國 立 清 華 大 學

博 士 論 文

重新探討語音情緒辨識的建模與評量方法：考慮標註者的主觀性與情緒的模糊性

Revisiting Modeling and Evaluation Approaches in Speech Emotion Recognition: Considering Subjectivity of Annotators and Ambiguity of Emotions

系所別：電機工程學系博士班　　組別: 系統A

學號姓名：104061701　周惶振 (Huang-Cheng Chou)

指導教授：李祈均　　博士 (Prof. Chi-Chun Lee)

中 華 民 國 一一三 年 七 月

摘要

語音情緒辨識在過去二十年裡獲得了越來越多的關注。建立語音情緒識別系統需要情緒數據庫，資料庫需要有人聲以及人類的情緒感受標記。研究人員們會訓練群眾標記者或內部標記者，在收聽或觀看情緒錄像後，通過選擇預先定義的情緒類別來描述和提供他們的情緒感知。然而，當研究人員們要求標記者們從預定義的情緒中選擇情緒時，觀察到標記者之間出現分歧是很常見的。爲了處理這種標記者之間的分歧，大部分專家學者們將分歧視爲雜訊，並使用標籤聚合方法來獲得單一的共識情緒標記，作爲訓練語音情緒識別系統的學習目標。雖然這種通行做法將任務簡化爲單一情緒標籤識別任務，但這個方法忽略了人類情緒感知的自然行爲。在本論文中，我們主張應重新檢視語音情感識別中的建模和評估方法。主要的研究問題是：(1) 我們是否應該移除少數的情感評分？(2) 我們是否應該只有讓語音情感識別系統學習少數人的情感感知？(3) 語音情感識別系統是否應該每次只預測一種情緒類別？

從心理學領域相關研究成果發現情緒感受是主觀的，每個個體對同一情感刺激的情緒感受可能有所不同。此外，人類感知中的情緒類別界限是重疊、混合且模糊的。這些情感的模糊性和情緒感受的主觀性的發現啓發我們重新審視在語音情緒辨識中的建模與評估方法。本博士論文探討了構建語音情緒識別系統的三個層面的新穎觀點。首先，我們接受情緒感受的主觀性，並考慮標記者的所有情緒標記。傳統的方法只允許每位標記者對每個樣本給予一票情緒標記，但我們藉由考慮所有標記者的所有標記，利用現有的軟標籤方法(soft-label)重新計算標籤表示方式。此外，我們直接利用個別標記者的情緒標記來訓練個別標記者的語音情緒識別系統，並聯合訓練個別標記者語音情緒識別系統和標準語音情緒識別系統(使用共識標籤)。在使用多數決所獲得的共識標籤當作最終情緒標籤進行測試時，個別標記者的建模方法提升了語音情緒辨識系統的性能。

其次，我們重新思考了評估語音情緒辨識系統的方法以及語音情緒辨識任務的制定和定義。我們主張在評估語音情緒辨識系統的性能時，不應該刪除任

何數據和情緒標記。此外，我們認爲語音情緒辨識任務的定義可以包含情緒的共現性（例如，悲傷和生氣）。因此，樣本的眞實標籤不應該是單一情緒標籤，而應該是包含更多情緒感知多樣性的分佈式標籤。我們提出了一種新的標籤聚合規則，稱爲「全包容規則」(all-inclusive rule)，用於選擇訓練集和測試集的數據和其情緒標記。對四個公開英文情緒數據庫的結果表明，使用「全包容規則」方法決定的訓練集所訓練的語音情緒辨識系統，在各種測試條件下，其性能優於使用傳統方法，包括絕對多數決和相對多數決訓練的語音情緒辨識系統。

最後但同樣重要的是，我們受到心理學研究關於情緒共現性的發現啓發。我們根據情緒資料庫訓練集中情緒標記來估計情緒共現性的頻率，並基於每種情緒類別的數量對矩陣進行標準化。接著，我們使用單位矩陣減去標準化矩陣作爲懲罰矩陣。我的想法是在訓練過程中當模型預測到罕見的情緒共現時對語音情緒辨識系統進行懲罰。因此，懲罰矩陣被整合進現有的目標函數，例如交叉熵損失(cross-entropy)。在最大的英語情緒數據庫結果顯示，即使在單一情緒標籤測試條件下，懲罰矩陣也提升了語音情緒辨識系統的性能。

致謝

我要向那些在我博士旅程中支持和指導我的人表達我眞誠感謝。我的指導教授，李祈均博士，從大學專題開始指導我，從2014年到2024年，老師的專業知識、寶貴的見解和不懈的支持對這篇博士論文有著非常大的貢獻。祈均老師的指導和支持讓我可以順利完成這篇博士論文以及旅程。

我也向我的博士學位考試委員會的委員們，林嘉文教授、馬席彬教授、陳柏琳教授、冀泰石教授，以及王新民博士，感謝各位委員們抽空參與，並提供建設性的回饋與建議，以及投入時間審閱我的博士論文。也謝謝指導過我的Carlos Busso教授、李宏毅教授、劉奕汶教授、Albert Ali Salah教授、阮大成博士，和Alexander Visheratin博士，他們不僅是合作者，還是我的指導員；有他們的帶領和建議，讓我有不斷解決問題的勇氣和想法。

在七年的博士生涯，有參與過數次的國際研討會，而「傑出人才發展基金會」與「計算語言學與中文語言處理學會」在我就讀博士班期間提供慷慨資金與鼓勵。在他們的財務支持下，這項工作得以完成。我同樣感謝聯詠博士獎學金、國家科學技術委員會博士生赴國外計畫、Google東亞學生旅行獎勵、中華扶輪獎學金以及國立清華大學校長博士生卓越獎學金；這些獎學金讓我能夠專心研究而無財務之憂。

我也感謝ACII 2017和國際口語溝通學會獎勵（INTERSPEECH 2022），讓我能夠在國際會議上親自到現場介紹我們的研究，這些機會提供了寶貴的經驗和新的視角。同時，能夠在享有盛譽的期刊上發表我的研究成果，包括APSIPA《Transactions on Signal and Information Processing》及IEEE《Transactions on Affective Computing》期刊，也是一個非常有價值的經歷，也謝謝ISCA Student Advisory Committee，讓我可以成爲委員，爲國際會議貢獻，也讓我認識到非常多來自世界各地和各領域的專家和學者們。

在此，我想向我的家人們、所有實驗室的夥伴們、朋友們（林維誠、陳思睿、Seong-Gyun Leem、Lucas Goncalves、吳姿瑩、Ali N. Salman、張凱爲、林

羿成、吳海濱、任文澤、Andrea Vidal、許德丞、陳志杰和林旻萱）和人生中的教練和老師們(徐碩鴻老師、林靜枝老師、曹昱博士、張俊盛教授、徐桂平老師、吳德成教練、黃錫瑜教授、陳榮順老師、劉素貌教練、黃老師、蔡文祺老師、范光榮老師、楊硯茗教練、Richard Lee老闆、洪光燦教練、張炳煌教練和王禮章老師)表示感謝，感謝他們無微不至的支持和鼓勵。他們的耐心和理解讓這段旅程變得可承受且充實。感謝你們成爲這段旅程的一部分。

Abstract

Over the past twenty years, there has been a growing focus on speech emotion recognition (SER). To develop SER systems capable of identifying emotions in speech, researchers need to gather emotional databases for training purposes. This process involves training crowdsourced raters or in-house annotators to express their emotional responses after experiencing emotional recordings by selecting from a set list of emotions. Nevertheless, it is common for raters to disagree on emotion selection from these predefined categories. To address this issue, many researchers consider such disagreements noise and apply label aggregation techniques to produce a unified consensus label, which serves as the target for training SER systems. While the common practice simplifies the task as a single-label recognition task, it ignores the natural behaviors of human emotion perception. In this dissertation, we contend that we should revisit the modeling and evaluation approaches in SER. The driving research questions are (1) Should we remove the minority of emotional ratings? (2) Should we only let the SER systems learn the emotional perceptions of a few people? (3) Should SER systems only predict one emotion per speech?

Based on the findings of psychological studies, emotion perception is subjective. Each individual could have varying responses to the same emotional stimulus. Additionally, boundaries of emotions in human perception are overlapped, blended, and ambiguous. Those ambiguities of emotions and subjectivity of emotion perceptions inspire us to revisit modeling and evaluation approaches in SER. This dissertation explores novel perspectives on three main views of building SER systems. First, we embrace the subjectivity of emotional perception and consider every emotional rating from annotators. Also, the conventional approach only allows each rater to provide one vote for each sample. Still, we re-calculate label representation in the distributional format with the existing soft-label method by considering all ratings from all raters. Moreover, we

directly utilize ratings of individual annotators to train SER systems and jointly train the individual SER systems and the standard SER systems. The modeling of individual annotators improves the performances of SER systems on the test sets with the consensus labels obtained by the majority vote.

Secondly, we rethink the determination of methods to evaluate SER systems and the formulation and definition of the SER task. We argue that we should not remove any data and emotional ratings when assessing the performances of SER systems. Also, we think the definition of SER task can have a co-occurrence of emotions (e.g., sad and angry). Therefore, the ground truth of samples should not be the one-hot single label, and it can be distributional to include more diversity of emotion perception. We propose a novel label aggregation rule, named the "all-inclusive rule," to use all data and include the maximum emotional rating for the test set. The results across 4 public English emotion databases show that the SER systems trained by the train set decided by the proposed method outperformed the ones trained by the conventional techniques, including majority rule and plurality rule on the various testing conditions.

Finally, we draw inspiration from psychological research on the co-occurrence of emotions. We assess the frequency with which different categorical emotions occur together, using emotional ratings from the training data of emotion databases. This matrix is then normalized, considering the frequency of each emotion class. We derive a penalization matrix by subtracting the normalized matrix from an identity matrix. We aim to apply penalties to SER systems during training when they predict rarely occurring combinations of emotions. This penalization matrix is integrated into objective functions like cross-entropy loss. The findings from the largest English emotion database indicate that using the penalization matrix enhances the performance of SER systems, even under single-label testing conditions.

With the extensive results, we conclude that (1) we should involve the minority of emotional ratings instead of removing them to build better-performance SER systems,

(2) we should consider emotional ratings from more people instead of fewer people during training SER systems to get better-performance SER systems; (3) we should allow SER systems to predict multiple emotions to handle the possibility of co-occurring emotions in the real-life scenarios. In future work, we plan to investigate training emotion recognition systems with multi-modalities (e.g., video, text, and audio) to process signals to improve the performance of SER systems. Also, we are interested in the relationship between the number of training human-labeled data and the performances of SER systems. Furthermore, we aim to understand the performance bias in the demographic groups, such as gender, race, and age. Last but not least, we plan to build a multi-lingual emotion recognition system.

Acknowledgements

I am deeply grateful to everyone who has supported and guided me during my Ph.D. journey. I extend special thanks to my advisor, Professor Chi-Chun Lee, who has been a guiding force since my undergraduate years, from 2014 to 2024. Professor Lee's expertise, invaluable insights, and unwavering support have been instrumental in completing this doctoral dissertation. His guidance has played a crucial role in the successful conclusion of my dissertation and my entire Ph.D. journey.

I am also profoundly grateful to the members of my dissertation committee, Professor Chia-Wen Lin, Professor Hsi-Pin Ma, Professor Berlin Chen, Professor Tai-Shih Chi, and Dr. Hsin-Min Wang. Thank you all for taking the time to participate, providing constructive criticism, and reviewing my work. Big thanks go to my co-advisors, Professor Carlos Busso, Professor Hung-yi Lee, Professor Yi-Wen Liu, Professor Albert Ali Salah, Dr. Da-Cheng Juan, and Dr. Alexander Visheratin, who have not only been collaborators but also respected mentors. Their support has created a positive and motivating research environment.

I am indebted to the Foundation for the Advancement of Outstanding Scholarship and the Association for Computational Linguistics and Chinese Language Processing for the generous funding and resources provided during my research. With their financial support, this work was possible. I acknowledge the scholarships I received from the NOVATEK Fellowship and the National Science and Technology Council Ph.D. Students Study Abroad Program, Google East Asia Student Travel Grants, The Rotary Foundation Excellence Scholarship, and National Tsing Hua University Dean Ph.D. Student Excellence Scholarship: These scholarships allowed me to concentrate fully on my research without financial worry.

I am also thankful for the opportunities to present my work at international conferences, such as the ACII 2017 and International Speech Communication Association

Grants (INTERSPEECH 2022), which provided valuable feedback and new perspectives. Publishing my work in esteemed journals, including APSIPA Transactions on Signal and Information Processing and IEEE Transactions on Affective Computing, has also been an enriching experience.

I want to express my gratitude to my family, my lab-mates (BIICers), and friends (Wei-Cheng Lin, Szu-Jui Chen, Seong-Gyun Leem, Lucas Goncalves, Tz-Ying Wu, Ali N. Salman, Kai-Wei Chang, Yi-Cheng Lin, Haibin Wu, Wenze Ren, Andrea Vidal, Te-Cheng Hsu, Jhih-Jie Chen, and Min-Hsuan Lin), and the coaches and teachers in my life (Professor Shawn S. H. Hsu, Teacher Ching-Chin Lin, Dr. Yu Tsao, Professor Jason S. Chang, Teacher Kuei-Ping Hsu, Coach Te-Cheng Wu, Professor Shi-Yu Huang, Teacher Rong-Shun Chen, Coach Su-Mao (Tammy) Liu, Teacher Huang, Teacher Wen-Qi Tsai, Teacher Guang-Rong Fan, Coach Yan-Ming Yang, CEO Richard Lee, Coach Guang-Can (Michael) Hung, Coach Bing-Huang Zhang, and Teacher Li-Zhang Wang) for their support and encouragement. Their help and understanding have made this journey bearable and fulfilling. Thank you all for being part of this journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Recent research has led to the development of automated systems that perform tasks traditionally done by humans like automatic speech recognition [3] and speech translation [4]. Moreover, Figure 1.1 shows increasing research studies on speech emotion recognition (SER) to understand emotions from human voice from 2001 through 2023. In recent years, SER has been a potential indicator to predict customer satisfac-

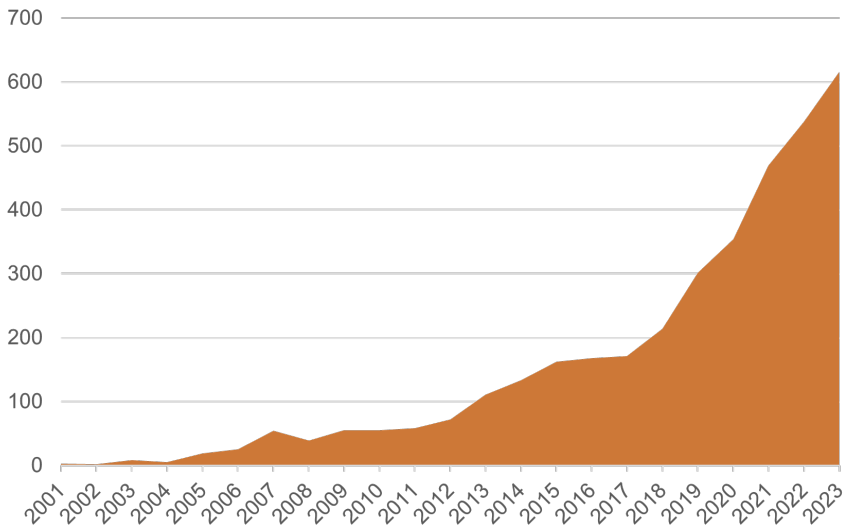


Figure 1.1: The figure illustrates the trends of papers whose title includes speech emotion recognition based on the Google Scholar website. The x-axis means years; the y-axis is the number of papers.

tion [5]. Moreover, SER is crucial for voice assistants as it enables them to generate speech with appropriate emotions by predicting the emotional tone of users' voices. I also observe some startups contribute to building SER solutions to improve users' experiences, such as HUME, BEHAVIORAL SIGNALS, UNIPHORE, and COGNOVI LABS. Those companies provide emotion-aware solutions for their clients to understand users' emotions and improve user experiences.

Additionally, SER needs interdisciplinary collaboration, such as studies of psychology, spoken language, and natural language understanding. Most researchers follow psychological studies to design and collect emotional corpus. In common practice, researchers provide pre-defined categorical emotions for raters to choose from after listening to or watching emotional stimuli. In most public emotion databases, each sample was annotated by at least 3 annotators. However, it is expected to observe disagreement among raters on the same sample. The prior SER studies utilize label aggregation methods, such as majority vote or plurality vote, to obtain a single consensus label as the ground truth to train SER systems. This common practice removes a minority of emotional perceptions and the data without a consensus label.

The recent findings of psychological studies [6, 7] reveal that the emotion perception

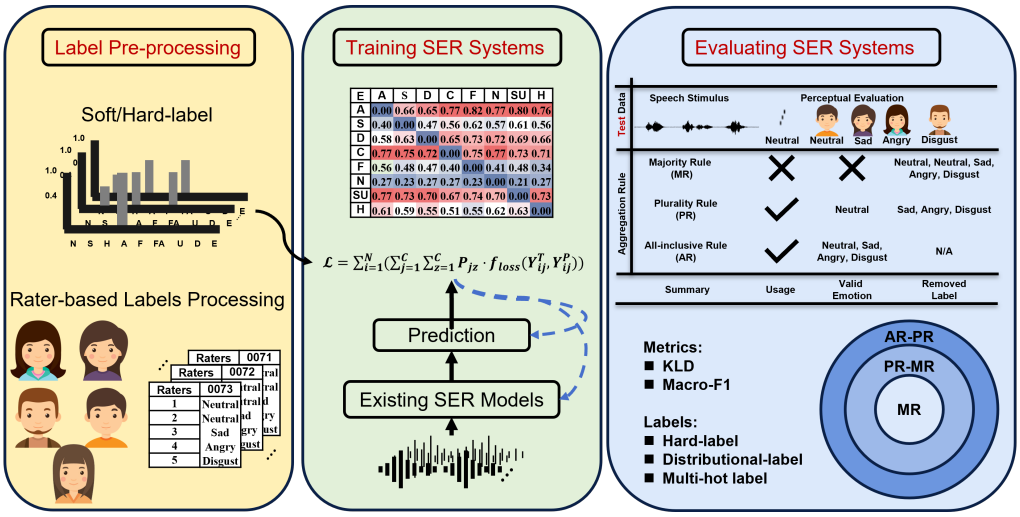


Figure 1.2: The figure illustrates three main contributions to modeling and evaluation approaches in SER systems.

of humans is high-dimensional, and boundaries of emotions among emotion perception of humans are blended and overlap. The critical findings inspire and motivate me to revisit standard SER systems' whole modeling and evaluation methods. We have come up with the following research questions to answer: (1) Should we remove those minority emotional ratings? (2) Should we only let SER systems learn the emotional perception of a few annotators? (3) Should SER systems only predict one single emotion for each sample? Those questions could be split into two main factors, the subjectivity of emotion perception and ambiguity of emotions, contributing to the disagreement of emotion perception among raters because of human bias, including gender [8], culture [9], and age [10]. This dissertation dives into the whole process of modeling and evaluating SER systems, and the entire process is threefold, as shown in Figure 1.2: (1) label preprocessing, (2) evaluation of SER systems, and (3) training of SER systems. We propose three novel methods to contribute to each process separately to improve the modeling and assessment of SER systems based on the existing SER frameworks.

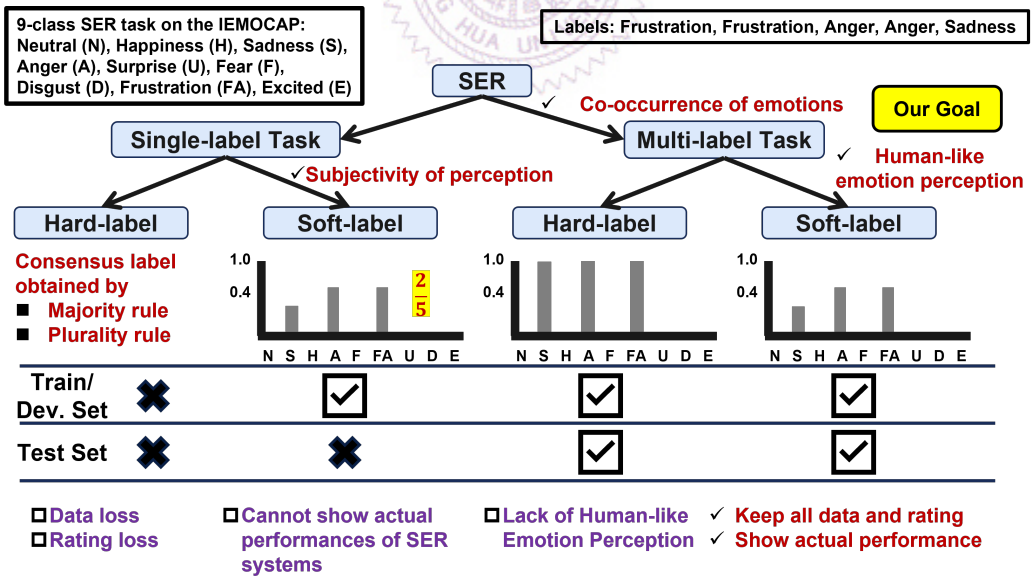


Figure 1.3: The figure illustrates the current trends of prior works on SER.

1.2 Background, Related Works, and Challenges

This section introduces an overview of the background and challenges of the prior SER studies.

1.2.1 Emotion Representations

There are two main ways to represent emotion perception. One is a dimensional attribute that assumes every attribute is independent of each other, like arousal [11] or valence [12]. The other one is categorical emotions [6,7], such as anger or sadness. This dissertation only focuses on categorical emotions since the perception of categorical emotions is better perceived across cultures than dimensional attributions [13].

1.2.2 Emotion Recognition Systems using Multi-/Uni-modality

Emotion can be expressed through various modalities, including facial expressions, body movements, vocal speech, and text. According to Cowen and Keltner [7], humans can recognize at least 27 distinct emotions from video, 24 from vocal expression, and 28 from facial and bodily cues. Furthermore, humans can discern approximately 13 emotions from music. Consequently, different modalities lead to varying emotional perceptions in humans.

Previous research has employed various modalities to develop emotion recognition systems. For example, Goncalves et al. [14] utilized audio-visual data to build such systems. Almedia et al. [15] focused on training systems to identify emotions from facial expressions. Additionally, studies [16, 17] have successfully detected emotions through music. This dissertation, however, is concerned solely with speech-based emotion recognition systems [18, 19].

1.2.3 Evaluation of SER Systems

Many research studies on SER primarily focus on predicting a single emotion, using methods like majority voting [1] or plurality rules [20] to derive a consensus label for evaluating SER systems. This approach often discards emotional ratings with minority opinions and data lacking a consensus label, resulting in a simpler test set. Consequently, SER system performance evaluations may only partially represent their true capabilities due to excluding some data and emotional ratings. Figure 1.3 showcases the trends in previous SER research, highlighting that many studies approach SER as a single-label task with hard-label (one-hot encoding) targets. For instance, in Figure 1.3, a sample from IEMOCAP was rated as frustration, frustration, anger, anger, and sadness. Since frustration and anger received equal votes, no consensus label was achieved, leading to the sample being removed from the training set (development set), resulting in data and emotional rating loss. Moreover, defining SER as a single-label task neglects the frequent co-occurrence of emotions in real-world situations [21]. Although some research considers the subjective nature of emotion perception by using all emotional ratings as soft labels for training [22], the "tie" samples are still excluded from the test set, so the evaluation of SER systems does not fully reflect their actual performance.

Few studies treat Speech Emotion Recognition (SER) tasks as multi-label tasks, contrary to other emotion recognition fields, which do. For example, in text emotion recognition [23, 24], image emotion classification [25], and audio-visual emotion recognition [26, 27], emotions are considered valid if they receive any vote. They use multi-hot labels, as illustrated in Figure 1.3. A close study by [28] on facial expression recognition calculated soft labels based on vote frequency for each emotion, converting these into binary vectors (either multi-hot or single-hot) based on the threshold $(1/(C-1)$, where C is the number of emotion classes). However, this misses key information about primary and secondary emotions. In label distribution learning [29], [30] used facial expression databases to collect emotion intensity scores from multiple annotators, av-

Table 1.1: Table summarizes the loss of data and emotion rating across the public emotion databases.

Label Aggregation Method	Majority Rule		Plurality Rule		**Our Goal**	
Emotion Database	Data	Rating	Data	Rating	Data	Rating
IMPROV (P)	9.18%	28.52%	4.63%	26.41%	0%	0%
CREMA-D	35.80%	52.96%	8.55%	40.57%	0%	0%
PODCAST (P)	44.81%	59.87%	19.85%	49.24%	0%	0%
IEMOCAP	31.37%	49.44%	25.32%	45.70%	0%	0%
IMPROV (S)	54.18%	76.91%	12.32%	56.70%	0%	0%
PODCAST (S)	92.01%	96.99%	33.72%	78.13%	0%	0%
Average	44.56%	60.78%	17.40%	49.46%	0%	0%

eraging these to normalize the distributional labels for system training and evaluation. Similarly, [31] did this for text emotion recognition. However, these methods are challenging to apply in SER due to the lack of emotional intensity scores, as SER databases typically ask raters to select pre-defined emotions. This is possible because querying raters on intensity could be overwhelming and bias-inducing [32, 33]. Our objective is to incorporate all data and emotional ratings in test sets to assess SER systems' performance accurately. We employ a soft-label format as ground truth, convertible to a binary vector to highlight traditional accuracy. Thus, we propose metrics to evaluate the distribution similarity between predictions and the actual data. This dissertation will report results using two such metrics.

1.2.4 Label Prepossessing for Training SER Systems

The most common approaches to obtain consensus labels are majority rule and plurality rule, whose definitions are introduced below. Table 1.1 summarizes the loss of data and emotional ratings using two conventional label aggregation methods, majority rule and plurality rule. We aim to try our best to retain all data and emotional ratings to train the SER systems.

- Majority Rule (MR): it selects one of the pre-defined emotion classes only if more than half of the votes select that class.
- Plurality Rule (PR): it selects a class if one emotional class obtains more votes

than others.

We argue that the two above conventional aggregation methods lose variations of emotion perception and the number of data during SER systems, leading to the poor ability to predict the samples that have co-occurrence of emotions. Therefore, we propose two ways to model the variations of emotion perception. (1) The first is to model individual raters' SER systems since each person has a different sensitivity to various emotions [34]. For instance, some raters are good at perceiving sad emotions; some can easily sense happy emotions. (2) The second approach proposes a novel label aggregation rule, named the "all-inclusive" rule, incorporating all the labeled data with the emotional ratings in the emotion databases.

1.2.5 Co-occurrence of Emotions

In everyday situations, emotions often occur simultaneously [21]. However, most previous research in SER does not account for the possibility of concurrent emotions like "contempt and anger," treating the task as one requiring a single label. Although soft-label or multi-label approaches can accommodate emotion co-occurrence, they do not effectively depict the relationships between different emotions. For example, a person is more likely to feel simultaneously sad and neutral than both sad and happy. In text emotion recognition, a study by [35] adapted a loss function initially proposed by [36] to account for label correlation among raters, thereby quantifying dependency between emotions. Additionally, research by [37] introduced a multi-label focal objective function designed to enhance the differentiation between positive and negative emotions to improve emotion recognition perceptivity in text systems. This approach, however, overlooks the possibility of both negative and positive emotions occurring together. Unlike the studies mentioned earlier, we directly examine the relationships between co-existing categories of emotions based on their co-occurrence frequencies, allowing for mixed emotional states such as happiness and anger. For instance, consider

a scenario where a girl was upset at her boyfriend for being over an hour late, but upon his arrival, she saw he was holding her a present, her favorite dress. By conceptualizing the co-occurrence frequency of emotions in a matrix and normalizing it to create a penalty matrix, we can integrate this into existing objective functions like cross-entropy to penalize models during training. The penalty matrix assigns greater loss values when SER systems predict rare emotional co-occurrences, as indicated by the training set annotations. Importantly, this approach allows for the recognition of both positive and negative emotions happening concurrently.

1.2.6 Disagreement between Raters on the Emotion Datasets

Unlike common classification tasks (e.g., speaker identification or laughter/cry detection) with well-established "gold standard" labels, subjective tasks like emotion recognition lack clear labels and often rely on perceptual evaluations. Researchers frequently use crowd-sourcing platforms such as Amazon Mechanical Turk to collect labels rapidly and extensively [38]. While cost-effective, this method often compromises label quality. This trade-off is particularly significant in subjective tasks, where the inherent ambiguity amplifies annotation variability. Subjective tasks, such as emotion perception [39] or hate speech tagging [40, 41], pose unique challenges due to their fundamentally subjective nature. Obtaining labels for these tasks is complex because they heavily depend on individual interpreters. Annotator disagreements can result from various factors [42–44], such as diverse backgrounds leading to different interpretations, lack of interest in providing accurate labels, emotional biases, and contextual differences [45]. These variances introduce substantial noise into the labeling process, particularly problematic in crowd-sourced evaluations [46].

Annotation noise poses a major challenge, and various strategies have been developed to lessen its effects. For speech emotion classification tasks, it's important to understand that while noise can lead to discrepancies in labels, it's not the only cause

of disagreement. In line with the methodologies applied in the MSP-PODCAST corpus [39], efficient noise reduction approaches include excluding evaluators with persistently low agreement rates, pausing crowd-sourcing efforts when agreement falls below a certain threshold, and employing in-house staff who can undergo specialized training to enhance label consistency. Yet, it's crucial to note that perceptual variations are not merely noise; they can offer valuable insights that a speech emotion recognition system should utilize.

This dissertation aims to demonstrate that traditional methods of label aggregation, such as majority or plurality voting, often overlook the nuanced nature of subjective perception and may not be suitable for speech emotion classification. We propose an alternative aggregation method for the speech emotion recognition task, aiming for a more comprehensive and inclusive approach to label aggregation.

1.3 Contributions

The research goals of this dissertation aims to answer three questions:

(1) Should we remove those minority emotional ratings?

(2) Should we only let SER systems learn the emotional perception of a few annotators?

(3) Should SER systems only predict one emotion for each sample?

This dissertation proposes three methods to improve the process for label preprocessing, evaluation of SER systems, and training of SER systems, respectively.

1.3.1 Every Rating Matters Considering the Subjectivity of Annotators

We first explore the inherent subjectivity of how each annotator perceives emotions to improve the performances of SER systems. We aim to maximize the emotional ratings using the existing soft-label method introduced by [22] for training SER systems. Additionally, we develop an individual SER model for each annotator based on their respective ratings. To integrate all possible emotional data, we merged the embeddings obtained from pre-trained SER models using the traditional hard-label and the existing soft-label methods across five individual annotator SER systems for late-fusion. The results indicate that the proposed framework improves performance when assessed on a test set determined by a majority vote.

1.3.2 Novel Evaluation Method by an All-Inclusive Aggregation Rule

In addition, we implemented an innovative label aggregation approach that incorporates all annotated data from the test phase, ensuring no data is overlooked. We propose maintaining all emotional ratings within test data samples to precisely evaluate the performance of SER systems, since determining consensus labels isn't practical in real-world situations. Our ground truth is structured as a distribution reflecting the frequency ratio of emotional votes for each emotion class. Like the study conducted by [22], we believe that using a distributional similarity metric offers a more accurate assessment of SER systems' performance compared to accuracy-based metrics like the macro-F1 score, due to the subjective nature of SER. Distributional metrics better match how humans perceive emotions [6, 7]. Nevertheless, since the SER field is more accustomed to accuracy-based metrics, we also provide results using those. However, we've noticed that transforming distributional labels into binary vectors for accuracy evaluation might lead to losing some emotional ratings, which is different from our goal. Still, accuracy-

based metrics give the community and reviewers a more precise understanding.

1.3.3 Training Loss Using Co-occurrence Frequency of Emotions

To model the connection between co-occurrence of emotions, we counted the counts of co-occurrence of emotions based on labeled data in the train set of the emotion database as a matrix and normalized the frequency matrix by the number of individual emotion classes. Then, we use the unit matrix to subtract the normalized frequency matrix as a penalization matrix. To penalize the SER systems during training when the models predict rare co-occurrence of emotions, we integrate the designed penalization matrix into the current common objective functions, e.g., cross-entropy. Considering the link to the co-occurrence of emotions, this method improves the SER performances on the test conditions, including single-label and multi-label tasks.

1.4 Outline of the Dissertation

The rest of the dissertation is structured as follows. It primarily focuses on three key aspects of modeling and evaluating SER systems. We begin by presenting the public emotion databases (Chapter 2), and then introduce the three proposed methods: modeling the subjectivity of annotators (Chapter 3), a novel evaluation method (Chapter 4), and a training loss that accounts for the co-occurrence of emotion classes (Chapter 5).

Chapter 2

Emotion Databases

The chapter discusses four public emotional databases leveraged in this dissertation. We might use one or multiple databases in different chapters, but we introduce them here.

2.1 IEMOCAP

The IEMOCAP dataset [47] was carefully constructed from motion capture, audio, and visual data. Ten professional actors speak English. This unique dataset includes scripted and spontaneous dialogues, primarily focusing on romantic relationship scenarios to evoke diverse emotions. Each session involved one male and one female actor. To guarantee an expressive variation, performers utilized scripts to elicit distinct feelings. The final recordings were segmented into 10,039 utterances. All utterances have human-typed transcripts. Raters watched segmented clips and selected emotions from a predefined list of ten categories: neutral, happy, sad, angry, surprised, fear, disgusted, frustrated, excited, and "other." Addressing issues related to results reproducibility, mentioned in prior research [48], we have provided meticulous details on dataset splits in Section 2.5.2. Given the original dataset's deficiency of standardized split sets, our documentation aims to bridge that gap, ensuring greater consistency and

reproducibility. The IEMOCAP is used in Chapter 3 and Chapter 4.

2.2 IMPROV

The MSP-IMPROV dataset, also known as IMPROV [49], collects audio-visual recordings, and 12 actors are English speakers. All dyadic interactions represent four distinct emotional states: angry, happy, sad, and neutral. The sessions are thoroughly segmented into 8,438 individual clips by humans, with each clip being assessed by a minimum of five annotators using crowdsourcing methods. To maintain the high quality of annotations, the dataset integrates a quality control technique outlined by [50], aimed at identifying and excluding unreliable annotators.

The labeling process for the dataset involves two distinct procedures: primary (P) and secondary (S) emotions. Raters are instructed to select one emotion from five predefined emotions in the primary procedure: angry, happy, sad, neutral, or ”other.” The secondary emotions span a wider spectrum, including frustrated, depressed, disgusted, excited, fear, and surprised. Since the dataset lacks predefined sets for training, development, and testing, we elucidate the suggested division of these sets in Section 2.5.2 to ensure clarity and maintain consistency in follow-up analyses. The IMPROV is only used in Chapter 4.

2.3 CREMA-D

The CREMA-D dataset [51] collects high-fidelity audio-visual recordings from 91 professional actors, comprising forty-three women and forty-eight men. They were directed to deliver one of six unique emotions with the given scripts: angry, disgusted, fearful, happy, sad, or neutral. A key highlight is the comprehensive labeling procedure; spanning over 7,442 segments, each segment was evaluated by more than two thousand distinct crowdsourcing raters. All data were subjected to assessments from

no fewer than six annotators, who identified one of the six defined emotions for each performance.

The perceptual annotation process operates within three contexts: voice- and face-only and audio-visual. Raters focus exclusively on listening to the segments' audio in the voice-only context. They watch the actors' faces without audio input in the face-only context. Lastly, in the audio-visual context, annotators evaluate both the facial expressions and the audio concurrently.

For this dissertation on SER, we specifically concentrate on the emotional labels gathered from the voice-only scenario. Unlike numerous previous SER studies that leveraged labels from the audio-visual context or omitted annotation specifics entirely, we decided to depend exclusively on voice-only labels for the evaluation. Furthermore, our paper includes detailed specifications regarding the dataset splits we utilized, which can be found in Section 2.5.3. The CREMA-D is only used in Chapter 4.

2.4 MSP-PODCAST

The MSP-PODCAST dataset [39] is a comprehensive, emotionally varied voice compiled from licensed podcast resources. These recordings are initially split into utterances and subsequently labeled through a crowdsourcing website. The labeling protocol encompasses primary (P) and secondary (S) methods. Raters choose from nine categorized emotions in the primary emotion: angry, sad, happy, surprised, fear, disgusted, contempt, neutral, and "other." The secondary emotion broadens this scope to the primary emotions, plus eight additional ones: amused, frustrated, depressed, concerned, disappointed, excited, confused, and annoyed, making a total of 17 emotional categories. At least five contributors meticulously annotate each utterance to ensure robust and reliable annotations. Different dataset versions certify various utterance quantities within the training, development, and complementary test groups (test1 and test2). The

MSP-PODCAST is used in Chapter 4 and Chapter 5.

2.5 Standard Partition

The preceding study [48] indicates that 80.77% of SER research papers produce irreproducible results with the widely recognized IEMOCAP dataset. The primary obstacle to reproducibility is the absence of standardized data splits (e.g., training, development, and test sets) within the database. Previous studies each defined their partitions; however, they often withheld detailed partitioning methodologies or source code, complicating repeatability. Consequently, this dissertation aims to make SER more transparent and reproducible for everyone. We establish and define standard partitions for four prominent and publicly available emotion databases, facilitating future SER advancements.

2.5.1 Standard Partition of the IEMOCAP

In a speaker-independent scenario, models are trained using data from specific individuals and evaluated using data from entirely different individuals who were not part of the training set. This approach ensures an unbiased and robust assessment of the model's performance. For instance, in the IEMOCAP study, we summarize the data partitioning approach in Table 2.1. Each session involves two speakers in interactive dialogues and allows us to define five speaker-independent splits, referred to as Ses. 1 through Ses. 5. We perform a five-fold cross-validation, as detailed in Table 2.1, whereby every fold

Table 2.1: Table overviews the defined standard partitions for the IEMOCAP and the **Ses.** means the session in the database.

Partition	Training Set	Development Set	Test Set
1	Ses. 1,2,3	Ses. 4	Ses. 5
2	Ses. 2,3,4	Ses. 5	Ses. 1
3	Ses. 3,4,5	Ses. 1	Ses. 2
4	Ses. 1,4,5	Ses. 2	Ses. 3
5	Ses. 1,2,4	Ses. 3	Ses. 4

Table 2.2: Table overviews the defined standard partitions for the IMPROV and the **Ses.** means the session in the database.

Partition	Training Set	Development Set	Test Set
1	Ses. 1,2,3,4	Ses. 5	Ses. 6
2	Ses. 1,2,3,6	Ses. 4	Ses. 5
3	Ses. 1,2,5,6	Ses. 3	Ses. 4
4	Ses. 1,4,5,6	Ses. 2	Ses. 3
5	Ses. 3,4,5,6	Ses. 1	Ses. 2
6	Ses. 2,3,4,5	Ses. 6	Ses. 1

Table 2.3: Tables summarize the session in the CREMA-D emotion database. Notice that the **M** and **F** mean the male and F, respectively.

Session	Gender	Speaker ID
1	7M;11F	1037-1054
2	12M;6F	1001-1018
3	13M;6F	1073-1091
4	9M;9F	1055-1072
5	15M;3F	1019-1036

comprises different sets for training, development, and testing to guarantee a thorough assessment of the model's effectiveness across sessions.

2.5.2 Standard Partition of the MSP-IMRPOV

The IMPROV dataset is segmented into 6 distinct folds for cross-validation purposes. All folds feature a specific blend of development, training, and test data as detailed in Table 2.2. The splitting method provides the SER system with information on interactions featuring various pairs of speakers and testing on completely new speaker pairings. Consequently, this strategy systematically evaluates the model's capacity to generalize to various dyadic exchanges within the IMPROV dataset.

2.5.3 Standard Partition of the CREMA-D

The CREMA-D database is split into 5 subsets according to speaker IDs for the speaker-independent context. Each subset comprises a unique blend of males and females and specific speaker IDs, elaborated in Table 2.3. The partitioning strategy aligns with the methodology employed for the IEMOCAP dataset discussed in section 2.5.1.

Chapter 3

Every Rating Matters Considering Subjectivity of Annotators

We present a methodology in which joint learning addresses emotional rating uncertainty and annotator idiosyncrasies by leveraging both hard and soft emotion label annotations and individual and crowd annotator modeling. Further analyses reveal that emotion perception heavily depends on raters. Combining hard labels with soft emotion distributions provides a well-rounded approach to affect modeling. Additionally, the integrated learning of both general emotional insights and specific rater profiles yields the highest accuracy in emotion recognition.

3.1 Motivation

Traditional SER systems [1, 2] typically use the plurality rule or majority rule from a group of raters as the learning targets, termed the hard label, to train emotion recognizers. However, factors like gender [8], culture [9], and age [10] significantly influence emotion perception, leading to natural disagreement and ambiguity in annotations [52, 53]. Consequently, the hard label approach overlooks emotion perception's diverse annotations and subjective nuances. To address this limitation, researchers [2, 22] have

proposed using soft labels—a distributional representation instead of a single definitive label—to capture blended emotion perceptions better. While the soft labeling method enhances flexibility in representing the variability of emotion perception, it still disregards the unique input of individual annotators because it creates the label distribution by aggregating inputs from all annotators. Therefore, we first build individual-based SER systems to model diverse and accurate subjectivity of emotion perception to improve aggregated emotion performance.

3.2 Background and Related Works

3.2.1 Subjectivity of Emotion Perception

Variability in how raters perceive emotions is not a new observation. Annotator modeling, which addresses this issue, has gained attention for years. For instance, the work [54] employed agreement between raters to assign weights to training instances, which facilitated the development of a speaker-dependent audio-visual emotion recognition system. Similarly, Han et al. [55] introduced a model that leverages inter-rater disagreement to estimate perception uncertainty, thereby enhancing performance in continuous dimensional emotion tracking from audio and video sources. However, they still consider the ratings of all raters at that same time, but our method directly models individual raters' emotion perception, which can preserve more subjectivity of emotion perception.

3.2.2 Mixture of Annotators

Disagreement is present not just in emotion perception but also in various other areas like medical image tagging. Yan et al. [56] suggested that discrepancies arise because each annotator possesses unique medical domain knowledge. To address this,

they proposed a method involving multiple annotators, which takes into account all available information by repeatedly using training data points until the models fully grasp each annotator's input. Their approach was found to be more effective than the traditional method that uses majority voting to determine ground truth.

3.2.3 Soft-label Training Method for SER Systems

Steidl et al. [22] initially argued that using only a single emotion label as ground truth in emotion recognition tasks might not be suitable due to the subjective nature of emotion perception. They proposed using soft labels as ground truth based on count data and utilized entropy loss as an evaluation criterion. However, they re-assigned categorical emotions to enhance inter-rater agreement. In contrast, we retained the original labels to reflect the original emotion perception. Furthermore, Fayek et al. [57] showed that training SER systems with soft labels can yield better performance than those trained with hard labels on test sets where the majority rule determines the consensus label. Additionally, Ando et al. [2] adjusted the soft labels with a small α value, creating an effect similar to label smoothing for training their SER systems. Their findings indicated that SER systems trained with these modified soft labels outperformed those trained with common soft labels and hard labels. Finally, Zhang [58] was the first to demonstrate the advantages of using soft labels in cross-corpus SER, through their proposed objective function.

3.3 Resource and Task Formulation

We employ the IEMOCAP database as referenced in section 2.1, encompassing emotional ratings by 12 distinct raters across 10 categorical emotion classes. For consistency with previous research, we utilize the identical evaluation data wherein the entry is tagged with a singular emotion state, determined by a majority vote from more than

three annotators. This study focuses on recognizing four primary emotion classes: sad, neutral, happy, and angry. Following the practices in [1,2], we consolidate the happiness and excitement categories into one: happiness. This aggregation includes 5,531 data samples used in the emotion recognition evaluation; this approach aligns with the traditional use of the IEMOCAP dataset as a benchmarking standard. The test set excludes the data without consensus labels. The emotion class distributions of data samples are sad: 19.60%, happy: 29.58%, neutral: 30.88%, and angry: 19.94%. Half of the raters are also actors, while the other half consists of in-house raters. We selected only 5 out of the 6 in-house raters (E1, E2, E4, E5, and E6) since these five provided annotations for samples across all 5 sessions. Therefore, the proposed method only builds those 5 individual annotators.

3.4 Speech Emotion Classifier

3.4.1 Input Features

We use the openSMILE toolkit [59] to derive frame-level acoustic features from utterances. Specifically, the "Emobase.config" configuration helps us obtain a 45-dimensional feature set. This set includes 12-dimensional Mel-Frequency Cepstral Coefficients (MFCCs), voice probability measures, zero-crossing rates, fundamental frequency (F0) values, loudness metrics, and their respective first-order derivatives. Additionally, the second-order derivatives of both loudness and MFCCs are featured. The feature extraction process is carried out with a frame length of 60ms and a step size of 10ms. These features are normalized for each speaker with z-score normalization and then downsampled by averaging over sets of five consecutive frames.

3.4.2 Model Structure

Utilizing the framework [1], all models in this study are designed. This framework comprises an input layer, a bidirectional long short-term memory (LSTM) layer, a fully connected layer, and an output layer. Mirsamadi et al. [1] investigated the various attention mechanisms, and their proposed weighted pooling layer considering the attention weights applied over the frame-level input features achieved the best performance when the input features extracted by the "Emobase.config" file in the OpenSMILE toolkit. Therefore, all models used the same structure, and we denoted the model as the **BLSTM-FC**.

3.4.3 Training Labels

All models utilize the BLSTM-FC architecture, consistent with the design proposed in prior research [1]. Our objective in this study is to address the subjectivity of emotional perception. To tackle the variability in emotion perception, we train BLSTM-FC models using two distinct label learning: hard-label (denoted as CROWD$_H$) and soft-label (denoted as CROWD$_S$). The hard label is the conventional way to obtain a single-label emotion as ground truth. To consider that emotion perception could be overlapped and blended, Steidl et al. [22] first propose the soft labels to calculate the distributional label based on the votes for each emotion class. Fayek et al. [57] integrated the inter-rater variability with soft-label to build SER systems. Also, Ando et al. [2] modified the formulation of the calculation of soft labels by introducing α to slightly change the distribution of conventional soft labels as below.

$$t(c_i) = \frac{\alpha + \sum_n^R v_i^n}{\alpha \times C + \sum_i^C \sum_n^R v_i^n}, \tag{3.1}$$

where c_i means the $i-$th emotion class, n represents the $n-$rater, v_i^n is the binary value to check whether $n-$rater chooses c_i emotions, C is how many categorical emo-

tions and the R represents the number of annotators for t samples. In this dissertation, we follow the study [2] to assign the α value as 0.75, and C is 4 since the number of emotion classes is 4.

3.5 Proposed Method

3.5.1 Rater-Modeling

Due to the inherent subjectivity and unique individual differences, different people may interpret the exact spoken phrase in varied ways. To enhance the SER systems, we incorporate rater modeling. We categorize annotators into two groups: **Crowd** and **E**. In this context, **Crowd** refers to their collective annotations in the used dataset, while E pertains to the annotations by each specific individual rater. For the **Crowd** category, we develop two distinct models based on whether the targets are hard or soft labels (as detailed in Section 3.4.3). Conversely, for the E category, models are trained using soft labels available to each annotator, implying that each annotator's quantity of annotated data varies. Table 3.1 and Table 3.2 show a summation of how many data points are used for training each model, and Table 3.2 overviews the number of data samples for models. With soft labels, **Crowd**$_S$ has over 3,185 more samples compared to **Crowd**$_H$.

Table 3.1: Table overviews the data samples used to train models.

Model	Total	Multiple	Single
Crowd$_H$	5531	0	5531
Crowd$_S$	7774	3185	4589
E1	5954	44	5910
E2	7845	38	7807
E4	6429	212	6217
E5	422	3	419
E6	773	20	753

Table 3.2: Table summarizes the label distribution for each emotion class of the models. Notice that each sample could have more than one emotional rating.

Model	Neutral	Anger	Sadness	Happiness
CrowdH	80.88%	19.94%	19.60%	29.58%
CrowdS	29.33%	17.77%	17.10%	35.79%
E1	8.49%	21.21%	20.64%	49.67%
E2	22.45%	26.58%	19.62%	31.35%
E4	52.88%	12.41%	10.95%	23.76%
E5	69.88%	15.29%	5.88%	8.94%
E6	26.73%	15.76%	14.22%	43.38%

3.5.2 Final Concatenation Layer

After successfully computing all model components, including the two Crowd models and 5 E_N models, we froze their respective model weights. Then, we concatenate the representation from the final layer before the softmax activation in each BLSTM-FC model (depicted within the square box in Fig. 3.1). This concatenated layer is followed by an additional softmax layer to output the last prediction. The entire architecture is presented in Fig. 3.1.

3.6 Experimental Setup

3.6.1 Foundational Component

The foundational component is the BLSTM-FC model equipped with an attention mechanism. This model architecture includes two fully connected (FC) layers with Rec-

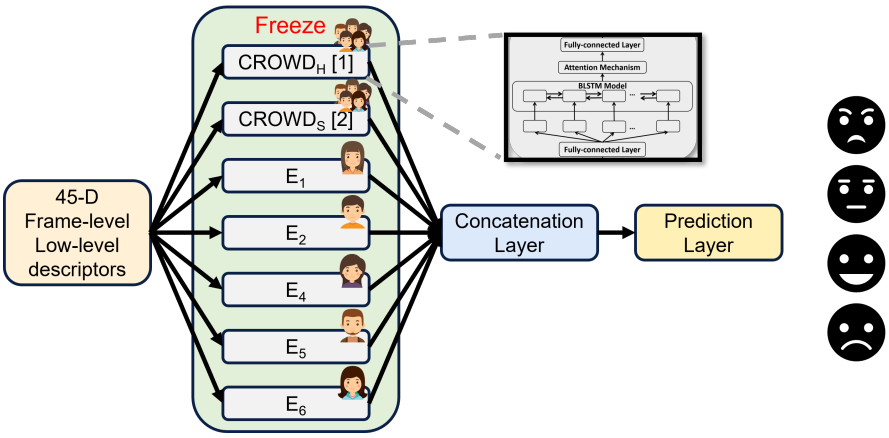


Figure 3.1: The figure illustrates the overall proposed model.

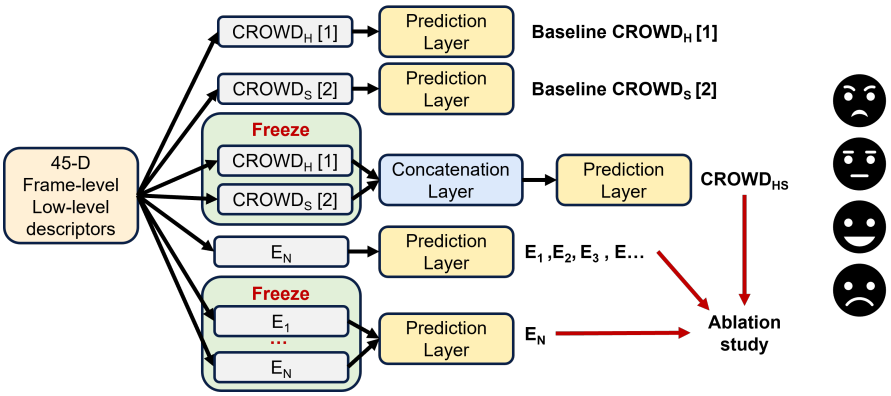


Figure 3.2: The figure illustrates the baselines and models for the ablation study.

tified Linear Unit (ReLU) activation functions, a BLSTM layer enhanced using attention weights, and concludes with a fully-connected layer employing a softmax function. Specifically, the model comprises 256 hidden units in the initial dense layer, 128 hidden units in the BLSTM layer with attention, 256 hidden units in the second dense layer, and the final softmax-enabled layer differentiates into four units. Each layer also integrates a dropout mechanism with a 50%

3.6.2 All Model Comparison

Figure 3.2 illustrates the various baselines and models considered in the ablation study. Subsequently, we evaluate and compare the performance outcomes for each component of the overall architecture, as detailed below.

- **Baseline CROWD**$_H$: This model closely parallels the previous proposed, but it utilizes a BLSTM-FC framework trained on hard labels in the study [1].

- **Baseline CROWD**$_S$: This model employs soft label training, a method proposed by the work [2], designed to utilize all labeled samples.

- **Baseline CROWD**$_{HS}$: The model represents a fusion of Crowd$_H$ and Crowd$_S$. It combines all Crowd-relevant information by concatenating the representations from Crowd$_H$ and Crowd$_S$ before feeding them into the final softmax layer.

- **Proposed Rater Model, $\mathbf{E}_N$**: Each of the E_N models is trained using soft label learning based on the annotations made by individual raters.

- **Proposed Fusion of $\mathbf{E}_N$**: The model integrates all individual E_N components (five separate annotators). It consolidates all rater-specific information by concatenating the representations from each E_N before passing them through the final softmax layer.

- **Proposed Model**: As depicted in Figure 3.1, this model represents our ultimate proposal and leverages all available Crowd and E_N information. It achieves this by concatenating the representations from both Crowd_H and Crowd_S before feeding them into the final softmax layer.

3.7 Other Training Details

We run the experiments on the cross-validation in the leave-one-session-out condition (as shown in Table 3.3), evaluating performance in an unweighted average recall (UAR). The batch size is 32; the learning rate is 0.0001; the number of epochs is 200. Early stopping criteria to minimize the loss value on the validation set are applied during training across all configurations to prevent overfitting and ensure model effectiveness. The optimizer utilized in this study is ADAMMAX [60].

Table 3.3: Table summarizes cross-validation details of the IEMOCAP, following [1,2].

Fold	Training Set	Development Set	Test Set
1	Ses. 1,2,3,4		Ses. 5
2	Ses. 2,3,4,5		Ses. 1
3	Ses. 3,4,5,1	Randomly 10% of training data	Ses. 2
4	Ses. 1,4,5,2		Ses. 3
5	Ses. 1,2,4,3		Ses. 4

Table 3.4: Table summarizes the results in unweighted average recall (UAR) of all models evaluated on the IEMOCAP.

Model	Overall	Neutral	Angry	Happy	Sad
Crowd$_H$ [1]	0.5745	0.5571	0.6329	0.4502	0.6577
Crowd$_S$ [2]	0.5712	0.4970	0.6298	0.6285	0.5314
Crowd$_{HS}$	0.5858	0.5966	0.5931	0.5363	0.6171
E1	0.5098	0.0804	0.6131	**0.7724**	0.5734
E2	0.5968	0.3878	0.6435	0.6425	0.6261
E4	0.4859	0.8129	0.4542	0.3820	0.2944
E5	0.3762	**0.8689**	0.4762	0.1121	0.0475
E6	0.4582	0.3685	0.4010	0.6039	0.4595
EN	0.6024	0.4964	0.6364	0.6148	0.6619
Proposed	**0.6148**	0.5455	**0.6451**	0.6032	**0.6656**

3.8 Experimental Results and Analyses

Table 3.4 provides an overview of the results across various comparative models. Our presented framework achieves the highest overall emotion classification performance, with an unweighted average recall (UAR) of 61.48%. This method exceeds the performance of the previously leading approaches, surpassing Crowd$_H$ proposed by [1] by 3.18% and Crowd$_S$ proposed by [2] by 4.36% in absolute terms. These results highlight the benefits of incorporating subjective annotator emotional information to enhance emotion recognition over previous methods.

One significant finding is the differential effectiveness of soft-label vs. hard-label learning: Crowd$_S$ demonstrates better performance on happy emotions than Crowd$_H$, which performs better with neutral and sad emotions. The complementary strengths of Crowd$_H$ and Crowd$_S$ underscore the importance of their integration for advanced emotion recognition results. Historically, happiness has been a challenging class to identify [1, 2], and it benefits from soft-label learning, suggesting that happiness has a more diffused presence within the acoustic spectrum than other emotions, such as anger and sadness.

Furthermore, individual models tend to have low recognition rates, probably because of the subjective perspectives of unique raters and the uneven distribution of emotion classes within each annotator's labeling. For instance, as Table 3.4 illustrates, the E_1 model shows low recognition accuracy for the neutral state but reasonable performance

for recognizing happy emotion, while it is reverse for the E_5 model. Examining the emotion distribution in Table 3.2 reveals that this discrepancy is linked to the variety and quantity of emotional data annotated by each rater. By integrating raters' models at the late-fusion level, E_N effectively taps into multiple complementary viewpoints, enabling it to learn a well-rounded and nuanced understanding of emotion perception from distinct individual perspectives.

Last but not least, by employing individual rater models, our proposed method can expand the dataset used for building SER systems, compared to the traditional hard-label approach that limits the number of data in the train set to instances where consensus among raters is achieved. This allows for a more robust and comprehensive understanding of emotional nuances.

3.9 Summary

Human perception of emotions is relatively subjective and differs greatly from person to person. In this study, we introduce a model that merges dominant sentiment annotations with individual subjectivity models to advance emotion classification accuracy. The framework achieves an impressive score of 61.48% on a task involving 4-class emotion categorization. Although numerous studies have tackled the issue of annotator subjectivity, this pioneering approach explicitly combines consensus and individual differences in emotion perception, leading to improved classification performance on a benchmark dataset.

Chapter 4

Novel Evaluation Method by an All-Inclusive Aggregation Rule

When choosing test data for subjective tasks, many studies form ground truth labels through aggregation techniques like the majority or plurality rules. Unfortunately, these techniques ignore data points lacking consensus, simplifying the test set compared to real-world tasks where predictions are necessary for every sample. These ignored data points often contain ambiguous signals with overlapping traits witnessed by annotators. We emphasize the need to account for all annotations and samples in the dataset, as focusing only on performance metrics derived from a test set filtered by majority or plurality rules may skew the model's performance evaluations. We specifically investigate SER tasks and note that traditional aggregation rules result in data loss ratios between 4.63% and 92.01%, as shown in Table 4.1. Based on this insight, we introduce a versatile, all-inclusive rule, a label aggregation approach to appraise SER systems using comprehensive test data. We differentiate the conventional single-label approach with a multi-label methodology catering to the coexistence of various emotions. Training an SER model with data chosen by the all-inclusive rule consistently achieves better macro-F1 scores when evaluated on the whole test set, including ambiguous, non-consensus

samples.

4.1 Motivation and Background

Given the inherently subjective nature of these tasks, models are often evaluated using labels derived from human perceptual assessments, where multiple raters annotate each data point. The common practice for processing these annotations and creating training and testing sets relies on majority or plurality aggregation methods. These methods ignore annotations that do not correspond with the consensus label. However, co-existing emotions are frequently observed in everyday interactions [21], making it challenging for a single label to represent a sample's emotional perception fully. Additionally, data points lacking agreement are excluded. According to the *majority rule* (MR), data is dismissed if no class achieves more than 50% of the votes. Similarly, the *plurality rule* (PR) disregards data if no single class receives more votes than others. This exclusion of ambiguous samples undermines the validity of systems meant for real-world applications since these samples are not represented in the test set. Previous studies have investigated the use of all existing annotations during training, adopting soft-label learning strategies to incorporate all samples [18, 57, 61–65]. Nonetheless, test sets remain *simplified*, considering only sentences that meet MR or PR criteria, thereby overlooking complex and ambiguous samples.

We propose a practical methodology that amalgamates all annotations gathered from subjective evaluations for training and test sets, thereby enhancing the practical applicability of these systems. Although this approach is compatible with any domain requiring labels derived from perceptual assessments, our focus is on the SER task, where co-occurring emotions are common in everyday interactions. Unlike traditional methods that neglect non-consensus labels, we retain all data points in the training and test sets. This strategy enables SER models to utilize comprehensive data during training

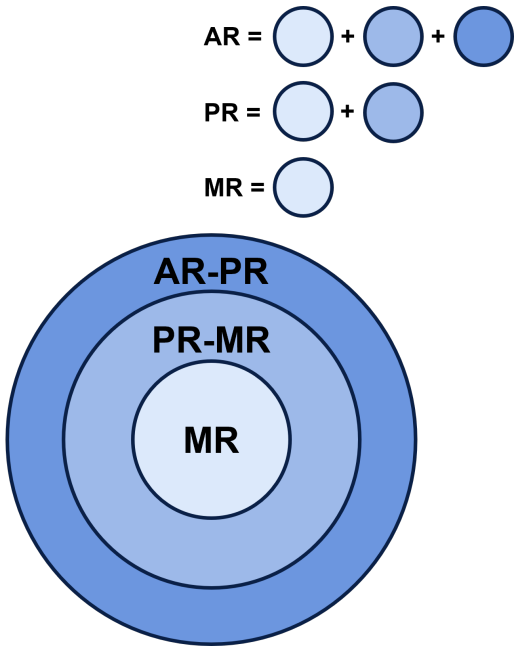


Figure 4.1: A diagram illustrating how much data and ratings are used in the final test set according to each aggregation method. MR contains the lowest amount of data, and AR always includes the entire test set available in the dataset.

and ensures they are evaluated on all samples in the test sets, including those with divergent labels. We call this framework the *all-inclusive rule* (AR) method. The ability to accommodate co-occurring emotions is a cornerstone of the approach. Figure 4.1 compares the all-inclusive aggregation strategy against the majority and plurality rules, typically applied in SER tasks to select data points for the test set. By implementing the AR method, every data point is included in the test set, thus underlining the exhaustive performance evaluation of SER systems. The main research questions driving the study are as follows.

- How is the performance of SER systems influenced by using different aggregation methods for the training set annotations?
- Does utilizing data from the all-inclusive rule in training an SER system enhance performance on ambiguous emotions compared to data processed with majority or plurality rules?
- Which label learning strategy should be employed for training SER systems to

achieve optimal performance when tested on the entire data set?

4.2 Previous Literature

4.2.1 Evaluation of SER Systems

Evaluating SER systems on a complete test set is crucial. However, the common approach involves discarding samples that lack consensus emotion labels. When gathering emotional annotations from multiple workers, significant disagreement often exists among annotators [66–68], leading many studies to eliminate numerous data points from the test set. For instance, the IEMOCAP and CREMA-D corpora use the majority rule (MR) to construct ground-truth labels, discarding approximately 31.37% and 35.8% of the data, respectively, as shown in Table 4.1 [47, 51, 69]. Researchers using these corpora often adhere to this rule for testing their models [70–73]. Similarly, the IMPROV and PODCAST corpora use the plurality rule (PR) to annotate primary and secondary emotional labels for each speaking turn [39, 49], and this default aggregation method has been widely adopted in subsequent studies [74–76].

The aforementioned studies assume that each speaking turn has only one emotional category, ignoring secondary emotions in the recordings. In reality, emotional states often co-exist (e.g., a person can be sad and angry) [21]. Therefore, consolidating multiple annotations into a single class and discarding non-consensus data points does not accurately capture whether SER system predictions reflect the complex emotional behaviors observed in daily interactions. Although some studies have explored using a "multiple-hot" vector to frame SER as a multi-label problem [23, 27, 77], this approach does not discern dominant emotions. It treats all annotations equally valid, even if a single annotator selected a class.

To our knowledge, Riera et al. [78] is the only study advocating for including all test samples in evaluating SER systems rather than discarding non-consensus data. How-

Table 4.1: Table is an overview of the number of utterances, emotion classes, and data loss ratios in several prominent emotional datasets. Here, P indicates primary emotions, and S indicates secondary emotions.

Label Aggregation		MR		PR		AR	
Database	# Utterances	Data	Rating	Data	Rating	Data	Rating
IMPROV (P)	8438	9.18%	28.52%	4.63%	26.41%	0.00%	5.12%
CREMA-D	7442	35.80%	52.96%	8.55%	40.57%	0.00%	0.00%
PODCAST (P)	90978	44.81%	59.87%	19.85%	49.24%	0.00%	6.15%
IEMOCAP	10039	31.37%	49.44%	25.32%	45.70%	0.00%	3.10%
IMPROV (S)	8438	54.18%	76.91%	12.32%	56.70%	0.00%	4.23%
PODCAST (S)	90978	92.01%	96.99%	33.72%	78.13%	0.00%	1.64%
Average		44.56%	60.78%	17.40%	49.46%	0.00%	3.37%

ever, their study did not explore training SER systems with various label learning methods. It involved relabeling some emotions (e.g., labeling excited as happy and surprised as ”other”), which is a significant limitation. Unlike that study, this dissertation retains the original emotional classes across all datasets. It involves empirical experiments with different label learning methods and test sets created using different aggregation methods.

4.2.2 Curriculum Learning for Emotion Recognition

Lotfian and Busso [79] rigorously employed curriculum learning, utilizing annotator disagreement on samples to enhance the performance of SER systems in predicting valence and arousal. Furthermore, Yang et al. [80] utilized emotion shift as difficulty scores to categorize samples as ”easy” or ”hard.” They trained text-based conversational emotion recognition systems progressively, starting with easy samples and moving to more challenging ones. Their findings show that curriculum learning boosts performance in emotion recognition within conversations, as incorporating harder samples during training increases the training loss, thereby refining the system’s accuracy. Similar findings have been reported in recent research [81].

4.3 Methodology

We present a new aggregation methodology called the *all-inclusive rule* (AR), designed to facilitate the training and evaluation of SER systems using an exhaustive test set. This includes data points lacking Majority Rule (MR) or Plurality Rule (PR) consensus. The definition, significance, and application of this rule are thoroughly explained.

4.3.1 Proposed All-inclusive Rule

The All-Inclusive Rule (AR) is an aggregation methodology that retains all annotated samples within a dataset, regardless of vote frequencies. This method ensures data points are never disregarded. Initially, AR collects all classifications assigned to each data point, creating the ground truth. When forming the training set, the ground truth is represented either by a one-hot encoding or the vote distribution based on the selected label learning strategy. For the hard-label approach, AR identifies the emotional class with the most votes as the ground truth—akin to the plurality rule. In cases where a clear majority is absent, one of the top-voted classes is randomly selected as the ground truth. An example is shown in Table 4.2 Case (C1), where the hard label could be (1,0,0,0) or (0,0,1,0). AR produces the ground truth for the soft-label or distribution-label approach by reflecting the vote distribution among emotional classes.

AR consistently utilizes the distributional ground truth when producing the test set, irrespective of the chosen label-learning strategy, which is highlighted in the rightmost column of Table 4.2. This comprehensive approach ensures every annotated data point and all its annotations are integral to the test set. The all-inclusive rule enhances the label descriptor, more accurately capturing the emotional nuances of data points by integrating sentences that reflect ambiguous emotions into the test set.

Table 4.2: Here is an overview of how label vectors are constructed for the *all-inclusive rule* (AR) using three examples, each with five annotations, in a four-class emotion classification task. The four emotions include neutral (N), happy (H), angry (A), and sad (S). Label vectors are created in the format: (N, H, A, S). We show three examples. For instance, (C1) N,N,A,A,S demonstrates that the five emotional annotations for Case (C1) selected two for neutral, two for angry, and one for sad.

Case	**Training Set**			**Test Set**
	Hard-label	Soft-label	Distribution-label	Label
(C1) N,N,A,A,S	(1,0,0,0) OR (0,0,1,0)	(0.4,0.0,0.4,0.2)	(0.4,0.0,0.4,0.2)	(0.4,0.0,0.4,0.2)
(C2) N,N,H,A,S	(1,0,0,0)	(0.4,0.2,0.2,0.2)	(0.4,0.2,0.2,0.2)	(0.4,0.2,0.2,0.2)
(C3) N,N,N,A,S	(1,0,0,0)	(0.6,0.0,0.2,0.2)	(0.6,0.0,0.2,0.2)	(0.6,0.0,0.2,0.2)

4.3.2 Employing the All-Inclusive Rule for Test Set Construction

We include all data samples and prioritize every available emotion as a learning target to capture the full range of opinions gathered during perceptual evaluation. For instance, previous research using the IEMOCAP corpus has aggregated annotated emotions into a 4-class emotion classification task (e.g., combining excitement and happiness while disregarding less frequent classes such as fear, surprise, and disgust). Unlike this methodology, we do not overlook any emotional states nor confine the SER models to be trained or tested solely on a few selected emotions.

In addition, our all-inclusive rule allows SER models to be tested on the entire test set, including secondary emotions. Previous studies have often disregarded secondary emotional annotations due to considerable data loss from standard aggregation methods (up to 92.01% as noted in Table 4.1). As the AR method utilizes the fully annotated test set, we can assess the SER model with secondary emotions, which have not been examined before. Table 4.1 highlights the proportion of data loss introduced by different aggregation methods across the four datasets used in this study for primary and secondary emotions—utilization of MR and PR results in discarding up to 92.01% and 33.72% of the data, respectively. The most significant data loss occurs when classifying secondary emotions in the MSP-Podcast corpus.

4.4 Experimental Setup

4.4.1 Resource

Four publicly available emotion databases, as detailed in Chapter 2, are utilized in our research. The corpus version 1.10, containing 104,267 annotated utterances, is employed. However, the “Test2” set is excluded, narrowing the dataset to 90,978 utterances, as shown in Table 4.1.

4.4.2 Speech Emotion Classifier

To assess the performance of various aggregation methods, we utilize the Wav2vec2.0 architecture [82], which has demonstrated strong performance in SER tasks in multiple studies [83, 84]. Specifically, we implement the ”wav2vec2-large-robust” variant, as proposed in Hsu et al. [85], which has proven to be the best-performing model in the study by Wagner et al. [84]. The model processes raw audio signals directly, preceding spectrograms or Mel-frequency cepstral coefficients (MFCC). The datasets are sampled at a rate of 16k Hz to match the sampling rate of the pre-trained model data.

For efficiency and reproducibility, we’ve optimized the model by removing the top 12 of the 24 transformer layers from the architecture, which preserves recognition performance while reducing the number of parameters [84]. Two hidden layers and a softmax output layer are attached to this trimmed Wav2vec2.0 model. Each hidden layer with the rectified linear unit (ReLU) activation function comprises 1,024 nodes. Implemented with average pooling per utterance, the outputs from the Wav2vec2.0 layers feed into the classification layers. We apply a dropout function with (p = 0.5) to the first and second classification layers for regularization.

Our implementation is based on the HuggingFace library [86] and uses a pre-trained ”wav2vec2-large-robust” model. During fine-tuning, convolutional and transformer layers of the Wav2vec2.0 model are frozen—a strategy that has shown better performance

Table 4.3: Here is an overview of the data loss ratios introduced by the label aggregation method on the PODCAST development, test, and train sets. P represents primary emotions, and S represents secondary emotions.

Rule	Set	PODCAST (P)	PODCAST (S)
MR	Training	47.95%	88.63%
	Development	47.90%	89.64%
	Test	47.08%	90.90%
PR	Training	18.76%	29.46%
	Development	19.62%	28.90%
	Test	17.14%	27.76%
AR	Training	0.00%	0.00%
	Development	0.00%	0.00%
	Test	0.00%	0.00%

than fully fine-tuning all parameters [83]. We employ the Adam optimizer [87] and set up a learning rate as 0.0001, structuring mini-batches from 32 utterances and training the model over 100 epochs. The best recognition performance model is chosen from the development set after the training epochs. The entire implementation is carried out in Pytorch [88] and executed on an NVIDIA Tesla V100 GPU.

For the POSCAST corpus, we utilize pre-defined training (63,076 samples), development (10,999 samples), and test (16,903 samples) sets. Due to the smaller nature of other corpora, we implement a cross-validation strategy. Speaker-independent sessions are used for IEMOCAP (five sessions) and IMPROV (six sessions). As the CREMA-D corpus lacks pre-defined sessions, we manually divide it into five speaker-independent sessions. For IEMOCAP, IMPROV, and CREMA-D, a K-fold cross-validation is applied, where K means how many sessions. Each fold includes one session as the test set, one as the development set, and the remaining is the training set. More details about partitions are in Section 2.5

Table 4.1 reflects data loss ratios derived from each label aggregation method across various databases. We evaluate data loss ratios in every partition (train, development, test set). Observations indicate that data loss trends are relatively consistent across all four datasets. Thus, Table 4.3 exemplifies these distributions specifically for the PODCAST database, and the trends are similar in Table 4.1.

4.4.3 Train/Test Set Defined by Aggregation Rules

In this evaluation, the models are trained and tested using both matching and mismatching aggregation rules. For both the training and development sets, ground truth is established using MR, PR, and AR, denoted as MR_{Train}, PR_{Train}, and AR_{Train}, respectively. The models are assessed on sets derived from MR, PR, and AR rules during testing. Additionally, two extra test conditions are defined, illustrated by the *donuts* in Figure 4.1: PR-MR includes test samples accepted by PR but not by MR, whereas AR-PR encompasses those accepted by AR but not by PR. The AR-PR condition is the most ambiguous set, as it includes samples with non-consensus annotations. To our knowledge, this study is the first to evaluate SER models using such non-consensus annotations.

4.4.4 Label Learning for SER

Alongside evaluating different aggregation methods, we assess the performance using various label-learning approaches. The experiments employ three label learning strategies: hard-, soft-, and distribution-label learning.

For hard-label learning, the ground truth is constructed with a one-hot encoding, where the class receiving the highest number of votes from annotators is represented as "1.0". When utilizing the training set aggregated with AR, if there is no clear consensus, one of the top-voted emotions is randomly chosen as the ground-truth emotion. We smooth the one-hot encoding ground truth vector using the smoothing strategy proposed by Szegedy et al. [89] with a parameter set of 0.05. This method slightly adjusts probabilities for classes assigned initially a zero value. The SER systems are then trained using the cross-entropy (CE) objective function.

For both soft-label learning and distribution-label learning, the ground-truth vector reflects the distribution of annotator votes. This is achieved by dividing the vote count

for each class by the total number of votes for each data point. We also apply the label-smoothing strategy used in hard-label learning. Soft-label learning targets are optimized using the CE loss function, while distribution-label learning uses the Kullback–Leibler divergence (KLD) as the cost function.

4.4.5 Evaluation Metrics and Statistical Significance

Hard-decision-based assessment

This study employs macro-F1 scores to evaluate SER performance, which involves calculating precision and recall rates. The MR, PR, and PR-MR test sets are structured by selecting a single class, making them compatible with macro-F1 scoring. The class receiving the most votes is chosen as the target for these test sets. An output is valid if the emotion category with the highest predicted probability matches the target class.

For test sets gathered under AR and AR-PR conditions, which contain non-consensus labels, we permit the coexistence of multiple emotions to compute the macro-F1 score. Targets are selected based on using the threshold to the binarized vectors. A prediction is valid if the proportion for a specific category exceeds $1/C$, with C representing how many emotion classes. This method is in line with those employed in previous studies [78, 90].

Consider an emotion recognition task that distinguishes between four emotions: neutral, anger, sadness, and happiness. In one instance, five different reviewers each gave their rating based on the sample, resulting in the following annotations: angry (A), sad (S), sad (S), neutral (N), and angry (A). To determine the label distributions, we categorize the data into (N, A, S, H) and get the proportions (0.2, 0.4, 0.4, 0.0).

The threshold is set to (1/4 = 0.25), so we convert the ground truth to (0,1,1,0). During the inference, suppose we have predictions from three different models: (0.1, 0.45, 0.45, 0.0), (0.2, 0.35, 0.35, 0.1), and (0.45, 0.1, 0, 0.45). Applying the (0.25) threshold,

these outputs are converted into (0,1,1,0), (0,1,1,0), and (1,0,0,1), respectively. In this example, the (0,1,1,0) and (0,1,1,0) fully match the ground truth.

Distribution-based Assessment

Following the approach proposed by Steidl et al. [22], where results are assessed using an entropy-based metric, we employ the *Kullback-Leibler divergence* (KLD) to determine the similarity between the model's predicted distribution and the subjective annotations. This method checks whether an SER model aligns with human emotional perception. Unlike the macro-F1 evaluations, which binarize the model's output for single-label or multi-label tasks, we utilize the model's probability distributions across all test sets and measure them using KLD. As illustrated in Table 4.6, lower KLD values indicate better performance in this context.

Distribution-based Assessment Versus Hard-decision-based Assessment

Distribution-based assessments, such as KLD, are suitable metrics for evaluating distributions. In contrast, hard-decision-based assessments like the F1-score necessitate categorical decisions, making them less appropriate for comparing distribution similarities. Distribution-based assessments keep all data and maximum usage of emotional ratings while evaluating the SER performances. Every assessment has different advantages and disadvantages, so both are presented in the paper. Table 4.4 outlines the benefits and limitations of using KLD versus traditional accuracy metrics (e.g., macro-F1, micro-F1, and weighted-F1). Reporting both metrics offers complementary and more detailed insights.

Statistical Significance

We assess the statistical significance of the results per each aggregation method utilized. For cross-validation experiments (IEMOCAP, CREMA-D, and IMPROV), the

Table 4.4: Comparison of distribution-based and hard-decision-based assessment metrics.

Metric	Distribution-based assessment	Hard-decision-based assessment
Advantages	• We can directly compare the models' predictions to human perception without needing extra thresholding methods. • It is tuned to the overall shape and structure of the distributions, reflecting both the accuracy and confidence of the predictions.	• The macro-F1 score has a defined range between 0 and 1, facilitating straightforward interpretation and comparison across various models and datasets. • By applying thresholds to the labels, we can mitigate the influence of noisy annotations from raters who were not attentive.
Limitations	• The performance differences between the baseline and proposed models are difficult to interpret because the scale differences are minimal. • It lacks a fixed range, which makes it challenging to interpret and compare absolute values across different datasets or models. • It is susceptible to slight distribution variations, especially when dealing with sparse or high-dimensional data.	• All predictions and ground truth need to be binary (0 or 1)—this requirement can lead to information loss and reduced granularity, as it oversimplifies the distributional nature of the ground truth. • It is challenging to assess model performance on less common emotions, such as fear, due to limited available data.

predictions for each condition across all folds are concatenated, ensuring that every piece of data is considered (i.e., each sample appears in one fold of the test set). Predictions from all pre-defined test sets using a single model are employed directly for the PODCAST experiments. After collecting these predictions, they are segmented into 40 folds to average the macro-F1 score. A two-tailed t-test is then conducted to determine statistical significance and to confirm whether the p-value is below 0.05. The notations $*$, †, and $\star$ are used to indicate when a model's performance is significantly superior to models trained using MR_{Train}, PR_{Train}, and AR_{Train} sets, respectively.

4.5 Experimental Results and Analyses

The experimental analysis begins by benchmarking our presented method against state-of-the-art (SOTA) baselines, showcasing the advantages of the SER strategy utilized in this research (Section 4.5.1). Following that, we address the three research questions outlined in Section 4.1 (Sections 4.5.2, 4.5.3, and 4.5.4).

4.5.1 Comparison of Results with Prior SOTA Methods

We assess the SER model's performance against three existing SOTA benchmarks using the IMPROV(P), CREMA-D, PODCAST(P), and IEMOCAP corpora. As discussed by Li et al. [91], the first reference model establishes an end-to-end system that converts speech into spectrograms and utilizes a self-attention mechanism to highlight emotional elements within sentences. At its time of publication, this model set the SOTA performance for identifying four primary emotions from the IEMOCAP database. The second baseline is presented in the work by Pepino et al. [92], which leverages wav2vec 2.0 for extracting speech characteristics and combines them with hand-crafted features from eGeMAPS [93], achieving top-tier classification outcomes on the IEMOCAP collection. The third model framework introduced by Goncalves and Busso [75] highlights a transformer-based architecture with multimodal losses, setting new SOTA records on the CREMA-D and IMPROV(P) databases. For a fair comparison, we have focused only on their "audio-only" mode version, using 65 discrete acoustic low-level descriptors for input.

All these baseline models were reconstructed as described in their respective research studies and evaluated under the same experimental conditions as the model. Following previous research, the investigations treat SER as a single-label problem, providing performance results under MR or PR test paradigms while training and evaluating models across all primary emotion categories. Table 4.5 details the comparative

Table 4.5: Table shows the comparative evaluation between our proposed model and existing SOTA baselines for the IMPROV(P), CREMA-D, IEMOCAP, and PODCAST(P) databases. The performance measurements are in the macro-F1 score, capturing the effectiveness of models by grouping labels in the test sets following the "majority rule" (MR) or "plurality rule" (PR).

Aggregation	Method	MR			PR
		IMPROV(P)	CREMA-D	IEMOCAP	PODCAST(P)
MR_{Train}/PR_{Train}	Li et al. [91]	0.398	0.311	0.256	0.150
	Pepino et al. [92]	0.331	0.223	0.191	0.142
	Goncalves et al. [75]	**0.539**	0.574	0.261	0.161
	The proposed	0.512	**0.591**	**0.269**	**0.184**
AR_{Train}	The proposed	**0.562**	**0.585**	**0.279**	**0.166**

Table 4.6: Table illustrates the Kullback-Leibler divergence (KLD) when training and testing with each aggregation method under each label-learning strategy for each database. We highlight in bold the best performance for each condition. We denote $*$, †, and $\star$ when a model has significantly better performance than a model training with MR_{Train}, PR_{Train}, and AR_{Train}, respectively.

Database	Aggregation (training)	Hard-label learning					Soft-label learning					Distributional-label learning				
		MR	PR	AR	PR - MR	AR - PR	MR	PR	AR	PR - MR	AR - PR	MR	PR	AR	PR - MR	AR - PR
IMPROV(P)	MR_{Train}	0.235†	0.231†	0.229†	0.143	0.190	0.175	0.172	0.171	**0.108**	**0.150**	**0.232**	**0.233**	**0.236**	**0.271**	0.285
	PR_{Train}	0.256	0.252	0.249	**0.138**	0.193	**0.172**	**0.169**	**0.169**	0.117	0.162	0.240	0.242	0.244	0.289	0.286
	AR_{Train}	**0.211**$*$†	**0.208**$*$†	**0.206**$*$†	0.139	**0.180**	0.182	0.180	0.178	0.115	0.157	0.237	0.238	0.241	0.277	**0.283**
CREMA-D	MR_{Train}	0.078	0.091	0.094	0.122	0.129	**0.055**	0.058	0.059	0.066	0.066	0.112	0.136	0.142	0.192	0.203
	PR_{Train}	**0.064**$*$	**0.073**$*$	**0.075**$*$	0.094$*$	0.099$*$	0.058	0.057	0.057	0.056$*$	0.059$*$	0.109	0.131	0.136	0.185	0.188$*$
	AR_{Train}	0.068$*$	0.075$*$	0.076$*$	**0.092**$*$	**0.095**$*$	0.058	**0.056**	**0.056**$*$	**0.053**$*$	**0.055**$*$	**0.103**$*$	**0.122**$*$†	**0.126**$*$†	**0.166**$*$†	**0.178**$*$
PODCAST (P)	MR_{Train}	**0.150**	0.142	0.141	0.128	0.136	**0.157**	**0.142**	**0.139**	0.114	0.125	0.203	0.219	0.223	0.248	**0.242**
	PR_{Train}	0.150	**0.137**	**0.135**	**0.112**$**$	**0.125**$**$	0.164	0.144	0.140	**0.110**	**0.122**	0.203	0.218	0.222	0.244	0.243
	AR_{Train}	0.170	0.153	0.150	0.123	0.134	0.172	0.151	0.146	0.113	0.125	**0.200**	**0.214**	**0.220**	**0.240**	0.245
IEMOCAP	MR_{Train}	0.203	0.201	0.202	0.180	0.205	0.162	0.160	0.156	0.128	0.147	0.211	0.212	0.221	0.223	0.245
	PR_{Train}	0.202	0.201	0.201	0.186	0.201	0.150$*$	0.148$*$	0.145$*$	0.125	0.135$*$	**0.204**	**0.206**	0.214	0.225	0.237
	AR_{Train}	**0.185**	**0.183**$*$†	**0.182**$*$†	**0.159**$*$†	**0.178**$*$†	**0.143**$*$	**0.141**$*$	**0.138**$*$	**0.120**	**0.130**$*$	0.208	0.209	**0.214**	**0.217**	**0.227**$*$
IMPROV (S)	MR_{Train}	0.118	0.128	0.130	0.139	0.147	**0.090**†$\star$	0.088	0.089	0.087	0.094	**0.116**	0.158	0.163	0.205	0.196
	PR_{Train}	**0.103**$**$	**0.103**$*$	**0.104**$*$	0.102$*$	**0.109**$*$	0.103	0.089	0.088	0.074$*$	0.082$*$	0.119	0.150	0.155	0.184$*$	0.190
	AR_{Train}	0.116	0.108$*$	0.108$*$	**0.099**$*$	0.110$*$	0.100	**0.087**	**0.086**	**0.072**$*$	**0.080**$*$	0.120	**0.146**$*$	**0.150**$*$	**0.174**$*$	**0.180**$*$
PODCAST (S)	MR_{Train}	0.085	0.098	0.102	0.100	0.113	**0.074**†$\star$	0.071	0.073	0.071	0.078	0.079	0.144	0.151	0.153	0.171
	PR_{Train}	**0.075**	**0.062**$*$	**0.063**$*$	**0.060**$*$	0.067$*$	0.091	**0.060**$*$	**0.060**$*$	**0.056**$*$	**0.058**$*$	**0.067**	0.123$*$	0.132$*$	0.131$*$	0.156$*$
	AR_{Train}	0.084	0.064$*$	0.064$*$	0.061$*$	**0.065**$*$	0.097	0.062$*$	0.061$*$	0.057$*$	0.058$*$	0.073	**0.117**$*$	**0.126**$*$	**0.124**$*$	**0.149**$*$

results. We outlined comparisons with the other SOTA models trained on MR$_{Train}$ datasets for the CREMA-D and IEMOCAP repositories and PR$_{Train}$ datasets for the IMPROV(P) and PODCAST(P). The model, trained on the AR$_{Train}$ set, surpasses all examined SOTA results on the IMPROV(P), CREMA-D, IEMOCAP, and PODCAST(P) databases. These findings substantiate that the SER approach is robust against the SOTA references.

Specifically, on the IMPROV(P) corpus, our method achieves a macro-F1 score of 0.562 with AR$Train$, outstripping the result from Goncalves and Busso [75] which had a score of 0.539. Nevertheless, training with PR$Train$ resulted in a lower macro-F1 score of 0.512 compared to 0.539. For the CREMA-D set, we report a result of 0.591 using the MR$_{Train}$ set and 0.585 with the AR$_{Train}$ scenario in a macro-F1 score, both excelling beyond the SOTA score of 0.574. Moreover, our approach similarly outperforms other SOTA methodologies [75,91,92] in both the IEMOCAP and PODCAST(P) datasets.

4.5.2 Assessment with Full and Partial Test Data

Table 4.7 and Table 4.6 present the macro-F1 scores and KLD values for the different combinations of aggregation methods and label-learning strategies, respectively. These scores are derived from 18 experiments using various databases, where models were trained with the MR_{Train}, PR_{Train}, or AR_{Train} sets (6 databases × 3 learning strategies). Figures 4.2 and 4.3 illustrate the average performance for each evaluation set (MR, PR, AR, PR-MR, or AR-PR). A small-sample test of the hypothesis (matched pairs) was conducted on these results. Table 4.8 summarizes the overall averaged results from 18 experiments, and the bold numbers mean the better performance on each test set. Based on the KLD metric, the SER system trained with the training set selected by the proposed AR can perform overall better than those trained with MR_{Train} and PR_{Train}.

Our first research question investigates: **How is the performance of SER systems influenced by different aggregation methods used in the training set?** We address this question by evaluating the models trained under varying conditions on complete

Table 4.7: The table presents the macro-F1 scores achieved when models are trained and tested using each aggregation method across various label-learning strategies for each database. The highest performance for each condition is highlighted in bold. Symbols such as $*$, †, and $\star$ are used to indicate when a model significantly surpasses the performance of those trained with MR_{Train}, PR_{Train}, and AR_{Train}, respectively.

	Aggregation	Hard-label learning					Soft-label learning					Distributional-label learning				
Database	(train/test set)	MR	PR	AR	PR - MR	AR - PR	MR	PR	AR	PR - MR	AR - PR	MR	PR	AR	PR - MR	AR - PR
IMPROV(P)	MR_{Train}	0.512†	0.507†	0.555†	0.300	**0.516**†	0.595	0.587	0.613	**0.346**	0.530	**0.612**	**0.604**	0.599	**0.401**	0.440
	PR_{Train}	0.450	0.448	0.513	0.305	0.465	**0.600**	**0.593**	**0.623**	0.341	**0.531**	0.601	0.596	0.590	0.359	0.436
	AR_{Train}	**0.562**$*$†	**0.555**$*$†	**0.593**$*$†	**0.335**	0.498	0.576	0.569	0.602	0.339	0.518	0.602	0.594	**0.600**	0.340	**0.441**
CREMA-D	MR_{Train}	0.591	0.532	0.551	0.381	0.500	0.640	0.575	0.671	0.409	0.651	0.518$\star$	**0.474**$\star$	0.411	**0.357**	0.368
	PR_{Train}	**0.600**	**0.545**	0.595$*$	**0.390**	0.572$*$	0.667	0.594	0.699$*$	0.416	0.688	**0.518**$\star$	0.473$\star$	**0.419**	0.357	**0.374**
	AR_{Train}	0.585	0.528	**0.607**$*$	0.386	**0.593**$*$	**0.673**	**0.615**$*$	**0.710**$*$	**0.444**	**0.706**$*$	0.486	0.442	0.414	0.340	0.370
PODCAST (P)	MR_{Train}	0.214$\star$	0.184$\star$	0.303	0.143	0.300	0.215	0.185	0.326	0.145	0.328	0.161	0.137	0.162	0.102	0.159
	PR_{Train}	**0.259**$**$	**0.232**$**$	**0.403**$**$	**0.187**$**$	**0.420**$**$	**0.241**$\star$	**0.207**$**$	**0.397**$**$	**0.160**$\star$	**0.408**$**$	0.195$*$	0.166$*$	0.192$*$	0.126$*$	0.184$*$
	AR_{Train}	0.192	0.166	0.330$*$	0.129	0.351$*$	0.199	0.174	0.355$*$	0.138	0.367$*$	**0.204**$*$	**0.175**$*$	**0.200**$*$	**0.139**$*$	**0.192**$*$
IEMOCAP	MR_{Train}	0.269	0.260	0.339	0.203	0.351	0.346	0.343	0.412	0.257	0.426	0.354	0.341	0.299	0.253	0.287
	PR_{Train}	0.259	0.254	0.345	0.186	0.355	0.369	0.359	0.433	**0.279**	0.453	**0.377**	0.361	0.320	0.253	0.306
	AR_{Train}	**0.279**	**0.268**	**0.365**	**0.238**†	**0.378**	**0.390**$*$	**0.383**$*$	**0.464**$*$†	0.266	**0.479**$*$	0.369	**0.361**	**0.325**$*$	**0.265**	**0.317**
IMPROV (S)	MR_{Train}	0.424	0.254	0.229	0.234	0.245	**0.451**	0.299	0.379	0.278	0.386	0.361	0.185	0.137	0.149	0.150
	PR_{Train}	**0.455**$\star$	**0.340**$*$	0.328$*$	**0.318**$*$	0.360$*$	0.433	0.353$*$	0.483$*$	0.342$*$	0.505$*$	0.397	0.248$*$	0.181$*$	0.219$*$	0.189$*$
	AR_{Train}	0.391	0.315$*$	**0.337**$*$	0.311$*$	**0.365**$*$	0.410	**0.360**$*$	**0.491**$*$	**0.343**$*$	**0.522**$*$	**0.431**$*$	**0.306**$*$†	**0.216**$*$†	**0.282**$*$†	**0.227**$*$†
PODCAST (S)	MR_{Train}	0.344	0.078	0.138	0.076	0.141	**0.389**$\star$	0.080	0.199	0.076	0.198	0.352	0.051	0.060	0.047	0.059
	PR_{Train}	**0.392**$\star$	0.113$*$	0.327$*$	0.111$*$	0.328$*$	0.321	0.122$*$	0.450$*$	0.122$*$	0.457$*$	0.412	0.076$*$	0.078$*$	0.072$*$	0.074$*$
	AR_{Train}	0.283	**0.125**$*$	**0.352**$*$†	**0.124**$*$†	**0.357**$*$†	0.237	**0.139**$*$	**0.457**$*$	**0.142**$*$	**0.466**$*$	**0.425**	**0.078**$*$	**0.091**$*$†	**0.075**$*$	**0.088**$*$†

Table 4.8: The table shows averaged macro-F1 scores across 18 experiments listed in Table 4.7 and Table 4.6 with different databases and label-learning strategies on the different evaluation sets generated by three rules, *majority rule* (MR), *plurality rule* (PR), and *all-inclusive rule* (AR). The ↓ means the lower values mean the higher performance; the ↑ is the opposite.

Metric	**KLD ↓**					**Macro-F1 Score ↑**				
Train\Test	**MR**	**PR**	**AR**	**PR-MR**	**AR-PR**	**MR**	**PR**	**AR**	**PR-MR**	**AR-PR**
MR$_{Train}$	0.1408	0.1491	0.1512	0.1488	0.1623	0.4082	0.3153	0.3546	0.2309	0.3353
PR$_{Train}$	0.1406	0.1425	0.1433	0.1382	0.1507	**0.4192**	0.3378	0.4098	0.2524	0.3947
AR$_{Train}$	**0.1404**	**0.1397**	**0.1402**	**0.1334**	**0.1461**	0.4052	**0.3418**	**0.4172**	**0.2576**	**0.4019**

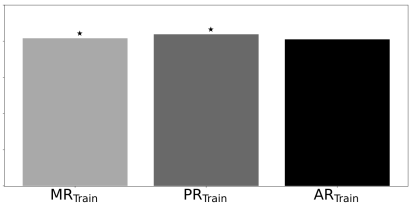
and incomplete test data in a cross-corpus setting.

Assessment on the Complete Test Set (AR)

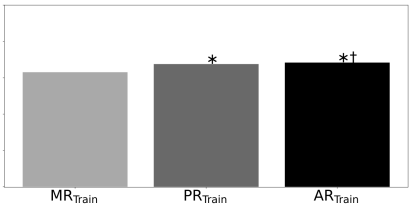
When evaluating using the AR approach with all annotated data in the test set, Figure 4.2c demonstrates that the macro-F1 score obtained with the AR_{Train} set is significantly higher than those with the MR_{Train} and PR_{Train} sets across the 18 conditions. Remarkably, models trained with the AR_{Train} set achieved the highest macro-F1 scores in 14 of 18 experiments, as detailed in Table 4.7. These results indicate that training with the AR approach can substantially enhance performance compared to the MR or PR criteria. This underscores the value of incorporating more annotated samples during the training process of SER tasks.

Assessment of the Incomplete Test Sets (MR & PR)

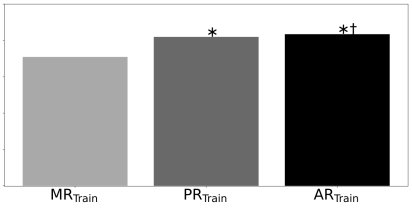
The single-label SER performance was assessed under the MR and PR test conditions. As detailed in Table 4.7, testing with the PR set consistently yielded lower performance than testing with the MR set, as the MR set omits more ambiguous samples. Similarly, performance in the PR-MR condition was generally worse than in the PR conditions. These results highlight that including more ambiguous samples (lacking majority consensus) in the test set—common in practical scenarios—can degrade SER model performance. Therefore, using only PR or MR to define the test set might not accurately reflect the outcomes likely to be encountered in real-world deployments where



(a) Macro-F1 scores on MR set.



(b) Macro-F1 scores on PR set.



(c) Macro-F1 scores on AR set.

Figure 4.2: Averaged macro-F1 scores across 18 experiments (as shown in Table 4.7) involving different databases and label-learning strategies on various evaluation sets generated by the three rules: *majority rule* (MR), *plurality rule* (PR), and *all-inclusive rule* (AR). The notations $*$, †, and $\star$ are used to indicate when a model achieves significantly better performance than models trained with MR_{Train}, PR_{Train}, and AR_{Train}, respectively.

recognition of every sentence is necessary.

Out of the 18 conditions analyzed, Table 4.7 indicates that when tested with the PR set, training with the AR_{Train} configuration resulted in the best performance in 11 out of 18 cases (approximately 61%), and 14 out of 18 cases (approximately 78%) when tested with the AR set. Figure 4.2b presents statistically significant evidence that the average macro-F1 scores for models trained with AR_{Train} were superior to those trained with the PR_{Train} set. These findings suggest that using the AR approach for training aggregation may enhance SER performance on samples with lower annotation agreement.

For tests evaluated on the MR condition, models trained with the AR_{Train} set outperformed others in only 7 out of 18 experiments (approximately 39%). Figure 4.2a shows a decline in performance for the AR_{Train} trained model compared to those trained with

either the MR_{Train} or PR_{Train} sets.

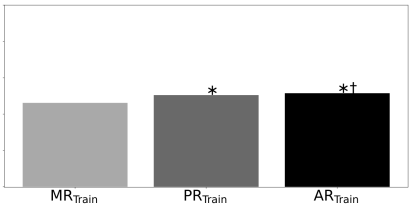
Including more complex samples in the training set (AR_{Train}) appears to reduce accuracy for the most straightforward samples; however, this trade-off bolsters the model's resilience in real-world contexts where both ambiguous and unambiguous samples are commonly encountered. This suggests that training SER systems with the AR_{Train} set could be more efficient for real-life applications, reflecting the genuine mix of sample types in practical environments. Additionally, increased training samples with the PR_{Train} set displayed better performance than those trained with MR_{Train}, aligning with findings reported by Chou et al. [18].

Assessment in Cross-Corpus Scenarios

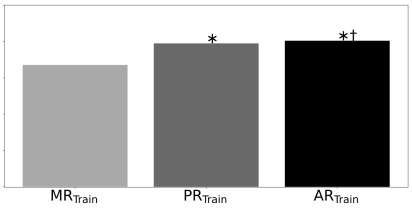
The past experimental outcomes were obtained using within-corpus settings. Our interest is to investigate the impact of training models with various aggregation strategies in cross-corpus settings. We've opted to train the model with the PODCAST (P) corpus and test its performance using the IMPROV (P) corpus to showcase the advantages of the AR method in cross-corpus experiments. The IMPROV (P) corpus comprises anger, sadness, happiness, and neutral emotions. On the other hand, the PODCAST (P) corpus includes the additional emotions of surprise, fear, disgust, and contempt, along with the emotions present in IMPROV (P). Given the emotional overlap, we can perform this cross-corpus evaluation where a SER model trained with the PODCAST (P) corpus aims to predict emotions within the IMPROV (P) corpus. Our approach involves directly utilizing the models trained on the PODCAST (P) corpus and assessing their efficacy on the IMPROV (P) set. We specifically focus on predictions for anger, sadness, happiness, and the neutral state. We transform the distribution predictions into binary labels by applying a threshold and then compute the results using the macro F1 score. For instance, let's consider a sample prediction for the IMPROV (P) dataset as: (angry, sad, happy, surprised, fear, disgusted, contempt, neutral) = (0.2, 0.2, 0.1, 0.1, 0.2,

Table 4.9: The table presents the cross-corpus macro-F1 scores for models trained using the 8-class MSP-PODCAST (P) dataset, applied to predict emotions in the 4-class IMPROV (P) dataset.

Test Set	MR	PR	AR	PR-MR	AR-PR
MR_{Train}	0.445	0.441	0.520	0.271	0.506
PR_{Train}	0.445	0.448	0.521	**0.295**	0.495
AR_{Train}	**0.458**	**0.459**	**0.523**	0.276	**0.520**



(a) Macro-F1 scores on PR-MR set.



(b) Macro-F1 scores on AR-PR set.

Figure 4.3: Averaged macro-F1 scores across 18 experiments (detailed in Table 4.7) with different databases and label-learning strategies on the varied evaluation sets generated by the PR-MR and AR-PR rules. The notations $*$, †, and $\star$ indicate significantly better performance compared to models trained using the MR_{Train}, PR_{Train}, and AR_{Train} sets, respectively.

0.1, 0.0, 0.1). We extract the four key predictions without renormalizing them: (angry, sad, happy, neutral) = (0.2, 0.2, 0.1, 0.1). The next stage involves using the threshold $1/C$ =1/8 to convert predictions into a binary form: (1, 1, 0, 0). Assuming the ground truth of the sample is: (anger, sadness, happiness, neutral) = (0.4, 0.4, 0.1, 0.1). Given a threshold of 1/4 for the IMPROV (P) corpus, the binary conversion is (1, 1, 0, 0). In this case, the prediction accuracy equates to 100%.

Table 4.9 presents the macro-F1 scores for cross-corpus testing, generated using the MR, PR, AR, PR-MR, and AR-PR labels. The results indicate that using the AR_{Train} set for training yields superior performance on the MR, PR, AR, and AR-PR test sets. This assessment also highlights the presented approach's efficacy for cross-corpus evaluations.

4.5.3 Evaluation on the Ambiguous Set

We address the second research question: **is the performance of an SER system on ambiguous emotions enhanced when trained with data derived from the all-inclusive rule as opposed to data obtained using the majority or plurality rules?**

Performance on the $AR - PR$ Condition

We examine the results under the $AR - PR$ test condition, which includes only the samples from the AR dataset that are not part of the PR dataset. As detailed in Table 4.7, models trained with the MR_{Train} set did not yield the best performances in 17 out of 18 cases (approximately 94%) when evaluating this condition. In these evaluations, models trained with the PR_{Train} set demonstrated significantly better performance in 10 out of 18 experiments (approximately 56%). Further, Figure 4.3b indicates that the model trained with the MR_{Train} set produced the lowest classification performance in the averaged macro-F1 score. These findings suggest that including more ambiguous data in the training set (i.e., models trained with either the PR_{Train} or AR_{Train} sets) enhances performance on sentences embodying more unclear emotions. Consequently, we conclude that training SER models exclusively with the MR dataset is ineffective in predicting ambiguous emotions.

Additionally, Figure 4.3 reveals that models trained with the AR_{Train} set achieve significantly higher averaged macro-F1 scores compared to those taught with either the MR_{Train} set or the PR_{Train} set on both $AR - PR$ and $PR - MR$ test conditions. Therefore, we recommend employing the AR approach to select training data for SER tasks.

Analysis of the Feature Embeddings

The goal is to display model embeddings trained with different sentence aggregation methods, as illustrated in Figure 4.4. The investigation utilizes the PODCAST (P)

dataset, concentrating on anger, sadness, happiness, and neutral emotions for clearer visual representation. Our focus is on test set segments with either high or low levels of agreement—Cohen's Kappa statistic [94] is used to define high and low agreement groups. From the test samples, the top 2% showing high agreement is chosen: 21 samples for "sadness," 33 for "anger," 97 for "happiness," and 139 for "neutral," as these are assumed to represent speech with emotional solid consensus.

Additionally, we explore ambiguous cases by selecting the top 2% of test samples with low agreement, meaning these samples largely lack clear consensus. We analyze sentences containing a mix of two emotions along with their respective quantities in brackets: anger-neutral (30), sadness-happiness (18), neutral-happiness (67), and anger-sadness (30). These cases are indicative of low agreement among annotators.

We utilized T-SNE (*T-distributed stochastic neighbor embedding*) to project the 1,024-dimensional feature vectors onto two-dimensional plots, aiming to visualize the data distribution. The visualizations include two primary emotions along with one composite emotion; for instance, the plot may feature "anger," "sadness," and the composite "anger-sadness" emotions. Each plot centers the emotional label around the mean values of the sentences from each class. Figure 4.5 illustrates the embeddings produced by models trained on the AR_{Train} and MR_{Train} sets. For conciseness, analogs for the PR_{Train}, $AR-PR_{Train}$, and $PR-MR_{Train}$ sets are omitted. The T-SNE plots reflect distinct separations between high-agreement emotions, with the two-emotion complex

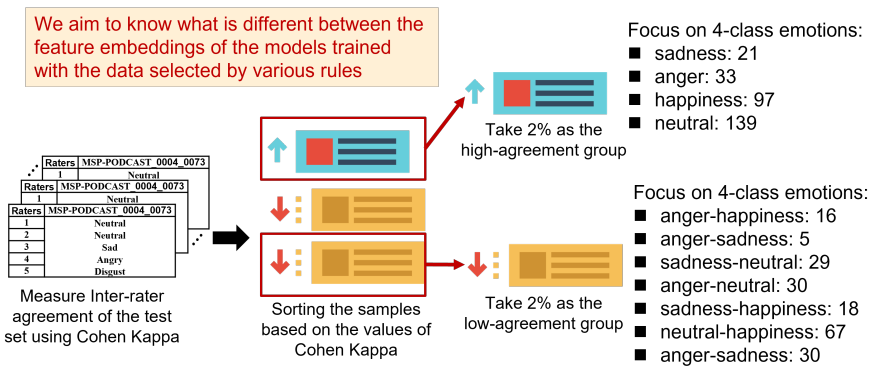
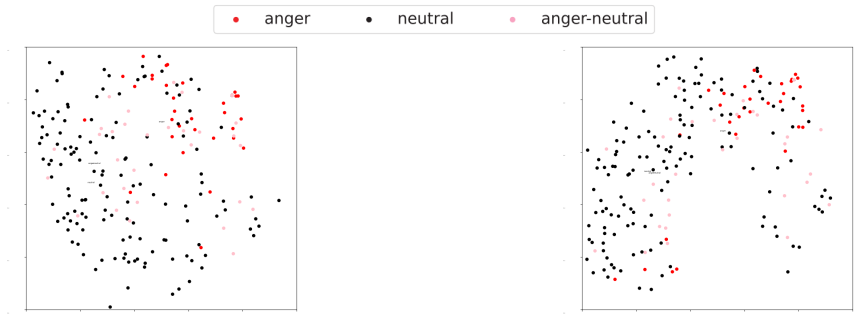
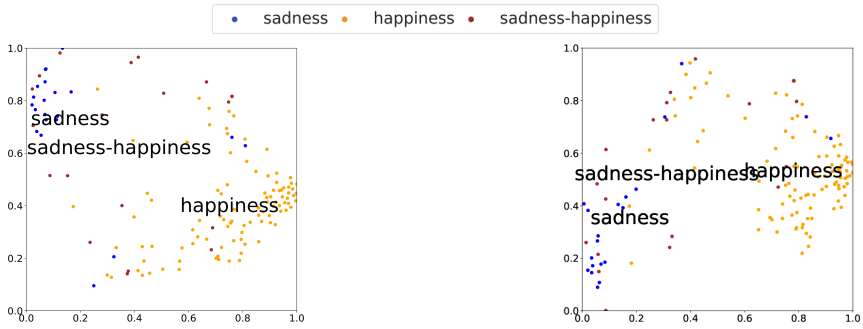


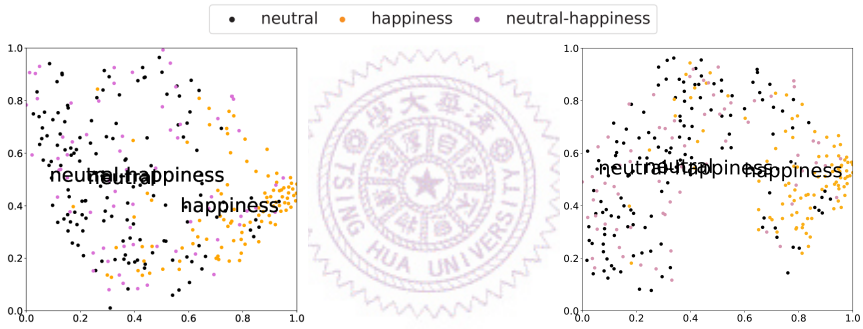
Figure 4.4: The figure depicts the procedure of visualizing feature embeddings.



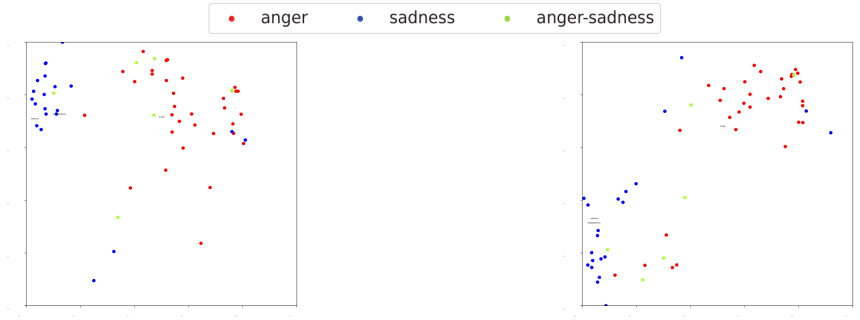
(a) anger-neutral,AR$_{Train}$ (b) anger-neutral,MR$_{Train}$



(c) sadness-happiness,AR$_{Train}$ (d) sadness-happiness,MR$_{Train}$



(e) neutral-happiness,AR$_{Train}$ (f) neutral-happiness,MR$_{Train}$



(g) anger-sadness,AR$_{Train}$ (h) anger-sadness,MR$_{Train}$

Figure 4.5: T-SNE visualizations using embeddings generated from models trained with the MR$Train$ and AR$Train$ sets show the distribution of feature embeddings. This analysis includes the following emotion pairs: anger-neutral, sadness-happiness, neutral-happiness, and anger-sadness.

samples often appearing between pure emotions with high agreement.

Comparing embeddings from models trained on the AR_{Train} versus MR_{Train} sets,

Table 4.10: Silhouette scores for emotional clusters observed in the embeddings are analyzed. The feature embedding analysis includes the following emotion pairs: anger-neutral (ang.-neu.), sadness-happiness (sad.-hap.), neutral-happiness (neu.-hap.), and anger-sadness (ang.-sad.). The highest silhouette score for each pair is emphasized in bold.

Case	ang.-neu.	sad.-hap.	neu.-hap.	ang.-sad.
MR_{Train}	-0.0366	0.4085	0.0819	0.0819
PR_{Train}	0.0502	0.3994	0.1291	0.0492
AR_{Train}	**0.0618**	**0.4571**	**0.1369**	0.1627
AR-PR_{Train}	-0.1371	0.1597	0.0166	0.2695
PR-MR_{Train}	-0.1379	0.17	0.0108	**0.4395**

it is apparent that the AR_{Train} set offers superior separation between classes, as evidenced by the centralized labels of emotions in the plots. Further validation comes from the silhouette score [95], a metric assessing how healthy clusters are formed within the embedding space (ranging from -1 for poor clustering to +1 for ideal clustering). We derived 1,024-dimensional feature representations using models trained under differing consensus-agreement conditions. Table 4.10 provides the silhouette scores across three clusters: the first emotion, the second emotion, and their respective composite cases. This analysis incorporates embeddings generated from models trained on PR_{Train}, $AR - PR_{Train}$, and $PR - MR_{Train}$ sets.

According to the table, the model trained with the AR_{Train} set ranks highest in silhouette score for "anger-neutral," "sadness-happiness," and "neutral-happiness" clusters. Intriguingly, the model trained with $PR - MR_{Train}$ had the top silhouette score for the "anger-sadness" cluster. Generally, models trained on datasets containing more ambiguous samples are better able to cluster responses to complex emotion scenarios than models trained solely on the MR_{Train} engagement.

Impact of Extra Data Incorporated by the AR Approach

One advantage of using the AR_{Train} set is the larger amount of data incorporated in training, as it utilizes all available samples, unlike the MR_{Train} or PR_{Train} sets. However, this extra data isn't the only contributor to the effectiveness of this strategy.

We performed experiments comparing models trained on datasets of similar size using both oversampling and undersampling strategies.

For the oversampling approach, we generated synthetic data following the method proposed by Pappagari et al. [76]. The data generation continued until it matched the quantity used in the AR_{Train} set. Table 4.11 presents the results, with the "Real Data" column indicating the number of data samples in the training set and the "Synthetic Data" column showing how many utterances were generated as synthetic samples. Table 4.11 illustrates that the models trained with the AR_{Train} set consistently outperform both the "MR_{Train} + synthetic data" and "PR_{Train} + synthetic data" models, underscoring the advantage of including additional samples from the $AR - PR_{Train}$ set in predicting ambiguous samples.

For the undersampling strategy, we conducted experiments to reduce training set sizes for PR_{Train} and AR_{Train} to match the sample size of MR_{Train}. The samples to be removed were selected randomly. Table 4.11 summarizes the macro-F1 scores for this approach, showing that models trained with the AR_{Train} set deliver superior performance on the AR, $PR - MR$, and $AR - PR$ test sets.

Additionally, a secondary undersampling strategy was implemented, training models under two conditions of uniform size across MR_{Train}, PR_{Train}, and AR_{Train}. The first condition included 20,000 randomly chosen training points under their respective aggregation rules (MR, PR, or AR). The second condition added 12,831 random points

Table 4.11: Analysis of the impact of additional data introduced through the AR approach. The table outlines two strategies: the oversampling approach, which leverages data augmentation, and the undersampling approach, which involves randomly removing samples to achieve consistency in the dataset.

Experiments	Train	Real Data	Synthetic Data	Reduce Data	MR	PR	AR	PR-MR	AR-PR
Oversampling	MR_{Train}	32,831	30,245	0	0.217	0.188	0.343	0.145	0.345
	PR_{Train}	51,243	11,833	0	0.206	0.178	0.368	0.136	0.368
	AR_{Train}	63,076	0	0	**0.234**	**0.211**	**0.398**	**0.172**	**0.407**
Undersampling	MR_{Train}	32,831	0	0	**0.237**	**0.207**	0.366	0.164	0.372
	PR_{Train}	32,831	0	18,412	0.214	0.186	0.380	0.142	0.388
	AR_{Train}	32,831	0	30,245	0.227	0.201	**0.387**	**0.164**	**0.399**

to result in 32,831 total samples. These extra points, picked randomly from the rest of the training set, do not meet consensus criteria. Since not all 32,831 samples have consensus for MR and PR rules, a soft-label learning strategy was applied. Table 4.12 lists the macro-F1 scores for both conditions, revealing that adding more samples is consistently beneficial. Notably, the AR_{Train} models frequently achieved the highest performance in both conditions (20,000 and 32,831 samples). These findings suggest that the AR method boosts the performance of SER models by including ambiguous data, with its benefits extending beyond simply enlarging the training set.

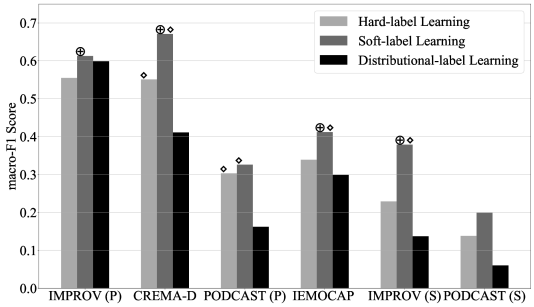
4.5.4 What is the most effective label learning method for SER?

We focus on the third research question: **Which label learning strategy is the most effective for training SER systems when evaluated on the entirety of the test set?**

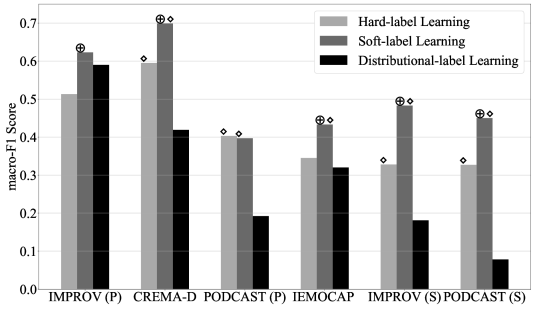
Figure 4.6 illustrates the performance of SER systems trained with various aggregation techniques and label learning methods. These results are evaluated on the entire test set utilizing the AR method. Among the label-learning techniques, the strategy that employs KLD as the cost function performs the worst. Utilizing the CE loss function yields superior SER performance compared to KLD. Within the methods that use CE, soft-label learning outperforms hard-label learning. Table 4.7 shows that SER systems using soft-label learning surpassed those using hard-label learning in 17 out of 18 cases (around 94%). This finding is consistent with prior studies, which have demonstrated

Table 4.12: The results for the undersampling strategy compare training sets consisting of either 20,000 samples, following the designated aggregation rules, or 32,831 samples formed by randomly adding 12,831 samples regardless of consensus. This table shows results in a macro-F1 score, underscoring the advantages of using the AR_{Train} set.

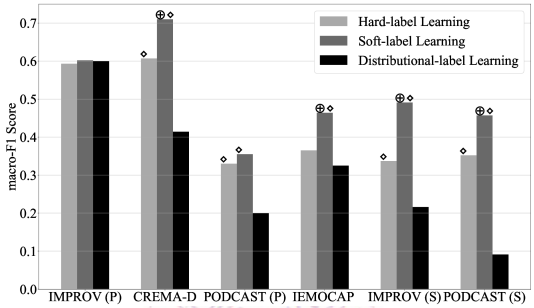
Train Set	#	MR	PR	AR	PR-MR	AR-PR
MR_{Train}	20,000	0.303	0.320	0.317	0.334	0.309
	32,831	**0.333**	**0.350**	**0.349**	**0.362**	**0.345**
PR_{Train}	20,000	0.336	0.375	0.377	0.404	0.382
	32,831	**0.350**	**0.391**	**0.394**	**0.424**	**0.404**
AR_{Train}	20,000	0.353	0.400	0.402	0.438	0.408
	32,831	**0.367**	**0.415**	**0.418**	**0.454**	**0.430**



(a) Training with MR.



(b) Training with PR.

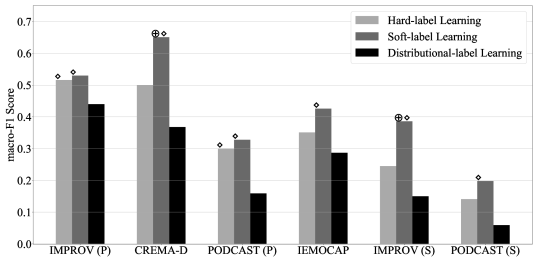


(c) Training with AR.

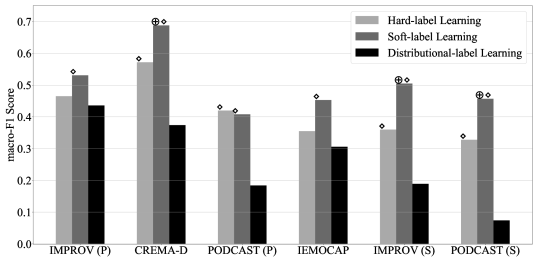
Figure 4.6: The bar plots depict the macro-F1 scores achieved with distributional-label learning, hard-label learning, and soft-label learning strategies. All models are assessed on the entire test set and aggregated by the AR strategy for each database. We use the symbols $\oplus$, ‡, and $\diamond$ to indicate when a model significantly outperforms those trained with distributional, soft, and hard-label learning strategies, respectively.

that representing emotions with soft-encoding and employing the CE loss function is a more effective label learning strategy for training SER models [2, 55, 57, 61, 65].

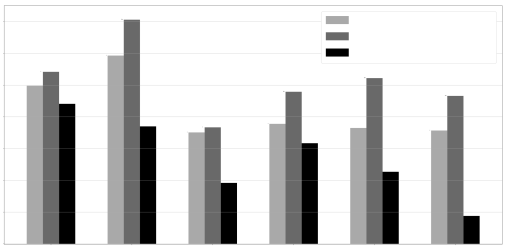
Additionally, Figure 4.7 provides an overview of the macro-F1 scores for each database, comparing three different label learning methods. We focus solely on the $AR - PR$ test set for more precise interpretation, as these samples are the most emotionally ambiguous. Figures 4.7a (training with MR), 4.7b (training with PL), and 4.7c (training with AR) demonstrate that the soft-label learning strategy is the most appropriate method among the existing learning approaches for training SER systems to rec-



(a) Training with MR.



(b) Training with PR.



(c) Training with AR.

Figure 4.7: The macro-F1 scores for each database are provided for distributional-label learning, hard-label learning, and soft-label learning strategies. All models have been evaluated on the **AR-PR** test set, which includes samples that do not achieve MR or PR consensus. Symbols $\oplus$, ‡, and $\diamond$ denote instances where a model significantly outperforms those trained using distributional, hard, and soft-label learning strategies, respectively.

ognize mixed emotions from the ambiguous samples within the $AR - PR$ set.

4.6 Summary

This paper examined speaker-independent categorical SER systems' performance using an all-inclusive test set, where no data was excluded, and our aggregation rule was applied. Compared to traditional label aggregation methods like the majority rule and plurality rule, the all-inclusive rule allows for the retention of all annotated data and emotional ratings, thereby making it possible to train and evaluate SER systems that are

tailored to real-world scenarios. The initial examination revealed that adhering to the majority or plurality rule excludes a notable portion of the annotated test samples, resulting in a poor representation of expected SER performance in realistic scenarios. In real-world applications, the classifier must recognize emotions in all sentences, regardless of consensus. Experiments with the comprehensive test set demonstrated that employing the all-inclusive aggregation rule to define the ground truth yields more reliable SER performance by incorporating more speech samples with low-agreement annotations. Our findings also indicated a decline in SER model performance as more ambiguous samples were included in the test set, highlighting the significance of using the complete test set. Additionally, we found that training exclusively with high-agreement data is insufficient for predicting ambiguous emotions. Moreover, the results showed that soft-label learning leads to the best performance among label-learning strategies when applied to the entire test set. The average SER performance of models trained with data selected through the all-inclusive aggregation rule consistently surpassed those trained with data specified by the majority or plurality rule on incomplete and comprehensive test sets.

Building upon the findings of this dissertation, future research could broaden the application of the all-inclusive rule to encompass other subjective tasks. Its effectiveness could mainly be assessed in areas like *text-to-speech* (TTS) and textless *speech-to-speech translation* (S2ST) systems. For instance, Zhou et al. [96] introduced a system that synthesizes human voices with mixed emotions; however, the limited range of emotions indicates the potential for enhancing emotion embedding. By employing the all-inclusive rule, which utilizes the entire dataset, it is anticipated that more authentic and varied emotional expressions will be achieved in TTS systems, surpassing current approaches. Moreover, despite its critical role in natural human interaction, present S2ST systems still need to incorporate emotional information [97, 98]. Acknowledging the significance of emotions in effective communication, integrating the all-inclusive rule

into S2ST systems is believed to significantly enhance their realism in speech conversion.

Chapter 5

Training Loss by Using Co-occurrence Frequency of Emotions

Previous research in SER focusing on categorical emotions has traditionally concentrated on identifying a single emotion per utterance based on the assumption that each emotion is distinct. However, it has emerged that emotions may overlap, particularly in ambiguous emotional expressions that combine elements, such as happiness mixed with surprise. As a result, emotion recognition systems have been updated to predict multiple emotional categories. This approach, though, typically neglects the relationships among different emotions during the training process, treating them independently. This study investigates the interconnected nature of emotional categories and how these relationships impact the training of SER models. Specifically, the co-occurrence frequencies of emotions are assessed based on perceptual evaluations within the training data. This information creates a matrix that assigns penalties according to class dependencies, enforcing harsher penalties for errors between more dissimilar emotions. This matrix combines three established label learning techniques using the modified loss function. Subsequently, SER models are developed using both the newly integrated penalization matrix and the traditional cost functions typically utilized. The

newly introduced penalization matrix significantly improves the macro F1-score on the PODCAST dataset, showing increases of 17.12%, 12.79%, and 25.8% for hard-label, multi-label, and distribution-label learning methods, respectively.

5.1 Motivation

SER is crucial for enhancing human-centered computer interactions. Emotional labels for training SER systems are generally obtained from perceptual evaluations. Nonetheless, subjectivity in emotion perception leads to interpreting the same speech, and often the same speech differently [6, 68]. Conventionally, SER studies handle the variability in emotional annotations as noise, using label aggregation methods to produce a consensus for training SER systems [99–103]. This methodology tends to disregard the possibility of co-occurring emotions frequently present in emotional expressions (e.g., a sentence might convey both happiness and surprise). Although multi-label learning [27, 77, 104] permits ground truths with several valid classes, it still overlooks the relationships among emotions by assuming all emotion classes are disconnected. The dissertation posits that categorical emotions are interrelated and suggests that considering the commonness of emotion co-occurrence in emotional ratings can create better SER systems.

Examining simultaneous emotions that appear in everyday life [21], Xu et al. [105] capitalized on the prevalence of concurrent emotions in perceptual assessments to forge initial linkages within their presented graph-based DNN. Meanwhile, the work [22] leveraged frequent emotional misclassifications by evaluators to evaluate an SER system's performance. Furthermore, some research works have utilized the soft-label method to capture secondary emotions present in voice [57, 61, 64]. This body of work highlights the importance of accounting for feedback from all evaluators in perceptual evaluations, even when there is variation from agreed-upon labels.

This research explores how incorporating emotional co-occurrence data can improve the training process of a SER system. It involves creating a co-occurrence frequency matrix that captures the relationships between different emotions as determined by perceptual evaluations. This matrix is subsequently converted into a penalization matrix, which adjusts the loss functions by assigning greater penalties for predicting combinations of emotions that seldom occur together. Our approach folds the penalization matrix into existing cost functions as a ”penalty loss,” thereby increasing the loss value if the model forecasts emotions with low co-occurrence frequencies. For instance, since anger and contempt are seldom co-selected in perceptual evaluations, predicting these two emotions jointly will incur a higher penalty than predicting commonly co-occurring emotions like anger and sadness. This strategy acknowledges the interdependence of emotions, leveraging their relationships to enhance model performance.

5.2 Background and Related Works

5.2.1 Contrastive Learning in Emotion Recognition

To enhance SER systems, Wang et al. [106] designed an objective function aimed at maximizing inter-class distance while minimizing intra-class distances. Their findings demonstrate the advantages of this approach. Additionally, Zhao [107] introduced a nearest neighbor contrastive learning technique to emphasize distinctions between various emotions by leveraging local neighborhood information, which subsequently boosted SER system performance under cross-corpus testing conditions. Different from the previously mentioned studies, the frequency of co-occurrence of emotions in the training set is used by us to model the relationship between emotion classes.

5.2.2 Label Learning in Emotion Recognition

This work explores the application of co-occurrence frequencies of emotional classes derived from perceptual evaluations to enhance the training process of an SER system. It examines related methods such as distribution-, hard-, multi-label learning, and emotion classification. To illustrate, we use a 4-class emotion classification task that includes angry (A), neutral (N), happy (H), and sad (S) emotions. In this scenario, five annotators independently assess a sentence, each assigning one label. For a single sample, an example set of labels might look like this: "S, N, S, N, S."

Emotion Recognition using Hard-label Learning

In SER studies, emotion databases annotated via perceptual evaluations often employ consensus labels derived from collating individual annotations through techniques like majority voting or the plurality rule. Such methods produce a definitive hard-label vector indicating a single emotional class, thereby excluding secondary emotions noted by raters who did not align with the majority group. For instance, the constructed one-hot vector could be (0, 0, 1, 0). However, few works have introduced soft labels [57, 61, 64], enabling each sample to be linked with multiple emotions. The work [18] built each annotator individually to capture the subjective nature of perceptual evaluations. These approaches permit the inclusion of sentences in the training dataset even when annotators have differing emotional perceptions. Contrary to systems trained with hard-label learning—which predict a single emotion per utterance and assume that emotions are independent—these advanced techniques consider emotional co-occurrence. My findings reveal that adopting the "penalty loss" method improves performance.

Emotion Recognition using Multi-label Learning

Multi-label learning empowers emotion classifiers to detect several emotions in a single data instance, generating multiple hard vectors that may highlight more than one emotion. For example, the emotions identified could be represented as (1, 0, 1, 0). Prominent examples of this approach in emotion recognition are found in studies like [27, 77, 108]. Although multi-label learning facilitates recording various emotions, it assumes these emotions manifest independently. This traditional methodology also fails to signify the relative importance of emotions, such as discerning primary emotions from secondary ones. In the research, we utilize the approach described by the work [104], applying multi-label learning to display enhancements in SER systems through the ”penalty loss” concept.

Emotion Recognition using Distribution-label Learning

Distribution-label learning [29] aims to produce a distributional output under the assumption that labels are spread across categories so that their probabilities add up to 1. For example, the distributional label might be (0.4, 0.0, 0.6, 0.0). During the training of the SER systems, the objective function can let the model minimize the distance between the actual and predicted distributions. A couple of notable applications of distribution-label learning to SER include the studies by [90] and [109]. Chou et al. [90] tailored distribution-label learning by translating emotional ratings into distribution labels, and we adopted a similar method to train the model. The KLD is typically used as the loss function in traditional distribution-label learning. Our research examines the impact of introducing a ”penalty loss” on the accuracy of classification results.

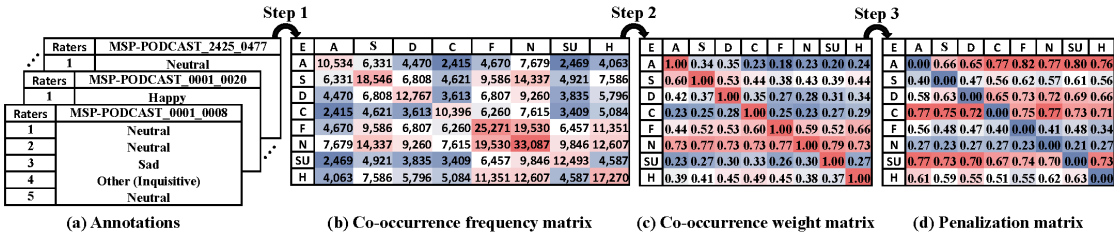


Figure 5.1: Illustration of a process for generating the presented penalization matrix. The 8-class emotions involved include contempt (C), neutral (N), sad (S), happy (H), fear (F), disgusted (D), angry (A), and surprised (SU). The procedure in detail can be found in Section 5.3.1.

5.3 Proposed Method

5.3.1 Penalization Weights based on the Counts of Co-Existing Emotions

Figure 5.1 details the three-phase process of forming the penalization matrix. In the initial phase, we adopt the method outlined by [105] to generate the co-occurring matrix from emotional labels in the training dataset. The proposed matrix indicates how frequently different pairs of emotions appear together, resulting in an **eight times eight symmetric matrix**. Take Figure 5.1 (b) as an example. The entry (”S”, ”S”) has the value 18,546, indicating that *angry* was chosen across 18,546 utterances. Among these occurrences, *neutral* was chosen 14,337 times. Consequently, the entries at positions (”S”, ”N”) and (”N”, ”S”) are populated with 14,337, respectively.

In phase 2, we create a matrix indicating the probability of co-existing emotions by dividing each element by the total count of instances marked with the corresponding emotion in each column. Take the second column labeled ”S” from Fig. 5.1 (b) as an example—the co-occurrence frequencies of *anger* with other emotions are 6,331, 18,456, 12,767, 3,613, 6,807, 9,260, 3,835, and 5,796. By dividing each of these frequencies by the total occurrences of *sad* (6,331), we obtain co-occurrence probabilities: 0.34, 1.00, 0.37, 0.25, 0.52, 0.77, 0.27, and 0.41. As an example, the co-existing possibility of disgust and sad (0.37) is greater than that of sad and contempt (0.25). This normalized

matrix is termed the co-existing weight matrix (Figure 5.1 (c)), which is asymmetric owing to the normalization performed column-wise.

In phase 3, the co-occurrence weight matrix converts into a ”penalization matrix.” This conversion is essential as it penalizes the SER models for forecasting rare emotion combinations. The conversion process is simple: Each element in the co-existing emotion weight matrix is subtracted from one, resulting in the penalization matrix (Figure 5.1 (d)). Employing the method causes an increase in the training loss if the model predicts infrequent combinations of emotions.

5.3.2 Label Processing to Train SER Systems

We incorporate the penalization matrix into the three labeling approaches outlined in Section 5.2: hard-, multi-, and distribution-label learning. Each technique uses a specific cost function. Hard-label learning relies on cross-entropy (CE) loss, multi-one on binary cross-entropy (BCE) loss, and distribution-label learning on the KLD loss. Furthermore, we apply the label smoothing strategy suggested by [89] to both hard-label and distribution-label learning, setting the parameter at 0.05. A slight value (10^{-6}) is also added to entries initially quantified as zero for the multi-label vector.

5.3.3 Loss Functions Integrated by the Proposed Penalization Matrix

The technique of integrating a penalization matrix into loss functions is proposed. This method requires the definition of $N \times K$ matrices for both the model's predictions (Y^P) and the actual labels (Y^T). Here, C signifies the number of emotional categories, while N denotes the number of samples under consideration. The variation in each row of Y^T depends on the label learning approach in use, whether it is distribution-, multi-, or hard-label learning. Afterwards, the loss value matrix ($L \in R^{N \times C}$) is determined

using the following approach:

$$L = f_{loss}(Y^T, Y^P) \tag{5.1}$$

The loss function, represented by (f_{loss}), could be the cross-entropy (CE). The elements of (L) are calculated as ($f_{loss}(Y^T ij, Y^P ij)$), where ($j \in 1, \ldots, K$) and ($i \in 1, \ldots, N$), to estimate the categorical emotion loss for each utterance. Subsequently, the proposed matrix has been incorporated into Equation 5.1. The matrix introduced, denoted as (P), belongs to ($R^{K \times K}$) (refer to Fig. 5.1). The integration of the loss function is achieved through the penalization matrix, represented by (L_{P+loss}), as follows:

$$\begin{aligned} L_{P+loss} &= \textstyle\sum_{i=1}^{N}(L_i \cdot P) \\ &= \textstyle\sum_{i=1}^{N}(\sum_{j=1}^{C}\sum_{z=1}^{C} P_{jz} \cdot f_{loss}(Y_{ij}^T, Y_{ij}^P)). \end{aligned} \tag{5.2}$$

When we replace f_{loss} with the objective functions, the equations will be:

$$L_{P+CE} = -\sum_{i=1}^{N}(\sum_{j=1}^{C}\sum_{z=1}^{C} P_{jz} \cdot Y_{ij}^T \cdot \log(Y_{ij}^P)), \tag{5.3}$$

$$\begin{aligned} L_{P+BCE} = -\sum_{i=1}^{N}(\sum_{j=1}^{C}\sum_{z=1}^{C}(P_{jz} \cdot Y_{ij}^T \cdot \log(Y_{ij}^P) \\ + P_{jz} \cdot (1 - Y_{ij}^T) \cdot \log(1 - Y_{ij}^P))), \end{aligned} \tag{5.4}$$

$$L_{P+KLD} = \sum_{i=1}^{N}(\sum_{j=1}^{C}\sum_{z=1}^{C} P_{jz} \cdot Y_{ij}^T \cdot \log(\frac{Y_{ij}^T}{Y_{ij}^P})). \tag{5.5}$$

In training, the initial loss is modified by the proposed loss, resulting in the total loss as described by Equation 5.6. In this equation, $\alpha \in R$ and β are assigned a value of

either 1 or 0.

$$\mathscr{L}_P^{loss} = \beta L_{loss} + \alpha \cdot L_{P+loss}. \tag{5.6}$$

5.4 Experimental Setup

5.4.1 Resource

We evaluate the proposed method using the Podcast corpus [39], using version 1.9 in this ongoing collection effort. The training set consists of 55,283 speech utterances, the development set includes 9,546, and the test set comprises 16,570 utterances—primary emotions in the perceptual evaluation cover 8 classes mentioned in Section 2.4. For hard-label learning, we only utilize speaking turns with agreed-upon consensus labels and discard segments lacking agreement. Further details can be found in Section 2.4.

5.4.2 Acoustic Features

In the research conducted by [110], the effectiveness of various attributes in SER was assessed across multiple publicly accessible emotional datasets. This analysis identified the wav2vec feature set, introduced by the study [111], as one of the most robust techniques for feature extraction. Consequently, the models incorporate the 512-dimensional wav2vec feature as input. To prepare the data for training, we standardized all features using z-normalization, which was calculated according to the mean and standard deviation from the training dataset.

5.4.3 SER Models and Other Details

We adopted the chunk-level SER models proposed by [112] as the primary model. This approach systematically processes sentences of varied lengths by converting them

into a specified number of uniformly sized chunks through overlapping adjustments. Based on the paper's suggestions, we utilized LSTM as the feature encoder at the chunk level, along with the RNN-AttenVec model for chunk-level attention [112]. This combination allows us to capture emotion-related information at different levels. For further details on the network architecture, see Lin and Busso [112]. Model parameters adhered to those in Chou et al. [19]. Based on previous studies, for the output layer, we used softmax activation for CE [99, 101, 103] and KLD [19, 29], with sigmoid activation for BCE [104, 113]. We used Adam optimizer and set the learning rate to 0.0001; we set the batch size to 128. Then, we train all models with the epochs of 25. Finally, we saved the best-performing model according to their minimum loss on the development set. To analyze the influence of the suggested objective function on the accuracy of SER systems, the parameter α in Equ. 5.6 was set to either 0.5 or 1.0. We performed experiments using only L_{P+loss} and using only L_{loss}.

5.4.4 Evaluation Metrics

To gauge the accuracy of prediction labels against ground truth, we resort to a suite of tailored metrics based on the type of label learning method in use. For hard-label learning, I used macro-, micro-, and weighted-F1 scores, unweighted average recall, and unweighted average precision. In the case of multi-label and distribution-label learning, I employ measures similar to those referenced in the study by [113]: ranking and hamming loss, coverage error, alongside macro F1. Moreover, micro- and weighted-F1 scores are integrated to estimate multi-label tasks specifically. For a performance assessment of multi-label and distribution-label methods, binarization of predictions is necessary. For multi-label learning, predictions' probabilities are changed into "multi-hot" binary vectors using a threshold of 0.5, following [104, 113]'s methodology. For distribution-label learning, a threshold is fixed at 1/convertingion of probability predictions into binary vectors, in line with Chou et al. [19].

5.4.5 Statistical Significance

The methodology from Lin and Busso [112] is employed, where the original test set is divided at random into 30 smaller subsets of roughly equal size. The average results for all metrics are then reported. A two-tailed t-test is utilized to conduct the statistical significance test, evaluating the difference between the proposed approach and the baseline models. A result is considered statistically significant if the p-value is less than 0.05.

5.5 Experimental Results and Analyses

Table 5.1 summarizes the average outcomes across various configurations. Evaluation metrics are marked with $\uparrow$ to show that higher values indicate better performance, while $\downarrow$ signifies that lower values are better. The $*$ symbol is used to flag results that show statistically significant improvements over baseline SER systems where the proposed loss is not applied (i.e., $\alpha = 0$ in Eq. 5.6). The table is divided into three sections corresponding to different objective functions: CE, BCE, and KLD for hard-, multi-, and distribution-label learning, respectively. The column labeled "β" denotes whether the loss L (with $\beta = 1$) was accounted for in the experiments as opposed to excluding it ($\beta = 0$). The α column states the value assigned to the PL loss weight. Specifically, the top-performing values for each loss function are emphasized in bold.

5.5.1 Does incorporating the penalty loss (L_{P+loss}) benefit SER Systems?

Table 5.1 reveals that models utilizing the proposed penalty loss (L_{P+loss}) typically achieve the best outcomes across different label learning strategies. For example, when the results are evaluated on the single-emotion utterances, the model with L_{P+loss} where

Table 5.1: The table overviews the results on distributional-label, multi-hard-label, and single-label tasks for the primary emotion classification task. The mark $*$ denotes that the outcomes for SER systems utilizing the presented matrix have statistical significance compared to the baseline ($\alpha = 0$; $\beta = 1$).

Single-label Task						
f_{Loss}	β α	**Unweighted Average Recall ↑**	**Unweighted Average Precision ↑**	**Macro F1 ↑**	**Micro F1 ↑**	**Weighted F1 ↑**
CE	**1 0**	0.144	0.133	0.111	0.424	0.318
	1 0.5	**0.156***	**0.137***	0.129*	**0.425**	**0.347**
	1 1	0.155*	0.136	**0.130***	0.408	0.346
	0 1	0.154*	0.136	0.128*	0.396	0.341
Muli-label Task						
f_{Loss}	β α **Hamming Loss ↓**	**Ranking Loss ↓**	**Coverage Error↓**	**Macro F1 ↑**	**Micro F1 ↑**	**Weighted F1 ↑**
BCE	**1 0** 0.304	0.603	6.899	0.219	0.466	0.352
	1 0.5 0.303	0.608	6.928	0.215	0.462	0.348
	1 1 **0.303**	**0.587**	**6.837**	0.235	**0.482**	0.370
	0 1 0.305	0.597	6.871	**0.247***	0.477	**0.378**
Distribution-label Task						
f_{Loss}	β α **Hamming Loss ↓**	**Ranking Loss ↓**	**Coverage Error↓**	**Macro F1 ↑**	**Micro F1 ↑**	**Weighted F1 ↑**
KLD	**1 0** **0.294**	0.511	6.279	0.283	0.522	0.431
	1 0.5 0.308	**0.507**	6.220	0.322*	**0.533**	0.471*
	1 1 0.315	0.509	**6.214**	0.330*	0.532	0.475*
	0 1 0.337	0.530	6.284	**0.356***	0.526	**0.496***

$\alpha = 0.5$ topped four out of five evaluation metrics and exhibited a 16.22% relative gain in macro F1-score (maF1) compared to the baseline. In the multi-label classification setting, the model using L_{P+loss} with $\alpha = 1.0$ led in four out of six metrics, showing a 7.31% relative gain in maF1 over the baseline performance. In terms of the distribution-label task, the majority of the best performances resonate from models integrated with L_{P+loss}. Explicitly focusing on maF1, the leading model with KLD reported a remarkable 25.8% relative enhancement over the baseline and surpasses SOTA results documented by the work [19] (31.6% maF1). This suggests that the presented penalty loss (L_{P+loss}) significantly improves model accuracy in primary emotion classification tasks. Furthermore, the results exhibit that aggregating the primary loss L_{loss} with the presented loss L_{P+loss} tends to boost overall effectiveness.

5.5.2 Effect of Co-occurrence Matrix

The aim is to determine if the proposed methodology effectively allows systems to capture the intended co-existing emotions matrix accurately. This is evaluated by measuring the distance between the co-existing emotions matrices derived from the learn-

ing targets in the training set and those obtained from model predictions, utilizing the Frobenius norm as the distance metric. The focus is on models that use either L_{P+loss} ($\alpha = 1; \beta = 0$) or L_{loss} ($\alpha = 0; \beta = 1$). When L_{P+loss} is implemented, a decrease in the distance was observed from 4.27 to 4.00 using the multi-label approach and from 4.13 to 3.39 using the distribution-label approach. This reduction signifies that the co-existing emotion matrix, as the proposed penalization method predicted, is more closely aligned with the target co-occurrence matrix.

5.6 Summary

This dissertation utilized co-existing emotion counts to create a penalization weight for training a speech emotion recognition (SER) model. The proposed penalty loss considers the relationships between emotions, applying more significant penalties for inaccuracies between more distinct emotions. The effectiveness of this penalty loss was assessed through three distinct label learning approaches: distribution-based, hard-label, and multi-label learning. The results suggest that the proposed loss generally improves recognition performance across most evaluation metrics in eight classes' primary emotion classification tasks.

Chapter 6

Conclusion

This dissertation introduces three proposed methods to optimize the pipeline for constructing speech emotion recognition (SER) systems. These methods have been documented in our studies, including rater-modeling [18], all-inclusive [114], and the penalization objective function that accounts for co-currencies of emotions [90]. The findings from this dissertation and the respective studies [18, 90, 114] have facilitated answering the primary research questions.

(1) Should we remove the minority of emotional ratings? The minority of emotional ratings should not be discarded, as retaining them can enhance the recognition of ambiguous utterances with complex emotional perceptions. Adding these minority ratings also boosts the performance of SER systems on comprehensive test sets that mirror real-world conditions. Further analyses reveal that standard approaches, which neglect these minority ratings, show reduced capabilities in clustering samples with dual emotions. Therefore, including the minority of emotional ratings is essential to account for the subjectivity inherent in emotion perception.

(2) Should we only let the SER systems learn the emotional perceptions of a few people? The findings of this dissertation indicate that allowing SER systems to learn emotion perception from a larger group can enhance recognition capabilities through the proposed all-inclusive rule or individual-rater modeling method. This not only im-

proves the performance of SER systems but also offers a unique opportunity for personal growth. By incorporating emotional ratings from a broader range of people, each person's unique sensitivity and emotional background can contribute valuable insights. This should inspire and motivate researchers, developers, and practitioners in the field of SER to continue their work and strive for excellence.

(3) Should SER systems only predict one emotion per speech? To address real-world conditions that involve the simultaneous occurrence of various emotion classes, SER systems need training to predict multiple emotions simultaneously. This dissertation convincingly illustrates that multi-label SER systems can achieve superior performance on test sets featuring a single consensus label determined by the plurality rule and the all-inclusive rule. This underlines the effectiveness of multi-label emotion prediction, which should be incorporated into SER systems rather than focusing on a single emotion.

6.1 Discussion and Limitation

While the dissertation proposes three methods to address challenges in SER, several other factors need consideration. For instance, the effect of various inputs, such as handcrafted features or raw audio signals, on prediction outcomes should have been analyzed. Additionally, the relationships between layer embeddings and label spaces should be explored. Visualizing these relationships could provide fascinating insights. Furthermore, all the suggested methods are designed for traditional SER emotion databases, while the latest dataset [115] includes intensity scores for each categorical emotion. Using this type of emotion database might mean our methods must be optimized for training and evaluating SER systems. It would be intriguing to see if our methods could be adapted for samples with intensity scores.

Furthermore, the proposed comprehensive rule recommends incorporating all data

into the emotion dataset. However, an examination of how the dataset size impacts SER system performance was not conducted. This omission stems from the fact that the largest existing emotion dataset comprises merely around 200,000 utterances, significantly less than databases used in other speech tasks such as Automatic Speech Recognition (ASR). Additionally, extensive experiments under cross-domain conditions were not undertaken, and only a single experiment was performed. It is essential to note that cross-domain experiments are crucial for real-world application of SER systems.

In conclusion, this dissertation recommends utilizing the original emotion classifications within the database. A pertinent question that arises is how many emotion classes are sufficient to accurately capture emotional perception in real-world scenarios. According to [7], humans are capable of detecting 24 distinct emotions through vocal expressions. Therefore, it would be intriguing to combine multiple emotion databases to encompass these 24 emotions.

6.2 Future Works

All data and emotional ratings in the emotion databases can be utilized with the proposed methods during inference. In future studies, the first benchmark for SER will be proposed based on the partitions defined in this dissertation, allowing for a standardized evaluation of models and easy comparison of performances across different SER models. The intention is to examine performance disparities caused by natural biases in emotional perceptions, including gender, race, and ethnicity. Additionally, the impact of missing modality on SER system performance will be explored, considering that real-world conditions might lead to signal loss. Noisy label learning, such as facial expression recognition [116], is worth investigating to check whether it can be useful in improving SER systems. Multilingual SER systems will be developed in at least ten languages due to slight variations in emotion perception across languages, which might

offer complementary information to enhance overall system performance. Finally, personalized SER systems will be constructed based on speaker or user profiles to improve user experience in real-world applications, recognizing that each individual could have unique emotional responses.

Bibliography

[1] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.

[2] A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, Y. Ijima, and Y. 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)*, 2018, pp. 4964–4968.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[4] S. Communication, L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim, J. Hoffman, M.-J. Hwang, H. Inaguma, C. Klaiber, I. Kulikov, P. Li, D. Licht, J. Maillard, R. Mavlyutov, A. Rakotoarison, K. R. Sadagopan, A. Ramakrishnan, T. Tran, G. Wenzek, Y. Yang, E. Ye, I. Evtimov, P. Fernandez, C. Gao, P. Hansanti,

E. Kalbassi, A. Kallet, A. Kozhevnikov, G. M. Gonzalez, R. S. Roman, C. Touret, C. Wong, C. Wood, B. Yu, P. Andrews, C. Balioglu, P.-J. Chen, M. R. Costa-jussà, M. Elbayad, H. Gong, F. Guzmán, K. Heffernan, S. Jain, J. Kao, A. Lee, X. Ma, A. Mourachko, B. Peloquin, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, A. Sun, P. Tomasello, C. Wang, J. Wang, S. Wang, and M. Williamson, "Seamless: Multilingual Expressive and Streaming Speech Translation," 2023. [Online]. Available: https://arxiv.org/abs/2312.05187

[5] L. F. Parra-Gallego and J. R. Orozco-Arroyave, "Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments," *Digital Signal Processing*, vol. 120, p. 103286, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200421003250

[6] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1702247114

[7] ——, "Semantic Space Theory: A Computational Approach to Emotion," *Trends in Cognitive Sciences*, vol. 25, no. 2, pp. 124–136, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136466132030276X

[8] J. A. Hall and D. Matsumoto, "Gender differences in judgments of multiple emotions from facial expressions," *Emotion (Washington, D.C.)*, vol. 4, no. 2, p. 201—206, June 2004. [Online]. Available: https://doi.org/10.1037/1528-3542.4.2.201

[9] D. Matsumoto, "American-Japanese Cultural Differences in the Recognition of Universal Facial Expressions," *Journal of Cross-Cultural Psychology*,

vol. 23, no. 1, pp. 72–84, 1992. [Online]. Available: https://doi.org/10.1177/0022022192231005

[10] A. Suzuki, T. Hoshino, K. Shigemasu, and M. Kawamura, "Decline or improvement?: Age-related differences in facial expression recognition," *Biological Psychology*, vol. 74, no. 1, pp. 75–84, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301051106001669

[11] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychological review*, vol. 110, no. 1, p. 145—172, January 2003. [Online]. Available: https://doi.org/10.1037/0033-295x.110.1.145

[12] L. F. Barrett, "Valence is a basic building block of emotional life," *Journal of Research in Personality*, vol. 40, no. 1, pp. 35–55, 2006, proceedings of the 2005 Meeting of the Association of Research in Personality. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092656605000590

[13] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner, "The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures," *Nature human behaviour*, vol. 3, no. 4, pp. 369–382, 2019.

[14] L. Goncalves and C. Busso, "AuxFormer: Robust Approach to Audiovisual Emotion Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7357–7361.

[15] J. Almeida, L. Vilaça, I. N. Teixeira, and P. Viana, "Emotion Identification in Movies through Facial Expression Recognition," *Applied Sciences*, vol. 11, no. 15, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/15/6827

[16] J. S. Gómez-Cañón, E. Cano, Y.-H. Yang, P. Herrera, and E. Gomez, "Let's agree to disagree: Consensus Entropy Active Learning for Personalized Music

Emotion Recognition," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. ISMIR, Oct. 2021, pp. 237–245. [Online]. Available: https://doi.org/10.5281/zenodo.5624399

[17] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music Emotion Recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021.

[18] H.-C. Chou and C.-C. 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)*, 2019, pp. 5886–5890.

[19] H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. 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)*, 2022, pp. 7717–7721.

[20] S. Parthasarathy and C. Busso, "Semi-Supervised Speech Emotion Recognition With Ladder Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, 2020.

[21] K. Vansteelandt, I. Van Mechelen, and J. B. Nezlek, "The co-occurrence of emotions in daily life: A multilevel approach," *Journal of Research in Personality*, vol. 39, no. 3, pp. 325–335, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092656604000431

[22] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, ""Of all things the measure is man" automatic classification of emotions and inter-labeler consistency [speech-based emotion recognition]," in *Proceedings. (ICASSP '05). IEEE*

International Conference on Acoustics, Speech, and Signal Processing, 2005., vol. 1, 2005, pp. I/317–I/320 Vol. 1.

[23] D. Zhang, X. Ju, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal Multi-label Emotion Detection with Modality and Label Dependence," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 3584–3593. [Online]. Available: https://aclanthology.org/2020.emnlp-main.291

[24] X. Kang, X. Shi, Y. Wu, and F. Ren, "Active Learning With Complementary Sampling for Instructing Class-Biased Multi-Label Text Emotion Classification," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 523–536, 2023.

[25] M. Wang, Y. Zhao, Y. Wang, T. Xu, and Y. Sun, "Image emotion multi-label classification based on multi-graph learning," *Expert Systems with Applications*, vol. 231, p. 120641, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423011430

[26] X. Ju, D. Zhang, J. Li, and G. Zhou, "Transformer-based Label Set Generation for Multi-modal Multi-label Emotion Detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 512–520. [Online]. Available: https://doi.org/10.1145/3394171.3413577

[27] D. Zhang, X. Ju, W. Zhang, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal Multi-label Emotion Recognition with Heterogeneous Hierarchical Message Passing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp. 14 338–14 346, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17686

[28] S. Li and W. Deng, "Blended Emotion in-the-Wild: Multi-label Facial Expression Recognition Using Crowdsourced Annotations and Deep Locality Feature Learning," *Int. J. Comput. Vision*, vol. 127, no. 6–7, p. 884–906, jun 2019. [Online]. Available: https://doi.org/10.1007/s11263-018-1131-1

[29] X. Geng, "Label Distribution Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[30] Y. Zhou, H. Xue, and X. Geng, "Emotion Distribution Recognition from Facial Expressions," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1247–1250. [Online]. Available: https://doi.org/10.1145/2733373.2806328

[31] D. Zhou, X. Zhang, Y. Zhou, Q. Zhao, and X. Geng, "Emotion Distribution Learning from Texts," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 638–647. [Online]. Available: https://aclanthology.org/D16-1061

[32] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 248–255.

[33] ——, "The Ordinal Nature of Emotions: An Emerging Approach," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 16–35, 2021.

[34] R. A. Martin, G. E. Berry, T. Dobranski, M. Horne, and P. G. Dodgson, "Emotion Perception Threshold: Individual Differences in Emotional Sensitivity," *Journal of Research in Personality*, vol. 30, no. 2, pp. 290–305, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0092656696900197

[35] H. Alhuzali and S. Ananiadou, "SpanEmo: Casting Multi-label Emotion Classification as Span-prediction," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Online: Association for Computational Linguistics, Apr. 2021, pp. 1573–1584. [Online]. Available: https://aclanthology.org/2021.eacl-main.135

[36] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning Deep Latent Space for Multi-Label Classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10769

[37] J. Deng and F. Ren, "Multi-Label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 475–486, 2023.

[38] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 859–866. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf

[39] R. Lotfian and C. 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, 2019.

[40] Z. Waseem, "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter," in *Proceedings of the First Workshop on NLP and Computational Social Science*, D. Bamman, A. S. Doğruöz, J. Eisenstein, D. Hovy, D. Jurgens, B. O'Connor, A. Oh, O. Tsur, and S. Volkova, Eds. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 138–142. [Online]. Available: https://aclanthology.org/W16-5618

[41] A. M. Davani, M. Díaz, and V. Prabhakaran, "Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 92–110, 01 2022. [Online]. Available: https://doi.org/10.1162/tacl_a_00449

[42] V. Prabhakaran, A. Mostafazadeh Davani, and M. Diaz, "On Releasing Annotator-Level Labels and Information in Datasets," in *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, C. Bonial and N. Xue, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 133–138. [Online]. Available: https://aclanthology.org/2021.law-1.14

[43] M. Sandri, E. Leonardelli, S. Tonelli, and E. Jezek, "Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2428–2441. [Online]. Available: https://aclanthology.org/2023.eacl-main.178

[44] S. Oluyemi, B. Neuendorf, J. Plepi, L. Flek, J. Schlötterer, and C. Welch, "Corpus Considerations for Annotator Modeling and Scaling," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long*

Papers), K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 1029–1040. [Online]. Available: https://aclanthology.org/2024.naacl-long.59

[45] L. Martinez-Lucas, A. Salman, S.-G. Leem, S. G. Upadhyay, C.-C. Lee, and C. Busso, "Analyzing the Effect of Affective Priming on Emotional Annotations," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2023, pp. 1–8.

[46] C. Hube, B. Fetahu, and U. Gadiraju, "Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–12. [Online]. Available: https://doi.org/10.1145/3290605.3300637

[47] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 25, 2008.

[48] N. Antoniou, A. Katsamanis, T. Giannakopoulos, and S. Narayanan, "Designing and Evaluating Speech Emotion Recognition Systems: A Reality Check Case Study with IEMOCAP," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[49] 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 Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[50] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, 2016.

[51] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[52] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–8.

[53] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The Ambiguous World of Emotion Representation," 2019. [Online]. Available: https://arxiv.org/abs/1909.00360

[54] Y. Kim and E. M. Provost, "Leveraging inter-rater agreement for audio-visual emotion recognition," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 553–559.

[55] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 890–897. [Online]. Available: https://doi.org/10.1145/3123266.3123383

[56] Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, and J. Dy, "Modeling annotator expertise: Learning when everybody knows a bit of something," in *Proceedings of the Thirteenth International Conference*

on Artificial Intelligence and Statistics, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 932–939. [Online]. Available: https://proceedings.mlr.press/v9/yan10a.html

[57] H. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 566–570.

[58] B. Zhang, Y. Kong, G. Essl, and E. M. Provost, "f-similarity preservation loss for soft labels: A demonstration on cross-corpus speech emotion recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 5725–5732, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/4518

[59] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[60] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[61] R. Lotfian and C. 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)*, 2017, pp. 415–420.

[62] Y. Kim and J. 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)*, 2018, pp. 5104–5108.

[63] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Speech Emotion Recognition Based on Multi-Label Emotion Existence Model," in *Proc. Interspeech 2019*, 2019, pp. 2818–2822.

[64] K. Sridhar, W.-C. Lin, and C. Busso, "Generative Approach Using Soft-Labels to Learn Uncertainty in Predicting Emotional Attributes," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2021, pp. 1–8.

[65] J. Li, Y. Chen, X. Zhang, J. Nie, Z. Li, Y. Yu, Y. Zhang, R. Hong, and M. Wang, "Multimodal Feature Extraction and Fusion for Emotional Reaction Intensity Estimation and Expression Classification in Videos With Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023, pp. 5838–5844.

[66] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005, emotion and Brain. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608005000407

[67] E. Mower, M. J. Mataric, and S. Narayanan, "Human Perception of Audio-Visual Synthetic Character Emotion Expression in the Presence of Ambiguous and Conflicting Information," *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 843–855, 2009.

[68] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *ArXiv e-prints (arXiv:1909.00360)*, pp. 1–19, May 2019.

[69] C. Busso and S. S. Narayanan, "Scripted dialogs versus improvisation: lessons learned about emotional elicitation techniques from the IEMOCAP database," in *Proc. Interspeech 2008*, 2008, pp. 1670–1673.

[70] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017, advances in Cognitive Engineering Using Neural Networks. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089360801730059X

[71] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, and B. W. Schuller, "Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition," *Neural Networks*, vol. 141, pp. 52–60, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608021000939

[72] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, "Negative Emotion Recognition using Deep Learning for Thai Language," in *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, 2020, pp. 71–74.

[73] B. Mocanu, R. Tapu, and T. Zaharia, "Utterance Level Feature Aggregation with Deep Metric Learning for Speech Emotion Recognition," *Sensors*, vol. 21, no. 12, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/12/4233

[74] M. Neumann and N. T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," in *ICASSP 2019 -*

2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 7390–7394.

[75] L. Goncalves and C. Busso, "Improving Speech Emotion Recognition Using Self-Supervised Learning with Domain-Specific Audiovisual Tasks," in *Proc. Interspeech 2022*, 2022, pp. 1168–1172.

[76] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copy-paste: An augmentation method for speech emotion recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6324–6328.

[77] X. Ju, D. Zhang, J. Li, and G. Zhou, "Transformer-based Label Set Generation for Multi-modal Multi-label Emotion Detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 512–520. [Online]. Available: https://doi.org/10.1145/3394171.3413577

[78] P. Riera, L. Ferrer, A. Gravano, and L. Gauder, "No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems," in *Workshop on Speech, Music and Mind (SMM 2019)*, Graz, Austria, September 2019, pp. 11–15.

[79] R. Lotfian and C. Busso, "Curriculum Learning for Speech Emotion Recognition From Crowdsourced Labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 815–826, 2019.

[80] L. Yang, Y. Shen, Y. Mao, and L. Cai, "Hybrid Curriculum Learning for Emotion Recognition in Conversation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11 595–11 603, Jun. 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/21413

[81] J. Li, X. Wang, Y. Liu, and Z. Zeng, "ERNetCL: A novel emotion recognition network in textual conversation based on curriculum learning strategy," *Knowledge-Based Systems*, vol. 286, p. 111434, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705124000698

[82] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[83] X. Wang, H. Liu, C. Shi, and C. Yang, "Be Confident! Towards Trustworthy Graph Neural Networks via Confidence Calibration," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 23 768–23 779. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/c7a9f13a6c0940277d46706c7ca32601-Paper.pdf

[84] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[85] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Proc. Interspeech 2021*, 2021, pp. 721–725.

[86] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, and Q. L. amd A.M. Rush, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.

[87] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015, pp. 1–13.

[88] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, BC, Canada, December 2019, pp. 1–12.

[89] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[90] H.-C. Chou, C.-C. Lee, and C. Busso, "Exploiting Co-occurrence Frequency of Emotions in Perceptual Evaluations To Train A Speech Emotion Classifier," in *Proc. Interspeech 2022*, 2022, pp. 161–165.

[91] Y. Li, T. Zhao, and T. Kawahara, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning," in *Proc. Interspeech 2019*, 2019, pp. 2803–2807.

[92] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[93] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

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

[95] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[96] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech Synthesis With Mixed Emotions," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3120–3134, 2023.

[97] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, Y. Adi, J. Pino, J. Gu, and W.-N. Hsu, "Textless Speech-to-Speech Translation on Real Data," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 860–872. [Online]. Available: https://aclanthology.org/2022.naacl-main.63

[98] X. Li, Y. Jia, and C.-C. Chiu, "Textless Direct Speech-to-Speech Translation with Discrete Speech Representation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[99] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4749–4753.

[100] S. Parthasarathy and C. Busso, "Ladder Networks for Emotion Recognition: Using Unsupervised Auxiliary Tasks to Improve Predictions of Emotional Attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[101] Z. Aldeneh and E. Mower Provost, "Using regional saliency for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2741–2745.

[102] M. Abdelwahab and C. Busso, "Study Of Dense Network Approaches For Speech Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*. Calgary, AB, Canada: IEEE, April 2018, pp. 5084–5088.

[103] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech Emotion Recognition Using Multi-hop Attention Mechanism," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2822–2826.

[104] X. Kang, X. Shi, Y. Wu, and F. Ren, "Active Learning With Complementary Sampling for Instructing Class-Biased Multi-Label Text Emotion Classification," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 523–536, 2023.

[105] P. Xu, Z. Liu, G. I. Winata, Z. Lin, and P. Fung, "EmoGraph: Capturing Emotion Correlations using Graph Networks," *CoRR*, vol. abs/2008.09378, 2020.

[106] X. Wang, S. Zhao, and Y. Qin, "Supervised Contrastive Learning with Nearest Neighbor Search for Speech Emotion Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 1913–1917.

[107] Y. Zhao, J. Wang, C. Lu, S. Li, B. W. Schuller, Y. Zong, and W. Zheng, "Emotion-Aware Contrastive Adaptation Network for Source-Free Cross-Corpus Speech Emotion Recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 846–11 850.

[108] D. Zhang, X. Ju, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal multi-label emotion detection with modality and label dependence," in *Empirical Methods in Natural Language Processing (EMNLP 2020)*, Virtual Conference, November 2020, pp. 3584–3593.

[109] W. Wu, C. Zhang, X. Wu, and P. C. Woodland, "Estimating the Uncertainty in Emotion Class Labels With Utterance-Specific Dirichlet Priors," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2810–2822, 2023.

[110] A. Keesing, Y. S. Koh, and M. Witbrock, "Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech," in *Proc. Interspeech 2021*, 2021, pp. 3415–3419.

[111] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[112] W.-C. Lin and C. Busso, "Chunk-Level Speech Emotion Recognition: A General Framework of Sequence-to-One Dynamic Temporal Modeling," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1215–1227, 2023.

[113] H. Fei, Y. Zhang, Y. Ren, and D. Ji, "Latent Emotion Memory for Multi-Label Emotion Classification," *Proceedings of the AAAI Conference on Artificial*

Intelligence, vol. 34, no. 05, pp. 7692–7699, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6271

[114] H. Chou, L. Goncalves, S. Leem, A. N. Salman, C. Lee, and C. Busso, "Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule," *IEEE Transactions on Affective Computing*, no. 01, pp. 1–15, jun 2024.

[115] E. Zhang, R. Trujillo, and C. Poellabauer, "The MERSA Dataset and a Transformer-Based Approach for Speech Emotion Recognition," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 13 960–13 970. [Online]. Available: https://aclanthology.org/2024.acl-long.752

[116] S. Zhang, Z. Huang, D. P. Paudel, and L. Van Gool, "Facial emotion recognition with noisy multi-task annotations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 21–31.