EMBRACING AMBIGUITY AND SUBJECTIVITY USING THE ALL-INCLUSIVE AGGREGATION RULE FOR EVALUATING MULTI-LABEL SPEECH EMOTION RECOGNITION SYSTEMS

Huang-Cheng Chou[1], *Haibin Wu*[2], *Lucas Goncalves*[3], *Seong-Gyun Leem*[3], *Ali Salman*[3], Carlos Busso[3], Hung-yi Lee[2], and Chi-Chun Lee[1]

[1]National Tsing Hua University, Taiwan
[2]National Taiwan University, Taiwan
[3]The University of Texas at Dallas, USA

ABSTRACT

Speech Emotion Recognition (SER) faces a distinct challenge compared to other speech-related tasks because the annotations will show the subjective emotional perceptions of different annotators. Previous SER studies often view the subjectivity of emotion perception as noise by using the majority rule or plurality rule to obtain the consensus labels. However, these standard approaches overlook the valuable information of labels that do not agree with the consensus and make it easier for the test set. Emotion perception can have co-occurring emotions in realistic conditions, and it is unnecessary to regard the disagreement between raters as noise. To bridge the SER into a multi-label task, we introduced an "all-inclusive rule," which considers all available data, ratings, and distributional labels as multi-label targets and a complete test set. We demonstrated that models trained with multi-label targets generated by the proposed AR outperform conventional single-label methods across incomplete and complete test sets.

Index Terms— speech emotion recognition, label aggregation method, multi-label learning, the subjectivity of emotion perception, the ambiguity of emotions

1. INTRODUCTION

SER is an essential technology in human-computer interaction (HCI) systems [1], and the SER systems training relies on data and collected emotional ratings. Each data point is typically annotated by multiple annotators, with most emotion databases requiring at least three annotators per data point. However, researchers frequently encounter disagreements among annotators in most public emotion databases [2, 3, 4]. For example, it is not uncommon for three different annotators to select three distinct emotional options after listening to the same speech recordings. This variability highlights a fundamental challenge in SER: **the subjectivity of emotion perception**. Such disagreements underscore the complexity of accurately identifying emotions from speech, as individuals' interpretations of emotional content can vary widely based on their experiences, biases, and cultural backgrounds [5, 3, 6]. However, the standard approaches regard the disagreement as noise and use the majority rule (MR) or plurality rule (PR) to find the consensus labels. If the data has no consensus labels, those data are removed from the test set. The process cannot reveal the actual performance of SER systems since the test cannot reflect a realistic scenario. Fig. 1 illustrates the different amounts of data used based on each aggregation method. The MR has the lowest amount of data, making the test easier.

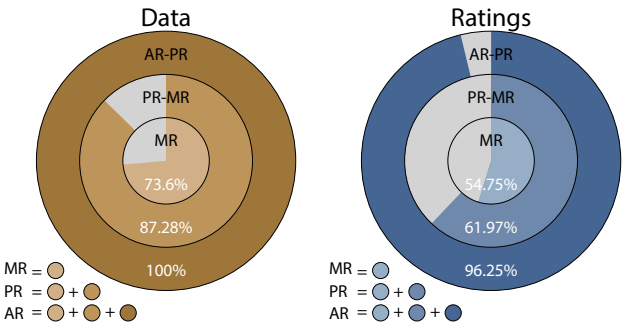

Fig. 1. Overview of the averaged usage ratio of the data and ratings generated by three rules, *majority rule* (MR), *plurality rule* (PR), and *all-inclusive rule* (AR). A diagram that illustrates how much data and ratings are used in the final test set according to each aggregation method. MR contains the lowest amount of data, and AR always includes the entire test set available in the dataset.

Previous studies have noticed the disagreement between annotators and revealed the benefits of utilizing all existing annotations for training SER models. For instance, studies have investigated the benefits and usage of a soft-label learning strategy to include all the samples during training SER models [7, 8, 9, 10, 11, 12, 13, 14, 15]. Chou and Lee [16, 17] also show the effectiveness of modeling individual annotators' SER systems for modeling the subjectivity of emotion perception. However, those studies still regard the SER task as a single-label task, so they only allow each data to have only one emotion. Also, the test set is still *simplified* by only considering sentences with consensus labels obtained by the MR or PR label aggregation rules, removing a minority of emotional ratings, and discarding complex and ambiguous data samples that could have more than one emotion. However, it is important to acknowledge that perceptual differences are not necessarily noisy. The detailed definitions of the MR and PR are in section 4.3.

Additionally, most prior studies often select a limited number of emotions as their focus. For example, the well-known IEMOCAP corpus [18], encompasses ten distinct emotions; however, studies

frequently narrow its focus to just four emotions: neutral, anger, sadness, and happiness [19, 20], despite frustration being one of the most frequently annotated emotions within the dataset. Those standard processes leave inevitable critical issues as below:

(1) More than 12% of data and 38% of emotional annotations are discarded by the common label aggregation methods.

(2) Most prior studies never reveal the actual performances of SER systems since many data and ratings in the test set are removed.

(3) Mixed emotions (co-occurrence of emotions) have not been considered in the evaluation of the conventional SER systems.

To address the three issues above, we propose an all-inclusive label aggregation rule (AR) to maximize the usage of all emotional ratings, utilize all data samples in the datasets for training and evaluating the SER systems, and employ distributional labels as multi-label targets. With the proposed aggregation rule, we want to address the open question: **Should the SER task be approached as a multi-label recognition task?** To answer the question, we train and evaluate the existing state-of-the-art SER framework [21] with the data sets defined by the conventional and the proposed aggregation rules across four public emotion datasets, the IEMOCAP [18], MSP-IMPROV [22], MSP-PODCAST [23], and BIIC-PODCAST [24]. We found that the SER systems using conventional rules (e.g., MR and PR) perform worse on the complete test set than on the incomplete one. Our results also show that training with the proposed AR leads to overall better performances than using the MR or PR when testing with complete and incomplete test data.

2. BACKGROUND AND RELATED WORK

2.1. Label Representations and Learning Strategies

Consensus is required to generate labels for emotion recognition tasks. The three most common approaches are outlined below, with Table 1 summarizing the labels according to these three definitions.

- **Single-label**: each sample can only be mapped to a single emotion. Most SER researchers use MR [19, 25] or PR [9, 26] to generate one consensus emotional label as the learning target. This method drops the data without consensus emotions. While some studies [7, 8, 9, 10, 11, 16, 12, 13, 14, 15] utilized the soft label as labels of training data during training SER models, the ground truth of data is still single-label. However, different from the soft label, we allow the test data to have more than one emotion and use an ad-hoc threshold to convert distributional predictions into multiple emotions.
- **Multi-hot label**: each sample can be associated with multiple perceived emotions. The multiple-hot label is mainly used in text emotion recognition [27]. They take all emotional classes as target emotions even if only one annotator chooses that emotion on the given sample.
- **Distribution-label**: each sample can carry more than one emotion. The distribution label is mainly used in facial expression emotion recognition [28, 29]. They calculate the distribution based on the number of votes for each emotion.

In this work, we use an alternative way to define the labels for each sample in the two stages. In the training phase, we adopt a distribution label as our learning objective, mirroring the concept introduced in [30]. This approach is derived from the understanding that assessing the consistency between a model's predicted distribution and subjective annotations is an effective method for evaluating if an SER model aligns with human emotional perception. Furthermore, the work of psychologists [6, 31] supports the idea that emotion perception is not only high-dimensional but also blended in nature. For the evaluation phase, inspired by [32], we employ a threshold technique [33, 34, 35] to transform the distribution label into a binary vector, which serves as the basis for emotion decision-making for each sample—similar to a multi-hot encoding scheme. Detailed information on this method can be found in Section 4.3. This threshold approach aids in excluding infrequent emotions, thereby enhancing the robustness of SER systems as noted by [32].

Let us consider a hypothetical scenario within a four-class emotion recognition task involving neutral (N), anger (A), sadness (S), and happiness (H) classes. In this evaluation, we use the MR, PR, and AR to define the learning targets for the train sets. We define them as MR_{Train}, PR_{Train}, and AR_{Train}, respectively. For testing, We denote them as MR_{Test}, PR_{Test}, and AR_{Test}. Table 2 highlights the operational distinctions among the aggregation rules across three cases. Within Case (1), AR_{Train} is capable of incorporating the minority perspectives, such as sadness, which MR_{Train} and PR_{Train} overlook, opting instead to dismiss these minority annotations. Moving to Case (2), which presents a scenario with equally predominant emotions, anger and sadness, both MR and PR rules exclude this data from the training and test sets due to their inability to resolve the "tie" case.

Table 1. Overview of the label vectors for the various definitions of emotion recognition with three cases. Each example has five annotations. We illustrate the tree definitions with a four-class emotion classification task. The four emotions include neutral (N), anger (A), sadness (S), and happiness (H). A label vector is created as follows: (N,A,S,H). For instance, N,A,A,S,S indicates that the five emotional annotations for Case 2 selected two for neutral, two for anger, and two for sadness.

Case	(1) A,A,A,S,S	(2) N,A,A,S,S	(3) N,N,H,A,S
Single-label	(0,1,0,0)	Ignored	PR: (1,0,0,0) (MR: Ignored)
Multi-hot label	(0,1,1,0)	(1,1,1,0)	(1,1,1,1)
Distribution-label	(0.0,0.6,0.4,0.0)	(0.2,0.4,0.4,0.0)	(0.4,0.2,0.2,0.2)

Table 2. Overview of the label vectors for the various aggregation rules across three distinct cases, each with five annotations. We illustrate the rules with a four-class emotion classification task, including neutral (N), anger (A), sadness (S), and happiness (H). We employ label vectors to represent the distribution of annotations (N, A, S, H). For example, consider the label vector for Case 2 indicated as N,A,A,S,S; this signifies that out of the five emotional annotations, one was neutral, two were anger, and two were sadness.

Case	(1) A,A,A,S,S	(2) N,A,A,S,S	(3) N,N,H,A,S
MR_{Train}	(0,1,0,0)	Removed	Removed
PR_{Train}	(0,1,0,0)	Removed	(1,0,0,0)
AR_{Train}	(0.0,0.6,**0.4**,0.0)	**(0.2,0.4,0.4,0.0)**	(0.4,0.2,0.2,0.2)
MR_{Test}	(0,1,0,0)	Removed	Removed
PR_{Test}	(0,1,0,0)	Removed	(1,0,0,0)
AR_{Test}	(0,1,**1**,0)	**(0,1,1,0)**	(1,0,0,0)

Table 3. The table provides a summary of the data and ratings loss ratios associated with Majority Rule (MR), Plurality Rule (PR), and All-inclusive Rule (AR).

Aggregation Rule	MR		PR		AR	
Database	**Data**	**Rating**	**Data**	**Rating**	**Data**	**Rating**
IEMOCAP	31.37%	49.44%	25.32%	45.70%	**0.00%**	**3.10%**
IMPROV (P)	9.18%	28.52%	5.63%	26.41%	**0.00%**	**5.12%**
PODCAST (P)	44.81%	59.87%	19.85%	49.24%	**0.00%**	**6.15%**
B-PODCAST (P)	20.25%	43.18%	0.06%	30.76%	**0.00%**	**0.61%**
Average	26.40%	45.25%	12.72%	38.03%	**0.00%**	**3.75%**

2.2. Selection of Test Set for Evaluating SER Systems

Using the complete test set to evaluate the performances of SER systems is very essential. However, the common practice to handling the data without consensus is to remove them from the test set. For example, the IEMOCAP corpora use MR for constructing the ground-truth labels [18, 36], discarding approximately 31.37% of the data and 49.44% of the ratings shown in Table 3. However, real-life emotional states can co-occur in many situations (e.g., sad and angry) [31, 6, 37]. Previous studies discarded many data points in the test set since they assumed each sample had only one emotional category. Then, the ground-truth category does not reflect secondary emotions that are also conveyed in the utterance. Therefore, aggregating multiple annotations into a single class and discarding non-consensus data points of the test set is not appropriate to accurately evaluate whether the predictions of SER systems can represent emotional behaviors observed in daily interactions. In this work, we first show the better performances of the SER systems trained with labels using the proposed label aggregation method across the conventional incomplete test sets and the proposed complete test sets. To our best knowledge, this is the first study that evaluates SER models with data samples with non-consensus annotations.

3. METHODOLOGY

To bridge to the multi-label SER, we introduce an alternative aggregation rule named the ***All-inclusive rule*** **(AR)**, maximizing the usage of annotated ratings within a corpus, ensuring that no data points are discarded. Initially, the AR compiles all classes attributed to each data point to establish the ground truth and utilizes the distributional ground truth to train SER systems to form the SER task's training set. For instance, Table 2 shows different ground truth formats according to aggregation rules. The $Train$ means the training phase; The $Test$ is the testing phase. The AR uses the count for each emotion to calculate the distributional label. Also, the AR ensures that every annotated data point and all usable emotional annotations are included in the test set. For instance, in Table 3, compared to the MR method, the multi-label approach retains an additional 26.40% of the data and 41.45% more emotional ratings. Furthermore, it preserves 12.72% more data and 34.28% more emotional ratings than the PR method.

An interesting aspect of AR in the testing phase (AR_{Test} in Table 2) involves adopting a threshold method. This method converts distributional labels into binary vectors based on a defined threshold. Emotions are considered positive if the proportion of a class surpasses the threshold of $1/C$, where C is the total number of emotion classes involved. Therefore, with four emotion categories, the threshold is set at 0.25. This approach allows Case (2) in Table 1 to acknowledge anger and sadness as part of the ground truth, demonstrating AR's nuanced capacity to capture a spectrum of emotional states in contrast to the binary resolutions offered by MR and PR.

Notice that the AR allows for including sentences with ambiguous emotions in the test set, ensuring a comprehensive and naturalistically accurate approach to SER. This AR aims to capture the high-dimensional nature of emotion perception during training and facilitates a more accurate and comprehensive representation of emotional states, aligning the SER system's capabilities with the complex nature of human emotions across various test sets.

4. EXPERIMENTAL SETTINGS

We introduce the resources, preprocessing procedures, objective function, evaluation process, and the SER framework as follows. For the databases without predefined training, development, and testing splits, we describe the details about the split sets in the supplementary material[1] (section A) to ensure a thorough evaluation process. The details about the splits can resolve the reproducibility issue mentioned in [38] that 80.77% results using the IEMOCAP dataset that cannot be reproduced.

4.1. Resources

In this study, we utilize four public emotion databases to demonstrate the performance of our SER framework. Contrary to the common practice that discards emotional ratings and data samples as illustrated in Table 3, our approach is focused on leveraging the entirety of available emotional annotations and data to precisely aim at pre-defined emotion categories in order to refine our recognition capabilities.

The SAIL-IEMOCAP [18], referred here to as IEMOCAP, encompasses a collection of recordings derived from five dyadic conversations, performed by ten professional actors in English. These recorded sessions have been meticulously segmented into 10,039 utterances. The corpus is annotated with ten distinct emotions: neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, and "other." In this study, we omit the "other"" category. Annotators were permitted to assign multiple emotions to each utterance, reflecting the complex nature of human emotion. Our experiments adopt a 5-fold cross-validation defined in the supplementary material, section A.1.

The MSP-IMPROV [22], referred here to as IMPROV, comprises high-quality audio-video recordings, featuring performances by 12 actors in English, encapsulated within five primary emotions: anger, happiness, sadness, a neutral state, and "other." The six dyadic sessions are diligently segmented into 8,438 clips. We exclude the "other" category to focus on four-class emotions. In the experiments, we adopt a 6-fold cross-validation defined in section A.2.

The MSP-PODCAST [23], referred here to as PODCAST, offers a collection of naturalistic emotional speech segments, selected from a broad spectrum of real-world podcast recordings. The annotation includes both primary and secondary emotional scenarios. We only employ the primary emotions, which encompass nine categories:

[1] Supplementary Material

anger, sadness, happiness, surprise, fear, disgust, contempt, neutral, and "other." We opt to exclude the "other" category in the study. We utilize version 1.11 of the PODCAST dataset. This version incorporates a comprehensive collection of 84,030 utterances for training, 19,815 for development, and a combined total of 45,462 utterances for testing, achieved by merging 30,647 from test set 1 and 14,815 from test set 2.

The BIIC-PODCAST [24], referred here to as B-PODCAST, serves as a Mandarin Chinese variant of the original PODCAST collection. We utilized release version 1.01 of the corpus, and our study concentrated on primary emotions. The dataset composition includes 48,815 utterances designated for training, 10,845 for development, and 10,340 for testing.

4.2. Selection Emotion Classes

We use the pre-defined emotions while excluding the "other" category in some datasets to maximize the usage of data samples and reflect the original behaviors of raters. The count of target emotions analyzed in our study stands at 4 for IMPROV, 8 for PODCAST and B-PODCAST, and 9 for IEMOCAP.

4.3. Aggregation Rules Comparison

We implement conventional and novel aggregation methods for label generation to evaluate and discern the performance disparities between single-label and multi-label SER systems. These methods are pivotal in preparing and assessing the data within the SER systems.

- **Majority Rule (MR)** selects the emotion class designated by over half of the annotations as the definitive target emotion. This approach disregards less frequently annotated emotions, excluding and wasting instances where annotations do not converge on a majority consensus.
- **Plurality Rule (PR)** focuses on the emotion with the most annotations. It highlights the emotion that appears more often in the annotations. This strategy defines an emotional label even if it does not make up more than half of all annotations.
- **All-inclusive Rule (AR)** incorporates every annotation to formulate a distribution-like representation based on the number of annotations for each emotional class. This method ensures that no samples are excluded, maintaining an inclusive dataset that reflects the diversity of emotional expressions indicated by the annotations.

4.4. Class-balanced Objective Function

We follow [35] to employ the Class-Balanced Cross-Entropy Loss (CBCE), as introduced by [39], to address the challenge of imbalance in the annotation distributions observed across all utilized databases, which is a concern also highlighted in [35]. The foundational concept of CBCE revolves around introducing a weighting factor designed to adjust the loss function values. The inverse frequencies of the classes in the training dataset directly influence this adjustment. The factor is $\frac{1-\beta}{1-\beta^{n_j}}$, where n_j is the number of positive samples in the j^{th} emotion class in the train set, and $\beta \in (0, 1]$ is a hyperparameter. The number of factors to weigh the loss values equals the number of target emotions. The CBCE value can be calculated using Eq. 1:

$$\mathcal{L}_{CBCE} = \sum_{j=1}^{K} \left(\frac{1-\beta}{1-\beta^{n_j}} \cdot {\mathcal{L}_{CE}}^{(j)}\right), \tag{1}$$

where ${\mathcal{L}_{CE}}^{(j)}$ is the value of cross-entropy loss [40] for the j^{th} emotion. We set the beta value to 0.99.

4.5. Evaluation Metrics and Confidence Interval

We deploy the macro-F1 score [41], a comprehensive metric that considers recall and precision rates concurrently. This evaluation approach is implemented using Scikit-learn [42]. For multi-label classifications, the evaluation process involves the application of thresholds to the ground truth data to delineate the target classes. Specifically, a prediction for a given class is considered accurate if its proportional representation in the predictions exceeds $(1/C)$, where C denotes the total count of emotional classes involved. This threshold-based approach, resonating with the methodologies outlined in existing literature [34, 33], underpins our calculation of macro-F1 scores.

Inspired by the methodology of Steidl et al. [30], where results are assessed with an entropy-based metric, we utilize the *Kullback-Leibler divergence* (KLD) and *Jensen-Shannon Divergence* (JSD) to quantify the similarity between the model's prediction distribution and subjective annotations. This approach helps determine whether an SER model aligns with human emotion perception.

Notice that we collect the predictions from each partition defined in section A, and then measure the performances in macro-F1 score with the average and the lower and upper bound of the confidence interval (CI) between 2.75% and 97.5% using the toolkit [43]. All results are single-run with the fixed random seed number.

4.6. SER Framewrok

We employed the SER framework initially introduced by [21], which builds upon the foundational "wav2vec2-large-robust" model as proposed by [44]. Following their pioneering methodology, we tailored the architecture to enhance efficiency and maintain high recognition performance by removing the top 12 transformer layers from the original 24-layer structure. Following [21][2], our model configuration includes adding two hidden layers, each containing 1,024 nodes, atop the modified "wav2vec2-large-robust" backbone. These layers are activated using the rectified linear unit (ReLU) activation function. A softmax output layer follows these hidden layers, providing a probabilistic distribution over the target emotion classes. Furthermore, we applied average pooling to the resulting representations, feeding it into the classification layers. We applied a dropout function with a probability of 0.5 to the first and second layers of the classification architecture to regularize the model, following the work [21]. The number of model parameters is around 317 million.

4.7. Models Training and Choice

We use the AdamW optimizer [45] with a 0.0001 learning rate. The batch size and epoch are set as 32 and 50, respectively. We choose

[2]https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

Table 4. The table summarizes the overall performance across the four public emotion datasets in macro-F1 scores. We also indicate the lower and upper bound of the confidence interval (CI) between 2.75% and 97.5% for each result (lower bound, upper bound).

Database	Train\Test	Average	$\mathbf{MR}_{Test}$	$\mathbf{PR}_{Test}$	$\mathbf{AR}_{Test}$
IEMOCAP	$\mathbf{MR}_{Train}$	0.391 (0.384,0.398)	0.366 (0.358,0.373)	0.365 (0.358,0.373)	0.442 (0.435,0.448)
	$\mathbf{PR}_{Train}$	0.377 (0.371,0.384)	0.348 (0.341,0.354)	0.347 (0.340,0.354)	0.437 (0.431,0.443)
	$\mathbf{AR}_{Train}$	**0.421 (0.414,0.429)**	**0.378 (0.371,0.385)**	**0.376 (0.369,0.384)**	**0.510 (0.503,0.518)**
IMPROV	$\mathbf{MR}_{Train}$	0.607 (0.597,0.616)	0.603 (0.593,0.613)	0.597 (0.588,0.607)	0.620 (0.612,0.629)
	$\mathbf{PR}_{Train}$	0.609 (0.600,0.619)	0.607 (0.596,0.617)	0.599 (0.589,0.609)	0.622 (0.614,0.631)
	$\mathbf{AR}_{Train}$	**0.638 (0.629,0.647)**	**0.634 (0.624,0.643)**	**0.627 (0.618,0.635)**	**0.654 (0.646,0.662)**
PODCAST	$\mathbf{MR}_{Train}$	0.344 (0.341,0.347)	0.300 (0.297,0.304)	0.297 (0.294,0.301)	0.434 (0.431,0.436)
	$\mathbf{PR}_{Train}$	0.349 (0.347,0.352)	0.293 (0.290,0.296)	0.292 (0.290,0.295)	0.463 (0.460,0.466)
	$\mathbf{AR}_{Train}$	**0.359 (0.356,0.362)**	**0.306 (0.302,0.310)**	**0.306 (0.303,0.310)**	**0.464 (0.461,0.467)**
B-PODCAST	$\mathbf{MR}_{Train}$	0.266 (0.256,**0.280**)	0.234 (0.222,**0.252**)	0.235 (0.225,**0.250**)	0.329 (0.323,0.337)
	$\mathbf{PR}_{Train}$	0.264 (0.259,0.271)	0.231 (0.225,0.237)	0.233 (0.228,0.238)	0.329 (0.323,0.337)
	$\mathbf{AR}_{Train}$	**0.271** (**0.264**,0.278)	**0.239** (**0.232**,0.247)	**0.241** (**0.234**,0.247)	**0.333 (0.327,0.340)**

the best models according to the lowest value of the class-balanced cross-entropy loss (Equation 1) on the development set. We use two Nvidia Tesla v100 GPUs with 32 GB memory. The total number of GPU hours is around 50.

5. RESULTS AND ANALYSIS

5.1. Impact of Aggregation Rules on Data and Ratings

The MR and PR methods demonstrate significant data loss, an aspect outlined in Table 3. Across the four emotion databases, MR and PR contribute to an average data loss of 26.40% and 12.72%, respectively, illustrating a large reduction in data usage for SER system training and evaluation. In contrast, the AR retains all data points, with the minimal exclusion of 3.75% representing the "other" annotations, while the MR and PR lose 45.25% and 38.03% of ratings, respectively. This difference in data retention between the methods underscores a critical limitation of traditional SER systems, which may not leverage the full spectrum of available emotional information due to the inherent data loss associated with MR and PR methods. Consequently, these traditional systems risk being evaluated on a diminished dataset, potentially impairing the robustness and generalizability of SER applications.

5.2. Impact of Aggregation Rules on SER Performances

Table 4 summarizes the performance of the SER models trained with datasets selected by different aggregation rules in macro-F1 scores. We also indicate the lower and upper bound macro-f1 scores in the confidence interval (CI) between 2.75% and 97.5% for each result (lower bound, upper bound) using the toolkit [43]. In the column, **Average**, the models trained with the data set selected with AR, AR_{Train}, perform the best among the three rules on average and across 4 public emotion databases. The models trained by AR_{Train} led to 5.13% and 5.58% relative improvement than the models trained by MR_{Train} and PR_{Train}, respectively, when all models are trained with the mentioned training information in section 4.7.

5.3. Evaluation with Complete and Incomplete Test Sets

We evaluate the performances of SER systems in three different testing conditions. The $\mathbf{MR}_{Test}$ and $\mathbf{PR}_{Test}$ are the test data sets selected by MR and PR, respectively, and they are incomplete test sets. Instead, the $\mathbf{AR}_{Test}$ contains all data samples and allows data points to have more than one emotion. We showed the examples in section 2.1. Additionally, the amount of test sets differ, as shown in Fig. 1. The AR_{Test} contains more data samples that have mixed emotions than the PR_{Test} and the MR_{Test}.

By employing the AR rule, SER systems can be trained with more data samples and various emotional ratings and better recognize ambiguous samples containing mixed emotions. Also, the performances of the SER systems trained by AR_{Train} significantly outperform the models trained with the conventional methods, MR or PR. Therefore, we suggest that training SER systems with the AR_{Train} set is more effective for real-life deployments where the test set includes a mix of ambiguous and unambiguous data, reflecting the true complexity of real-world scenarios.

More specifically, using the proposed *all-inclusive* rule has one significant advantage the "complete" test set can serve as a benchmark for evaluating SER systems. By using the complete test, the community is provided with a uniform standard for assessing and comparing SER systems since all data are used directly without the need for selection or filtration, unlike conventional methods, such as EMO-SUPERB [46]. For example, some studies exclude data without a majority consensus label, or they only focus on specific emotion classes, such as four or six emotions out of nine emotions in the IEMOCAP corpus [18]. The *all-inclusive* rule also takes into account samples with co-existing emotions, which is more reflective of real-world scenarios and a crucial step toward practical applications of SER systems. The rationale behind this method is that it effectively handles the inherent subjectivity and variability in emotional labeling. Unlike plurality and majority rules, which often treat these variations as noise and consequently exclude samples and disregard emotional ratings, the all-inclusive rule incorporates the entire set of samples and a broader range of emotional ratings. This thorough approach is essential for developing more effective SER systems compared to traditional aggregation techniques.

5.4. Differences in Distribution Between Predictions and Human Perception

Unlike converting the model's output into binary labels for single-label or multi-label tasks during macro-F1 evaluations, we employ the model's probability distributions directly for all test sets using

Table 5. The results are presented using Kullback–Leibler Divergence (KLD), following the same format as shown in Table 4.

Database	Test\Train	Average	$\mathbf{MR}_{Test}$	$\mathbf{PR}_{Test}$	$\mathbf{AR}_{Test}$
IEMOCAP	$\mathbf{MR}_{Train}$	1.183 (1.157,1.209)	1.226 (1.195, 1.257)	1.257 (1.229, 1.285)	1.066 (1.048, 1.084)
	$\mathbf{PR}_{Train}$	1.178 (1.156,1.202)	1.242 (1.216, 1.269)	1.274 (1.25, 1.301)	1.019 (1.002, 1.036)
	$\mathbf{AR}_{Train}$	**0.984 (0.965,1.005)**	**1.063 (1.041, 1.086)**	**1.096 (1.075, 1.119)**	**0.794 (0.778, 0.809)**
IMPROV	$\mathbf{MR}_{Train}$	0.781 (0.758,0.802)	0.816 (0.792, 0.839)	0.835 (0.810, 0.860)	0.691 (0.673, 0.707)
	$\mathbf{PR}_{Train}$	0.785 (0.762,0.806)	0.819 (0.793, 0.843)	0.839 (0.815, 0.863)	0.696 (0.679, 0.713)
	$\mathbf{AR}_{Train}$	**0.589 (0.573,0.604)**	**0.640 (0.622, 0.658)**	**0.654 (0.637, 0.67)**	**0.472 (0.459, 0.483)**
PODCAST	$\mathbf{MR}_{Train}$	1.074 (1.064,1.085)	1.083 (1.071, 1.096)	1.240 (1.229, 1.251)	0.900 (0.893, 0.907)
	$\mathbf{PR}_{Train}$	1.107 (1.099,1.114)	1.18 (1.172, 1.19)	1.295 (1.287, 1.303)	0.845 (0.839, 0.85)
	$\mathbf{AR}_{Train}$	**0.931 (0.924,0.938)**	**0.988 (0.98, 0.997)**	**1.111 (1.104, 1.118)**	**0.694 (0.689, 0.698)**
B-PODCAST	$\mathbf{MR}_{Train}$	1.181 (1.165,1.197)	1.252 (1.233, 1.269)	1.310 (1.293, 1.328)	0.982 (0.970, 0.994)
	$\mathbf{PR}_{Train}$	1.069 (1.052,1.085)	1.122 (1.102, 1.14)	1.196 (1.178, 1.214)	0.889 (0.876, 0.902)
	$\mathbf{AR}_{Train}$	**0.988 (0.974,1.002)**	**1.048 (1.032, 1.064)**	**1.117 (1.101, 1.132)**	**0.799 (0.788, 0.809)**

Table 6. Results are displayed using Jensen-Shannon Divergence (JSD), adhering to the same format as Table 4.

Database	Test\Train	Average	$\mathbf{MR}_{Test}$	$\mathbf{PR}_{Test}$	$\mathbf{AR}_{Test}$
IEMOCAP	$\mathbf{MR}_{Train}$	0.252 (0.248,0.256)	0.261 (0.257, 0.266)	0.281 (0.277, 0.285)	0.213 (0.210, 0.216)
	$\mathbf{PR}_{Train}$	0.253 (0.249,0.257)	0.276 (0.272, 0.280)	0.266 (0.262, 0.271)	0.217 (0.214, 0.220)
	$\mathbf{AR}_{Train}$	**0.227 (0.223,0.230)**	**0.250 (0.246, 0.254)**	**0.255 (0.251, 0.258)**	**0.175 (0.172, 0.178)**
IMPROV	$\mathbf{MR}_{Train}$	0.165 (0.162,0.169)	0.177 (0.173, 0.181)	0.180 (0.176, 0.185)	0.139 (0.136, 0.142)
	$\mathbf{PR}_{Train}$	0.164 (0.159,0.168)	0.175 (0.170, 0.179)	0.179 (0.174, 0.183)	0.138 (0.134, 0.141)
	$\mathbf{AR}_{Train}$	**0.142 (0.139,0.145)**	**0.157 (0.154, 0.161)**	**0.160 (0.157, 0.163)**	**0.109 (0.107, 0.111)**
PODCAST	$\mathbf{MR}_{Train}$	0.239 (0.237,0.241)	0.248 (0.246, 0.250)	0.274 (0.272, 0.276)	0.195 (0.194, 0.197)
	$\mathbf{PR}_{Train}$	0.259 (0.258,0.261)	0.279 (0.278, 0.281)	0.299 (0.298, 0.301)	0.200 (0.199, 0.201)
	$\mathbf{AR}_{Train}$	**0.228 (0.227,0.229)**	**0.245 (0.244, 0.247)**	**0.268 (0.267, 0.269)**	**0.171 (0.170, 0.172)**
B-PODCAST	$\mathbf{MR}_{Train}$	0.275 (0.272,0.278)	0.292 (0.288, 0.295)	0.301 (0.298, 0.304)	0.232 (0.229, 0.234)
	$\mathbf{PR}_{Train}$	0.248 (0.245,0.250)	0.262 (0.259, 0.265)	0.275 (0.272, 0.278)	0.206 (0.204, 0.208)
	$\mathbf{AR}_{Train}$	**0.239 (0.236,0.242)**	**0.254 (0.252, 0.257)**	**0.267 (0.264, 0.270)**	**0.196 (0.193, 0.198)**

the *Kullback-Leibler divergence* (KLD) and *Jensen-Shannon Divergence* (JSD). The results, presented in Table 5 and 6, respectively, indicate that a lower KLD and JSD value corresponds to better performance. The observed patterns in these results are consistent with those found in the macro-F1 score, as shown in Table 4.

6. DISCUSSION AND LIMITATIONS

With the compressive and empirical experiments in Table 4, the AR rule represents an alternative and potentially practical aggregation approach for the SER task. Our findings align with the recent emerging emotion theory, the **semantics space theory** [6]. The core of the theory is that emotions are blended and high-dimensional. Also, the semantics space theory emphasizes that boundaries of categorical emotions are not discrete but blended, and it explains that emotion perception could have more than one emotion (mixed emotion). With the proposed all-inclusive rule, we can make the SER systems have the ability to recognize actual human emotions from speech, reflecting the findings of the emerging semantics space theory [6]. However, we only employ one model and evaluate the proposed method on the four public emotion datasets in English and Chinese.

7. CONCLUSION AND FUTURE WORK

We introduce the novel all-inclusive rule (AR), designed to deploy all data, maximize the usage of the emotional ratings available in these datasets to avoid data wasting, and generate multi-target labels throughout both the training and evaluation phases of SER systems. To demonstrate our approach's effectiveness, we employed one of the latest SER models [21], evaluating it under three different data aggregation rules: two conventional methods (majority (MR) and plurality rules (PR)) and our proposed AR rule. Models trained with the AR rule outperformed those trained with the conventional rules, showing 5.13% and 5.58% relative improvements in performance over the MR and PR, respectively. The improvement indicates that the AR rule can significantly improve the performance of SER systems. We suggest using the AR approach to select the training data for SER tasks. Such an improvement aligns with the emerging semantics space emotion theory, suggesting that speech emotion recognition should inherently be approached as a multi-label task. Besides, given that the current loss functions are not optimized for the SER task, we will develop a novel loss function explicitly tailored for SER in our future work. Furthermore, the models will be evaluated using the proposed comprehensive AR on bias and fairness to thoroughly understand potential issues, as highlighted in the recent study by [47].

8. ACKNOWLEDGMENTS

This research was supported by NSTC under Grants 112-2634-F-002-005 and the NSF under Grant CNS-2016719. We also thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

9. REFERENCES

[1] S Ramakrishnan and Ibrahiem MM El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2011.

[2] Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Interpreting ambiguous emotional expressions," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–8.

[3] Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth Narayanan, "The Ambiguous World of Emotion Representation," 2019.

[4] Huang-Cheng Chou, Lucas Goncalves, Seong-Gyun Leem, Ali N. Salman, Chi-Chun Lee, and Carlos Busso, "Minority Views Matter: Evaluating Speech Emotion Classifiers with Human Subjective Annotations by an All-Inclusive Aggregation Rule," *IEEE Transactions on Affective Computing*, pp. 1–15, 2024.

[5] David 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.

[6] Alan S. Cowen and Dacher Keltner, "Semantic Space Theory: A Computational Approach to Emotion," *Trends in Cognitive Sciences*, vol. 25, no. 2, pp. 124–136, 2021.

[7] H.M. 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.

[8] 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 *ACM international conference on Multimedia (MM 2017)*, Mountain View, CA, USA, October 2017, pp. 890–897.

[9] Reza Lotfian and Carlos Busso, "Formulating Emotion Perception as a Probabilistic Model with Application to Categorical Emotion Classification," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 415–420.

[10] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 4964–4968.

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

[12] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Speech emotion recognition based on multi-label emotion existence model," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2818–2822.

[13] K. Sridhar, W.-C. Lin, and C. Busso, "Generative approach using soft-labels to learn uncertainty in predicting emotional attributes," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2021)*, Nara, Japan, September-October 2021, pp. 1–8.

[14] X. Li, Z. Zhang, C. Gan, and Y. Xiang, "Multi-Label Speech Emotion Recognition via Inter-Class Difference Loss Under Response Residual Network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3230–3244, 2023.

[15] X. Li, Z. Zhang, C. Gan, and Y. Xiang, "Multi-Label Speech Emotion Recognition via Inter-Class Difference Loss Under Response Residual Network," *IEEE Transactions on Multimedia*, vol. 25, pp. 3230–3244, 2023.

[16] H.-C. Chou and C.-C. Lee, "Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 5886–5890.

[17] Huang-Cheng Chou and Chi-Chun Lee, "Learning to Recognize Per-Rater's Emotion Perception Using Co-Rater Training Strategy with Soft and Hard Labels," in *Interspeech 2020*, 2020, pp. 4108–4112.

[18] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

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

[20] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha 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.

[21] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn 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. 10745–10759, 2023.

[22] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower 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.

[23] Reza Lotfian and Carlos Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[24] Shreya G. Upadhyay, Woan-Shiuan Chien, Bo-Hao Su, Lucas Goncalves, Ya-Tse Wu, Ali N. Salman, Carlos Busso, and Chi-Chun Lee, "An Intelligent Infrastructure Toward Large Scale Naturalistic Affective Speech Corpora Collection," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2023, pp. 1–8.

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

[26] Wenbo Li, Yaodong Cui, Yintao Ma, Xingxin Chen, Guofa Li, Guanzhong Zeng, Gang Guo, and Dongpu Cao, "A Spontaneous Driver Emotion Facial Expression (DEFE) Dataset for Intelligent Vehicles: Emotions Triggered by Video-Audio Clips in Driving Scenarios," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 747–760, 2023.

[27] Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji, "Latent Emotion Memory for Multi-Label Emotion Classification," in *AAAI Conference on Artificial Intelligence (AAAI 2020)*, New York, NY, USA, February 2020, vol. 34, pp. 7692–7699.

[28] S. Li and W. Deng, "Blended Emotion in-the-Wild: Multi-label Facial Expression Recognition Using Crowdsourced Annotations and Deep Locality Feature Learning," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 884–906, June 2019.

[29] Ning Xu, Yun-Peng Liu, and Xin Geng, "Partial Multi-Label Learning with Label Distribution," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 6510–6517, Apr. 2020.

[30] 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.*, 2005, vol. 1, pp. I/317–I/320 Vol. 1.

[31] Alan S. Cowen and Dacher 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.

[32] Kusha Sridhar and Carlos Busso, "Speech Emotion Recognition with a Reject Option," in *Proc. Interspeech 2019*, 2019, pp. 3272–3276.

[33] Pablo Riera, Luciana Ferrer, Agustín Gravano, and Lara Gauder, "No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems," in *Proc. SMM19, Workshop on Speech, Music and Mind 2019*, Graz, Austria, September 2019, pp. 11–15.

[34] Huang-Cheng Chou, Wei-Cheng Lin, Chi-Chun Lee, and Carlos Busso, "Exploiting Annotators' Typed Description of Emotion Perception to Maximize Utilization of Ratings for Speech Emotion Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7717–7721.

[35] Huang-Cheng Chou, Lucas Goncalves, Seong-Gyun Leem, Chi-Chun Lee, and Carlos Busso, "The Importance of Calibration: Rethinking Confidence and Performance of Speech Multi-label Emotion Classifiers," in *INTERSPEECH 2023*, 2023, pp. 641–645.

[36] 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.

[37] Huang-Cheng Chou, Chi-Chun Lee, and Carlos Busso, "Exploiting Co-occurrence Frequency of Emotions in Perceptual Evaluations To Train A Speech Emotion Classifier," in *Interspeech 2022*, 2022, pp. 161–165.

[38] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[39] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[40] I. J. Good, "Rational Decisions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 14, no. 1, pp. 107–114, 1952.

[41] Juri Opitz and Sebastian Burst, "Macro f1 and macro f1," *arXiv preprint arXiv:1911.03347*, 2019.

[42] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[43] L. Ferrer and P. Riera, "Confidence Intervals for evaluation in machine learning," Computer software, 2024.

[44] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[45] Ilya Loshchilov and Frank Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[46] Haibin Wu, Huang-Cheng Chou, Kai-Wei Chang, Lucas Goncalves, Jiawei Du, Jyh-Shing Roger Jang, Chi-Chun Lee, and Hung-yi Lee, "Open-Emotion: A Reproducible EMO-SUPERB for Speech Emotion Recognition Systems," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024.

[47] Yi-Cheng Lin, Haibin Wu, Huang-Cheng Chou, Chi-Chun Lee, and Hung yi Lee, "Emo-bias: A Large Scale Evaluation of Social Bias on Speech Emotion Recognition," in *Interspeech 2024*, 2024, pp. 4633–4637.

Supplementary Material for Embracing Ambiguity And Subjectivity Using The All-inclusive Aggregation Rule For Evaluating Multi-label Speech Emotion Recognition Systems

A. SPLIT SETS

We meticulously delineate the specifics of the speaker-independent training, development, and test splits for the IMPROV [22] and IEMOCAP [18] datasets, acknowledging that these datasets are not publicly accessible. Our goal in presenting detailed descriptions of these split sets is to enhance the reproducibility of our research. By doing so, we aim to provide a comprehensive blueprint that allows replicating and validating our findings within the research community, thus fostering a deeper understanding and further exploration of the datasets in question. Notice that we collect the predictions from each partition, and then measure the performances in macro-F1 score using the toolkit [43]. All results are single-run with the fixed random seed number.

A.1. The IEMOCAP

Table 7 encapsulates the division of the IEMOCAP corpus for our study. We have split five independent splits, labeled as Dyad 1 through Dyad 5, corresponding to each session within the corpus. Notably, each session is characterized by a dyadic interaction between two speakers. To rigorously evaluate our model's performance across these interactions, we have employed a 5-fold cross-validation strategy. This method is graphically represented in Table 7, highlighting the unique configurations of training, development, and test sets for each fold. This approach ensures a detailed and unbiased assessment of model efficacy across different duos within the IEMOCAP dataset.

A.2. The IMPROV

In the speaker-independent scenario, the MSP-IMPROV corpus is partitioned into six folds for cross-validation. Each fold combines training, development, and test sets, as illustrated in Table 8. This partitioning strategy ensures that the model is trained on interactions involving different sets of speakers and evaluated on unseen speaker combinations, facilitating robust evaluation of model generalization across various dyadic conversations within the MSP-IMPROV corpus.

Table 7. IEMOCAP corpus partitions.

Partition	Training Set	Development Set	Test Set
1	Dyad 1,2,3	Dyad 4	Dyad 5
2	Dyad 2,3,4	Dyad 5	Dyad 1
3	Dyad 3,4,5	Dyad 1	Dyad 2
4	Dyad 1,4,5	Dyad 2	Dyad 3
5	Dyad 1,2,4	Dyad 3	Dyad 4

Table 8. MSP-IMPROV corpus partitions.

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