Minority Views Matter: Evaluating Speech Emotion Classifiers with Human Subjective Annotations by an All-Inclusive Aggregation Rule

Huang-Cheng Chou, Lucas Goncalves, Seong-Gyun Leem, Ali N. Salman, *Student Member, IEEE,* Chi-Chun Lee, *Senior Member, IEEE*, and Carlos Busso, *Fellow, IEEE*

Abstract—When selecting test data for subjective tasks, most studies define ground truth labels using aggregation methods such as the majority or plurality rules. These methods discard data points without consensus, making the test set easier than practical tasks where a prediction is needed for each sample. However, the discarded data points often express ambiguous cues that elicit coexisting traits perceived by annotators. This paper addresses the importance of considering all the annotations and samples in the data, highlighting that only showing the model's performance on an incomplete test set selected by using the majority or plurality rules can lead to bias in the models' performances. We focus on *speech-emotion recognition* (SER) tasks. We observe that traditional aggregation rules have a data loss ratio ranging from 5.63% to 89.17%. From this observation, we propose a flexible method named the all-inclusive aggregation rule to evaluate SER systems on the complete test data. We contrast traditional single-label formulations with a multi-label formulation to consider the coexistence of emotions. We show that training an SER model with the data selected by the all-inclusive aggregation rule shows consistently higher macro-F1 scores when tested in the entire test set, including ambiguous samples without agreement.

Index Terms—Speech Emotion Recognition, Learning from Disagreement, Subjective Perception, Multi-label Emotion Classification

1 Introduction

Recent developments in *affective computing* have attracted increasing attention to human perception systems. Current systems utilize outputs from ubiquitous multi-modality sensors (e.g., cameras and audio recording devices) to recognize states or traits describing human behaviors. These efforts have resulted in human-centered solutions for problems such as depression detection [1], pain recognition [2], deception detection [3], and emotion recognition [4]. Given the subjective nature of these tasks, these models are commonly tested with labels derived from human perceptual evaluations, where each data point is annotated by multiple raters. The standard practice to process these annotations and generate the training and testing sets for these models is to use the majority or plurality aggregation methods, as illustrated in Figure 1a. These methods discard annotations that disagree with the consensus label. However, it is common to have co-existing emotions in daily interactions [5], so a single label does not properly describe the emotional perception of a sample. Also, these methods discard data points without agreement. The *majority rule* (MR) discards the data if a class does not achieve more than 50% of the votes. The *plurality rule* (PR) discards the data if one class does not have more votes than the other classes. The issue with discarding samples without consensus is that it reduces the validity of systems intended for practical applications, as these ambiguous samples are not considered in the test set. Previous studies have only focused on how to utilize all existing annotations during training. For instance, studies have investigated the use of a soft-label learning strategy to include all the samples during training [6], [7], [8], [9], [10], [11], [12]. However, the test set is still *simplified* by only considering sentences with MR or PR agreement, discarding complex and ambiguous samples.

We explore an effective formulation that combines annotations, using all the annotations provided by subjective evaluations for both the train and test sets, making systems more suitable for practical applications. While the formulation is suitable for any problem using labels derived from perceptual evaluations, we focus on *speech emotion recognition* (SER), where emotions generally co-occur in daily interactions. Instead of discarding non-consensus labels, we include all the data points in the train and test sets, allowing SER models to use valuable information during training while being tested on all samples in the test set, even the data points without consensus agreement. We refer to our approach as the *all-inclusive rule* (AR) method. The flexibility to have co-occurred emotions is critical in our formulation. Figure 1b compares our all-inclusive aggregation with the majority and plurality rules generally used in SER tasks for deciding whether or not to include a data point in the test set. By using the AR method, we consider all the data in the test set, which allows us to show the complete performance of SER systems. The driving questions for our study are as follows:

- How is the performance of SER systems affected by

- *H.-C. Chou and C.-C. Lee are with the Department of Electrical Engineering, National Tsing Hua University, Taiwan. E-mail: hc.chou@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw*
- *L. Goncalves, S.-G. Leem, A. N. Salman, and C. Busso are with the Erik Jonsson School of Engineering & Commputer Science, The University of Texas at Dallas, Richardson TX 75080. E-mail: {goncalves, seonggyun.leem, ali.salman, busso}@utdallas.edu*

Manuscript received July 14, 2023; revised xxx xx, 2023.

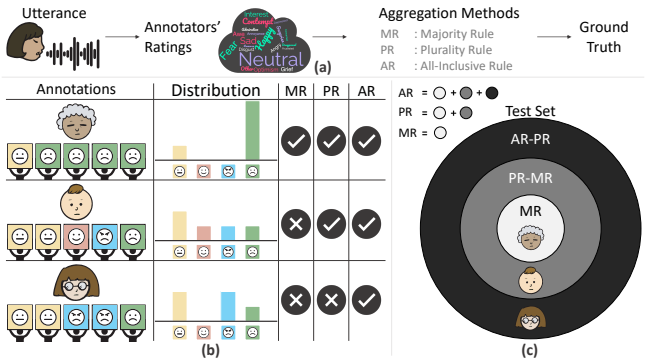


Fig. 1: Overview of the difference between the ground-truth generated by three rules, *majority rule* (MR), *plurality rule* (PR), and *all-inclusive rule* (AR). (a) Perceptual evaluation to obtain ground-truth labels. Each utterance is rated by several annotators, obtaining the consensus label using aggregation methods. (b) An illustration of how aggregation methods discard the data depending on the annotation distribution. The check mark means the data point is kept and the symbol "x" means the data point is discarded. (c) A diagram that illustrates how much data is included 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.

using different aggregation methods for the training set?

- Does training an SER system with data derived from the all-inclusive rule improve the performance on ambiguous emotions compared to data using the majority or plurality rules?
- What is the best label learning strategy for training SER systems when evaluated on the complete test set?

To investigate the aforementioned questions, we use common emotional speech databases generally used in SER tasks: the MSP-Podcast, USC-IEMOCAP, CREMA-D, and MSP-IMPROV corpora. Based on our experiments, we find that SER systems using conventional training strategies (i.e., majority or plurality rules) perform worse on the complete test set than on the incomplete test set. This result shows that these aggregation methods cannot properly handle ambiguous samples, which are much more difficult to predict. Our results also show that training with the all-inclusive rule leads to overall better performances than using the majority or plurality rule when testing with complete and incomplete test data. Additionally, we find that soft-label learning is the best training strategy for building SER systems, even if the system is evaluated on the complete test set.

2 Background and Related Work

This paper focuses on building categorical *speech emotion recognition* (SER) systems. We aim at developing formulations to evaluate the models on the complete test set of datasets, including the ambiguous samples. This section describes disagreement between raters and existing aggregation methods used to select the data samples.

2.1 Disagreement among Raters

In contrast to common classification tasks that have well-established 'gold standard,' subjective tasks such as emotion recognition do not have clear labels, which are often obtained with perceptual evaluations. Researchers frequently turn to crowd-sourcing platforms like Amazon Mechanical Turk for rapid and extensive label collection [13]. While this approach is cost-effective, it invariably leads to a compromise in label quality. This trade-off is particularly pronounced in subjective tasks, where the ambiguity inherent in the task itself compounds the variability in annotations.

Subjective tasks, such as emotion perception [14] or hate speech tagging [15], present unique challenges due to their inherently subjective nature. Labels in these tasks are not straightforward to obtain, as they depend heavily on individual interpreters. Disagreements among annotators can arise from various factors [16]: diverse backgrounds leading to different interpretations, lack of interest in providing correct labels, emotional priming, and contextual differences [17]. These variances introduce a significant amount of noise into the labeling process, which is especially problematic in crowd-sourced evaluations [18].

Noise in annotations is a critical concern, and various strategies have been devised to mitigate its impact. In the context of speech emotion classification tasks, it is important to acknowledge that while noise contributes to label discrepancies, it is not the sole reason for disagreement. Drawing from the methodologies employed in the MSP-Podcast corpus [14], several approaches can be effective in reducing noise. These strategies include filtering out evaluators with consistently low agreement scores, stopping crowdsourcing efforts when a threshold of agreement is not met, and relying on in-house workers, who can receive targeted training to improve label agreement. However, it is important to acknowledge that perceptual differences are not necessarily noisy. They can provide information that a SER system should leverage.

Our paper aims to demonstrate that traditional methods of aggregating labels (e.g., majority or plurality voting), which often overlook the nuanced nature of subjective tasks, may not be suitable for speech emotion classification. We propose an alternative aggregation method for the speech emotion recognition task, moving towards a more comprehensive and inclusive approach to label aggregation.

2.2 Existing Aggregation Method Definition

The two most common aggregation rules are:

- **Majority Rule (MR)**: it selects a class only if more than half of the votes select that class. If the frequency of any category does not reach half or more of the votes, the data point is discarded (see illustration in Figure 1b)
- **Plurality Rule (PR)**: it selects a class if one emotional class obtains more votes than other classes (i.e., receives a plurality). This rule does not require that the selected class has more than half of the votes. Data points with ties are discarded.

MR and PR assume that the ground truth is a single class. In reality, emotional classes often co-occur (e.g., excited+happy, surprised+happy, angry+sad). Therefore, they

are inadequate to accurately represent real-world emotional states. We present the *all-inclusive rule* that is more suitable for training and evaluating the performance of SER systems in practical applications where ambiguous expressions are expected.

2.3 Selection of Test Set for Evaluating SER Systems

Evaluating SER systems on the complete test set is very important. However, the standard approach to dealing with samples without consensus is to discard them from the test set. When collecting the emotional annotations from multiple workers, there is often a high disagreement between the annotators [19], [20], [21]. Therefore, many previous studies discarded many data points in the test set. For example, the IEMOCAP and CREMA-D corpora use MR for constructing the ground-truth labels [22], [23], [24], discarding approximately 31.37% and 35.8% of the data, respectively. Most studies using these corpora have followed the same rule to construct the ground-truth label for testing their models [25], [26], [27], [28]. The MSP-IMPROV and MSP-Podcast corpora use PR to annotate primary and secondary emotional labels for each speaking turn [14], [29], and studies using these corpora have kept this default aggregation rule to evaluate their models [30], [31], [32].

The studies mentioned in the previous paragraph have assumed that each speaking turn has only one emotional category, so the ground-truth category does not reflect secondary emotions also conveyed in the recordings. However, real-life emotional states can co-exist in many situations (e.g., a person can be simultaneously sad and angry) [5]. 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. Although some studies have investigated the use of the "multiple-hot" vector to define SER as a multi-label problem [33], [34], [35], this approach still cannot determine if some emotions are more dominant than others. It considers all the annotations provided to the speaking turn as ground truth, even if a class was only selected by a single annotator. To the best of our knowledge, Riera et al. [36] is the only study suggesting that we should use all test samples to evaluate SER systems, instead of discarding non-consensus data. However, they did not explore SER systems trained with various label learning methods. Additionally, they relabeled some of the emotions (e.g., excited as happy; surprised as "other"), which is an important limitation. Unlike this study, we use the original emotional classes in all the datasets, providing empirical experiments using different label learning methods and various test sets created by considering different aggregation methods.

2.4 Label Learning Methods for SER

Most SER studies have trained their models by using consensus labels obtained by aggregating individual perceptual emotional annotations with MR or PR. A common approach for training an SER model with a single consensus label is to minimize the *cross-entropy* (CE) between the prediction and the one-hot encoding extracted from an aggregated ground-truth. We refer to this approach as **hard-label learning**. Due to its simple formulation, many studies have used hard-label learning to train their SER model [25], [26], [27], [28], [30], [37], [38].

Although the hard-label learning strategy simplifies the representation of the ground-truth label, it does not take into account the subjectivity of perceptual evaluations. For example, the hard-label learning strategy cannot differentiate cases when the consensus label is barely reached (i.e., ambiguous cases) from unanimous cases (i.e., clear cases). To address this problem, some studies have used a soft-encoding to improve the one-hot vector [6], [7], [8], [9], [10], [11], [12]. Those studies have used the CE between the prediction and the ground-truth vector formed by estimating the proportions of annotations assigned to each class. We refer to this approach as **soft-label learning**, and we use CE as the cost function.

Instead of using hard-label and soft-label learning methods, few studies on SER have regarded the soft-label vector as a distribution, using the *Kullback-Leibler divergence* (KLD) as the objective function to train SER models. We refer to this approach as **distribution-label learning** [39], and we use KLD as the cost function. Those studies represented their ground-truth emotion with a multi-class soft-label encoding, training the model to minimize the KLD between the soft-encoded ground truth and the predicted distributions [40], [41], [42].

3 Methodology

We introduce an alternative aggregation rule, named the *all-inclusive rule* (AR), to train and evaluate the performance of SER systems on a complete test set, including data points without MR or PR consensus. We define this rule and explain its importance along with directions on how to use it.

3.1 Definition of All-inclusive Rule

The AR is an aggregation method that keeps all the annotated samples within a corpus regardless of the frequency of the votes. Data points are never discarded. In this rule, the first step is to gather all the classes given to each data point. At this point, AR generates the ground truth for the data. When creating the train set, the representation of the ground truth is defined as a one-hot encoding or as the distribution of the votes, depending on the desired label learning strategy. For the hard-label learning strategy, AR chooses the emotional class that has the highest votes as the ground truth, similar to the plurality rule. However, if no plurality is achieved within the votes, we randomly choose as the ground truth one of the classes that received the most votes. We illustrate this scenario in Table 1 Case (C1), where the hard-label can be either (1,0,0,0) or (0,0,1,0). For the soft-label or distribution-label learning strategy, AR generates the distributional ground truth based on the frequency of votes assigned to the emotional classes.

When generating the test set, AR always uses the distributional ground truth regardless of the label-learning strategy, shown in the rightmost column of Table 1. This method ensures that every annotated data point and all of its annotations are considered for the test set. The all-inclusive rule provides a label descriptor that better captures

TABLE 1: Overview of the label vectors for the *all-inclusive rule* (AR) with three examples. Each example has five annotations. We illustrate the rules with a four-class emotion classification task. The four emotions include neutral (N), happiness (H), anger (A), and sadness (S). A label vector is created as follows: (N,H,A,S). We list three examples. For instance, (C1) N,N,A,A,S indicates that the five emotional annotations for Case (C1) selected two votes for neutral, two votes for anger, and one vote for sadness.

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)

TABLE 2: Overview of the number of utterances, emotion classes, and data loss ratio in the datasets. P represents primary emotion, and S represents secondary emotions.

Database	Utterance	Emotion	Choice	MR	PR	AR
IMPROV (P)	8,438	4	Single	9.20%	5.63%	0%
CREMA-D	7,442	6	Single	35.80%	8.55%	0%
PODCAST (P)	90,978	8	Single	47.78%	18.56%	0%
IEMOCAP	10,039	9	Multiple	31.37%	25.32%	0%
IMPROV (S)	8,438	10	Multiple	54.17%	12.34%	0%
PODCAST (S)	90,978	16	Multiple	89.17%	29.08%	0%

the emotional content of the data points by incorporating sentences with ambiguous emotions in the test set.

3.2 Usage of All-inclusive Rule for Test Set Preparation

We utilize all data samples and consider every available emotion as a learning target to incorporate all opinions collected during the perceptual evaluation. Previous studies using the IEMOCAP corpus, for example, have aggregated all the annotated emotions into a 4-class emotion classification task (e.g., merging excitement and happiness, and discarding minor classes such as fearful, surprise, and disgusted). Unlike this approach, we neither ignore other emotional states present in a corpus nor limit the SER models to be trained or tested only on a few selected emotions.

Additionally, our all-inclusive rule enables SER models to be tested with the secondary emotions on the whole test set. The utilization of secondary emotional annotations has been ignored by previous studies due to the data loss caused by standard aggregation methods (up to 89.17% in Table 2). Since our AR utilizes the entire annotated test set, we can test our SER model with secondary emotions which have not been studied before. Table 2 illustrates the ratio of data loss imposed by the different aggregation rules on the four datasets used in this study for the primary and secondary emotions. The use of MR and PR discards up to 89.17% and 29.08% of the entire data, respectively. The worst case is classifying secondary emotions on the MSP-Podcast corpus.

4 Experimental Settings

4.1 Resources

We check if the all-inclusive aggregation method works well over four popular databases. We maximize the usage of the emotional annotations by using the all-inclusive aggregation rule. Some databases provide the annotators with the option to select the class "other", allowing the annotators to type their own emotional descriptions. In our experiments, we remove the annotations labeled as "other" except for cases when they provide descriptions that are equivalent to the pre-defined emotions (e.g., from "slightly happy" to "happiness"). For these cases, we aggregate the descriptions that are similar to the pre-defined emotions following one of the pre-processing steps presented in Chou et al. [41].

4.1.1 The MSP-Podcast Corpus

The MSP-Podcast corpus, referred here to as PODCAST, contains spontaneous and diverse emotional speech samples collected from various podcast recordings, which are split into speaking turns to form a speech repository. Several SER algorithms are used to retrieve speaking turns that are expected to be emotional by using the approach presented in Mariooryad et al. [43]. The annotation process uses a crowdsourcing protocol inspired by the work of Burmania et al. [44]. The perceptual evaluation includes the *primary emotions* (P) and *secondary emotions* (S). The annotators choose a single primary emotion, but they can select multiple secondary emotions for each sample. The primary emotions contain nine options: anger, sadness, happiness, surprise, fear, disgust, contempt, neutral, and "other". The secondary emotions consist of the primary emotions and eight more classes: amusement, frustration, depression, concern, disappointment, excitement, confusion, and annoyance (17 options in total). Each speaking turn is annotated by at least five different workers. This paper uses version 1.10 of the corpus, which consists of 104,267 annotated utterances. We exclude the "Test2" set in this paper, resulting in 90,978 utterances as listed in Table 2.

4.1.2 The USC-IEMOCAP Corpus

The USC-IEMOCAP corpus [23], referred here to as IEMOCAP, consists of motion capture, audio, and video recordings from five sessions of dyadic conversations collected from 10 professional actors. The corpus includes scripted and spontaneous spoken communication scenarios. Each dyad is provided with selected scripts to elicit emotional states (e.g., neutral, angry, sad, and happy emotions). All recorded conversations were manually segmented into 10,039 utterances, and at least three evaluators annotated the emotional categories of each utterance. The IEMOCAP has ten different emotion categories (neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, and "other"). Each annotator can provide more than one emotion and/or choose "other," if none of the classes are appropriate. If "other" is selected, they can type their own descriptions.

4.1.3 The CREMA-D Corpus

The CREMA-D corpus, referred here to as CREMA-D, is an audiovisual dataset containing high-quality recordings

collected from a racially and ethnically diverse group of 91 professional adult actors. The actors were given a set of sentences and asked to say every sentence targeting a specific emotional state. At least seven annotators rated the emotional labels under different conditions: audio-only, video-only, and audiovisual recordings. In total, 7,442 clips were collected and rated by 2,443 raters via a crowdsourcing platform. We only use the emotional annotations collected using the voice-only condition, because this paper focuses on building SER systems. The CREMA-D corpus was annotated with six different emotions (anger, disgust, fear, happy, neutral, and sad), where each annotator selected only one emotion. The option "other" was not allowed.

4.1.4 The MSP-IMPROV Corpus

The MSP-IMPROV corpus [29], referred here to as IMPROV, includes acted, elicited, and spontaneous emotional speech to explore the perception of emotion [45]. The MSP-IMPROV corpus consists of 8,438 utterances in total. Those recorded samples are annotated with a crowd-sourcing protocol that tracked the quality of the workers in real-time, stopping the evaluation when their quality dropped below an acceptable threshold [44]. Similar to the MSP-Podcast corpus, it allows annotators to choose a single *primary emotion* (P) and multiple *secondary emotions* (S). The primary emotions contain five emotion classes: anger, sadness, happiness, neutral, and "other." The secondary emotions included the primary emotions plus six more emotions: frustration, depression, disgust, excitement, fear, and surprise.

4.2 Speech Emotion Classifier

To assess the performance of various aggregation methods, we use the Wav2vec2.0 architecture [46] that has shown good performance for SER tasks in various studies [47], [48]. Among the variants of the Wav2vec2.0 model, we use the "wav2vec2-large-robust" architecture, proposed in Hsu et al. [49], that showed the best recognition performance in the study of Wagner et al. [48]. The model accepts raw signal as its input, instead of using spectrogram or *Mel-frequency cepstral coefficient* (MFCC). We use a 16kHz sampling rate for our dataset to match the sampling rate of its pre-trained data. For efficiency and reproducibility, we remove the 12 transformer layers from the top of the 24 transformer layers used in this architecture, which is shown to preserve the recognition performance with fewer parameters [48]. We attach two hidden layers and the softmax output layer on top of the Wav2vec2.0 model, where each hidden layer has 1,024 nodes. They are implemented with the *rectified linear unit* (ReLU) activation function. We aggregate the outputs of the Wav2vec2.0 model by using average pooling per utterance, then feed it to the classification layers. We apply the dropout function using $p = 0.5$ to the first and second layers of the classification layers to regularize the model.

When implementing the model, we utilize the pre-trained "wav2vec2-large-robust" model from the HuggingFace library [50], and attach the classification layers to the pruned Wav2vec2.0 model. Then, we fine-tune the model for each task. During the fine-tuning step, we freeze the convolutional layers, followed by the transformer layers of the Wav2vec2.0 model. This partial fine-tuning strategy has shown better performance than fine-tuning the entire parameters [47]. We use the Adam optimizer [51] with a 0.0001 learning rate. We group 32 utterances to construct each mini-batch and update the model for 100 epochs. After the 100 epochs, we select the model with the best recognition performance in the development set. We implement the code in Pytorch [52] and run the code on an NVIDIA Tesla V100 GPU.

When training our model on the MSP-Podcast corpus, we use the pre-defined train (63,076 samples), development (10,999 samples), and test (16,903 samples) sets of the corpus. The other corpora are smaller, so we use a cross-validation strategy. We use the speaker-independent sessions for the IEMOCAP (five sessions) and the MSP-IMPROV (six sessions). The CREMA-D corpus does not provide pre-defined sessions, so we manually split them into five speaker-independent sessions. For the IEMOCAP, MSP-IMPROV and CREMA-D corpora, we conduct a K-fold cross validation, where K denotes the number of sessions for each corpus. Each fold is organized as follows: one session for the test set, one session for the development set, and the remaining sessions for the train set.

Table 2 illustrates the data loss ratio introduced by each label aggregation method in the databases considered in this study. We evaluate the data loss ratio in each partition (i.e., train, development, or test set). We noticed that the data loss trends are similar across the four datasets, so we only present the distribution for the PODCAST database as an example. Table 3 shows the ratios. Overall, the data loss ratio of each partition is very similar to the data loss ratio for the entire database.

TABLE 3: The data loss ratio imposed by the label aggregation method on the training, development, and test sets of the PODCAST corpus. P represents primary emotion, 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%

4.3 Training/Test Set Selected by Aggregation Rules

In this evaluation, we train and test the models with match and mismatched aggregation rules. For the train and development sets, we use the MR, PR, and AR to define the ground truth. We denote them as MR_{train}, PR_{train}, and AR_{train}, respectively. For testing, we evaluate the models with the MR, PR, and AR sets. In addition, we define two extra test conditions, which are illustrated by the *donuts* in Figure 1c: PR-MR, the test set accepted by the PR, but discarded by the MR, and AR-PR, the test set accepted by the AR but discarded by the PR. The condition AR-PR repre-

sents the most ambiguous set with samples receiving non-consensus annotations. To our best knowledge, this is the first study that evaluates SER models with non-consensus annotations.

4.4 Label Learning for SER

In addition to the different aggregation methods, we also evaluate the performance using different label learning methods. In our experiments, we consider three different label learning strategies: hard-label learning, soft-label learning, and distribution-label learning. For the hard-label learning, we construct ground truth by using a one-hot encoding that has "1.0" for the class that received the maximum number of votes from the annotators. When we use the train set aggregated with the AR, we randomly choose one of the emotions that received the most votes as the ground-truth emotion if the sample does not have a consensus. We smooth the ground truth vector of this one-hot encoding by using the smoothing strategy proposed by Szegedy et al. [53] which utilizes a smoothing parameter set to 0.05. This smoothing strategy adds a small probability to emotional classes with zero value. We use the CE objective function to train the SER systems. For soft-label learning and distribution-label learning, we represent the ground-truth vector by using the distribution of the annotator's votes. We divide the number of votes for each class by the number of total votes for each data point. We also implement the label smoothing strategy used for the hard-label learning. The cost function for the soft-label learning is the CE loss, and for the distribution-label learning is the KLD.

4.5 Evaluation Metrics

This paper uses the macro-F1 scores to evaluate the SER performance, which requires estimating precision and recall rates. The MR, PR, and PR-MR test sets are formed by selecting a single class, so they are suitable for the macro-F1 score. The class that receives the maximum number of votes is selected as the target. We consider a prediction a success if the class with the maximum predicted probability agrees with the target class. In contrast, the test sets collected with the AR and AR-PR conditions contain samples with non-consensus labels. We allow co-existing emotions to estimate the macro-F1 score for these experiments. The target classes are selected by applying thresholds over the ground truth. We consider a prediction a success if the proportion for a class is above $1/C$, where C is the number of emotional classes, following the approach adopted by previous studies [36], [41]. For instance, consider a four-class emotion recognition task, and the emotion classes contain neutral, anger, sadness, and happiness. Assume we collect five annotations from five different unique raters for one sample, and the annotations contain neutral (N), anger (A), anger (A), sadness (S), and sadness (S). We first calculate the label distributions, which for this case is (N, A, S, H) = (0.2, 0.4, 0.4, 0.0). The threshold is 1/4=0.25, and the ground truth is converted to (0,1,1,0). During inference, we consider the predictions for three different models: (0.2,0.35,0.35,0.1), (0.1,0.45,0.45,0.0), and (0.45,0.1,0.0,0.45). The three predictions are transformed into (0,1,1,0), (0,1,1,0), and (1,0,0,1), respectively, using the threshold. In these cases, only the first two predictions are fully corrected.

We check the statistical significance of the results using each aggregation method. For the cross-validation experiments (IEMOCAP, CREMA-D, and IMPROV), we first concatenate all the predictions for each condition across all the folds, so the results consider all the data (i.e., each sample appears in one fold on the test set). For the PODCAST experiments, we directly use the predictions of all the predefined test sets from a single model. After collecting all the predictions, we split those predictions into 40 folds to evaluate the average of the macro-F1 score. We perform a two-tailed t-test to assign statistical significance if the $p-value$ is less than 0.05. We denote $*$, †, and $\star$ when a model has significantly better performance than a model training with the MR_{train}, PR_{train}, and AR_{train} sets, respectively.

5 Results and Analysis

The experimental evaluation starts by comparing our proposed framework to evaluate the aggregation rules with *state-of-the-art* (SOTA) baselines to demonstrate the merits of the SER strategy used in this study (Sec. 5.1). Then, we evaluate the three research questions listed in Section 1 (Secs. 5.2, 5.3 and 5.4).

5.1 Comparison of Results with Prior SOTA Methods

We compare the performance of our SER model with three existing SOTA approaches using the IMPROV(P), CREMA-D, PODCAST(P), and IEMOCAP corpora. The first baseline is the model proposed by Li et al. [54], which built an end-to-end framework that extracts a spectrogram from the input speech and integrates a self-attention mechanism to emphasize the emotional frames of the utterances. At the time this paper was published, this model achieved SOTA performance on the IEMOCAP database with four primary emotions. The second baseline is the model proposed by Pepino et al. [55]. This study used wav2vec 2.0 to extract speech representations, combining them with hand-crafted features (i.e., eGeMAPS [56]). The study achieved SOTA classification performance on the IEMOCAP corpus. The third baseline was proposed by Goncalves and Busso [31], which proposed a transformer architecture network trained with multimodal losses, achieving SOTA performance on the CREMA-D and IMPROV(P) corpora. To fairly compare this model with our approach, we only use the network under the "audio-only" scenario with 65 acoustic low-level descriptors as input. The three baselines are implemented following the description provided in their corresponding papers, evaluating the models under the same testing conditions as our model. In our experiments, we treat SER as a single-label task, following previous studies, and report performance on the MR or PR test conditions. We train and test these models using all the primary emotion classes.

Table 4 lists the results of our comparison. We compare our model to the SOTA models trained on the MR_{Train} data for the CREMA-D and IEMOCAP corpora and the PR_{Train} data for the IMPROV(P) and PODCAST corpora. Our model trained on the AR_{Train} set outperforms all three SOTA methods on the IMPROV(P), CREMA-D, IEMOCAP, and

TABLE 4: Macro-F1 score of existing SOTA baselines and our proposed model on the IMPROV(P), CREMA-D, IEMOCAP, and PODCAST(P) databases. The results are evaluated by aggregating the labels in the test set using either the *majority rule* (MR) or *plurality rule* (PR).

Aggregation	Method	MR IMPROV(P)	MR CREMA-D	MR IEMOCAP	PR PODCAST(P)
MR_{Train}/PR_{Train}	Li et al. [54]	0.398	0.311	0.256	0.150
	Pepino et al. [55]	0.331	0.223	0.191	0.142
	Goncalves et al. [31]	**0.539**	0.574	0.261	0.161
	Ours	0.512	**0.591**	**0.269**	**0.184**
AR_{Train}	Ours	**0.562**	**0.585**	**0.279**	**0.166**

PODCAST(P) corpora. Most of the best results came from our model, demonstrating that our SER models are competitive when compared with previous SOTA approaches. The baseline with the best results for all the conditions is the framework proposed by Goncalves and Busso [31]. On the IMPROV (P) corpus, we obtain a macro-F1 of 0.562 using AR_{Train}, outperforming this SOTA method (macro-F1 score: 0.539). However, when we train with the PR_{Train} set, we obtain a macro-F1 of 0.512, which is worse than the model trained with the approach proposed by Goncalves and Busso [31] (macro-F1 score: 0.539). On the CREMA-D corpus, to the best of our knowledge, we are the first to use annotations obtained with the voice-only condition to train SER systems. Our method obtains a macro-F1 of 0.591 using the MR_{Train} set, and 0.585 using the AR_{Train} set. This performance is better than the SOTA performance [31] (macro-F1 score: 0.574). On the IEMOCAP and PODCAST(P) corpora, our SER model outperforms all other SOTA methods [31], [54], [55].

5.2 Evaluation with Complete and Incomplete Test Data

Table 5 shows the macro-F1 scores for each combination of aggregation method and label-learning strategy. The results are based on 18 experiments with different databases where the models were trained with either the MR_{train}, PR_{train}, or AR_{train} set (6 databases $\times$ 3 learning strategies). Figures 2 and 3 present the average performance for each evaluation set (MR, PR, AR, PR-MR, or AR-PR). We perform a small-sample test of the hypothesis (matched pairs) on those results.

We consider our first research question: **how is the performance of SER systems affected by using different aggregation methods for the training set?** We evaluate this research question by assessing the models trained in different conditions in the complete test data, incomplete test data, and cross-corpus setting.

5.2.1 Evaluation on the Complete Test Set (AR)

When testing with the AR approach using all the annotated data in the test set, Figure 2c indicates that the macro-F1 score using the AR_{train} set is significantly higher than using the MR_{train} and PR_{train} sets across the 18 conditions. In fact, the best overall macro-F1 score in 14 out of 18 experiments was achieved by models trained with the AR_{train} set, as shown in Table 5. These findings suggest that incorporating the AR approach during training can enhance performance, compared to models trained with either the MR or the PR criterion. Adding more annotated samples in the training process of SER tasks is beneficial.

5.2.2 Evaluation on the Incomplete Test Sets (MR & PR)

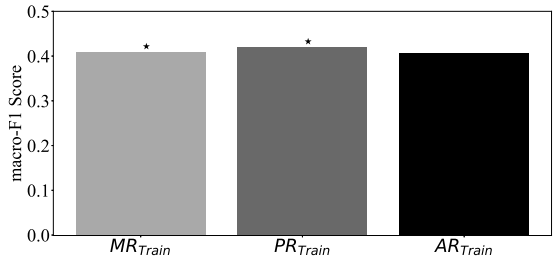
The single-label task SER performance was evaluated by considering the MR and PR test conditions. As shown in Table 5, testing with the PR set resulted in consistently lower performance compared to testing with the MR set, since more ambiguous samples were discarded in the MR set. Similarly, the performance in the PR-MR condition was generally lower than in the PR conditions. These results indicate that including more ambiguous samples in the test set (e.g., more samples without majority consensus), which are commonly seen in practical scenarios, can decrease the performance of SER models. Therefore, using either PR or MR to define the test set may not provide a representative picture of the realistic results that would be observed during the deployment of SER systems in real-world situations, where every sentence is expected to be recognized.

Out of 18 conditions, Table 5 shows that training with AR_{train} achieved the best performance in 11 cases when tested with the PR set (approximately 61%) and 14 cases when tested with the AR set (approximately 78%). Figure 2b shows that the average macro-F1 scores of the models trained with the AR_{train} are statistically significantly better than the ones obtained when training with the PR_{train} set. These results demonstrate that aggregating the annotations with the AR approach for the training set can improve the SER performance on samples with lower-agreement annotations.

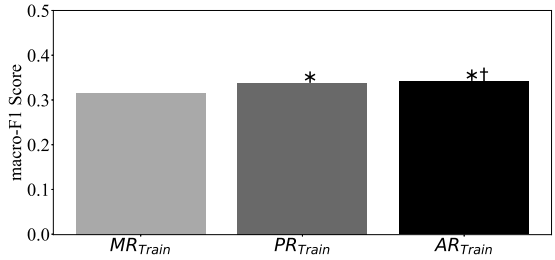
When focusing on the results evaluated on the MR test condition, the model trained with the AR_{train} set outperformed other methods in only 7 out of 18 experiments (approximately 39%). Figure 2a shows that the model trained with AR_{train} performed worse than the models trained with the MR_{train} or PR_{train} set. Incorporating more challenging samples in the training set (AR_{train}) seems to decrease its accuracy on the most straightforward (unambiguous) samples. However, this trade-off enhances the model's robustness in real-world scenarios, where ambiguous and unambiguous samples are inevitably encountered. Therefore, we argue that training SER systems with the AR_{train} set is, in general, 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." We also find that the average performance of the model trained with PR_{train} was higher than the results of the model trained with MR_{train} when the number of training samples was increased. This finding is consistent with the results reported in Chou et al. [9].

TABLE 5: The table illustrates the macro-F1 score 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.

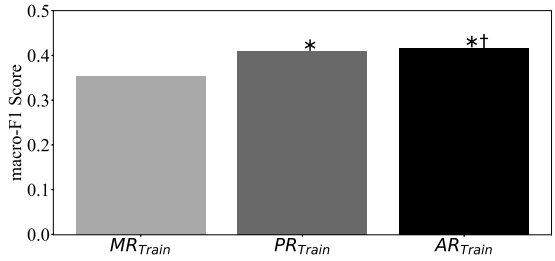
Database	Aggregation (train/test set)	Hard-label learning					Soft-label learning					Distributional-label learning				
		MR	PR	AR	PR - MR	AR - PR	MR	PR	AR	PR - MR	AR - PR	MR	PR	AR	PR - MR	AR - PR
IMPROV(P)	MR_{Train}	0.512†	0.507†	0.555†	0.300	**0.516**†	0.595	0.587	0.613	**0.346**	0.530	**0.612**	**0.604**	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**$*$†



(a) Macro-F1 scores on MR set.

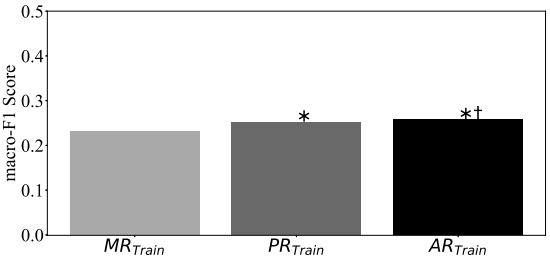


(b) Macro-F1 scores on PR set.

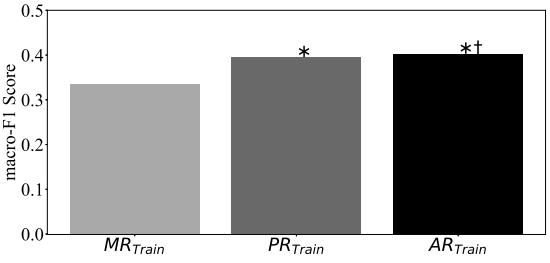


(c) Macro-F1 scores on AR set.

Fig. 2: Averaged macro-F1 scores across 18 experiments listed in Table 5 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). We denote $*$, †, and $\star$ when a model has significantly better performance than a model training with MR_{train}, PR_{train}, and AR_{train}, respectively.



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



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

Fig. 3: Averaged macro-F1 scores across 18 experiments listed in Table 5 with different databases and label-learning strategies on the different evaluation sets generated by the PR-MR and AR-PR sets. We denote $*$, †, and $\star$ when a model has significantly better performance than a model training with the MR_{train}, PR_{train}, and AR_{train} sets, respectively.

5.2.3 Evaluation in Cross-Corpus Settings

The previous experimental results are conducted in within-corpus settings. We are interested in studying the effect of training the models with different aggregation strategies in cross-corpus settings. We choose to train the model with the PODCAST (P) corpus and test the models with the IMPROV (P) corpus to demonstrate the benefits of the AR method in cross-corpus experiments. The IMPROV (P) corpus has the following emotions: anger, sadness, happiness, and neutral

emotions. The PODCAST (P) corpus includes the same four categories in addition to surprise, fear, disgust, and contempt. Given the overlap between emotions, we can conduct this cross-corpus evaluation, where a SER model is trained with the PODCAST (P), and the objective is to predict the emotions in the IMPROV (P) corpus. Our strategy is to use the models trained with the PODCAST (P) corpus as they are, evaluating their performance on the IMPROV (P) set. Then, we select the predictions on anger, sadness, happiness, and neutral state. We use the threshold to convert the distribution predictions into binary labels and calculate the results using the macro F1 score. For instance, consider the prediction of one sample of the IMPROV (P) as: (anger, sadness, happiness, surprise, fear, disgust, contempt, neutral) = (0.2, 0.2, 0.1, 0.1, 0.2, 0.1, 0.0, 0.1). We only select the four predictions without renormalizing the predictions: (anger, sadness, happiness, the neutral) = (0.2,0.2,0.1,0.1). In the next step, we use the threshold $1/C$ =1/8 to obtain the following binary format: (1,1,0,0). Consider that the ground truth of the sample is: (anger, sadness, happiness, the neutral) = (0.4,0.4,0.1,0.1). Since the threshold for the IMPROV (P) corpus is 1/4, the binary format is (1,1,0,0). In this scenario, the prediction is 100% correct.

Table 6 summarizes the cross-corpus macro-F1 results when the test set is created using the MR, PR, AR, PR-MR, and AR-PR labels. The table shows that training the models with the AR$_{Train}$ set leads to better performance on the MR, PR, AR, and AR-PR sets. This evaluation demonstrates that the proposed approach is also beneficial for cross-corpus evaluations.

5.3 Evaluation on the Ambiguous Set

We consider our second research question: **does training an SER system with data derived from the all-inclusive rule improve the performance on ambiguous emotions compared to data using the majority or plurality rules?**

5.3.1 Performance on the $AR - PR$ Condition

We analyze the results on the $AR - PR$ test condition, which only considers the samples in the AR set that are not included in the PR set. When exclusively using the samples from the $AR{-}PR$ test condition, Table 5 shows that training with MR_{train} does not show the best performances in 17 out of 18 cases (approximately 94%). We observe significantly better performance for models trained with the PR_{train} set in 10 out of 18 experiments (approximately 56%) when using the $AR - PR$ test condition. Moreover, Figure 3b shows the averaged macro-F1 score of the model trained with the MR_{train} set achieves the worst classification performance. These results show that containing more ambiguous data in the train set (i.e., the model trained with either the PR_{train} or AR_{train} set) can improve the performance on sentences with more ambiguous emotions. Therefore, we conclude that only training SER models with the MR set does not help predict ambiguous emotions. Besides, Figure 3 shows that the averaged macro-F1 scores of models trained with the AR_{train} set are significantly better than the ones trained with either the MR_{train} set or the PR_{train} set on both the $AR - PR$ and $PR - MR$ test conditions. We suggest using the AR approach to select the training data for SER tasks.

TABLE 6: The table shows cross-corpus macro-F1 results of the models trained with the 8-class MSP-PODCAST (P) set used to predict the emotions on the 4-class IMPROV (P) set.

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**

5.3.2 Analysis of the Feature Embeddings

We aim to visualize the embeddings of the models trained by various aggregation rules for sentences with high and low agreements. We take the PODCAST (P) to explore this question. We restrict the analysis to only four emotions for better visualizations: anger, sadness, happiness, and neutral state. We are interested in segments that have either low or high agreements from the test set for the visualization. We define the low and high agreement groups by selecting samples from the test set using the Cohen Kappa statistic [57]. We select the top 2% of the test samples with high agreement, obtaining 21 samples for "sadness," 33 samples for "anger," 97 samples for "happiness," and 139 samples for "neutral." We consider these samples to be representative of emotional speech with high agreement. We also consider ambiguous cases. Our approach is to select the top 2% of the test samples with low agreement. The majority of these samples do not have a clear consensus. We consider sentences that have the following two emotions, indicating in brackets the number of selected samples: anger-neutral (30), sadness-happiness (18), neutral-happiness (67), and anger-sadness (30).

We utilized the *T-distributed stochastic neighbor embedding* (T- SNE) to visualize the data distribution in the feature representation (1,024-dimensional vector) in two-dimensional plots. We visualize the figures with two emotions plus one complex emotion. For example, "anger," "sadness," plus "anger-sadness" (complex emotions). We print the name of the emotion centered around the average values of the sentences for the given class. Figure 4 shows plots for the embedding obtained with the models trained with the AR_{train} and MR_{train} sets. The results for PR_{train}, $AR - PR_{train}$, and $PR - MR_{train}$ are omitted for space limitations. The T- SNE plots show some separations between the emotions with high agreements. The samples from the complex samples with two emotions are often located between the emotional samples with high agreements. When we compare the embeddings created with models trained with the AR_{train} and MR_{train} sets, we observe a higher separation between the classes when using the AR_{train} set (see the location of the *name* of the emotions in the plots). We provide further evidence of this result with the silhouette score [58], which is a metric used to compare the quality of the clusters generated in the embeddings. It evaluates how well-separated and distinct the clusters are in the data, ranging from -1 (poor cluster) to +1 (perfect clusters). We extracted the 1,024-dimensional feature representation created by our models when they were trained with sets created with different types of consensus agreement. Table 7 lists the estimated

TABLE 7: Silhouette score of emotional clusters observed on the embeddings. The analysis of the feature embeddings includes the emotion pairs: anger-neutral (ang.-neu.), sadness-happiness (sad.-hap.), neutral-happiness (neu.-hap.), and anger-sadness (ang.-sad.). We highlight in bold the highest silhouette score for each case.

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**

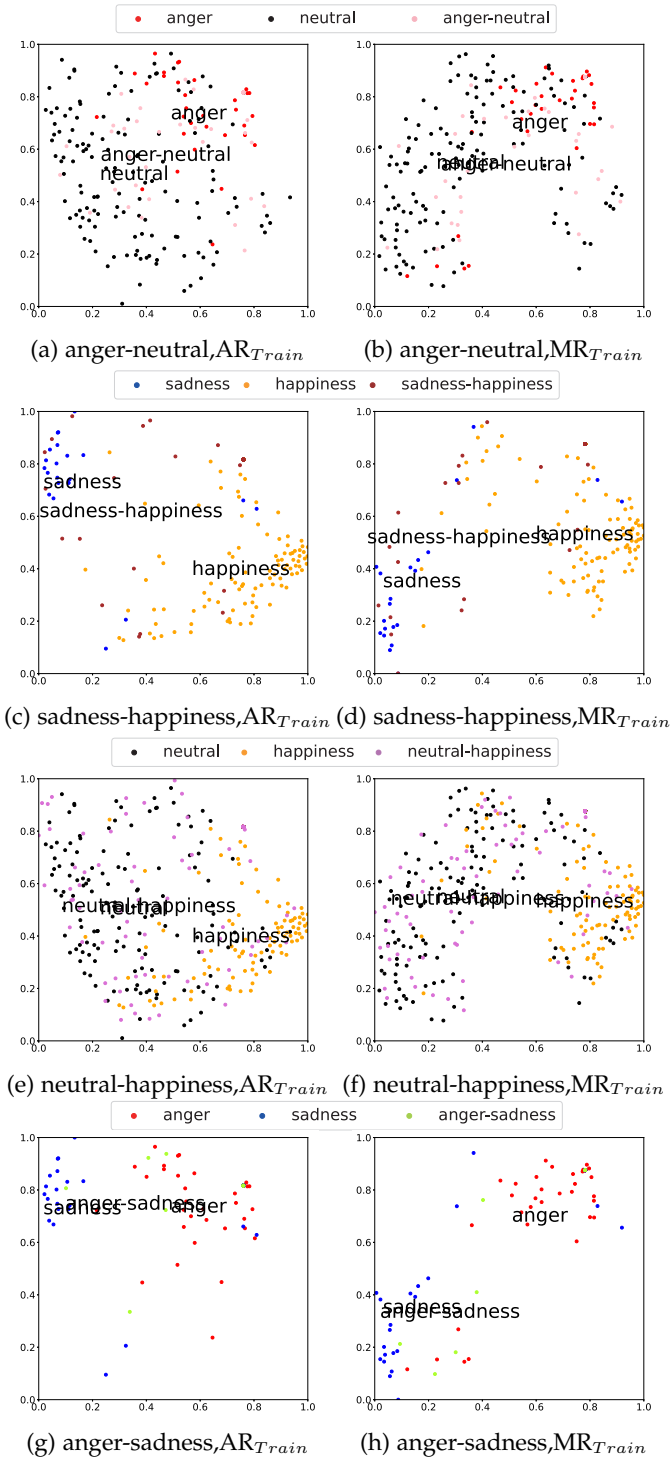
silhouette score with the three clusters: first emotion, second emotion, and the complex sentences with the two emotions. For this analysis, we also include the embedding generated by models trained with the PR_{train}, $AR - PR_{train}$, and $PR - MR_{train}$ sets. The Table shows that the model trained with the AR_{train} set achieves the highest silhouette score for the "anger-neutral", "sadness-happiness", and "neutral-happiness" cases. Surprisingly, the model trained with the $PR - MR_{train}$ set achieves the highest silhouette score on the "anger-sadness" case. Overall, the embeddings from the models trained with more ambiguous samples have a higher capability to cluster the complex emotion samples than the model trained with the MR_{train} set.

5.3.3 Role of Extra Data Added by the AR Approach

One of the benefits of using the AR_{train} set is the extra amount of data used during training, since it uses every sample in the data, in contrast to the MR_{train} or PR_{train} sets. However, adding extra data is not the only reason for the benefits of this strategy. We conducted experiments to compare models under training sets of similar size using oversampling and undersampling strategies.

We implement the oversampling approach by generating synthetic data. We follow the approach proposed by Pappagari et al. [32] to generate the data until the number of the data is equal to the number used in the AR_{train} set. Table 8 shows the results. The column "Real Data" means the number of original training sets; the column "Synthetic Data" indicates the number of generated synthetic training data. The table reports the performance in each testing set under this setting. The performance of the model trained with the AR_{train} set constantly outperforms the "MR_{train} + synthetic data" and the "PR_{train} + synthetic data." Therefore, the introduction of the samples in the $AR - PR_{train}$ set is indeed helpful in predicting ambiguous samples.

For the undersampling strategy, we also conduct experiments that reduce the training set until the number of samples in the PR_{train} and AR_{train} sets equals the number of data selected by the MR_{train} set. The samples to be removed are randomly selected. Table 8 summarizes the macro-F1 score for each condition. The table shows that training with the AR_{train} set leads to the best performance on the AR, $PR - MR$, and $AR - PR$ test sets. We conducted a second undersampling strategy where all the models are trained in two conditions with consistent size across the MR_{train}, PR_{train} and AR_{train} strategies. The first condition consists of 20K random data points in the training set selected following the corresponding rules (MR, PR, or AR). The second condition adds 12,831 data points, leading to 32,831 samples in the training set. These 12,831 samples are randomly selected from the rest of the training set, and they do not have to satisfy any consensus criteria. Since not all the 32,831 samples have consensus for



(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}

Fig. 4: T-SNE using embeddings generated by the models trained with the MR_{Train} and AR_{Train} sets. The feature embedding analysis includes the emotion pairs: anger-neutral, sadness-happiness, neutral-happiness, and anger-sadness.

TABLE 8: Analysis of the role of extra data added by the AR approach. The table reports the oversampling approach, which uses data augmentation, and the undersampling approach, which randomly removes samples until the data is consistent.

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**

TABLE 9: Results for the undersampling strategy that compares training sets with either 20,000 samples, following the corresponding aggregation rules, or 32,831 samples that are formed by randomly adding 12,831 samples regardless of whether they reach or do not reach consensus. The table shows the macro-F1 scores, highlighting the benefit 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**

the MR and PR rules, we train the models using a soft-label learning strategy, using all the samples in the set. Table 9 provides the macro-F1 score when training with both conditions. The table shows that adding more data is always helpful. Interestingly, we consistently observe the highest performance across test settings using the AR_{train} set for both conditions (20,000 and 32,831 sets). These results indicate that the AR method can improve the performance of the SER models by including ambiguous data, where the benefits are not only due to adding more samples.

5.4 What is the best label learning for SER?

We consider our third research question: **what is the best label learning strategy for training SER systems when evaluated on the complete test set?**

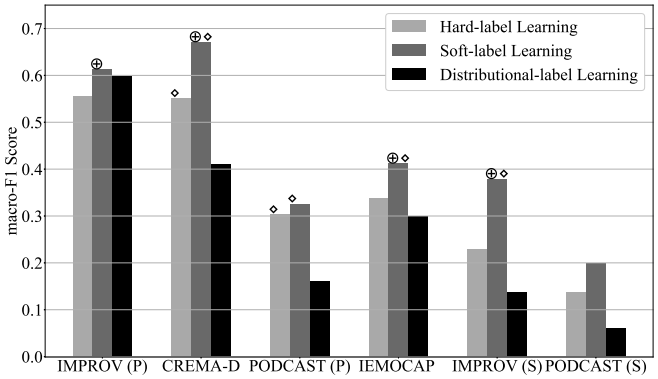
Figure 5 shows the performance of SER systems trained using various aggregation methods and different label-learning methods. We present these results on the entire test set using the AR approach. Among the label-learning methods, the distribution-label learning strategy, which uses KLD as the cost function, has the lowest performance. Using CE as the loss function results in better SER performance than using KLD. Among the label-learning strategies that use CE, soft-label learning shows better results than hard-label learning. Table 5 reveals that SER systems using soft-label learning outperformed systems using hard-label learning in 17 out of 18 cases (approximately 94%). This result aligns with previous studies, which have shown that representing emotions with soft-encoding and using the CE loss function is a more appropriate label learning strategy for training SER models [6], [7], [12], [59], [60].

Additionally, Figure 6 summarizes the macro-F1 scores for each database when using hard-label learning, soft-label learning, and distributional-label learning strategies. We only consider the results on the $AR - PR$ test set for better interpretation since these samples are the most emotionally ambiguous. Figures 6a (training with MR), 6b (training with PL), and 6c (training with AR) reveals that the soft-label learning strategy is the most suitable learning method to train SER systems from the existing learning methods to recognize mixed emotions from the ambiguous samples in the $AR - PR$ set.

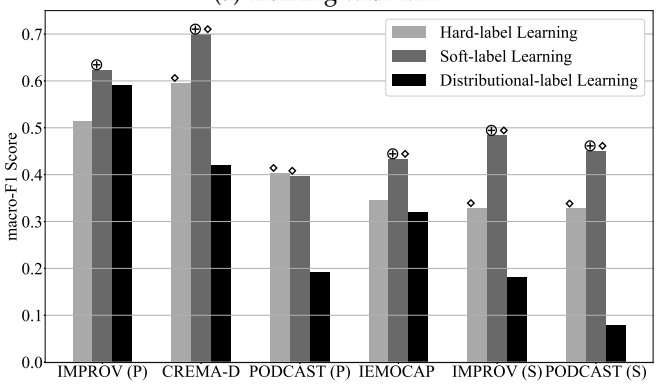
6 Conclusion and Future Work

This paper investigated the performance of speaker-independent categorical SER systems evaluated on an all-inclusive test set without discarding any data, using our aggregation rule. Our preliminary investigation showed that following the majority or plurality rule discards a significant portion of the annotated test samples, resulting in poor representations of the expected SER performance in realistic scenarios where the classifier must recognize the emotions in all the sentences, with or without consensus. The experiments with the all-inclusive test set showed that using the all-inclusive aggregation rule for defining the ground truth leads to more reliable SER performance, as the test includes more speech samples with lower-agreement annotations. Our results also indicated that the performance of SER models decreases as more ambiguous samples are included in the test set, emphasizing the importance of using the complete test set. Additionally, we found that training with high-agreement data alone cannot help to predict ambiguous emotions. Lastly, our findings showed that among label-learning strategies, soft-label learning leads to the best performance in the entire test set. The averaged SER performance of the model trained with data selected by the all-inclusive aggregation rule is consistently higher than those trained with data selected by the majority or plurality rule on both the incomplete and entire test sets.

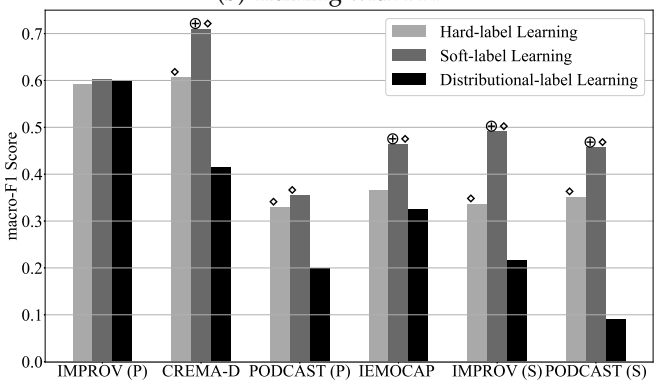
Building upon the findings presented in this study, future directions can expand the application of the all-inclusive rule to other subjective tasks. Specifically, we can investigate its effectiveness in the context of *text-to-speech* (TTS) and textless *speech-to-speech translation* (S2ST) systems. For instance, Zhou et al. [61] developed a system to synthesize human voices with mixed emotions. However, the number of emotions is limited, so their emotion embedding has room to improve. By applying the all-inclusive rule, which considers the entire dataset, we anticipate that the TTS system will become more realistic with a broader range of emotional expressions compared to existing methods. Likewise, current S2ST systems do not consider emotional



(a) Training with MR.



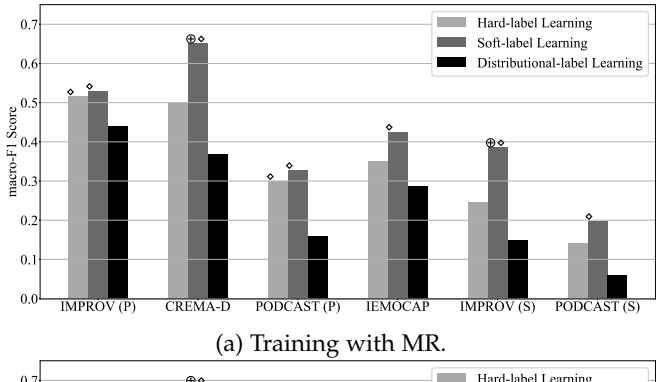
(b) Training with PR.



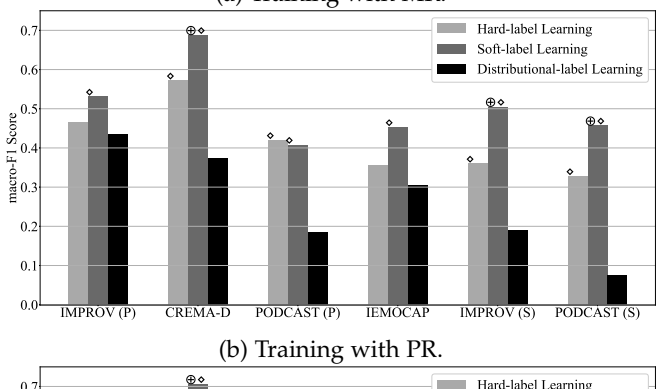
(c) Training with AR.

Fig. 5: Bar plots illustrate the macro-F1 scores when using hard-label learning, soft-label learning, and distributional-label learning strategies. All models are evaluated with the complete test set aggregated by the AR strategy for each database. We denote $\oplus$, ‡, and $\diamond$ when a model has significantly better performance than a model training with the hard-label learning, soft-label learning, and distributional-label learning strategies, respectively.

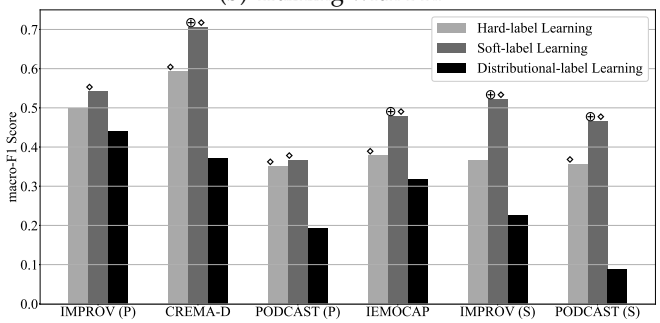
information [62], [63], which is a critical aspect of natural human conversation. Recognizing the significance of emotions in effective communication, we believe that integrating the all-inclusive rule into S2ST systems can significantly enhance their realism during speech conversion. We intend to investigate all-inclusive approaches for incorporating emotional information into S2ST systems and evaluate the impact when compared to traditional methods such as the majority rule. Besides, we will explore other important



(a) Training with MR.



(b) Training with PR.



(c) Training with AR.

Fig. 6: Macro-F1 scores for each database when using hard-label learning, soft-label learning, and distributional-label learning strategies. All models are evaluated with the **AR-PR** test set, which includes samples without MR or PR consensus. We denote $\oplus$, ‡, and $\diamond$ when a model performs significantly better than a model training with the hard-label learning, soft-label learning, and distributional-label learning strategies, respectively.

dimensions in addition to accuracy to derive better labels, including reduction of biases, improvement of fairness, and reduction of uncertainty [64], [65]. We expect that by considering all the annotations provided by the workers, we can derive better labels that also address these important dimensions.

ACKNOWLEDGMENTS

This research was supported by the NSF under Grant CNS-2016719. We also thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources. We also thank Andrea Vidal and Te-Cheng Hsu for their valuable comments.

REFERENCES

[1] J. Yoon, C. Kang, S. Kim, and J. Han, "D-vlog: Multimodal vlog dataset for depression detection," in *AAAI Conference on Artificial Intelligence (AAAI 2022)*, vol. 36, Virtual Conference, June 2022, pp. 12 226–12 234.

[2] P. Thiam, V. Kessler, M. Amirian, P. Bellmann, G. Layher, Y. Zhang, M. Velana, S. Gruss, S. Walter, H. Traue, D. Schork, J. Kim, E. André, H. Neumann, and F. Schwenker, "Multi-modal pain intensity recognition based on the SenseEmotion database," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 743–760, July-September 2021.

[3] H.-C. Chou, W.-S. Chien, D.-C. Juan, and C.-C. Lee, ""Does it matter when i think you are lying?" improving deception detection by integrating interlocutor's judgements in conversations," in *Findings of the Association for Computational Linguistics (ACL-IJCNLP 2021)*, Virtual Conference, August 2021, pp. 1846–1860.

[4] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 6447–6451.

[5] K. Vansteelandt, I. Van Mechelen, and J. Nezlek, "The co-occurrence of emotions in daily life: A multilevel approach," *Journal of Research in Personality*, vol. 39, no. 3, pp. 325–335, June 2005.

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

[7] R. Lotfian and C. 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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. Early Access, 2023.

[13] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines," in *International Conference on Language Resources and Evaluation*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:13015530

[14] 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, October-December 2019.

[15] 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*, 2016, pp. 138–142.

[16] 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

[17] L. Martinez-Lucas, A. Salman, S.-G. Leem, S. Upadhyay, C.-C. Lee, and C. Busso, "Analyzing the effect of affective priming on emotional annotations," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, September 2023.

[18] 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

[19] 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, May 2005.

[20] E. Mower, M. Matarić, 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, August 2009.

[21] 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.

[22] C. Busso and S. Narayanan, "Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, September 2008, pp. 1670–1673.

[23] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[24] H. Cao, D. Cooper, M. Keutmann, R. 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, October-December 2014.

[25] H. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, August 2017.

[26] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, and B. 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, September 2021.

[27] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, "Negative emotion recognition using deep learning for Thai language," in *Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, Pattaya, Thailand, March 2020, pp. 71–74.

[28] B. Mocanu, R. Tapu, and T. Zaharia, "Utterance level feature aggregation with deep metric learning for speech emotion recognition," *Sensors*, vol. 21, no. 12, p. 4233, June 2021.

[29] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. 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, January-March 2017.

[30] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP, 2019)*, Brighton, UK, May 2019, pp. 7390–7394.

[31] L. Goncalves and C. Busso, "AuxFormer: Robust approach to audiovisual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7357–7361.

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

[33] X. Ju, D. Zhang, J. Li, and G. Zhou, "Transformer-based label set generation for multi-modal multi-label emotion detection," in *ACM International Conference on Multimedia (MM 2020)*, Seattle, WA, USA, October 2020, pp. 512–520.

[34] 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.

[35] 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," in *AAAI Conference on Artificial Intelligence (AAAI 2021)*, vol. 35, Virtual Conference, February 2021, pp. 14 338–14 346.

[36] 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.

[37] 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.

[38] ——, "Active learning for speech emotion recognition using deep neural network," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*, Cambridge, UK, September 2019, pp. 441–447.

[39] X. Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, July 2016.

[40] W. Wu, C. Zhang, X. Wu, and P. Woodland, "Estimating the uncertainty in emotion class labels with utterance-specific Dirichlet priors," *ArXiv e-prints (arXiv:2203.04443)*, pp. 1–11, March 2022.

[41] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 7717–7721.

[42] 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 *Interspeech 2022*, Incheon, South Korea, September 2022, pp. 161–165.

[43] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[44] 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, October-December 2016.

[45] E. Mower Provost, Y. Shangguan, and C. Busso, "UMEME: University of Michigan emotional McGurk effect data set," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 395–409, October-December 2015.

[46] 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 (NeurIPS 2020)*, vol. 33, Virtual, December 2020, pp. 12 449–12 460.

[47] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned Wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding," *ArXiv e-prints (arXiv:2111.02735)*, pp. 1–7, November 2021.

[48] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. Early Access, 2023.

[49] W.-N. Hsu, Y.-H. H. T. B. Bolte, K. Lakhotia, R. Salakhutdinov, and A. 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.

[50] 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.

[51] 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.

[52] 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.

[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, June 2016, pp. 2818–2826.

[54] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2803–2807.

[55] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using Wav2vec 2.0 embeddings," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 3400–3404.

[56] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. 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, April-June 2016.

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

[58] 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

[59] 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.

[60] 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.

[61] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Speech synthesis with mixed emotions," *IEEE Transactions on Affective Computing*, pp. 1–16, October 2022.

[62] A. Lee, H. Gong, P. 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 (NAACL-HLT 2022)*, Seattle, United States, July 2022, pp. 860–872.

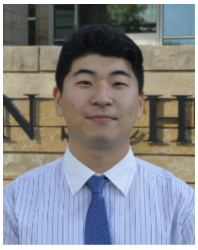
[63] X. Li, Y. Jia, and C.-C. Chiu, "Textless direct speech-to-speech translation with discrete speech representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[64] C.-C. Lee, K. Sridhar, J.-L. Li, W.-C.Lin, B.-H. Su, and C. Busso, "Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 22–38, November 2021.

[65] C.-C. Lee, T. Chaspari, E. M. Provost, and S. S. Narayanan, "PAn Engineering View on Emotions and Speech: From Analysis and Predictive Models to Responsible Human-Centered Applications," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1142–1158, 2023.

Huang-Cheng Chou (S'19) received a B.S. degree in electrical engineering (EE) from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2016. He is working toward a Ph.D. with the EE Department, NTHU, Hsinchu, Taiwan. His research interests are automatic speech emotion recognition and automatic deception detection. He won the Merry Electroacoustics Award 2021 (Bronze Award), the Best Regular Paper Award on the APSIPA ASC 2019, and the FUJI Xerox Research Award (2016). He is a co-author on the recipient of the 2019 MOST Futuretek Breakthrough Award. He was the recipient of the NOVATEK MICROELECTRONICS CORP. Ph.D. Excellence Scholarship (2022-2023), the International Speech Communication Association (ISCA) Grants (2022), the Rotary Foundation Excellence Scholarship (2021), the Graduate Students Study Abroad Program sponsored by the Taiwan Ministry of Science and Technology (MOST) (2020), the NTHU President's Scholarship (2016-2020), Taiwan Imaging Tek Corporation (TITC) Excellence Scholarship (2015), Taiwan Semiconductor Manufacturing Company (TSMC) Excellence Scholarship (2014), and the travel grant sponsored by Alphabet/Google East Asia Student Travel Grants (2022), the Foundation for the Advancement of Outstanding Scholarship (FAOS) (2019, 2022, 2023), and the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) (2019, 2022), and ACII 2017. He is a Student Member of the AAAC, ACL, ISCA, APSIPA, ACLCLP, and IEEE Signal Processing Society.

Lucas Goncalves (S'22) received his BS in Electrical Engineering from University of Wisconsin - Platteville, in 2018. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas. At UTD, he is a Research Assistant at the Multimodal Signal Processing (MSP) laboratory. In 2022 and 2023, he was awarded the Excellence in Education Doctoral Fellowship from the Erik Jonsson School of Engineering and Computer Science. His research interests include areas related to affective computing, deep learning, and multimodal signal processing. He is also a student member of the IEEE Signal Processing Society (SPS) and International Speech Communication Association (ISCA).



Seong-Gyun Leem (S'21) received his B.S. and M.S. degree in Computer Science and Engineering at Korea University, Seoul, South Korea in 2018 and 2020, respectively. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas. His current research interests include speech emotion recognition, noisy speech processing, and machine learning.

Ali N. Salman Received his B.S. and M.S. Degrees in Computer Science at Indiana State University in 2015 and 2017, respectively. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas. His current research interests include affective computing, deep learning, and facial analysis.

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

Carlos Busso (S'02-M'09-SM'13-F'23) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. Carlos Busso is a Professor at the University of Texas at Dallas in the Electrical and Computer Engineering Department, where he is also the director of the Multimodal Signal Processing (MSP) Laboratory [http://msp.utdallas.edu]. His research interest is in human-centered multimodal machine intelligence and application, focusing on the broad areas of speech processing, affective computing, and machine learning methods for multimodal processing. He has worked on speech emotion recognition, multimodal behavior modeling for socially interactive agents, in-vehicle active safety systems, and robust multimodal speech processing. He was selected by the School of Engineering of Chile as the best electrical engineer who graduated in 2003 from Chilean universities. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. In 2015, his student received the third prize IEEE ITSS Best Dissertation Award (N. Li). He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and the Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie). In 2023, he received the Distinguished Alumni Award in the Mid-Career/Academia category by the Signal and Image Processing Institute (SIPI) at the University of Southern California. He received the 2023 ACM ICMI Community Service Award. He is currently an associate editor of the IEEE Transactions on Affective Computing. He is a member of AAAC and a senior member of ACM. He is an IEEE Fellow and an ISCA Fellow.