Learning with Rater-Expanded Label Space to Improve Speech Emotion Recognition

Shreya G. Upadhyay, Woan-Shiuan Chien, Bo-Hao Su, *Student Member, IEEE,* Chi-Chun Lee, *Senior Member, IEEE*

Abstract—Automatic sensing of emotional information in speech is important for numerous everyday applications. Conventional Speech Emotion Recognition (SER) models rely on averaging or consensus of human annotations for training, but emotions and raters' interpretations are subjective in nature, leading to diverse variations in perceptions. To address this, our proposed approach integrates the rater's subjectivity by forming the Perception-Coherent Clusters (PCC) of raters to be used to derive expanded label space for learning to improve SER. We evaluate our method on the IEMOCAP and MSP-Podcast corpora, considering scenarios of fixed and variable raters, respectively. The proposed architecture, Rater Perception Coherency (RPC)-based SER surpasses single-task models with consensus labels by achieving UAR improvements of 3.39% for IEMOCAP and 2.03% for MSP-Podcast. Further analysis provides comprehensive insights into the contributions of these perception consistency clusters in SER learning.

Index Terms—speech emotion recognition, multi-tasking, rater subjectivity, perception consistency clusters

1 Introduction

Speech Emotion Recognition (SER) has found many applications in domains such as customer call centers [1], voice analysis [2], and spoken dialogue systems [3], to incorporate emotional intelligence for recognizing user emotions. Over the years, many solutions have been proposed for SER training, taking into account different aspects of comprehensive modeling. However, most of these studies tend to follow a conventional approach, training the model with aggregated values or consensus labels [4]–[6], which uses a single point (consensus) approach. Recently, there has been a focus on label ambiguity by adopting a distributed emotion learning approach, such as soft-labelling [7], [8], and multi-labelling [9]–[11]. These studies have shown promising results and are better than the conventional approach, but they emphasize only the label ambiguity by learning the dominant emotions and treating the ratings as independent and identically distributed.

Psychological research suggests that there exists an individual difference in emotional sensitivity [12] and personality dimensions (such as introversion-extraversion, motivation, anxiety, etc.) that can influence the emotional constructs of rater during information processing, with situational moderators such as time pressure, time of day, and incentives also having an impact [13]. These factors can affect the moment-to-moment report of perceived emotions for judgment of dynamics and naturalistic expressions in speech prosody [14]. Also, some researchers argue that emotions, both positive and negative, are subjective and can have varying meanings for each individual [15], [16]. Also, emotions can encompass different emotional profiles, and their prevalence may vary across cultures [17] and age groups [18]. Positive emotions like *Happiness* can be classified into different profiles, such as excitement and enthusiasm, which are characterized by high arousal levels, and calmness and joy, which are characterized by low arousal levels [19], [20]. On the other hand, negative emotions can be classified based on between-category differentiation, where emotions like *Anger* and *Sadness* fall into high- or low-intensity categories [21], [22]. But all these broad category emotions such as *Happiness, Anger, Sadness* have fine-grained categories that distinguish closely related emotions, for example, *Anger* and *Frustration* [16]. This underscores the importance of considering rater ambiguity in SER training and specifically modeling it along with consensus labels to include the *raters emotional subjectivity* in learning.

The standard approach to examining rater subjectivity is outlined in recent studies [23]–[25] typically involves either averaging the outputs of multiple rater-specific models or learning the individual rater perception independently in a multi-task setting. However, these methods may give rise to two key concerns. Firstly, such learning can lead to highly complex and branched multi-task architectures that make it challenging to model individual perceptions. Secondly, these individual multi-task approaches raise issues with missing labels since it is rare to have a corpus where all samples are rated by each rater. Moreover, this architecture does not take into account corpora with variable raters or workers (e.g., MSP-Podcast). In this work, our proposed method aims to tackle these issues by introducing a novel approach that clusters raters into coherent groups based on their consistency in perception. This allows the integration of raters' subjectivity into the SER task while avoiding the complexity of learning individual rater perceptions. We evaluate the perception consistency correlations on the perceptual scales of inter-rater consistency. Our method involves simultaneous modeling of the *perception-coherent clusters* PCC of raters, enabling us to address the two aforementioned concerns by integrating rater subjectivity in a controllable manner in speech emotion learning.

Specifically, this study proposes a method to integrate

Department of Electrical Engineering, National Tsing Hua University, Taiwan.
E-mail:{shreya,wschien,borrissu}@gapp.nthu.edu.tw,cclee@ee.nthu.edu.tw

the rater ambiguity in SER by leveraging the emotional subjectivity of raters through a set of PCC. The homogeneous clustering is based on the inter-rater similarities, evaluated on perceptual scales of inter-rater consistency (IRC) and rater consistency with ground truth (RC-GT). Rather than using the conventional approach of averaging the outputs of multiple rater ratings (e.g., consensus) or considering just the label subjectivity (e.g., soft labeling) or independently learning individual rater's perception, our work takes a joint approach. That is, we propose a method of modeling the raters' PCC clusters, consisting of majority (*MajP*) and minority (*MinP*) raters' perceptions with the consensus-based emotion classifier (*MV*). This technique reduces the complexity in modeling by clustering and centering the raters' subjectivity into *MajP* and *MinP*, leading to improved performance of the SER task and also addressing the issue of missing labels. Moreover, In order to minimize the differences between feature representations originating from the same stimuli, we also incorporate the Maximum Mean Discrepancy (MMD) loss in our learning. The proposed rater-perception coherency-based (RPC) multi-perception learning method is tested and analyzed on two corpora: IEMOCAP [26] (with a fixed number of rater) and MSP-Podcast [27] (with the variable number of rater), resulting in the unweighted average recall (UAR) of 63.92% and 57.95%, respectively. This work introduces the concept of learning in a rater-expanded label space, which integrates the rater's emotional subjectivity and considers their perception homogeneity to achieve better SER performance.

2 Related Works

SER can be approached from various perspectives, with a focus on either the speaker's side or the rater's side. In this study, our emphasis lies on the rater's side of SER. Conventionally, affect-related recognition models have relied on single-label setup, where labels from multiple raters are combined using methods like majority voting or averaging [28], [29]. However, these consensus labels have limitations as they may introduce biases and fail to capture the subjective nature of emotions as well as individual variations in perception [12], [30], [31]. Hence, the objective of our work is to advance SER modeling by leveraging a large rater-expanded label space.

Emotional ambiguity presents a common challenge in emotion recognition due to the subjective nature of emotions and variations in human interpretations, as shown in previous studies [12], [17]. Recently, the field of SER has placed greater emphasis on mitigating these disparities stemming from the subjectivity of emotions and individual behaviors. Within the scope of this SER study, rater ambiguity addresses inconsistencies among raters interpretations when assigning emotional labels, while label ambiguity is concerned with the inherent ambiguity in emotion labels that originate from the diverse and subjective nature of emotions. While recent studies have explored rating ambiguity through adaptive learning, multi-task learning, and personal profiles [32]–[35], they do not explicitly account for the subjectivity for each rater. Despite the known impact of rater ambiguity on emotion perception [36], [37], previous research in emotion recognition has often overlooked this factor, with only a few recent studies specifically considering rater's emotion perception.

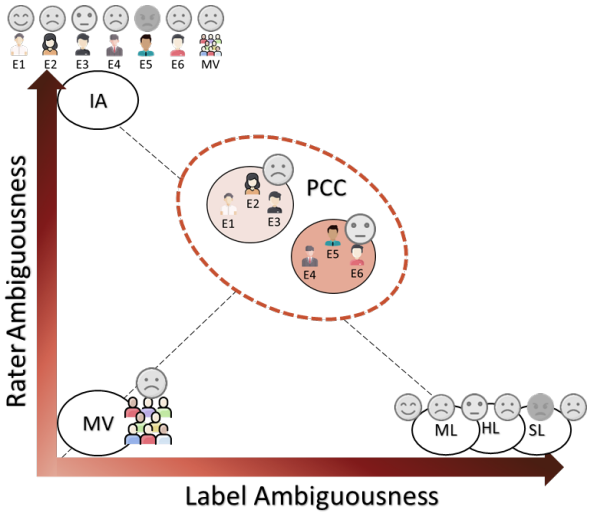


Fig. 1: Literature technique incorporating subjectivity in SER and our proposed approach present on the spectrum of rater ambiguity and label ambiguity; majority vote as MV, individual-rater learning as IA, and soft, hard, and muti-labeling as SL, HL, and ML, respectively.

Most of the work that accounts for the rater's ambiguity uses label ambiguity (soft labels, multi-label, etc.) to determine the underlying emotion distributions by employing techniques such as rater reliability or at the extreme, incorporating each rater's perceptions directly through a multi-task approach. Also, there has been a growing emphasis on utilizing ordinal regression techniques [38], [39]. These approaches have proven advantageous in effectively capturing and modeling the ordinal nature of emotion labels. However, there has been no consideration given to each rater's own characteristics who assigned the labels to the samples. Our proposed method fills this gap by explicitly accounting for rater ambiguity and leveraging a rater-expanded label space in SER modeling. Fig. 1 depicts the literature methods and proposed method on the spectrum of label and rater ambiguity. The literature on incorporating ambiguity in SER learning can be divided into two main categories. The first category focuses on label ambiguity, which involves the ambiguity in emotion labels assigned to a given stimulus without considering any specific rater behavior. The second category addresses rater ambiguity, which not only considers the ambiguity in the annotations but also takes into account the subjective nature of the raters in learning.

2.1 Studies Exploring Emotional Label Ambiguity

Several approaches have been proposed to deal with label ambiguity in learning to improve SER. For instance, soft-label learning is one of these methods, which takes into account all the ratings provided by multiple raters instead of using a one-hot vector to represent the consensus label. Also, multi-label learning is another technique [10], [11], which considers the co-occurrence of multiple emotions in speech. These approaches consider the ambiguity and

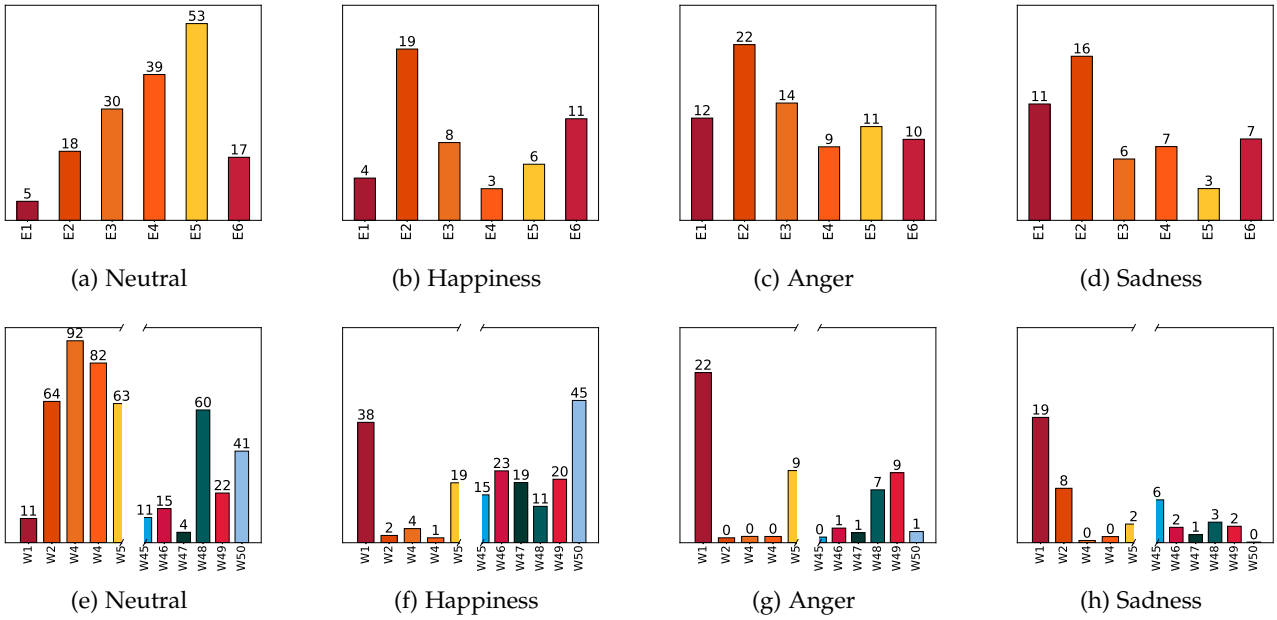


Fig. 2: Rater-wise emotional perception variation present in IEMOCAP ((a)-(d)) and MSP-Podcast ((e)-(h)) corpora over primary emotions; for IEMOCAP, six raters (E1, E2,.., E6) and for the MSP-Podcast corpus, we selected the top 50 raters (W1, W2, ..., W50) based on higher annotation frequency which are consistent to represent the range of rating ambiguity.

variability among the ratings during training and have been demonstrated to achieve better performance than using consensus labels alone [7], [8], [40], [41]. However, these methods do not include any rater behavior in learning by considering ratings as independent and identically distributed.

Similarly, few studies [8] have proposed approaches that address emotional subjectivity by using both hard and soft labels in a joint manner by involving multiple models in learning. Also, another work [42] utilizes a multi-task model that integrates disagreement over ratings as information in a multi-task modeling approach by measuring the degree of dissimilarity between the model's predictions and the soft-label targets. However, all these approaches incorporate label ambiguity to the extent of learning emotions but do not explicitly consider the source of the ratings at all.

2.2 Studies Exploring Rater Ambiguity

In recent years, there has been a growing emphasis on addressing rater ambiguity in addition to label ambiguity in the field of emotion recognition. Researchers have proposed several methods to mitigate the impact of rater ambiguity on the performance of emotion recognition systems. One approach [43], [44] involves utilizing multiple ratings to identify accurate or correct ratings by considering the disagreements among raters. These studies also take into account the expertise level of raters in the learning process [43], [44], where the error rate of each rater is modeled to depend on the data samples they annotate [43]. Another study [45] introduces a Gaussian process (GP)-based method to handle multiple annotations independently of the raters' expertise level. Another approach considers both the expertise and reliability of raters in a multitask manner [46], [47], assuming independence among raters. Also, studies like [7], [48] focus on individual learning for every single rater in a multitask setting to address the subjectivity of emotions. More recently, an idea proposed in [49] suggests modeling subjective affect-related tasks by considering scenarios where multiple raters provide labels for an input sample in an affect-related task.

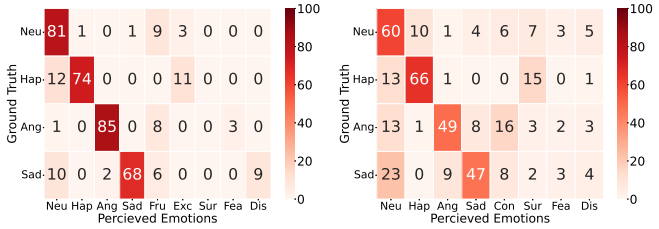
However, most of these approaches either depend on averaging across multiple raters' learning in multi-modal tasks or independently modeling each rater's perception in a multi-task manner. These methods can lead to complex architecture and an inability to handle missing labels, especially when there exist multiple and inconsistent raters. To overcome these limitations, we propose a new approach that considers the coherency of rater perceptions as PCC, which enables us to incorporate rater subjectivity but in a controllable manner into the learning process of rater ambiguous multi-perception SER.

3 Rater Ambiguity Over Corpora

3.1 Corpora

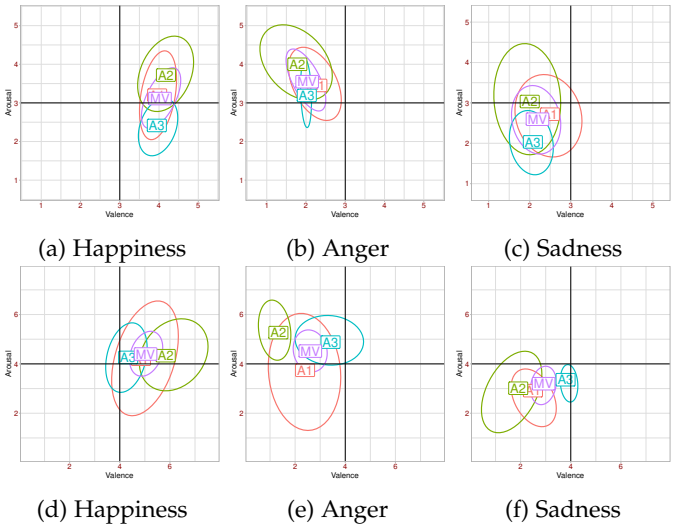
This work considers the IEMOCAP [26] and MSP-Podcast [27] SER corpora to evaluate our proposed method. Both the IEMOCAP and MSP-Podcast corpora have been extensively annotated by human raters, providing ground truth labels for emotions. These corpora serve as reliable sources for training and evaluating SER models, enabling researchers to assess the effectiveness and generalizability of their proposed approaches.

IEMOCAP: The IEMOCAP corpus is a dataset of approximately 12 hours of dyadic audio-visual interactions in the English language. It features 10 professional actors (five male and five female) acting out scripts and realistic scenarios over five sessions. The corpus includes both discrete



(a) IEMOCAP (b) MSP-Podcast

Fig. 3: The emotion profiles showing raters label ambiguity (%) in corpora; for each Ground Truth label, the variation in emotion rating over some frequent emotion list is shown with an average over number of raters.



(a) Happiness (b) Anger (c) Sadness

(d) Happiness (e) Anger (f) Sadness

Fig. 4: Rater-wise emotional perception variation present in IEMOCAP ((a)-(c)) and MSP-Podcast ((d)-(f)) corpora for major emotions (*Happinesss*, *Anger*, *Sadness*); *A1*, *A2*, and *A3* shows standard names for raters annotating for same samples in respective corpora and *MV* is majority vote.

and continuous annotations, which are provided by 2 to 4 raters for each stimulus. Overall, this corpus contains 10,039 utterances with an average duration of 4.5 seconds. The annotations were completed by a total of 12 raters, with six annotating for primitive attributes and six for emotional categories. This study specifically focuses on the emotional category labels and their corresponding raters, namely *E1*, *E2*, *E3*, *E4*, *E5*, and *E6*, to analyze the emotional subjectivity of the raters. The evaluation for this corpus is conducted using leave-one-session-out cross-validation over 4490 samples (consisting of four major emotions).

MSP-Podcast: The MSP-Podcast corpus (v.1.10) [27] is a rich database of diverse emotional naturalistic speech samples collected from various podcast recordings. It is increasingly being used for research on SER due to its scale and availability of emotionally balanced dialogues from a large number of speakers. Each sample in the corpus is rated by at least five different workers with primary emotions, secondary emotions, and emotional attributes. The database contains a total of 166 hours of data. To analyze the raters' ambiguity over this large and variable raters corpus, we consider the workers who have annotated at least 10% samples of the corpus, and the samples that have been rated by these selected raters only are chosen for further experiments. This way, we have selected around 25,014 samples with predefined train-val-test splits.

The aim of this work is to evaluate the performance of SER on four primary emotions as *Neutral*, *Happiness*, *Anger*, and *Sadness* over the IEMOCAP and MSP-Podcast corpora. Additionally, we also consider some frequent emotions like *Frustration* and *Excitement* in IEMOCAP and *Contempt* and *Surprise* in MSP-Podcast to investigate the rater ambiguity in majority and minority perception learning tasks. We selected these fine-grained distinguishable emotions based on their higher frequency of occurrence in the corpora. To observe the emotional perception variation over corpora, we investigated raters' emotional perceptions in both the IEMOCAP and MSP-Podcast corpora. The analysis plots are presented in Fig. 2, where we plot the distribution of each rater's annotations across all samples. In the case of IEMOCAP, we examined the annotations of six raters. Conversely, for the MSP-Podcast corpus, we selected the top 50 raters based on their higher frequency of ratings, effectively covering a range of rating ambiguities within the corpus and are consistent in this analysis. Fig. 2 shows a holistic view of emotional perception dynamics in corpora indicating the presence of ambiguity in distribution and thus the presence of raters' ambiguity and emotional subjectivity in corpora.

3.2 Rater Ambiguities in Emotional Corpora

To examine the presence of emotional ambiguity in the corpus, we compute the emotion profiles (descriptions of different emotions present) for each primary emotion in the respective corpora. These profiles are obtained by aggregating the perceptions of all raters for each emotion category across all samples to create a consensus. Fig. 3 shows the emotion ambiguity present in each emotion for both corpora. These profiles capture the variation in emotion ratings for each Ground Truth label across a set of most prominent emotions including *Neutral, Happiness, Anger, Sadness, Surprise, Fear, Disgust* over both corpora and *Frustration, Excitement* for IEMOCAP and *Contempt* for MSP-P. The estimation involves taking the majority vote from the ratings provided by multiple raters. It is important to note that the emotion profiles are calculated for the entire database and encompass all emotions and while plotting only considered emotions are shown; for example the "others" category is not included. Therefore, the summation of the profiles may not necessarily equal 100 due to the exclusion of a few categories. This emotion profile demonstrates the existence of ambiguous emotional subjectivity in the ratings provided by the raters in both corpora. Here, we can observe that the Ground truth emotions mostly vary among fine-grained distinguishable emotions, for example, for *Anger* emotion, the ambiguity present over *Frustration* and *Contempt* in IEMOCAP and MSP-Podcast corpora, respectively.

To observe the presence of ambiguity in the raters who actually provide these ratings, we generate the label space with arousal-valence (AV) plot for different raters of each corpus depicted in Fig. 4. In each corpus, we consider the three raters that annotate the same samples and are

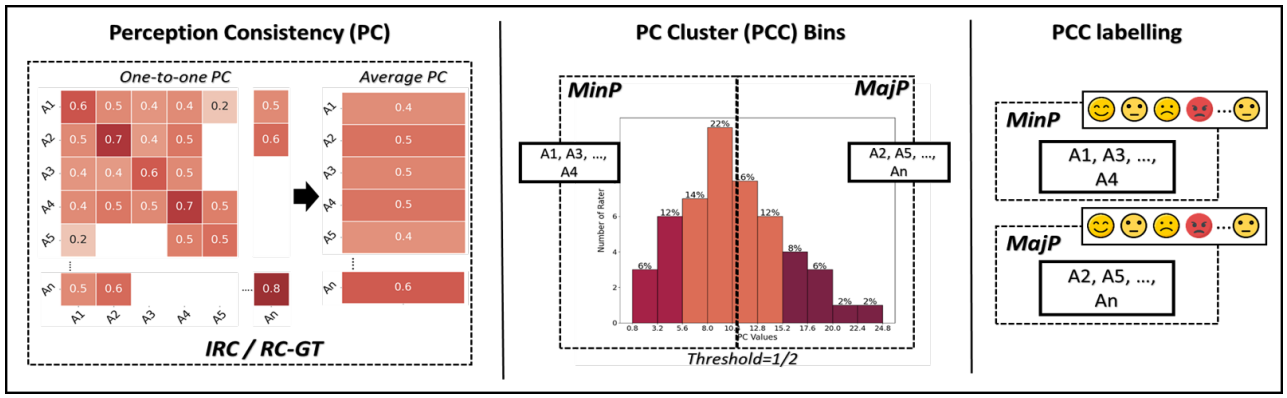


Fig. 5: Figure illustrates the clustering procedure employed in our novel approach, comprising three key components. The initial component is dedicated to evaluating perception coherency, followed by the estimation of PC cluster bins in the second component, and finally, the third component is focused on PCC label centering.

referred to as *A1*, *A2*, and *A3*, while *MV* represents the majority vote in Fig. 4. In Fig. 4, distinct colors are used to represent each label space for *A1*, *A2*, *A3*, and *MV*. Within the plot, 90% confidence ellipses show the distribution of labels from both raters and the MV on the AV scale. These label space plots show that both the IEMOCAP corpus ((a)-(c)) and the MSP-Podcast corpus ((d)-(f)) exhibit variations in emotional perception among raters for the four major emotions (*Happiness*, *Anger*, *Sadness*). This analysis reveals that different raters have varying perceptions of different emotions for the same set of samples. Furthermore, we see that some raters overlap more with others, indicating a scaled level of rater ambiguity that varies from low to high. Based on these analysis observations, it is evident that integrating raters' subjectivity can be important for learning a reliable SER model. In the following sections, we will provide a detailed analysis of raters' subjectivity based on their perception consistency, as well as the perception-coherent cluster definition used in this study.

4 Perception-Coherent Cluster Definition

Perception-coherent cluster (PCC) is a concept we use to define clusters of raters with similar perception consistency in rating emotions. In other words, raters in the same cluster exhibit a similar level of agreement or consistency in their ratings across different emotional categories. This concept is based on the assumption that raters with similar perception consistency are more likely to rate emotions in a similar way and their ratings are more reliable than those of raters in different homogeneous clusters. By using PCC to group rater with similar perception consistency, it can be easier to integrate the rater ambiguity as information on the emotion recognition models. The clustering process involved in our proposed approach is depicted in Fig. 5. It comprises three primary components. The initial segment concentrates on gauging perception coherency through the analysis of raters' perception consistency. The subsequent segment involves the estimation of PC cluster bins to organize these raters into the corresponding clusters. The final part is dedicated to PCC label centering to have labels for each cluster. The following section will elaborate on each of these processes.

4.1 Perception Coherency Estimating Methods

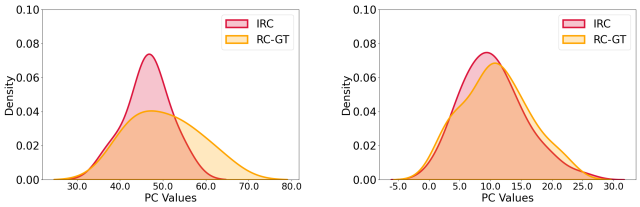
In order to capture the diversity of speech emotions across a large rater-label space, we employ a PCC approach to group individual raters with similar perceptions. To achieve this, we analyze the rater's perceptual consistency from two perspectives: 1) inter-rater consistency (IRC), and 2) rater consistency with ground truth (RC-GT), by analyzing the training portion of the corpora (excluding testing samples). This method is used as an approach to create homogeneous clusters of consistent perceptions among the raters.

Inter-Rater Consistency (IRC): We use the agreement between raters' perceptions by comparing their ratings to see how closely their emotions perceptions are aligned. For example, *A1* is perceiving *Anger* emotion for a sample, then, how similar the other rater's perception is to *A1*. To measure this perceptual similarity between raters, we utilize the Cohen's Kappa (κ) [50] statistic, which calculates the perceptual consistency between each pair of raters, such as between *A1*$\sim$*A2*, in terms of the perception consistency (PC) value.

Rater Consistency with Ground Truth (RC-GT): In addition to examining the raters' perception consistency with each other, we also investigate their agreement with the voted consensus ground truth labels to estimate the rater consistency with ground truth. To accomplish this, we evaluate the consistency of each rater's emotional perception with both other raters and with the ground truth. Specifically, we consider the agreement between each pair of raters and the majority vote. For instance, we determine whether the perceived emotion of *Anger* by *A1* and *A2* is consistent with the majority vote (*MV*), i.e., *A1*$\sim$*A2*$\sim$*MV*. To quantify the perception consistency for this RC-GT analysis, we use the same Cohen's Kappa (κ) measure as in the previous analysis. By analyzing the raters' perception consistency with the ground truth labels, we aim to gain insights into the level of inter-rater agreement present in the corpus.

4.2 PC Cluster Bins

The analyses of IRC and RC-GT provide insights into the consistency of rater perceptions. Perception correlation (PC) matrices are computed for each perspective in both corpora,



(a) IEMOCAP (b) MSP-Podcast

Fig. 6: Raters distribution in different histogram bins with respect to perception correlation (PC) values for IEMOCAP and MSP-Podcast corpora.

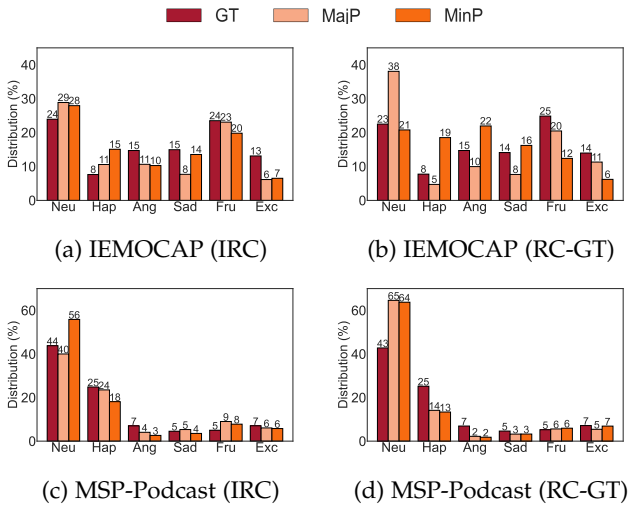
revealing that the RC-GT method yields higher PC values as it takes into account the raters' agreement with the majority vote. To cluster raters into homogeneous groups, we set the average PC values as a reference. We first assign raters of varying PC values to histogram bins and then divide them into two groups based on a threshold (default = number of bins * 1/2). The group with a lower mean is designated as the minority-perception (*MinP*) raters group, while the group with a higher mean is the majority-perception (*MajP*) raters group. To see the difference in PCC methods (IRC and RC-GT) for raters distribution in different bins with respect to perception correlation (PC) values, we plot the Kernel Density Estimation (KDE) shown in Fig. 6 for IEMOCAP and MSP-Podcast corpora. Here, we can observe noticeable changes in the distribution curves for both methods across both the IEMOCAP and MSP-Podcast corpora. Particularly, in the case of IEMOCAP, we observe more distinct differences in the bin distribution between the IRC and RC-GT methods.

4.3 PCC Label Centering

After obtaining the majority (*MajP*) and minority (*MinP*) PCC clusters of raters based on their perception consistency from Section 4.1, the next step is to centering the rater emotion ratings in these clusters to come up with the PCC labels. However, as not all samples are rated by each rater, several possible scenarios may arise while centering the cluster labels. Hence, we have established four rules for generating the PCC labels:

- If the majority of raters have a similar perception (the most probable case), we consider the majority vote as the label.
- If only one rater has rated a sample, we consider that rating as the cluster label since these groups of raters have high consistency in their perceptions.
- If all the ratings for a sample are different, we consider the rating with a high confidence rate with the ground truth as the cluster label (i.e., raters' overall consistency with ground truth labels).
- If there is no rating for a specific sample (unless all raters within the cluster have not provided a rating for that particular sample), we consider the cluster label as "NaN" and exclude it while modeling the SER.

Fig. 7 illustrates the distribution of labels for the consensus (*MV*), majority raters' perceptions (*MajP*), and minority raters' perceptions (*MinP*) in both corpora. It can be observed that in both corpora, the majority perceptions (*MajP*) intuitively exhibit high similarity to the consensus (*MV*) and predominantly peak towards major emotions. Conversely, the distributions of minority perceptions (*MinP*) tend to peak more towards fine-grained distinguishable emotions. This distribution highlights the raters' ambiguity over the label space and motivates our work of learning in a rater-expanded label space to improve emotion recognition.



(a) IEMOCAP (IRC) (b) IEMOCAP (RC-GT)

(c) MSP-Podcast (IRC) (d) MSP-Podcast (RC-GT)

Fig. 7: Emotion distribution of majority vote (*MV*) and different cluster labels (*MajP* to *MinP*) from PCC approach (*IRC* and *RC-GT*) for IEMOCAP and MSP-Podcast corpora; Since only the distribution of the considered emotions is presented here, the percentages may not be sum up to 100%.

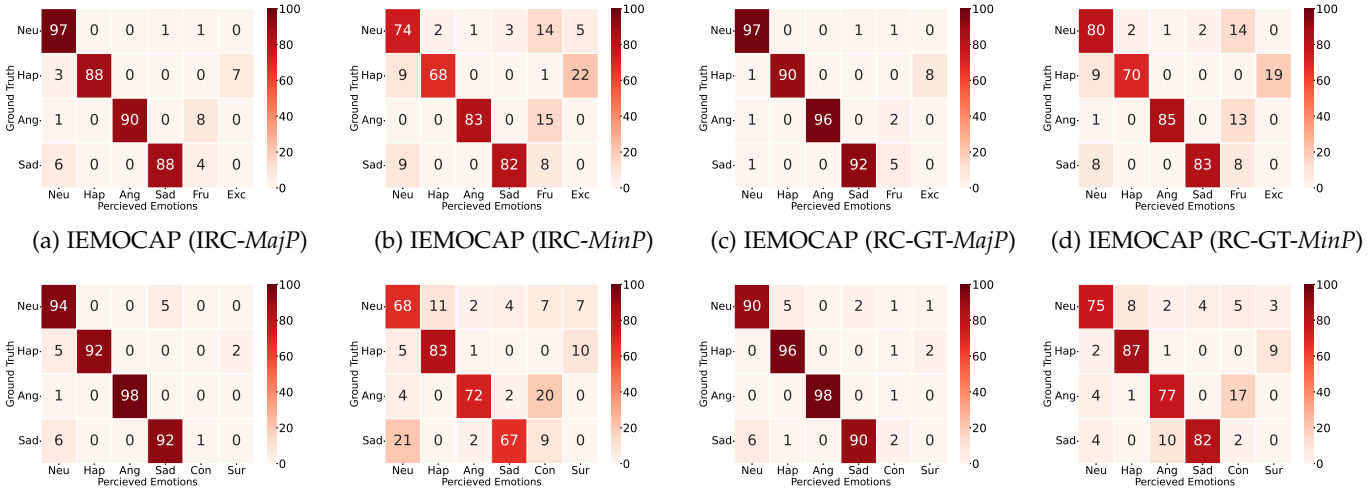
TABLE 1: PCC labels consistency table with *MV*, *MajP*, *MinP* for IEMOCAP and MSP-Podcast corpora in terms of Cohen's Kappa (κ).

	IEMOCAP		MSP-Podcast	
	IRC	RC-GT	IRC	RC-GT
MV-MajP	0.38	0.40	0.43	0.41
MV-MinP	0.25	0.28	0.30	0.34
MajP-MinP	0.52	0.50	0.49	0.47

4.4 PCC Label Analyses

To gain a deeper understanding of the PCC labels, which represent the centering among raters' emotions with similar perception consistency, related to their emotion profiles and cluster correlations, this section involves examining the relationship between the *perception-coherent cluster* (PCC) labels with their cluster-wise emotion profiles, and their coherency with each other and also with *MV*.

Table 1 shows the coherency table to analyze these clusters' consistency with *MajP* and *MinP*, and also with the consensus labels *MV* using Cohen's Kappa values. Based on the observations from Table 1, it is evident that the *MV* exhibits higher consistency with the majority-perception (*MajP*) with (0.38, 0.40) for IEMOCAP and (0.43, 0.41) for MSP-Podcast compared to the minority-perception (*MinP*) with (0.25, 0.28) for IEMOCAP and (0.30, 0.34) for MSP-Podcast over both IRC and RC-GT methods, respectively. This observation from Table 1 is expected since the majority perception aligns with the overall prevalent viewpoint.



(a) IEMOCAP (IRC-*MajP*) (b) IEMOCAP (IRC-*MinP*) (c) IEMOCAP (RC-GT-*MajP*) (d) IEMOCAP (RC-GT-*MinP*)

(e) MSP-P (IRC-*MajP*) (f) MSP-P (IRC-*MinP*) (g) MSP-P (RC-GT-*MajP*) (h) MSP-P (RC-GT-*MinP*)

Fig. 8: Emotion profiles for rater ambiguity (in %) of proposed PCC-based rater clusters (*MajP* to *MinP*) for IEMOCAP and MSP-Podcast corpora over both *IRC* and *RC-GT* methods.

Additionally, there is a notable correlation between *MajP* and *MinP* (IEMOCAP with 0.52 and 0.50; MSP-Podcast with 0.49 and 0.47 for IRC and RC-GT methods respectively), indicating that these perceptions are diverse yet exhibit a level of consistency. This consistent pattern can be observed in both corpora with both clustering methods, suggesting its generalizability.

Fig. 8 depicts the emotion profiles of the rater clusters (referred to as *MajP* to *MinP*) based on the PCC approach for the IEMOCAP and MSP-Podcast corpora, focusing on four major emotions (*Neutral, Happiness, Anger, and Sadness*). The analysis of the emotion profiles in Fig. 8 reveals interesting patterns. It shows that the ratings of the *MajP* cluster exhibit higher consistency towards these major emotions compared to the *MinP* cluster, where the raters tend to center their perceptions around a few fine-grained distinguishable emotions instead of just the primary emotions. For instance, the *Happiness* emotion profile also includes *Excited* and *Surprise* for the IEMOCAP and MSP-Podcast corpora, respectively. This observation shown in Fig. 8 demonstrates the presence of ambiguity among raters with minority perceptions regarding the primary emotions.

5 Research Methodology

5.1 Feature Extraction and Encoding

In this study, we utilize vq-wav2vec [51] as our feature set. This model provides speech representation by learning quantized features of raw audio for future time-step prediction. It incorporates multiple convolutional and residual blocks as a local encoder to encode the speech samples into a sequence of embeddings. The model is trained using self-supervised contrastive loss, which implicitly models the mutual information between the context and future audio samples. It also includes a quantization layer that uses k-means clustering constraints in vector quantized variational autoencoders. The resulting speech representations are then fed into a multi-head transformer, which employs self-attention to encode the input sequence.

This feature extraction block remains consistent throughout our experiments, with the input being the vq-wav2vec speech representations from audio samples in the corpora. The self-attended discrete latent embedding, acquired from this block is first averaged and subsequently undergoes further processing within the architecture for emotion category classification.

5.2 Multi-Perception SER Learning

This section describes our approach to incorporating a rater's emotional subjectivity into SER using a rater-expanded multi-perception label space (the PCC approach). We consider four major emotions, namely *Neutral*, *Anger*, *Happiness*, and *Sadness*. Fig. 9 depicts the model architecture for 4-category SER. It consists of two main components of our study: (a) the identification of *RPC* clusters based on the consistency of their perception, and (b) the training of the *RPC* multi-perception SER, which combines learning with respect to rater ambiguity and consensus label learning. Our proposed SER part of the architecture (b) has three branches designed to learn three different tasks: two for integrating raters' perceptions (*MajP* and *MinP*), and one for conventional consensus-based (*MV*) SER learning. To optimize the model, we use different loss functions during training. Specifically, we use Eq. 1 for *MajP*, Eq. 2 for *MinP*, and Eq. 3 for *MV*.

$$L_{MajP} = \mathbb{E}_{X_S, y_{MajP}}[\|CE(T(X_S), y_{MajP})\|] \quad (1)$$

$$L_{MinP} = \mathbb{E}_{X_S, y_{MinP}}[\|CE(T(X_S), y_{MinP})\|] \quad (2)$$

$$L_{MV} = \mathbb{E}_{X_S, y_C}[\|CE(T(X_S), y_{MV})\|] \quad (3)$$

where CE is the cross-entropy function, T is the transformer function, X_S is the source features, and $y_{MajP}, y_{MinP}, y_{MV}$ is the emotional labels from *MajP*, *MinP* and *MV*.

To improve the performance over *MV*, we also include the maximum mean discrepancy (MMD) loss in our proposed architecture. This loss is used to measure the differences between the encoded representations of *MajP* and

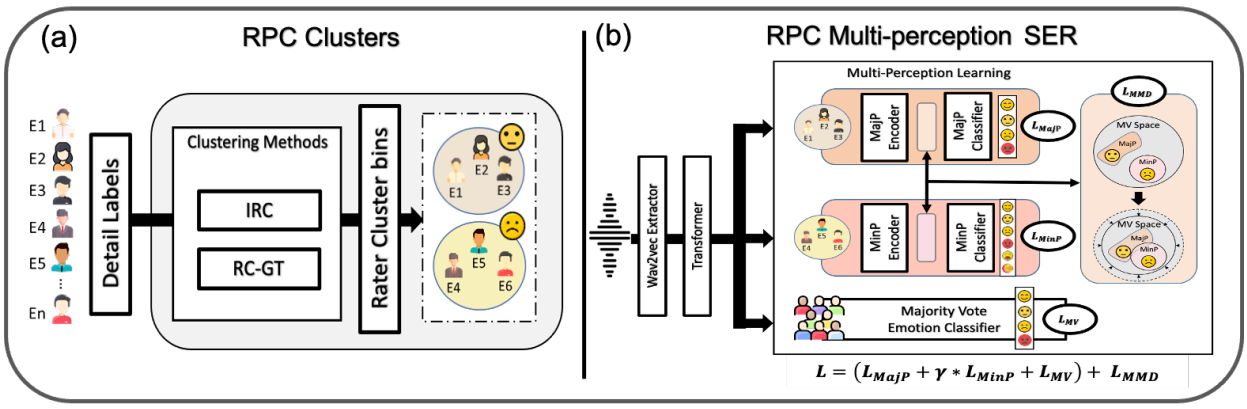


Fig. 9: This figure illustrates the model architecture for 4-category SER, which includes two parts: (a) the estimation of homogeneous clusters of raters based on inter-rater perception consistency, and (b) the multi-perception SER training that integrates rater-ambiguity conditioned learning jointly with consensus label learning using the clusters from part (a).

MinP in order to reduce their feature space discrepancies, as both of these representations are produced from the same stimuli. The equation for the *MajP-MinP* feature difference reduction loss is shown as Eq. 4.

$$L_D = MMD(MajP, MinP) \tag{4}$$

where MMD is MMD loss function, $MajP$ and $MinP$ is the encoded features for *MajP* and *MinP* respectively.

For each task, we employ four fully connected layers for emotion classification. Overall the complete loss for the proposed SER with rater-expanded label space learning with PCC cluster is shown below:

$$L = (L_{MajP} + \lambda * L_{MinP} + L_{MV}) + L_D \tag{5}$$

where L_{MajP} and L_{MinP} are the losses for the *MajP* and *MinP* SER tasks respectively, L_{MV} is for *MV* SER and L_D is for MMD loss. λ is the weight parameter and constant with λ= 0.6 for model with IEMOCAP and λ=0.8 for model with MSP-Podcast.

In this study, we propose that leveraging rater ambiguity through PCC clusters in the learning process can enhance the predictive effectiveness of the MV-based SER models. To maintain consistency with previous studies, in this context, the inference process is centered around the final branch, which is the MV classifier trained with both *MajP* and *MinP*, and serves as the primary component for making predictions.

6 Experimental Settings

6.1 Parameters

In all the conducted experiments, we utilize the Adam optimizer with a learning rate of 0.0001 along with a decaying factor. The proposed systems are trained using back-propagation, employing the loss function specified in Equation 5. On the other hand, the baseline systems are trained only with the cross entropy loss function. The network is trained for a maximum of 50 epochs, utilizing a batch size of 16 and implementing early stopping. To assess the performance of the models in multi-class classification tasks, the evaluation metrics used are the Unweighted Average Recall (UAR) and the Weighted F1 score (wF1). Furthermore, the Recall score (Recall) is employed to evaluate the specific emotion SER performances.

6.2 Experiments

6.2.1 Baseline Experiments

This study contains a range of baseline models that aim to explore and compare different methodologies related to rater-expanded learning for SER. The first experimental setup focuses on the single-task (*ST*) approach, which is the conventional consensus-based SER using a majority vote (*MV*). Another baseline approach addresses label ambiguity in SER learning through the use of soft-labelling (*Soft-L*, where each rated value is divided by the total number of annotations), hard+soft-labelling (*Hard+Soft*), and multi-labelling (*Multi-L*) models. For *Soft-L*, we adopt the same settings as in [40], while for *Hard+Soft* and *Multi-L*, we refer to [8] and [10] respectively. Additionally, we incorporate an individual-rater multi-task (*IA-MT*) SER baseline, which models each rater's perception independently, as described in [24]. By conducting these diverse experimental setups, we aim to gain valuable insights into the effectiveness and performance of each approach compared to our proposed method in handling rater subjectivity and label ambiguity in SER.

6.2.2 Ablation Experiments

In this study, we propose a novel approach called the rater-perception coherency-based (RPC) multi-perception learning method for integrating rater subjectivity in SER tasks. To gain a comprehensive understanding of the proposed approach and its variants, we conduct various ablation experiments to analyze further the effectiveness of different components in the proposed approach. Specifically, we investigate the performance of the only-*MajP* model, which focuses solely on the majority perceptions, and the only-*MinP* model, which emphasizes the minority perceptions. Additionally, as we also use the MMD loss in the proposed architecture, we investigate the impact of incorporating the MMD loss in the SER models by comparing the performance of the SER models with MMD (proposed *RPC*) and without MMD loss (*MajP+MinP*).

TABLE 2: Performance table with proposed methodology (RPC); considered baseline models (Soft-L, ST, IA-MT), and ablations (Ablation) on IEMOCAP and MSP-Podcast corpora in terms of UAR (%) and wF1 (%).

		IEMOCAP		MSP-Podcast	
		UAR	wF1	UAR	wF1
Baseline	ST	60.53	63.01	55.92	57.01
	Soft-L [40]	61.48	63.28	56.33	58.36
	Hard+Soft [8]	61.55	63.34	56.47	57.18
	Multi-L [10]	60.46	61.91	54.93	56.07
	IA-MT [24]	61.01	63.98	-	-
IRC	Only-*MajP*	60.09	61.72	56.71	57.27
	Only-*MinP*	58.71	59.16	54.66	55.34
	MajP+MinP	62.83	64.63	57.28	58.50
	RPC	**63.15**	**64.29**	**57.54**	**59.01**
RC-GT	Only-*MajP*	62.03	64.38	57.27	58.80
	Only-*MinP*	60.80	62.84	57.00	58.03
	MajP+MinP	62.65	65.34	57.39	59.33
	RPC	**63.92**	**65.98**	**57.95**	**60.27**

7 Experiment Results and Analysis

7.1 Experiment Performance Comparisons

7.1.1 Baseline Comparisons

To compare our proposed rater perception coherency-based *RPC* SER models with previous techniques, we opt for the baseline techniques that operate on the label space as detailed in Section 6.2.1. Our emphasis here is on the label side, we use a consistent backbone across all experiments as explained in Section 5.1. The performance results of the considered baseline models and *RPC* including both proposed methods(*IRC* and *RC-GT*) are presented in Table 2. From Table 2, we can see that the proposed *RPC* models perform better than the baseline models in both corpora over both the *IRC* and *RC-GT* methods. We can observe that these improvements are more significant with *RC-GT*, for instance, the proposed architecture *RPC* under *RC-GT* shows improvements over *IRC* with 0.77% of UAR and 1.69% of wF1 on the IEMOCAP corpus. It may be because of having higher consistency in the perceptions of the *RC-GT* method as discussed in Section 4.4. Also, these performance differences between the PCC methods are more significant in terms of the wF1 metric than UAR.

From Table 2 we can observe that the *RPC* model derived from the *RC-GT* method, achieves superior results compared to *ST*, with improvements of 3.45% and 2.03% in UAR, and 2.97% and 3.26% in wF1 on the IEMOCAP and MSP-Podcast corpora, respectively. In comparison to the *Soft-L* approach, the *RPC* model demonstrates improvements of 2.50% and 2.70% in UAR, and 1.62% and 1.91% in wF1 on IEMOCAP and MSP-Podcast corpora, respectively. Furthermore, when compared to *Hard+Soft* and *Multi-L*, *RPC* model exhibits significant improvements of 2.37% and 4.07%, as well as 3.46% and 2.64% in UAR and wF1 on IEMOCAP corpus. A similar improvement can be seen with the MSP-Podcast corpus. These results highlight the importance of integrating the rater's ambiguity information in the learning process, as even though these methods consider label ambiguity, they do not fully capture the impact of rater subjectivity. Additionally, our proposed *RPC* model outperforms *IA-MT* by 2.97% in UAR and 2.00% in wF1 on IEMOCAP, suggesting that the highly diverse label space in *IA-MT* can introduce complexity in learning, potentially hindering overall SER performance. However, it should be noted that this approach is not suitable for the MSP-Podcast corpus due to the presence of variable rates or workers.

TABLE 3: Baseline and proposed model *RPC* (best performing over both PCC methods) results shown in overall (in UAR) and specific emotion (in Recall) for both corpora.

	IEMOCAP				MSP-Podcast		
	ST	Soft-L	IA-MT	RPC	ST	Soft-L	RPC
All	60.53	61.48	61.01	**63.92**	55.92	56.33	**57.95**
Neu	71.07	63.98	64.35	64.22	57.13	55.05	53.76
Hap	62.95	63.41	59.23	**65.31**	61.56	55.92	**65.02**
Ang	48.34	61.58	65.46	**67.04**	56.03	61.14	**66.11**
Sad	59.73	56.95	55.11	59.13	48.98	53.27	54.01

7.1.2 Proposed Method Evaluation

To evaluate our proposed idea, this analysis encompasses investigations that consider the perceptions of Only-*MajP* and Only-*MinP* raters, as presented in Table 2. The results demonstrate that models utilizing Only-*MajP* exhibit higher UAR and wF1 performance on the IEMOCAP dataset, achieving 62.03% and 64.38% respectively, compared to the performance of Only-*MinP* model, which achieves 60.80% and 62.84% on IEMOCAP. However, when considering both *MajP* and *MinP* jointly (*MajP+MinP*), the UAR and wF1 performance improves to 63.92% and 65.98% on IEMOCAP. The similar trend we can see for MSP-Podcast corpus. These findings suggest that the minority perceptions (*MinP*) contain relevant emotional information that, when combined with the majority raters' perceptions (*MajP*), can enhance learning and ultimately improve the overall consensus-based SER performance. Additionally, the MMD loss function is employed to reduce the feature space of *MajP* and *MinP*. The results presented in Table 2 indicate that the *RPC* model outperforms the *MajP+MinP* model, with improvements of 1.27% and 0.64% in UAR, and 0.59% and 0.94% in wF1 for the IEMOCAP and MSP-Podcast datasets, respectively.

Table 3 shows the emotion-wise performance (in Recall %) comparison, indicating the notable improvements in the proposed *RPC* model's correct prediction of all emotions, particularly in *Happiness* and *Anger*. For IEMOCAP, there is a performance increase in *Happiness* by 2.36%, 1.90%, and 6.08% for *ST*, *Soft-L*, and *IA-MT*, respectively, while *Anger* shows improvements of 18.7%, 5.46%, and 1.58% for the same models. We have observed a similar pattern in the model performances across the MSP-Podcast corpus. Here, we can also observe that the *ST* model outperforms the *RPC* model for the *Neu* emotion, with recall scores of 71.07% and 64.22%, respectively. Similar trends are observed in other models (*Soft-L* and *IA-MT*) that integrate rater or label subjectivity. This performance variation highlights the nuanced influence of the modeling approach on specific emotional categories. It can be inferred that the *ST* model achieves higher recall for *Neu* by potentially generalizing well, benefiting from a straightforward definition of neutral

TABLE 4: Predicted emotion analysis (in %) for *MajP/MinP v/s. MV* in both corpus over 4 major emotions for proposed *RPC* SER model.

	IEMOCAP			MSP-Podcast		
	MajP	*MinP*	*MV*	*MajP*	*MinP*	*MV*
Neu	42.22	10.67	52.89	47.45	12.92	60.37
Hap	49.15	18.31	**67.46**	50.23	14.09	**64.22**
Ang	54.32	16.54	**70.36**	51.67	20.85	**72.52**
Sad	35.12	16.39	51.51	36.32	13.09	49.41

emotions. Conversely, the *RPC* model, with its consideration of subjectivity, exhibits a nuanced understanding but tends towards a more conservative prediction, impacting recall negatively.

7.2 *MajP/MinP* Perception Analyses

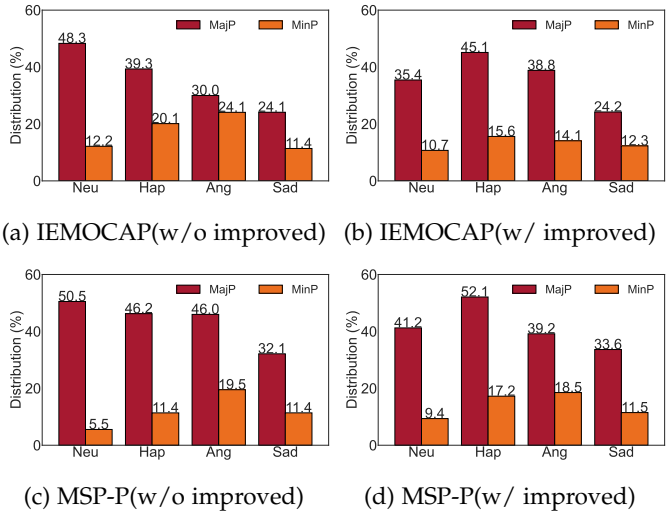
As demonstrated in the preceding Section 7.1, *RPC* outperformed the *non-RPC* baselines. To delve deeper into this concept, the subsequent analysis in this section compares the contributions of *MajP* and *MinP* across various emotions.

7.2.1 MajP/MinP v/s. MV

To obtain a comprehensive understanding of the contribution of rater PCC clusters (*MajP* and *MinP*) to the proposed method, we conduct an analysis of the prediction commonalities between *MajP* and *MinP* using *MV* (all from *RPC*). Only samples that are correctly predicted are considered for this analysis. Table 4 presents the distribution of correctly predicted samples by the *RPC* model across the corpora (in %). The analysis reveals that both *MajP* and *MinP* contribute to the *MV* predictions, with *MinP* exhibiting particular usefulness. In the case of *Happiness*, *Anger*, and *Sadness* emotions, *MinP* demonstrated commonalities of 18.31%, 16.54%, and 16.39% in IEMOCAP, respectively. Moreover, the analysis in Table 4 highlights that there are more commonalities observed for *Happiness* with 67.46% and *Anger* with 70.36%, while significant improvement is seen in *Sadness* as well with 51.51% for IEMOCAP corpus. These findings align with the recall scores for *Anger* with 67.04% and *Happiness* with 65.31% presented in Table 3 for the respective corpora. Similar patterns of performance trends are also observable in the MSP-Podcast corpus. These observations suggest that learning the emotional subjectivity of raters through the PCC approach, which favors the *MV* and specifically incorporates the perceptions of minority raters (*MinP*), contributes significantly to the learning process. Overall, the analysis shows that *Anger* and *Happiness* exhibit greater contributions, potentially attributed to the consideration of fine-grained distinguishable emotions such as *Excitement* and *Frustration*. Notably, *Frustration* is present in the *Anger* profile in Fig. 3.

7.2.2 MajP/MinP v/s. ST

In the previous section, we analyzed the contributions of *MajP/MinP* in the *RPC* model alongside *MV*. Now, we proceed with an analysis to explore the prediction commonalities between *MajP/MinP* of the *RPC* model and the single-task *ST* model. This examination aims to understand how much *MajP/MinP* contributes to the performance of



(a) IEMOCAP(w/o improved) (b) IEMOCAP(w/ improved)

(c) MSP-P(w/o improved) (d) MSP-P(w/ improved)

Fig. 10: Comparative analysis of predicted emotions for *MajP/MinP v/s. ST* is performed on both the IEMOCAP and MSP-Podcast datasets, considering the cases with and without enhanced predictions.

RPC compared to the *ST* case. To analyse this we focus on two sets of samples: "w/o improved" samples, which have similar correct predictions in both models, and "w/ improved" samples, which have correct predictions in the proposed *RPC* model but not in the *ST* model. The results, illustrated in Fig. 10, depict the prediction commonalities of *MajP* and *MinP* for major emotions in relation to the *ST* model predictions. Overall, each emotion category receives contributions from both *MajP* and *MinP*. However, the *Anger* and *Sadness* emotions exhibit the least difference in *MajP-MinP* contributions. Specifically, over the "w/o improved" samples, the *MajP-MinP* difference is 8.92% and 12.73% for IEMOCAP, and 26.47% and 20.75% for MSP-Podcast. For the "w/ improved" samples, the *MajP-MinP* difference is 16.75% and 11.89% for IEMOCAP, and 20.65% and 22.16% for MSP-Podcast. These results indicate that the contributions of *MajP* and *MinP* are higher in the "w/ improved" samples compared to the "w/o improved" case. Based on these experimental findings, It can be inferred that incorporating rater emotional subjectivity is a significant factor in improving SER. In addition, we argue that it is essential to consider both label ambiguity and raters' ambiguity jointly in learning because ambiguity in the annotation processes is inherently conditioned on individual rater ambiguity and emotion definitions.

7.3 Existing Literature Comparison

In earlier sections, we showcased the efficacy of *RPC*. Now, our attention shifts to a comparison with established methods outlined in Fig. 1 of Section 2. We emphasize that our method strategically positions itself between the extremes of label and rater ambiguity, contrasting with approaches solely focusing on one aspect. This comparison aims to underscore the superior performance of *RPC*, attributed to its controlled integration of rater subjectivity. In this section, we conduct two types of analyses. Firstly, we compare the performance on emotion-specific levels. Secondly, consid-

TABLE 5: Analysis of correct prediction accuracy (correctly predicted positive samples over the total true positive samples in %) for proposed model *RPC* and the considered models (*Soft-L* and *IA-MT*) on the scale of *Ambiguity* depicted in Figure 1 for both corpora.

	IEMOCAP				MSP-Podcast	
	RPC ↔ Soft-L		RPC ↔ IA-MT		RPC ↔ Soft-L	
Neu	61.47	58.29	61.47	58.08	59.33	56.9
	↑ 3.18		↑ 3.39		↑ 2.43	
Hap	68.39	64.03	62.39	68.01	60.32	56.48
	↑ **4.36**		↑ **5.62**		↑ **3.84**	
Ang	62.04	56.14	68.04	64.02	61.22	57.51
	↑ **5.90**		↑ **4.02**		↑ **3.71**	
Sad	58.68	57.56	58.68	56.85	54.72	52.79
	↑ 1.12		↑ 1.83		↑ 1.93	

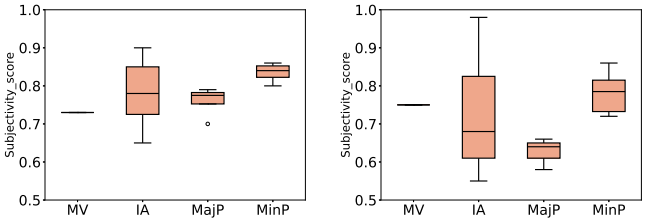
ering the controlled subjectivity aspect, we delve into the analysis of rater subjectivity.

7.3.1 Prediction Accuracy Analysis

To conduct this analysis, we analyze the correct prediction accuracy of *RPC* with two existing methods: *Soft-L* and *IA-MT* (Individual-rater Multi-Task). In this analysis, we evaluate the correct prediction accuracy by calculating the percentage of correctly predicted positive samples out of the total true positive samples. Table 5 provides the percentage increment in correct predictions achieved by our proposed *RPC* model compared to the extreme models (*Soft-L* and *IA-MT*) on the *Ambiguity* scale depicted in Fig. 1, considering both corpora. The results in Table 5 clearly demonstrate a significant improvement in prediction accuracy with our proposed *RPC* model compared to *Soft-L*, especially for the *Happiness* and *Anger* samples, with increments of 4.36% and 5.90% for IEMOCAP, and 3.84% and 3.71% for MSP-Podcast. Furthermore, when compared to the *IA-MT* model, our proposed *RPC* model also exhibits better prediction rates, achieving increments of 5.62% and 4.02%. These findings highlight the superiority of our proposed *RPC* model. Additionally, we observe the least improvement in the prediction accuracy for *Sadness* as compared to other emotions. For instance, in the case of IEMOCAP, there is an improvement of 1.12% and 1.83% for *Soft-L* and *IA-MT*, respectively. Similarly, in the case of MSP-Podcast, there is a 1.93% improvement in *Soft-L*. These findings are in line with the results presented in Section 7.2, which also highlight the impact of controlled rater ambiguity in training data on the prediction accuracy for *Sadness*.

7.3.2 Rater Subjectivity Analysis

The aforementioned observations highlight that integrating rater subjectivity into the learning process, as done in a controlled manner with *RPC*, enhances the performance of the *MV*-based SER system. To gauge the effectiveness of our *RPC* method in handling significant subjectivity, we introduce this analysis focusing on rater subjectivity across different methods. Here, subjectivity is characterized by lower agreement with other raters. To quantify this, we compute mean pairwise Cohen's kappa values (*Ck_value*) for each annotator with others and derive the subjectivity score as shown in Equation 6. For this analysis, we only choose methods that consider rater subjectivity in learning (only *IA-MT*) and compare with our *RPC* clusters. Fig. 11 shows methods on the scale of estimated subjectivity scores, providing insights into the level of subjectivity integration within each method. We have also presented the fixed subjectivity scores for corpora as illustrated in Fig. 11. These scores are estimated using the inter-rater agreement of the corpora to demonstrate no subjectivity integration in the context of *MV*.



(a) IEMOCAP (b) MSP-Podcast

Fig. 11: Depicts the integration of rater subjectivity in IEMOCAP and MSP-Podcast corpora through RPC clusters (MajP, MinP), MV, and individual annotation (IA) scenarios.

$$subjectivity\_score = 1 - Ck\_value \tag{6}$$

where Ck_value is the mean rater agreement using Cohen kappa values estimated for each rater.

In Fig. 11, it is evident that the consideration of individual rater subjectivity (*IA* for *IA-MT*) results in a high standard deviation of subjectivity scores. This signifies increased variability in subjectivity, potentially leading to complexity in learning. In contrast, our *RPC's* clusters, i.e., *MajP* and *MinP*, exhibit a well-controlled standard deviation compared to *IA*. In summary, the *MV*-based SER model, which excludes subjectivity, and the *IA-MT* model, covering a broad spectrum of subjectivity, represent two extremes. Our proposed *RPC* technique strikes a balance by integrating subjectivity in a controlled manner. Moreover, *IA-MT* methods are often challenging to implement or impractical for corpora with variable rater behavior, whereas *RPC* offers a more feasible solution.

7.4 RPC Analyses on Varying Coherence Threshold

To evaluate the impact of different rater selections on the effectiveness of *RPC*, we train the model using PCC clusters defined with various thresholds ranging from *MinP* to *MajP*. These clusters, denoted as PCC1, PCC2, PCC3, and PCC4, have thresholds of 20%, 40%, 60%, and 80%, respectively. By sorting the PC values scale, we allocate 20% of PC bins on the left side for *MinP* and 80% on the right side for *MajP*. To analyze the concept of clustering, we present the performance of our proposed PCC methodologies (RC and RC-GT) on the IEMOCAP and MSP-Podcast corpora in terms of UAR (%) and wF1 (%) in Table 6. The analysis reveals that the model's performance is only mediocre at the extremes, where either majority or minority perceptions are given more importance. For example, PCC1 shows UAR

TABLE 6: Performance table with different thresholding on proposed methodology (RPC) on IEMOCAP and MSP-Podcast corpora in terms of UAR (%) and wF1 (%); PCC1 is with threshold=20% (out of 100% of PC bins 20% left-side will consider for *MinP* and 80% of right-side will consider for *MajP*), PCC2, PCC3 and PCC4 with 40%, 60% and 80% respectively.

		IEMOCAP								MSP-Podcast							
		PCC1		PCC2		PCC3		PCC4		PCC1		PCC2		PCC3		PCC4	
		UAR	Wf1	UAR	Wf1	UAR	Wf1	UAR	Wf1	UAR	Wf1	UAR	Wf1	UAR	Wf1	UAR	Wf1
IRC	Only-*MajP*	61.16	63.07	61.27	63.63	60.34	63.22	59.02	61.25	52.37	53.36	53.71	54.19	52.12	53.05	50.55	51.21
	Only-*MinP*	54.16	56.06	54.24	57.15	53.55	56.40	51.81	55.18	51.43	52.27	51.01	52.89	52.17	53.33	49.33	51.53
	MajP+MinP	60.25	63.11	59.16	61.12	62.25	63.06	59.32	60.11	56.34	57.14	53.46	54.32	56.90	57.00	54.93	55.82
	RPC	61.21	61.56	**62.79**	**63.57**	**63.15**	**64.10**	59.15	61.03	56.67	58.33	**56.24**	**57.72**	**57.17**	**58.54**	55.88	56.35
RC-GT	Only-*MajP*	62.01	63.88	62.19	63.22	61.56	62.33	60.33	61.32	53.91	54.91	53.24	54.18	52.45	53.95	51.21	52.22
	Only-*MinP*	56.01	58.02	56.34	58.28	55.23	57.12	53.95	54.66	52.59	54.00	52.58	55.62	52.83	54.03	51.45	52.43
	MajP+MinP	62.50	64.14	62.74	63.99	62.23	63.59	60.12	61.93	56.68	67.45	53.34	55.17	56.98	57.81	54.21	56.01
	RPC	62.87	64.23	**63.98**	**65.06**	**63.19**	**66.76**	61.57	63.09	56.23	57.96	**57.19**	**58.90**	**58.13**	**60.11**	55.80	57.29

TABLE 7: Cluster consistency analysis for *MajP-MinP* and also with *MV* with a varying sliding threshold over *MinP* to *MajP*; *MA* and *MI* refers to *MajP* and *MinP*, respectively.

		IEMOCAP			MSP-Podcast		
		MV-MA	MV-MI	MA-MI	MV-MA	MV-MA	MA-MI
IRC	PCC1	0.27 ↑	0.05 ↓	0.31 ↓	0.24 ↑	0.19 ↓	0.27 ↓
	PCC2	0.18 ↑	0.10 ↓	0.45 ↑	0.25 ↑	0.11 ↓	0.36 ↑
	PCC3	0.14 ↓	0.27 ↑	0.42 ↑	0.12 ↓	0.24 ↑	0.47 ↑
	PCC4	0.12 ↓	0.28 ↑	0.33 ↓	0.06 ↓	0.23 ↑	0.29 ↓
RC-GT	PCC1	0.25 ↑	0.06 ↓	0.32 ↓	0.27 ↑	0.16 ↓	0.26 ↓
	PCC2	0.23 ↑	0.05 ↓	0.56 ↑	0.22 ↑	0.15 ↓	0.38 ↑
	PCC3	0.13 ↓	0.27 ↑	0.51 ↑	0.11 ↓	0.23 ↑	0.53 ↑
	PCC4	0.13 ↓	0.26 ↑	0.33 ↓	0.11 ↓	0.25 ↑	0.23 ↓

performances of 61.21% and 62.87% for IRC and RC-GT, respectively, in comparison to PCC2 with 62.79% and 63.98% of UAR, and PCC3 with 63.15% and 63.19% for IRC and RC-GT, respectively, for the IEMOCAP corpus. A similar trend is observed in the MSP-Podcast corpus models. These results indicate that the proposed method performs better when there is a balanced selection of *MajP* and *MinP* bins, as opposed to extreme prioritization of either group.

After observing the superior performance of balanced thresholded PCC in SER from Table 6, we also examine the consistency of clusters (*MajP-MinP*) and their correlation with the majority vote (*MV*). The coherence analysis with Cramer's correlation scores between the PCC labels and the consensus labels for all PCC combinations is presented in Table 7. The analysis reveals that as the threshold transitioned from *MinP* to *MajP*, the correlations between the *MV* labels and the corresponding clusters increased. For example, PCC1 exhibits stronger *MV-MajP* correlations for IEMOCAP (0.27 and 0.25) and MSP-Podcast (0.24 and 0.27) when utilizing both clustering methods (*IRC, RC-GT*), in comparison to *MV-MinP*, which shows lower correlations for IEMOCAP (0.05 and 0.06) and MSP-Podcast (0.19 and 0.16). Conversely, PCC2, PCC3, and PCC4 demonstrated different patterns, as shown in Table 7. Additionally, there is a higher level of coherence observed between the *MinP* and *MajP* clusters in PCC2 for IEMOCAP (0.45 and 0.56) and MSP-Podcast (0.36 and 0.38), as well as in PCC3 for IEMOCAP (0.42 and 0.51) and MSP-Podcast (0.47 and 0.53). It is worth noting that the *RC-GT* method has stronger correlations with the clusters compared to the *IRC* method. These findings align with the results of the SER modeling analysis presented in Table 6.

TABLE 8: Comparison of the *RPC* method in combination with other state-of-the-art deep learning techniques. The *Our* and *Our'* model shows the backbone architecture utilized in this study. Here, models denoted with (') represent the *RPC* models, while those without (') represent the *non-RPC* models.

	IEMOCAP		MSP-Podcast		BIIC-Podcast	
	UAR	wF1	UAR	wF1	UAR	wF1
Our	60.53	63.01	55.92	57.01	56.33	57.01
BiGRU [52]	62.02	64.15	56.38	57.56	58.33	60.34
BiLSTM [53]	62.33	63.10	57.40	59.25	59.44	61.02
CNN-BLSTM [33]	61.38	64.21	57.28	59.33	58.21	58.34
ConvLSTM [54]	62.02	65.73	56.90	58.41	59.38	60.05
Our'	**63.92** **	65.98	**57.95** **	60.27	**60.11** **	61.13
BiGRU'	**63.96** *	65.14	**57.35** *	59.37	**61.04** *	62.82
BiLSTM'	**64.20** **	66.83	**58.68** **	60.01	**61.65** **	63.07
CNN-BLSTM'	**63.12** **	66.04	**58.21** **	60.28	**60.53** *	61.45
ConvLSTM'	**64.38** *	67.65	**59.52** **	60.45	**61.92** **	62.90

7.5 Synergies with SOTA Deep-Learning Methods

From the preceding sections which utilize a consistent and a simple backbone structure shown in Section 5.1 for all scenarios, it is evident that our proposed RPC method exhibits significant effectiveness compared to existing strategies for handling label and rater ambiguity. We have further expanded our investigation by incorporating different deep-learning state-of-the-art (SOTA) techniques to examine how *RPC* performs in conjunction with the different backbone of deep-learning methods. We have now implemented some of the best-performing deep-learning strategies from the SER literature [33], [52]–[54] to assess their performance when combined with our *RPC* approach. These methods are evaluated both conventionally (only-*MV*) and integrated with our *RPC* method. The performances of these models across various corpora are presented in Table 8. The results from Table 8 clearly demonstrate that *RPC* versions consistently outperform their *non-RPC* counterparts. For instance, *BiGRU'* surpasses *BiGRU* by 1.94% in terms of UAR. Similar performance gains are observed in other *RPC* models as highlighted in Table 8. We also assess the statistical significance of these performance differences. We conduct paired t-tests with each *RPC* and *non-RPC* model performance differences, the results of this statistical test are annotated in Table 8 using asterisks (* for $p < 0.1$, ** for $p < 0.05$). These statistical test results shown in Table 8 affirm that the

observed performance differences are indeed significant.

Additionally, since we exclusively evaluated our RPC on two corpora (IEMOCAP and MSP-Podcast), we aim to assess its performance on another corpus with potentially diverse language characteristics. To conduct this analysis, we incorporate the BIIC-Podcast [55] corpus, which is publicly accessible and collected similarly to the MSP-Podcast but in the *Taiwanese Mandarin* language. The performances are shown in Table 8. We notice a consistent performance trend for this corpora too, better performance of *Our'* with *RPC* achieving 60.11% over *non-RPC Our* model obtaining 56.33% in UAR. This trend is evident with different deep-learning methods as well considered in Table 8. These results affirm that the *RPC* method consistently surpasses the baselines across all three corpora, underscoring the efficacy of our proposed *RPC* approach across diverse corpora.

8 Discussion and Conclusion

This study introduces a novel approach for improving SER by leveraging raters' emotional subjectivity through a rater-expanded label space using the PCC approach and multi-perception SER learning. The rater coherency is evaluated using two different viewpoints: inter-rater consistency (IRC) and rater consistency with ground truth (RC-GT). Our proposed SER approach rater perception coherency-based (*RPC*) model, outperforms other methods under the same setting (raters' side), achieving a 3.39% and 2.03% improvement in UAR over the *ST* (consensus) model for the IEMOCAP and MSP-Podcast corpora, respectively. This demonstrates the importance of incorporating raters' ambiguity as an emotional subjectivity and the value of learning in a rater-expanded label space for better SER.

This approach addresses some of the limitations of previous approaches, such as the ability to deal with missing labels by centering overall perception labels around majority raters' perception (*MajP*) and minority raters' perceptions (*MinP*), and the reliability to work for large corpora with a variable number of raters. Additionally, our method enables learning in a diverse space with a limited branch in multi-task learning, reducing the complexity and heaviness of the model and leading to better convergence in SER. In future work, we will explore how more efficiently this approach can be extended to larger corpora with a highly variable number of raters and also, explore other efficient methods to perceptually cluster the rater according to their perceptual behavior. Another avenue of exploration is devising strategies to mitigate missing label issues in any branch, potentially incorporating dynamic branch activation to address such scenarios.

Acknowledgements

This work was supported by the National Science and Technology Council Taiwan under Grants 110-2634-F-002-050 and 110-2221-E-007-067-MY3.

References

[1] Laurence Devillers, Christophe Vaudable, and Clément Chastagnol, "Real-life emotion-related states detection in call centers: a cross-corpora study," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[2] Ashish Tawari and Mohan Trivedi, "Speech based emotion classification framework for driver assistance system," in *2010 IEEE Intelligent Vehicles Symposium*. IEEE, 2010, pp. 174–178.

[3] Jaime Cesar Acosta, "Using emotion to gain rapport in a spoken dialog system," 2009.

[4] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[5] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa, "Speech emotion recognition using spectrogram & phoneme embedding.," in *Interspeech*, 2018, pp. 3688–3692.

[6] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[7] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016, pp. 566–570.

[8] Huang-Cheng Chou and Chi-Chun Lee, "Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5886–5890.

[9] Xin Kang, Xuefeng Shi, Yunong Wu, and Fuji Ren, "Active learning with complementary sampling for instructing class-biased multi-label text emotion classification," *IEEE Transactions on Affective Computing*, 2020.

[10] Atsushi Ando, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, and Yushi Aono, "Speech emotion recognition based on multi-label emotion existence model.," in *INTERSPEECH*, 2019, pp. 2818–2822.

[11] Yelin Kim and Jeesun Kim, "Human-like emotion recognition: Multi-label learning from noisy labeled audio-visual expressive speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5104–5108.

[12] Rod A Martin, Glen E Berry, Tobi Dobranski, Marilyn Horne, and Philip G Dodgson, "Emotion perception threshold: Individual differences in emotional sensitivity," *Journal of Research in Personality*, vol. 30, no. 2, pp. 290–305, 1996.

[13] Michael S Humphreys and William Revelle, "Personality, motivation, and performance: a theory of the relationship between individual differences and information processing.," *Psychological review*, vol. 91, no. 2, pp. 153, 1984.

[14] Nicola Dibben, Eduardo Coutinho, José A Vilar, and Graciela Estévez-Pérez, "Do individual differences influence moment-by-moment reports of emotion perceived in music and speech prosody?," *Frontiers in behavioral neuroscience*, p. 184, 2018.

[15] M Joseph Sirgy, "A review of "stumbling on happiness" authored by daniel gilbert: New york: Vintage books, 2005, isbn: 978-1-4000-7742-2," *Applied Research in Quality of Life*, vol. 2, pp. 141–143, 2007.

[16] Yali Wang, Chenyu Shangguan, Chuanhua Gu, and Biying Hu, "Individual differences in negative emotion differentiation predict resting-state spontaneous emotional regulatory processes," *Frontiers in Psychology*, vol. 11, pp. 576119, 2020.

[17] Jeanne L Tsai, Brian Knutson, and Helene H Fung, "Cultural variation in affect valuation.," *Journal of personality and social psychology*, vol. 90, no. 2, pp. 288, 2006.

[18] Cassie Mogilner, Sepandar D Kamvar, and Jennifer Aaker, "The shifting meaning of happiness," *Social Psychological and Personality Science*, vol. 2, no. 4, pp. 395–402, 2011.

[19] Lisa Feldman Barrett, "Discrete emotions or dimensions? the role of valence focus and arousal focus," *Cognition & Emotion*, vol. 12, no. 4, pp. 579–599, 1998.

[20] Margaret M Bradley and Peter J Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," Tech. Rep., Technical report C-1, the center for research in psychophysiology . . . , 1999.

[21] Yasemin Erbas, Eva Ceulemans, Madeline Lee Pe, Peter Koval, and Peter Kuppens, "Negative emotion differentiation: Its personality and well-being correlates and a comparison of different assessment methods," *Cognition and emotion*, vol. 28, no. 7, pp. 1196–1213, 2014.

[22] Hannah K Lennarz, Anna Lichtwarck-Aschoff, Marieke E Timmerman, and Isabela Granic, "Emotion differentiation and its relation

with emotional well-being in adolescents," *Cognition and Emotion*, vol. 32, no. 3, pp. 651–657, 2018.

[23] Huang-Cheng Chou, Wei-Cheng Lin, Chi-Chun Lee, and Carlos Busso, "Exploiting annotators'typed description of emotion perception to maximize utilization of ratings for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)(under review)*, 2022.

[24] Hassan Hayat, Carles Ventura, and Agata Lapedriza, "Recognizing emotions evoked by movies using multitask learning," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[25] Atsushi Ando, Takeshi Mori, Satoshi Kobashikawa, and Tomoki Toda, "Speech emotion recognition based on listener-dependent emotion perception models," *APSIPA Transactions on Signal and Information Processing*, vol. 10, 2021.

[26] 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*, vol. 42, no. 4, pp. 335–359, 2008.

[27] Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[28] Yu-Ting Lan, Wei Liu, and Bao-Liang Lu, "Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–6.

[29] Stefano Pini, Olfa Ben Ahmed, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, and Benoit Huet, "Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 536–543.

[30] Woan-Shiuan Chien and Chi-Chun Lee, "Achieving fair speech emotion recognition via perceptual fairness," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[31] Alexey Tarasov, Sarah Jane Delany, and Charlie Cullen, "Using crowdsourcing for labelling emotional speech assets," in *W3C workshop on Emotion ML*, 2010.

[32] Jae-Bok Kim, Jeong-Sik Park, and Yung-Hwan Oh, "Speaker-characterized emotion recognition using online and iterative speaker adaptation," *Cognitive Computation*, vol. 4, pp. 398–408, 2012.

[33] Yuanchao Li, Tianyu Zhao, Tatsuya Kawahara, et al., "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning.," in *Interspeech*, 2019, pp. 2803–2807.

[34] Anish Nediyanchath, Periyasamy Paramasivam, and Promod Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7179–7183.

[35] Jeng-Lin Li and Chi-Chun Lee, "Attentive to individual: A multimodal emotion recognition network with personalized attention profile.," in *Interspeech*, 2019, pp. 211–215.

[36] Boaz M Ben-David, Sarah Gal-Rosenblum, Pascal HHM van Lieshout, and Vered Shakuf, "Age-related differences in the perception of emotion in spoken language: The relative roles of prosody and semantics," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 4S, pp. 1188–1202, 2019.

[37] Jianwu Dang, Aijun Li, Donna Erickson, Atsuo Suemitsu, Masato Akagi, Kyoko Sakuraba, Nobuaki Minematsu, and Keikichi Hirose, "Comparison of emotion perception among different cultures," *Acoustical science and technology*, vol. 31, no. 6, pp. 394–402, 2010.

[38] Georgios N Yannakakis, Roddy Cowie, and Carlos Busso, "The ordinal nature of emotions," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 248–255.

[39] Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah, "A novel markovian framework for integrating absolute and relative ordinal emotion information," *IEEE Transactions on Affective Computing*, 2022.

[40] Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima, and Yushi Aono, "Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4964–4968.

[41] Reza Lotfian and Carlos Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 415–420.

[42] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al., "Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021.

[43] Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer Dy, "Modeling annotator expertise: Learning when everybody knows a bit of something," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 932–939.

[44] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy, "Learning from crowds.," *Journal of machine learning research*, vol. 11, no. 4, 2010.

[45] Pablo Morales-Álvarez, Pablo Ruiz, Raúl Santos-Rodríguez, Rafael Molina, and Aggelos K Katsaggelos, "Scalable and efficient learning from crowds with gaussian processes," *Information Fusion*, vol. 52, pp. 110–127, 2019.

[46] Trevor Cohn and Lucia Specia, "Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 32–42.

[47] Filipe Rodrigues and Francisco Pereira, "Deep learning from crowds," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[48] Christopher M Bishop and Nasser M Nasrabadi, *Pattern recognition and machine learning*, vol. 4, Springer, 2006.

[49] Hassan Hayat, Carles Ventura, and Agata Lapedriza, "Modeling subjective affect annotations with multi-task learning," *Sensors*, vol. 22, no. 14, pp. 5245, 2022.

[50] Jacob Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

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

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

[53] Shouyan Chen, Mingyan Zhang, Xiaofen Yang, Zhijia Zhao, Tao Zou, and Xinqi Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, pp. 7530, 2021.

[54] Mustaqeem and Soonil Kwon, "Clstm: Deep feature-based speech emotion recognition using the hierarchical convlstm network," *Mathematics*, vol. 8, no. 12, pp. 2133, 2020.

[55] Shreya G Upadhyay, Woan-Shiuan Chien, Bo-Hao Su, Lucas Goncalves, Ya-Tse Wu, Ali N Salman, Carlos Busso, and Chi-Chun Lee, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," .

Shreya G. Upadhyay (S'23) is currently pursuing a PhD degree in Electrical Engineering at National Tsing Hua University (NTHU), Taiwan. She obtained her BE degree in computer engineering from Mumbai University, India in 2013, and her MTech degree in computer engineering from K. J. Somaiya College of Engineering, India in 2018. Her research interests include behavioral speech signal processing, speech emotion recognition, automatic speech recognition, and acoustic sound event detection. She is a student member of ISCA, EURASIP, AAAC, and SPS.

Woan-Shiuan Chien (S'23) currently is a PhD student at the Electrical Engineering (EE) Department of National Tsing Hua University (NTHU), Taiwan. She received her B.S. degree in electrical engineering from Chung Yuan Christian University, Taiwan in 2015 and her M.S. degree in electrical engineering from the National Chung Cheng University (CCU), Taiwan in 2016. Her research interests are in human-centered behavioral signal processing and automatic speech emotion recognition. She was the recipient of the Outstanding Doctoral Students Program sponsored by the Taiwan Science and Technology Council (NSTC) (2022), and the travel grant sponsored by IEEE Signal Processing Society (2023) and the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) (2023). She is a Student Member of the AAAC, ACM, ACLCLP, and IEEE Signal Processing Society.

Bo-Hao Su (S'20) is currently pursuing his Ph.D. degree and received his B.S. degree in the Department of Electrical Engineering at National Tsing Hua University, Taiwan in 2017. He was awarded with NTHU Principal Outstanding Student Scholarship (2018 - 2022), and the Interspeech 2018 Sub Challenge Championship. His research field includes behavioral signal processing (BSP), cross-corpus speech emotion recognition, and machine learning. He is also a student member of ISCA.

Chi-Chun Lee (M'13, SM'20) is a Professor at the Department of Electrical Engineering of the National Tsing Hua University (NTHU), Taiwan. He received his B.S. and Ph.D. degrees both in Electrical Engineering from the University of Southern California, USA in 2007 and 2012. His research interests are in speech and language, affective computing, health analytics, and behavioral signal processing. He is an associate editor for the IEEE Transaction on Affective Computing (2020-), the IEEE Transaction on Multimedia (2019-2020), the Journal of Computer Speech and Language (2021-), the APSIPA Transactions on Signal and Information Processing and a TPC member for APSIPA IVM and MLDA committee. He serves as the general chair for ASRU 2023, an area chair for Interspeech 2016, 2018, 2019, senior program committee for ACII 2017, 2019, publicity chair for ACM ICMI 2018, late breaking result chair for ACM ICMI 2023, sponsorship and special session chair for ISCSLP 2018, 2020. He is the recipient of the Foundation of Outstanding Scholar's Young Innovator Award (2020), the CIEE Outstanding Young Electrical Engineer Award (2020), the IICM K. T. Li Young Researcher Award (2020), the NTHU Industry Collaboration Excellence Award (2021), and the MOST Futuretek Breakthrough Award (2018, 2019). He led a team to 1st place in the Emotion Challenge in Interspeech 2009, and with his students won 1st place in the Styrian Dialect and Baby Sound subchallenge in Interspeech 2019. He is a co-author on the best paper award/finalist in Interspeech 2008, Interspeech 2010, IEEE EMBC 2018, Interspeech 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in Journal of Speech Communication. He is also an ACM and ISCA member.