasets are presented in the following subsections.

1) IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is a multimodal database that contains 12 hours of recorded data [40]. The recordings are captured during dyadic interactions between five male and five female speakers. The interactions are almost five minutes long and are segregated into smaller utterances based on sentences, where each utterance is then assigned a label according to the emotion. Overall, the dataset contains nine different emotions. To be consistent with previous studies, four emotions including sad (1084 samples), happy (1636 samples), angry (1103 samples), and neutral (1708 samples) are used.

2) MSP-IMPROV

This corpus is a multimodal emotional database recorded from 12 actors performing dyadic interactions [41], similar to IEMOCAP [40]. The utterances in MSP-IMPROV are grouped into six sessions, and each session has recordings of one male and one female actor. The emotional labels are collected through perceptual evaluations using crowdsourcing [42]. The utterances in this corpus are annotated in four categorical emotions: angry, happy, neutral, and sad. To be consistent with previous studies [37], [43], this study uses all utterances with four emotions: anger (792 samples), sad (885 samples), neutral (3477 samples), and happy (2644 samples).

3) MELD

Multimodal EmotionLines Dataset [44] or MELD contains over 1400 dialogues and 13000 utterances and multiple speakers from the popular TV series Friends. The utterances are labelled with one of seven emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. To maintain consistency with the other datasets, this study includes four emotions: sadness (1002 samples), neutral (6436 samples), joy, and anger (1607 samples).

4) RECOLA

The RECOLA dataset, as introduced in Ringeval et al. [45], is a French multimodal corpus featuring spontaneous, collaborative, and emotional interactions. It includes contributions from 46 individuals (27 females and 19 males). In this work, the publicly accessible segment of RECOLA, comprising 1,308 utterances from 23 speakers, is utilised. The dataset is annotated with continuous labels for arousal and valence, spanning a range from -1 to 1. In our research, RECOLA is used for cross-corpus language SER, with a focus on the binary classification of arousal (low/high) and valence (negative/positive), in alignment with the approach in [46], [47]. Table I outlines the conversion of the original annotations of RECOLA and IEMOCAP into a binary format for analysis.

B. Speech Features

A consistent sampling rate of 16 kHz is used for utterances across all datasets, both for extracting the audio features and for converting the audio into Mel-spectrograms. The Mel-spectrograms are computed using a short-time Fourier transform of size 1024, a hop size of 256, and a window size of 1024. A total of 80 Mel-bands is specified for the output, with a cutoff frequency of 8 kHz. Each Mel-spectrogram is adjusted to a cutoff length of 256, resulting in a final shape of 80x256, where smaller samples are zero-padded. Finally, the Mel-spectrograms are normalised within the range of [-1, 1].

C. Hyperparameters

The VQ-VAE is trained using the following parameters: A batch size of 256 is chosen, and training is performed for a total of 1000 epochs with a learning rate of $1e^{-4}$. The convolution layers have a stride of 2 and a kernel size of 3, respectively. A total of 8192 token embeddings are selected, each with a dimensionality of 512. With this particular configuration, 64 codes are obtained for each given utterance. These codes, along with textual data, are passed to ChatGPT for the annotation of training data. Based on these annotations, the classifier is trained.

Our classifier comprises convolutional layers and a BLSTM-based classification network. Two CNN layers are used to generate high-level abstract feature representations. In line with previous studies [48], [49], a larger kernel size is used for the first convolutional layer and a smaller kernel size for the second layer in the proposed model. The representations learned by the CNN layers are passed to the BLSTM layer with 128 LSTM units for contextual representation learning. Following the BLSTM layer, an attention layer aggregates the emotional content spread across different parts of the utterance. The resulting attentive features are then fed into a dense layer with 128 hidden units to extract emotionally discriminative features for a softmax layer. The softmax layer uses the cross-entropy loss function to calculate posterior class probabilities, enabling the network to learn distinct features and achieve accurate emotion classification.

Prompts play a vital role in producing quality annotations. There is no standard way of designing a prompt for LLMs, and it requires a good amount of deliberation and trials to determine the best prompt for producing the annotations.

During the experiments, the Adam optimiser is used with its default parameters. The models are trained starting with a learning rate of 0.0001, and the validation accuracy is evaluated after each epoch. If the validation accuracy does not improve for five consecutive epochs, the learning rate is halved, and the model is reverted to the best-performing epoch. This process continues until the learning rate drops below 0.00001. Regarding the choice of the non-linear activation function, the rectified linear unit (ReLU) is selected due to its superior performance compared to leaky ReLU and hyperbolic tangent during the validation phase.

D. Prompt Design

Prompts play a vital role in producing quality annotations. There is no standard way of designing a prompt for LLMs, and it requires a good amount of deliberation and trials to determine the best prompt for producing the annotations. While there are a few guidelines provided in the literature about designing prompts, none of these guidelines directly address the annotation of speech emotion data. It is recommended to use as much context as possible with specific class labels and to also check for the explanations produced by the LLMs for a particular annotation. In this work, after careful consideration, different versions of the following prompt were used:

Task description: Given a text transcript of audio utterances, annotate the text with one of the following emotions: [happy, sad, angry, or neutral]. Use the additional context provided, including average energy and pitch of the utterance, gender information, and discrete audio codes, to make informed decisions on the annotation. The goal is to accurately label the emotions expressed in the audio.

Text transcript: (text transcript of spoken words)
Average energy: (numerical value)
Average Pitch: (numerical value)
Gender: (male or female)
Discrete audio codes: (numerical value)
Label set: [happy, sad, angry, neutral]
Output:
Label:
Explanation:

V. Experiments and Results

All experiments are conducted in a speaker-independent manner to ensure the generalisability of the findings. In this study, annotation is performed solely on the training data, and true labels are utilised for the testing sets in all experiments. Specifically, this research adopts an easily reproducible and widely used leave-one-speaker-out cross-validation scheme, as commonly employed in related literature [50], [51]. For cross-corpus SER, our methodology follows [51], [52], using IEMOCAP for training, while MSP-IMPROV is utilised for validation and testing. Each experiment is repeated 10 times, with the mean and standard deviation of the results calculated. The performance is presented in terms of the unweighted average recall rate (UAR), a widely accepted metric in the field that more accurately reflects classification accuracy across multiple emotion categories when the data is imbalanced.

A. Within Corpus Experiments

For the within-corpus experiments, the IEMOCAP data is selected, and the results are compared with the baseline UAR achieved using actual true labels. The classifier is trained in different settings: (1) true label settings, (2) zero-shot ChatGPT labels, and (3) few-shots ChatGPT labels. Initially, the CNN-BLSTM-based classifier is trained on true labels using the well-known leave-one-speaker-out scheme. In the subsequent experiments, the classifier is trained using the same scheme, but the training samples are annotated using ChatGPT according to the proposed approach. The second and third experiments are repeated using text only and text plus audio context. Results are presented in Figure 3, computed using the same testing data with actual true labels. Overall, results on data annotated using the few-shots approach show improved outcomes compared to the zero-shot scenario.

It is important to note that the emotion classification performance using training data annotated with only text is poor compared to the baseline. Here, baseline results represent when the classifier is trained using the original true annotations of IEMOCAP. This observation underscores the insufficiency of textual information alone to provide the necessary context for accurate annotation by ChatGPT. Consequently, additional context becomes essential to enable ChatGPT in effectively annotating the data. As previously found, for example, happy and angry voice samples often have high energy and pitch compared to a sad and neutral voice [53]. Building upon this insight, we incorporate the average energy and pitch values of a given utterance as additional contextual information for ChatGPT during the re-annotation process, both in zero-shot and few-shot settings. However, the performance improvement is not considerable, primarily due to the confounding factor of gender, as female voices typically exhibit higher pitch and energy compared to male voices [54]. To address this limitation, we extend the experiment by providing gender labels to ChatGPT, resulting in improved classification accuracy as illustrated in

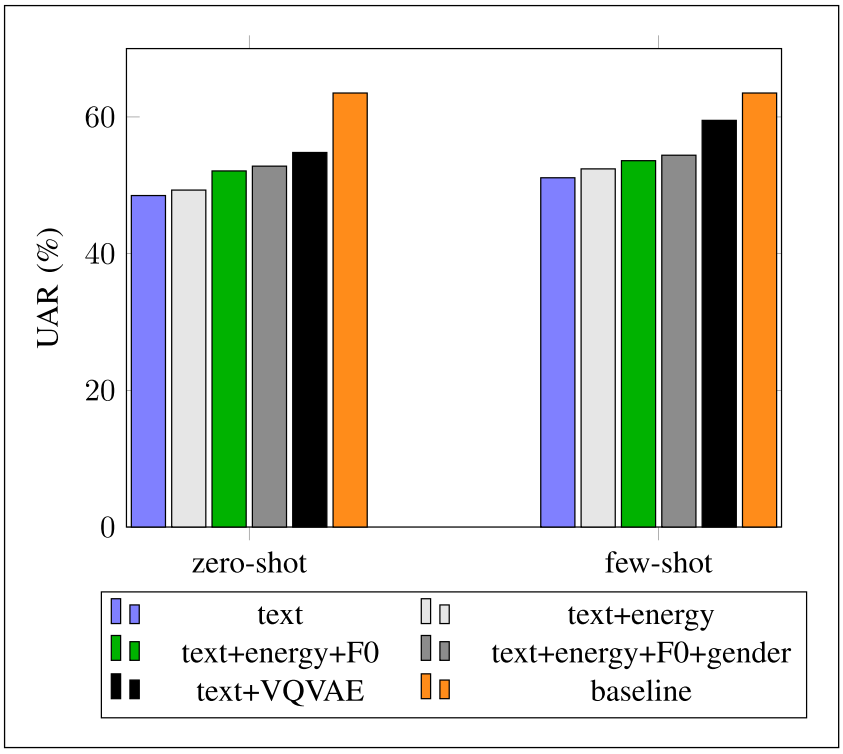


FIGURE 3 Comparing the classification performance (UAR %) using training data annotated by ChatGPT and original IEMOCAP labels. We use the same testing data with actual true labels in all the experiments.

Figure 3. In addition to average energy, pitch, and gender information, this paper further proposes the utilisation of audio patterns to provide enhanced audio context for annotation. To achieve this, a VQ-VAE model is used to encode the given utterance into discrete representations. These representations, along with the textual and other feature inputs, are employed in various experiments for annotation (refer to Figure 3). Notably, in the zero-shot scenario, no substantial improvements are observed. However, considerable improvements are achieved by incorporating the discrete codes generated by VQ-VAE, in conjunction with average energy, pitch, and gender information.

B. Cross-Corpus Evaluations

In this experiment, a cross-corpus analysis is performed to assess the generalisability of annotations performed using the proposed approach. Models are trained on IEMOCAP, and testing is carried out on MSP-IMPROV data. IEMOCAP is selected for training due to its more balanced data set, following previous studies [55], [56]. We randomly select 30.0 % of the MSP-IMPROV data for parameter tuning and 70.0 % of data as testing data. Results are reported using the few-shots annotation by ChatGPT, as it consistently demonstrated superior performance compared to the zero-shot setting (see Figure 3).

We compare our results with different studies in Table II. In [52], the authors used the CNN-LSTM model for cross-corpus evaluation. They showed that CNN-LSTM can learn emotional contexts and help achieve improved results for cross-corpus SER. In [55], the authors utilised the representations learnt from unlabelled data and fed it to an attention-based CNN classifier. They showed that the classifier's performance can be improved by augmenting the classifier with information from unlabelled data. We compare our results using the CNN-BLSTM-based classifier by using the IEMOCAP annotated by the ChatGPT model.

TABLE II Cross-corpus evaluation results using IEMOCAP and MSP-IMPROV datasets.

MODEL	UAR (%)
Attentive CNN [55]	**45.7**
CNN-BLSTM$_{(baseline)}$	45.4±0.83
text+energy+f0+gender	41.5±1.2
text+energy+f0+gender+VQ-VAE	42.4±0.9

This experiment demonstrates the generalisability of annotations performed by ChatGPT in cross-corpus settings. However, it is worth noting that our results do not surpass those of previous studies. In the subsequent experiment, the aim is to showcase the potential for enhancing the performance of SER using data annotations generated by ChatGPT, both within-corpus and cross-corpus settings.

To further explore cross-corpus SER, we extend our experimentation by training the classifier on the IEMOCAP dataset and subsequently testing it on the MELD dataset. In this process, the MSP-IMPROV dataset, annotated using ChatGPT, is utilised to augment the classifier. The approach is benchmarked against the findings of a recent study by Fu et al. [57], adhering to comparable experimental conditions and achieving superior results. In the referenced study [57], the authors employed adversarial training-based SER, training their model on IEMOCAP and evaluating it on MELD, attaining an unweighted accuracy (UA) of 50.97% in a single task learning framework. In contrast, our proposed methodology yielded a UA of 55.68%. This enhancement in performance not only underlines the effectiveness of our approach but also its generalisability to diverse datasets, such as those derived from TV shows, which differ from the IEMOCAP and MSP-IMPROV in terms of cross-corpus settings.

C. Cross-Language Evaluations

In this study, results for cross-language SER using the RECOLA data in dimensional emotion classification are presented. For dimensional emotion, the focus is on binary arousal and

TABLE III Dimensional cross-language SER results by UAR (%) using the IEMOCAP and RECOLA datasets.

MODEL	IEMOCAP (English) to RECOLA (French)		RECOLA (French) to IEMOCAP (English)	
	AROUSAL	VALENCE	AROUSAL	VALENCE
CNN [47]	59.2± 1.8	48.5± 1.5	60.7± 1.6	48.3± 2.0
ACNN [46]	59.3	49.1	61.2	47.5
CNN-LSTM$_{baseline}$	49.2±1.2	47.2± 1.3	52.2 ± 1.0	46.0± 1.4
CNN-LSTM$_{augmentation}$	**63.1±1.1**	**52.2± 1.4**	**64.2 ± 1.4**	**53.0± 1.1**

valence classification, employing a cross-language training and testing paradigm. Specifically, data from one language constitutes the training set while samples from the target language are reserved exclusively for testing. This mirrors the evaluation strategy employed in [47], [58], ensuring comparability. To demonstrate the effectiveness of the proposed approach, the classifier is initially trained with one-language data and tested on 50% of target language data. The remaining 50% of samples are annotated with the proposed approach and then used to augment the classifier. Results are compared to those presented in [47], [58] in Table III. In [58], an Attentive Convolutional Neural Network (ACNN) was utilised, achieving notable results through model fine-tuning on target language data. However, we are achieving better results without using any target labels. Similarly, better results are achieved compared to the CNN-based classifier used in [47]. This demonstrates that the proposed annotation technique helps achieve improved cross-language SER.

D. Augmentating the Training Data

In the previous two experiments, we demonstrate how new speech-emotional data can be annotated using a large language model like ChatGPT. However, the performance does not surpass the UAR achieved using actual labels. This experiment aims to address this limitation by showcasing the potential to improve SER performance through data augmentation using our proposed approach. For this, abundantly available audio data, such as from YouTube, can be annotated with our proposed approach. To validate this concept, the MELD dataset is selected, consisting of dialogue samples from the Friends TV series. We employ a few-shot approach, using samples from the IEMOCAP dataset for few-shots, and annotate the MELD data with four emotions: happy, anger, neutral, and sad. Both text and acoustic features are used for annotations, as this combination yields better performance. Results are presented in Figure 4, where the outcomes are compared with those of a CNN-BLSTM classifier using actual IEMOCAP labels and when data is augmented using the samples with ChatGPT annotations. This analysis provides insights into the effectiveness of data augmentation for enhancing the performance of the SER system.

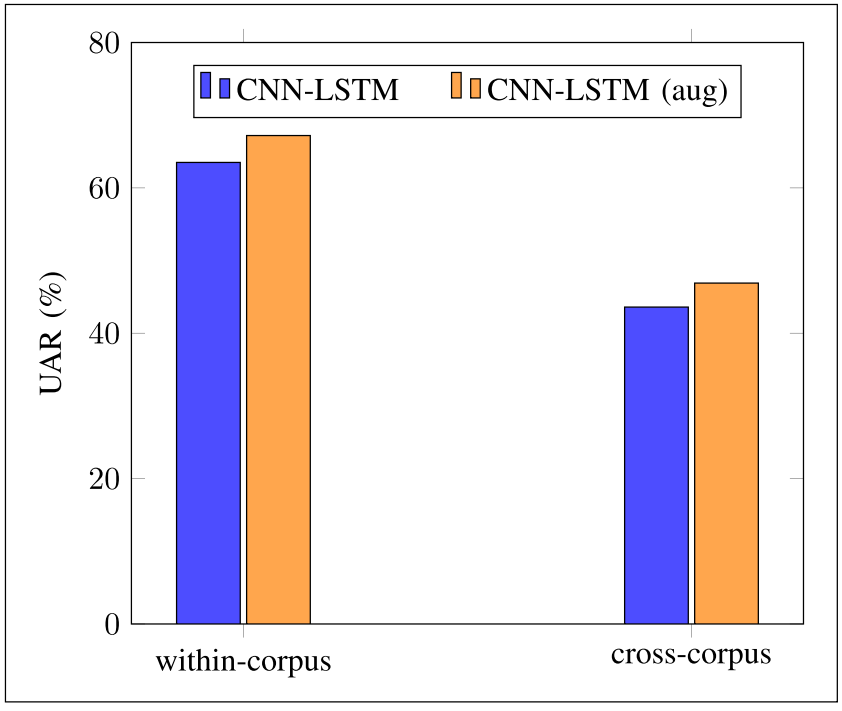


FIGURE 4 Comparing the baseline CNN-LSTM classifier performance (UAR %) with and without data augmentation. Here baseline represents when the model is trained on true labels and baseline+augmentation represents when the training data is augmented using MELD samples with annotations by ChatGPT.

Furthermore, this work provides a comprehensive comparison of the results with previous studies in both within-corpus and cross-corpus settings, as presented in Table IV. To ensure a fair comparison, the same evaluation strategies employed in these studies [59], [60] are adopted. In [59], the authors utilised DialogueRNN for speech emotion recognition using IEMOCAP data. Peng et al. [60] used an attention-based CNN network for emotion classification. We achieve better results compared to these studies by augmenting the classifier with additional data annotated by ChatGPT. One possible reason can be that these studies did not train the models with augmentation. In addition, the results are compared with [52], where the authors used different data augmentation techniques to augment the classifier and achieve improved results. In contrast, we use ChatGPT to annotate the

TABLE IV Comparison of results with previous studies.

MODEL	UAR (%)
within corpus	
DialogueRNN [59] (2019)	63.40
CNN-attention [60] (2021)	65.4
CNN-BLSTM (+ augmentation) (2022) [52]	65.1±1.8
Our work (+ augmentations) (2023)	**68.0± 1.4**
cross-corpus	
Cyclegan-DNN [51] (+ augmentations) (2019)	46.52±0.43
CNN-BLSTM (+ augmentations) [52] (2022)	46.2 ± 1.3
Our work (+ augmentations) (2023)	**48.1± 0.9**

TABLE V Unweighted accuracy (UA %) comparison with Wav2vec2 based studies.

STUDIES (YEAR)	UA (%)
Pepino et al. [62]	67.2 ± 0.7
Zou et al. [63]	69.80
Wav2Vec2 (+copypaste [11])	70.20±0.9
Wav2Vec2 (+proposed)	**72.61±0.7**

publicly available data and use it for the augmentation of the training set. Our proposed approach achieves considerably improved results compared to [52]. One possible reason is that we are adding new data in the classifiers' training set, however, authors in [52] employed perturbed versions of the same data, which can potentially lead to overfitting of the system. Similarly, considerably improved results for cross-corpus settings are achieved compared to previous studies [51], [52], where the authors augmented their classification models with either synthetic data or perturbed samples using audio-based data augmentation techniques like speed perturbation, SpecAugment, and mixup.

To further enhance our experiments, the fine-tuning of pre-trained models is conducted using our proposed data annotation approach. We use Wav2Vec2 [61] model and implement a simple classification head as used in [16]. We implement average pooling on the hidden states from the final transformer layer, followed by processing through a hidden and then an output layer. Results are compared with [62], [63] that also use Wav2Vec2 in SER and results are presented on the IEMOCAP dataset in Table V. We also fine-tune the wav2vec2 model on augmented data with some recently proposed augmentation techniques [11]. Consistent with our earlier findings, it is observed that the Wav2Vec2 model, when augmented with our proposed augmentation techniques, yields improved results. The observed improvement in performance is largely due to the enhanced data variety provided by our proposed augmentation methods. This contrasts with traditional "copypaste" techniques, which often augment the model with less varied data.

Overall, our results showcase the effectiveness of our approach in achieving superior performance compared to previous studies, both in within-corpus and cross-corpus settings. The utilisation of ChatGPT for data annotation and augmentation proves to be a promising strategy for enhancing SER systems.

E. ChatGPT vs. Human Annotation

This section presents an in-depth comparison of speech data annotations made by human experts and ChatGPT, examining both its free and paid versions. Our analysis identifies a significant performance gap between the unpaid and paid versions of ChatGPT in annotating emotional expressions in speech data. The paid version shows a high degree of consistency with human annotation, however, the free version displayed limitations in capturing the nuances of human emotions. The paid version of ChatGPT demonstrated the ability to reliably mirror human annotation performance. However, we also found significant disparities in scenarios involving complex emotional expressions or ambiguous contexts when relying solely on the unpaid version. These disparities underscore the limitations of LLMs in capturing the intricate aspects of human emotions conveyed through speech. To provide readers with a comprehensive understanding of these differences, we share some examples of annotations from both humans and ChatGPT.[1]

Despite these challenges, a significant finding from our research was the beneficial use of LLM-based annotations for augmenting existing datasets, thereby enhancing the overall performance of SER models. This improvement was consistently observed across various SER models included in our experiments. Integrating LLM annotations, even with their occasional discrepancies compared to human annotations, led to the creation of a more diverse and comprehensive dataset. This enriched dataset proved instrumental in strengthening the SER models' capability to interpret and understand a broader spectrum of emotional expressions and nuances in speech. Adopting this mixed-annotation approach, which combines human expertise with AI capabilities, appears to offer a promising pathway for advancing SER technology. It not only demonstrates the potential of LLMs like ChatGPT in SER but also emphasises their utility when complemented by human-annotated data.

F. Limitations

This section highlights the potential limitations of our work and, more broadly, the limitations of LLMs for data annotation. During the experiments, the following limitations are observed:

[1][Online]. Available: https://llm-annotations.github.io/chatgpt-ser/

- This work obtains promising results by augmenting the training data with samples annotated using ChatGPT. However, this approach proves ineffective when applied to corpora such as LibriSpeech [64], where the recordings lack emotional variations. Although we attempted to utilise LibriSpeech data (results are not shown here), the results are not as promising as those achieved with MELD.
- ChatGPT is known to be sensitive to prompt variability, which can lead to ambiguous and erroneous results if even slight changes are made to the prompt content. In order to address this issue, we suggest conducting experiments using different prompts to generate annotations (as presented in Section III-B). The inclusion of more context in the prompts has been shown to improve the quality of results. However, for SER annotation prompts, this can be particularly challenging due to the significant variability of human emotions within short time frames. This limitation stems from LLMs' reliance on training data.
- ChatGPT has not been trained particularly to annotate speech emotion data. While the emergent nature of ChatGPT has aided with annotation, relying exclusively on ChatGPT annotation is insufficient. Our research found that incorporating ChatGPT-based annotations alongside the training data leads to enhanced classification performance. Notably, when utilising multi-shot ChatGPT annotations instead of zero-shot annotations, we observe a substantial performance improvement.
- ChatGPT offers a significant cost reduction in data annotation. For instance, in our experiments, we are able to annotate IEMOCAP data examples using ChatGPT for approximately 30 USD, which is significantly lower than human annotations cost. However, it is paramount to note that the accuracy of ChatGPT-based annotations is not as good as human annotations because ChatGPT is not specifically trained for annotating speech emotion data. Therefore, it becomes a trade-off between cost and accuracy. Striking the right balance is crucial when utilising ChatGPT for data annotation to avoid potential inaccuracies in classification performance.

Despite the mentioned limitations, ChatGPT is found to be an invaluable tool for speech-emotion data annotation. It is believed that its capabilities will continue to evolve. Currently, generating annotations using ChatGPT and incorporating them to augment human-annotated data has demonstrated improved performance in speech emotion classification. This highlights the potential of ChatGPT as a valuable asset in advancing research in this field.

VI. Conclusions and Outlook

This paper conducted a comprehensive evaluation of ChatGPT's effectiveness in annotating speech emotion data. To the best of our knowledge, this study is the first of its kind to explore the capabilities of ChatGPT in the domain of speech emotion recognition. The results of our investigation have been encouraging, and we have discovered promising outcomes. Below are the key findings of our study:

- Based on our findings, we observed that text-based emotional annotations do not generalise effectively to speech data. To address this limitation, a novel approach was introduced that harnesses the audio context in annotating speech data. By incorporating the audio context, the performance of speech emotion recognition (SER) was successfully enhanced, yielding improved results compared to the text-based approach.
- We observed that the quality of annotations by ChatGPT considerably improved when using a few-shot approach compared to a zero-shot one. By incorporating a small number of annotated samples, we were able to achieve improved results in our evaluation.
- This paper introduced an effective technique to utilise large language models (LLMs) to augment the SER system with the annotated data by ChatGPT. The augmented system yielded improved results compared to the current state-of-the-art SER systems that utilise conventional augmentation techniques.

In our future works, the aim is to expand experimentation by applying the approach to new datasets and diverse contexts. This will allow for further validation of the effectiveness and generalizability of the proposed technique. Additionally, we plan to explore and compare the annotation abilities of different LLMs for speech emotion data, enabling us to gain insights into their respective strengths and weaknesses. We also intend to explore the use of LLMs for segment-wise annotation and continuous label annotation.

References

[1] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.

[2] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[3] J. Wei et al., "Emergent abilities of large language models," *IEEE Trans. Mach. Learn. Res.*, 2022.

[4] C. Cioffi-Revilla and C. Cioffi-Revilla, "Computation and social science," *Introduction Comput. Social Sci., Princ. Appl.*, pp. 35–102, 2017.

[5] P. Röttger, B. Vidgen, D. Hovy, and J. Pierrehumbert, "Two contrasting data annotation paradigms for subjective NLP tasks," in *Proc. 2022 Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2022, pp. 175–190.

[6] X. Liao and Z. Zhao, "Unsupervised approaches for textual semantic annotation, a survey," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1–45, 2019.

[7] C. Burns, H. Ye, D. Klein, and J. Steinhardt, "Discovering latent knowledge in language models without supervision," in *Proc. 11nth Int. Conf. Learn. Representations*, 2022.

[8] Y. Zhu, P. Zhang, E.-U. Haq, P. Hui, and G. Tyson, "Can ChatGPT reproduce human-generated labels? A study of social computing tasks," 2023, *arXiv:2304.10145*.

[9] F. Huang, H. Kwak, and J. An, "Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech," in *Proc. Companion Proc. ACM Web Conf. 2023*, 2023, pp. 294–297.

[10] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition,"

in *Proc. ICML Workshop Deep Learn. Audio, Speech, Lang.*, 2013, vol. 117, p. 21.
[11] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "CopyPaste: An augmentation method for speech emotion recognition," in *ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6324–6328.
[12] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *Proc. ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6319–6323.
[13] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "DST: Deformable speech transformer for emotion recognition," in *Proc. ICASSP 2023-2023 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
[14] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "SpeechFormer++: A hierarchical efficient framework for paralinguistic speech processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 775–788, 2023.
[15] Y. He, N. Minematsu, and D. Saito, "Multiple acoustic features speech emotion recognition using cross-attention transformer," in *Proc. ICASSP 2023-2023 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
[16] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, Sep. 2023.
[17] J. Pustejovsky and A. Stubbs, *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2012.
[18] E. Hoes, S. Altay, and J. Bermeo, "Using ChatGPT to fight misinformation: ChatGPT nails 72% of 12,000 verified claims," *PsyArXiv*, vol. 3, Apr., 2023.
[19] K.-C. Yang and F. Menczer, "Large language models can rate news outlet credibility," *arXiv:2304.00228*, 2023.
[20] P. Törnberg, "Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning," 2023, *arXiv:2304.06588*.
[21] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT outperforms crowd workers for text-annotation tasks," *Proc. Nat. Acad. Sci.*, vol. 120, no. 30, 2023, Art. no. e2305016120.
[22] T. Elmas and İ. Gül, "Opinion mining from youtube captions using ChatGPT: A case study of street interviews polling the 2023 Turkish elections," 2023, *arXiv:2304.03434*.
[23] J. Cegin, J. Simko, and P. Brusilovsky, "ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness," in *Proc. 2023 Conf. Empirical Methods Natural Lang. Process.*, 2023, pp. 1889–1905.
[24] T. Kuzman, I. Mozetic, and N. Ljubešic, "ChatGPT: Beginning of an end of manual linguistic data annotation? Use case of automatic genre identification," 2023, *arXiv:2303.03953*.
[25] M. Mets, A. Karjus, I. Ibrus, and M. Schich, "Automated stance detection in complex topics and small languages: The challenging case of immigration in polarizing news media," *PLoS One*, vol. 19, no. 4, 2024, Art. no. e0302380.
[26] Z. Wang, Q. Xie, Z. Ding, Y. Feng, and R. Xia, "Is ChatGPT a good sentiment analyzer? A preliminary study," 2023, *arXiv:2304.04339*.
[27] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can large language models transform computational social science?," *Comput. Linguistics*, vol. 50, no. 1, pp. 237–291, 2024.
[28] V. Veselovsky, M. H. Ribeiro, A. Arora, M. Josifoski, A. Anderson, and R. West, "Generating faithful synthetic data with large language models: A case study in computational social science," 2023, *arXiv:2305.15041*.
[29] Y. Mu et al., "Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science," in *Proc. Joint Int. Conf. Comput. Linguistics, Language Resources and Evaluation*, 2024, pp. 12074–12086.
[30] C. M. Rytting et al., "Towards coding social science datasets with language models," 2023, *arXiv:2306.02177*.
[31] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general AI? A first evaluation on chatGPT," *IEEE Intell. Syst.*, vol. 38, no. 2, pp. 15–23, Mar./Apr. 2023.
[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[33] Y. Liu et al., "Roberta: A robustly optimized BERT pretraining approach," *arXiv:1907.11692 364*.
[34] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want to reduce labeling cost? GPT-3 can help," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP 2021*, 2021, pp. 4195–4205.
[35] B. Guo et al., "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection," 2023, *arXiv:2301.07597*.
[36] A. Van Den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 30, 2017.
[37] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Proc. Interspeech*, 2019, pp. 3920–3924. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3252
[38] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Proc. Interspeech*, 2016, pp. 490–494. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1052
[39] Y. Kim and E. M. Provost, "Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions," in *Proc. 18th ACM Int. Conf. Multimodal Interact*, 2016, pp. 92–99.
[40] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, 2008, Art. no. 335.
[41] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan./Mar. 2017.
[42] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 374–388, Oct./Dec. 2016.
[43] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," in *Proc. Interspeech*, 2017, pp. 1098–1102.
[44] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2019, pp. 527–536. [Online]. Available: https://aclanthology.org/P19-1050
[45] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.
[46] M. Neumann et al., "Cross-lingual and multilingual speech emotion recognition on english and french," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5769–5773.
[47] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1912–1926, Jul.–Sep. 2023.
[48] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *Proc. ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 7405–7409.
[49] J. Gideon, M. G. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 1055–1068, Oct./Dec. 2021.
[50] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," in *Proc. Interspeech*, 2018, pp. 3107–3111.
[51] F. Bao, M. Neumann, and N. T. Vu, "Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 2828–2832.
[52] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Multitask learning from augmented auxiliary data for improving speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3164–3176, Oct./Dec. 2023.
[53] S. Yildirim et al., "An acoustic study of emotions expressed in speech," in *Proc. 8th Int. Conf. Spoken Lang. Process.*, 2004, pp. 2193–2196.
[54] P. J. Fraccaro et al., "Experimental evidence that women speak in a higher voice pitch to men they find attractive," *J. Evol. Psychol.*, vol. 9, no. 1, pp. 57–67, 2011.
[55] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 7390–7394.
[56] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," in *Proc. Interspeech*, 2018, pp. 3693–3697.
[57] C. Fu, C. Liu, C. Ishi, and H. Ishiguro, "An adversarial training based speech emotion classifier with isolated Gaussian regularization," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2361–2374, Jul./Sep. 2023.
[58] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Proc. Interspeech*, 2017, pp. 1263–1267.
[59] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 01, pp. 6818–6825.
[60] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale CNN and attention," in *Proc. ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 3020–3024.
[61] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 12449–12460.
[62] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*, 2021, pp. 3400–3404.
[63] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *Proc. ICASSP 2022-2022 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7367–7371.
[64] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.