The Importance of Calibration: Rethinking Confidence and Performance of Speech Multi-label Emotion Classifiers

Huang-Cheng Chou[1,2], *Lucas Goncalves*[1], *Seong-Gyun Leem*[1], *Chi-Chun Lee*[2], *Carlos Busso*[1]

[1]Multimodal Signal Processing (MSP) lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
[3]Department of Electrical Engineering, National Tsing Hua University, Taiwan

hc.chou@gapp.nthu.edu.tw, goncalves@utdallas.edu, SeongGyun.Leem@utdallas.edu,
cclee@ee.nthu.edu.tw, busso@utdallas.edu

Abstract

The uncertainty in modeling emotions makes *speech emotion recognition (SER) systems less reliable.* An intuitive way to increase trust in SER is to reject predictions with low confidence. This approach assumes that an SER system is well calibrated, where highly confident predictions are often right and low confident predictions are often wrong. Hence, it is desirable to calibrate the confidence of SER classifiers. We evaluate the reliability of SER systems by exploring the relationship between confidence and accuracy, using the *expected calibration error* (ECE) metric. We develop a multi-label variant of the post-hoc *temperature scaling* (TS) method to calibrate SER systems, while preserving their accuracy. The best method combines an emotion co-occurrence weight penalty function, a class-balanced objective function, and the proposed multi-label TS calibration method. The experiments show the effectiveness of our developed multi-label calibration method in terms of accuracy and ECE.

Index Terms: Speech emotion recognition, confidence calibration, class-balance loss, multi-label classification

1. Introduction

Speech emotion recognition (SER) is a key technology for advanced human-centered computing and human-machine interaction. Recently, the performances of SER systems have substantially improved with advances in speech representations with self-supervised pre-training models [1, 2]. Wagner et al. [2] achieved *state-of-the-art* (SOTA) performance in predicting sentiment on two well-known emotional databases (IEMOCAP [3] and MSP-Podcast [4]) using the Wav2vec2.0 architecture [5]. In fact, the top systems on the *emotion recognition* (ER) task in the SUPERB leaderboard [6] have utilized self-supervised pre-training models. In spite of the improved performance, the uncertainty associated with modeling emotions still opens questions about the reliability of SER predictions.

Guo et al. [7] discovered that predictions of modern neural networks are over-confident (e.g., high confidence for samples with low accuracies), which might affect the decision-making of machine learning systems. Studies have addressed this problem by developing confidence calibration methods [8, 9]. We use a *multi-label* SER model using the SOTA SER framework proposed by Wagner et al. [2] to assess, quantify, and address this calibration problem in SER systems. We surprisingly find that the predictions of SER systems are under-confident instead of over-confident. Having models' predictions that are under-confident is a problem. For example, strategies relying on a reject option, such as the work of Sridhar and Busso [10, 11], may reject too many "low confident" predictions for samples that are actually correctly predicted. It is important to develop strategies to calibrate the predictions of an SER model while improving or at least maintaining its performance.

This study analyzes the calibration level of SER systems, exploring potential approaches to calibrate their predictions. Specifically, we use the *expected calibration error* (ECE) [7] as the calibration metric to demonstrate the reliability of multi-label SER systems. We evaluate whether the class-balanced objective function and the emotion co-occurrence weight penalty function can improve the calibration and classification performance of SER systems. The class-balanced cost function [12] introduces a weight based on the number of samples for each class. We expect that this loss improves the calibration since models trained on imbalanced databases are more miscalibrated than the ones trained on balanced databases [13]. The emotion co-occurrence weight penalty function [14] considers frequent annotator disagreements between emotional classes, penalizing more infrequent emotions that do not co-occur. This loss function is expected to improve classification performance. Finally, we modify the well-known posthoc calibration method *temperature scaling* (TS) [7] to calibrate the proposed multi-label SER systems. The original TS method was designed for a single-label task. We develop an extension of the TS approach for multi-label systems by learning different emotion-dependent "Temperatures" to calibrate the confidence of the predictions for each individual emotion.

We validate our experiments on the MSP-Podcast corpus [4]. Compared to a baseline model, the class-balanced objective function leads to a 7.16% improvement gain in ECE, which shows better calibration, and an improvement gain of 4.26% in macro-F1 scores, which shows better performance. We observe the best performance when the class-balanced objective function is combined with the emotion co-occurrence weight penalty function. The proposed multi-label temperature scaling calibration method leads to clear improvements in ECE, with gains between 15.43% and 20%. We summarize the three main contributions of the paper: (1) we demonstrate the need for model calibration in SER, (2) we show that integrating a class-balanced objective function during the training process can improve the calibration and performance of a multi-label SER classifier; (3) we introduce a multi-label TS calibration method to improve the confidence of multi-label SER classifiers while preserving the classification performance.

2. Background

2.1. Multi-Label Emotion Classification

Multi-label emotion classification has been recently explored across many domains including facial expression classification [15], and text emotion classification [16]. Emotion perception is highly subjective, so annotator disagreements often exist in

annotations [17–19], which are then used as ground truth labels. It is also common to have a sentence conveying more than one emotion [20]. However, most previous SER studies regard disagreements as noise, relying on consensus labels to decide a single-label emotion class based on rules such as the majority vote or plurality rule [21–23]. Studies have shown the benefits of using labels that do not agree with the consensus label [18, 21, 24–26]. This study follows the work of Chou et al. [14, 27] to formulate the SER task as a multi-label problem.

2.2. Calibration of the Model Confidence

Deep neural networks are increasingly integrated into decision-making processes in many classification applications, such as speech recognition [28] and self-driving cars [29]. In classification tasks, the output probabilities of the model's predictions for each class can be regarded as confidence scores. The decision-making process might be affected if the confidence in the model's predictions cannot reflect the accuracy of the models. Guo et al. [7] observed that the predictions of modern neural networks are often over-confident in image and document classification tasks. The calibration of the confidence of a model has recently received researchers' attention to obtaining reliable systems [30]. However, most (if not all) existing calibration methods are designed for a single-label task instead of a multi-label task. Given our interest in multi-label SER, we choose to modify a well-known posthoc TS calibration method [7] to work for multi-label classification tasks.

3. Methodology

Our goal is not only to reveal the relationships between the confidence and the accuracy of multi-label SER systems, but also to explore approaches to calibrate confidence and improve the classification performance of SER systems. We add two existing objective functions and develop one calibration method to train and calibrate multi-label SER systems.

3.1. Task Definition and Multi-Hot Label Processing

While studies often define the objective of SER as a single-label recognition task, we formulate the problem as a multi-label recognition task. We calculate the proportion of the evaluations assigned to each emotional class by the annotators, forming a distribution. We select all the emotional classes with proportions above a given threshold. We use the threshold $1/K$ to binarize the distribution probabilities, which is the approach followed in previous studies [14, 27, 31]. K is the number of emotional classes. This step removes emotions that are not consistently provided by annotators, reducing label noise. As a result, we create a multi-hot vector representing multiple emotions.

We build a SOTA SER model implemented with the proposed losses. As a multi-label problem, this paper uses the *binary cross-entropy* (BCE) as the objective function. We define the $N \times K$ matrices for the ground truth and model prediction as $\boldsymbol{Y^T}$ and $\boldsymbol{Y^P}$, respectively. N is the number of samples. Notice that $\boldsymbol{Y^T}$ and $\boldsymbol{Y^P}$ are different. $\boldsymbol{Y^T}$ is a multi-hot vector, and $\boldsymbol{Y^P}$ is the prediction probability. The BCE loss value ($\mathcal{L}_{BCE}$) can be calculated as follows:

$$\mathcal{L}_{BCE} = -\sum_{j=1}^{K}\sum_{i=1}^{N}(Y_{ij}^{T} \cdot \log(Y_{ij}^{P}) + (1 - Y_{ij}^{T}) \cdot \log(1 - Y_{ij}^{P})). \tag{1}$$

3.2. Class-Balanced Objective Function

Spontaneous emotional databases collected in natural settings reflect the imbalance between emotions in their labels. This is the case for the MSP-Podcast corpus (Sec. 4.1), as can be seen in Figure 1. Most previous studies on SER have ignored imbalanced class distribution, which might cause miscalibration in the model predictions. This paper first explores whether considering *class-balanced loss* (CBL) during training SER systems can improve classification performance and get better calibrated SER systems. We integrate the "class-balanced sigmoid cross-entropy loss" in the study of Cui et al. [12], but we did not conduct other existing class-balance objective functions. The main idea is to add a weighting factor to adjust the values of the used loss function based on the inverses of the class frequency. The factor is $\frac{1-\beta}{1-\beta^{n_j}}$, where n_j is the number of positive samples in the j^{th} emotion class in the train set, and $\beta \in (0, 1]$ is a hyper-parameter. For our task, we have eight emotional classes, so we have eight factors to weigh the loss values. The CBL value can be calculated using Eq. 2, where ${\mathcal{L}_{BCE}}^{(j)}$ is the value of Eq. 1 for the j^{th} emotion.

$$\mathcal{L}_{CBL} = \sum_{j=1}^{K}(\frac{1-\beta}{1-\beta^{n_j}} \cdot {\mathcal{L}_{BCE}}^{(j)}). \tag{2}$$

3.3. Loss Function with the Co-occurrence Weight Matrix

Previous studies often use consensus labels in SER, ignoring the annotations that do not agree with the dominant emotional class. As a result, the relationships between emotions are hardly explored. Chou et al. [14] recently proposed an objective function that relies on the co-occurrence of emotional classes observed in the individual labels provided by the annotator. This cost function increases the penalty for cases when the models predict infrequent co-occurring emotions. The approach first calculates the $K \times K$ co-occurrence matrix by counting the frequencies that pairs of emotional classes appear in the annotations of a sentence. This matrix is set using the annotations in the train set. A variation from the approach presented by Chou et al. [14] to remove noise in the labels is that we use the co-occurring emotions after the label processing step presented in Section 3.1 (the original implementation used all annotations to estimate the co-occurrence matrix). The approach transforms the co-occurrence matrix into a penalization matrix, $\boldsymbol{P}$. Finally, $\boldsymbol{P}$ is integrated into the loss function. We use the BCE variant presented in Chou et al. [14] (Eq. 3), where P_{jz} are the entries in matrix $\boldsymbol{P}$,

$$\begin{aligned} PL = -\sum_{i=1}^{N}(\sum_{j=1}^{K}\sum_{z=1}^{K}(P_{jz} \cdot Y_{ij}^{T} \cdot \log(Y_{ij}^{P}) \\ + P_{jz} \cdot (1 - Y_{ij}^{T}) \cdot \log(1 - Y_{ij}^{P}))). \end{aligned} \tag{3}$$

Based on the analyses in Chou et al. [14], we also use a weight factor, α, in Eq. 4 to combine the penalized loss function (PL) and the BCE loss ($\mathcal{L}_{BCE}$), where $\alpha \in \mathbb{R}$ (e.g., 0.2) for better classification performance. We did not optimize the alpha value. The penalization matrix can be integrated into the *class-balanced loss* (CBL), and we denote it as the $\mathcal{L}_{P+CBL}$ in Eq. 5. The analysis will compare these cost functions.

$$\mathcal{L}_{P} = (1 - \alpha) \cdot \mathcal{L}_{CBL} + \alpha \cdot PL, \tag{4}$$

$$\mathcal{L}_{P+CBL} = (1 - \alpha) \cdot \mathcal{L}_{CBL} + \alpha \cdot \mathcal{L}_{P+CBL}. \tag{5}$$

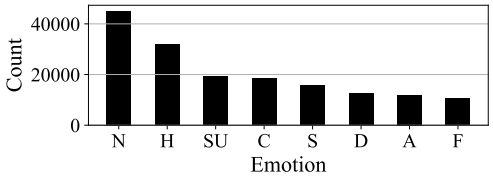


Figure 1: *The histogram shows the distribution of "multi-label" emotion classes for the MSP-Podcast corpus. Emotion includes anger (A), sadness (S), happiness (H), surprise (SU), fear (F), disgust (D), contempt (C), and neutral (N).*

3.4. Multi-Label Temperature Scaling Calibration

The *temperature scaling* (TS) calibration was originally used to calibrate the models' confidence on single-label tasks. We adapt the calibration method for a multi-label task. We follow the assumption proposed in Guo et al. [7] to regard the maximum probabilities of the predictions as the confidence values.

For a binary classification task, we obtain the predicted probabilities of the model, $\boldsymbol{Y^P}$, which is a $N \times 2$ matrix. The confidence value of each prediction is the maximum probability of the predictions. Then, we can use the TS value to calibrate the confidence of the predictions. The TS calibration method contains the following steps. First, we extract the prediction probabilities from the pre-trained models by freezing the models' weights. Then, we calculate the logits matrix by using the LogSoftmax activation function. We divide the logits vector by the learnable single scalar *Temperature* (T). Finally, the model learns the optimal value for T by minimizing the *negative log likelihood* (NLL) loss on the development set.

The task in this paper is a multi-label emotion classification task. We formulate the multi-label classification problem with multiple binary classification tasks. In this setting, we can use the TS calibration approach to learn the scalar parameter for each of the emotions (denoted $T(j)$ for emotion j). We estimate the predicted probabilities of the model, creating the $N \times K$ matrix $\boldsymbol{Y^P}$. Its j^{th} column corresponds to the model prediction $\boldsymbol{Y_j^P}$ ($N \times 1$) for the emotion j. We convert $\boldsymbol{Y_j^P}$ into a $N \times 2$ matrix denoted by $\boldsymbol{Z^P(j)}$. The first column of the matrix ($\boldsymbol{Z_0^P(j)}$) corresponds to $\boldsymbol{Y_j^P}$. The second column ($\boldsymbol{Z_1^P(j)}$) can be obtained by using $\mathbf{1}$-$\boldsymbol{Y_j^P}$. The confidence can be calibrated by using the following steps. First, we obtain the $N \times 2$ logit matrix ($\boldsymbol{a(j)}$ for the emotion j) by applying the LogSoftmax function, as shown in Eq. 6. Finally, the calibrated confidences ($\boldsymbol{c(j)}$ for emotion j) are calculated using Eq. 7. The result is a set of K temperatures (T), one for each emotion. When $T(j)$ equals 1, it maintains the original confidence values. When $T(j)$ is larger than 1, it "smooths" the confidence of the predictions.

$$\boldsymbol{a(j)} = LogSoftmax(\boldsymbol{Z^P}) = log(\frac{exp(\boldsymbol{Z^P})}{\sum_q^2 exp(\boldsymbol{Z_q^P})}), \quad (6)$$

$$\boldsymbol{c(j)} = Softmax(\frac{\boldsymbol{a(j)}}{T(j)}) = \frac{exp(\boldsymbol{a}(j)/T(j))}{\sum_q^2 exp(\boldsymbol{a_q(j)}/T(j))}. \quad (7)$$

4. Experimental Settings

4.1. The MSP-Podcast Corpus

We develop and evaluate the proposed calibration strategy using release 1.10 of the MSP-Podcast corpus [4], which has English speech. This version contains about 166 hrs of audio recordings, including 63,076 sentences in the train set, 10,999 sentences in the development set, and 16,903 in the test set. These sets aim to create speaker-independent partitions. Each sentence has at least five annotators collected with a crowdsourcing protocol adapted from the study of Burmania et al. [32]. This study focuses on the annotations of the primary emotions, where every annotator can only select one emotion from the following list: anger (A), sadness (S), happiness (H), surprise (SU), fear (F), disgust (D), contempt (C), neutral (N). They can also select "other." We show the emotion distribution of the corpus in Figure 1 based on the multi-label setting described in Section 3.1. One sentence might have more than one emotional label.

4.2. Speech Emotion Classifier and Implementation Details

To investigate whether modern SER systems need calibration, we employ the SER architecture proposed by Wagner et al. [2]. This framework obtained SOTA performance on valence prediction using release 1.7 of the MSP-Podcast corpus Based on results of the study [2], we choose the "wav2vec2-L-robust-12" model, which only uses the first 12 transformers of the full wav2vec2-large-robust model [2]. This method is an efficient approach that preserves the performance of the full model with half the transformers.

The model has one average pooling function for the outputs of the last transformer layer. Then, it has a two-hidden layer with batch normalization, and one output layer having a sigmoid function. We freeze the weights of the *convolutional neural network* (CNN) layers, and fine-tune the parameters of the transformer layers with the MSP-Podcast corpus to minimize the total cost function. We use the Adam optimizer with a learning rate of 0.0001 and a batch size of 32. We train for ten epochs, selecting the best model with the lowest loss value on the development set. All the other hyper-parameters for training the model are the same as the "wav2vec2-L-robust-12" model introduced by Wagner et al. [2]. We implement the SER architecture using the HuggingFace library [33]. The code was implemented on PyTorch using an NVIDIA RTX A6000 GPU.

4.3. Evaluation Metric and Statistical Analysis

This study analyzes and optimizes not only the performance of the models, but also their calibration. The following metrics allow us to study relationships between the accuracy and the confidence of models' predictions.

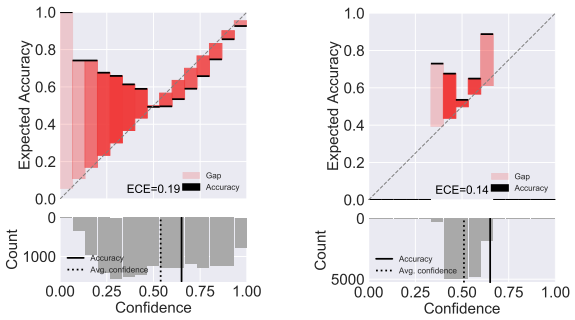
Calibration Metric: We use the *expected calibration error* (ECE) proposed by Guo et al. [7] as a metric. The ECE is a well-used metric to measure the calibration of a model. The ECE can be estimated by the weighted average of the difference between the accuracy observed in each bin and its confidence. Eq. 8 shows this process, where B is the number of bins, and N^b is the number of samples in the b^{th} bin.

$$ECE = \sum_{b=1}^{B} \frac{N^b}{N} |accuracy^b - confidence^b|. \quad (8)$$

The bins split the confidence scores of the predictions into 15 groups ($B = 15$) with the same width. For instance, the first bin includes the samples whose confidence scores range from 0 to $1/15$ shown in Figure 2. For a perfectly calibrated model, having a confidence of 0.6 implies that 60% of the predictions are correct. We calculate the ECE scores for the eight emotions, reporting the average across emotions as the final score.

Classification Performance Metric: We use the macro-F1 (**maF1**), micro-F1 (**miF1**), and weighted-F1 (**weF1**) scores to estimate classification performances of the models. These metrics consider the corresponding precision and recall rates.

Statistical Analysis: We split the test set into 40 subsets to assess the statistical significance of the experimental results. We



(a) Original (b) Calibrated

Figure 2: *Confidence versus accuracy diagrams of predictions for a model trained with $\mathcal{L}_{P+CB}$ ($\alpha = 0.2$) for happiness before and after the multi-label TS calibration method. The grey bars are the distribution of the data based on confidence.*

report the average predictions of the models across the 40 subsets. We perform a two-tailed t-test to assign statistical significance if the p-value≤ 0.05. The symbol $*$ denotes if a method is significantly better than the baseline model.

5. Evaluation

Our goal is to study the confidence of SER systems, and to explore whether considering the *class-balanced objective loss* (CBL), the emotion co-occurrence weight *penalty loss* (PL), and the proposed multi-label TS calibration method can improve both the classification performance and calibration of the SER system. We evaluate the following loss functions:

- $\mathcal{L}_{BCE}$ (Eq.1): We use the BCE loss to train the SER system.
- $\mathcal{L}_{CBL}$ (Eq.2): We use the CBL loss to train the SER system.
- $\mathcal{L}_{P}$ (Eq.4): We add the PL loss in the loss functions to train the SER systems. We set α in Eq. 4 to 0.0, 0.2, 0.5, and 0.8.
- $\mathcal{L}_{P+CB}$ (Eq.5): We use the same setting as the previous scenario, but we consider the CBL loss.

Table 1 summarizes the results of the SER systems trained with different loss functions on the MSP-Podcast corpus. The macro-F1 score on the 8-class primary emotion classification task is 0.352 using the $\mathcal{L}_{BCE}$ loss function, which is a SOTA performance on the MSP-Podcast corpus. However, we observe that the ECE value of the model trained with Eq. 1 is only 0.335. This value means that around 33% of the predictions are miscalibrated. Hence, the predictions of the SOTA SER systems need to be calibrated. Figure 2(a) shows the accuracy and confidence for happiness using the $\mathcal{L}_{P+CB}$ with $\alpha = 0.2$. The figure shows that the SER system is under-confident, since the accuracy is clearly higher than the confidence for the first seven bins. Other emotions have similar patterns.

Can CBL and PL improve the performance and calibration of an SER system? Table 1 shows that CBL (Eq. 2, $\mathcal{L}_{CBL}$) can improve both the classification performance and calibration of an SER system. However, PL only improves the classification performance ($\mathcal{L}_{P}$, Eq. 4).

Can we improve the confidence of the predictions of SER systems without performance drops? Due to the high ECE values in Table 1, we apply the proposed multi-class temperature scaling calibration method on all models. The column with the symbol $\star$ means that the confidence of the model's predictions is calibrated by the multi-label TS calibration method. We observe that classification performances for all the models are not changed, which is a similar finding reported by Gun et al. [7]. However, this approach improves the ECE values with gains between 15.4% and 20% (see column **Gain** in Tab. 1). The best model in terms of macro-F1 score is the system trained with both the co-occurrence weight penalty function and the class-balanced objective function (Eq. 5, $\mathcal{L}_{P+CB}$, with $\alpha = 0.2$). Figure 2(b) shows this model for happiness after the calibration. This model achieves the macro-F1 score of 0.401. After the multi-label TS calibration, the ECE value is 0.270, which is just slightly higher than the best ECE value reported in Table 1 (ECE = 0.263 using $\mathcal{L}_{CBL}$ and the multi-label TS calibration). This result suggests that we can achieve high performance and improved calibration by combining the CBL, PL, and multi-label TS calibration methods.

Table 1: *Performance and ECE results of the models. The column α provides the weight used in either Eqs. 4 or 5. "✓" indicates the use of the class-balanced objective functions. The column for **ECE($\star$)** indicates the use of a multi-label TS calibration method. $\uparrow$ signals that a metric is better if it is higher. Otherwise, we use $\downarrow$. The symbol $*$ indicates that the values of the models are statistically significant over the baseline (Eq. 1).*

Loss	α	**CB**	**maF1**$\uparrow$	**miF1**$\uparrow$	**weF1**$\uparrow$	**ECE**$\downarrow$	**ECE($\star$)**$\downarrow$	**Gain**
$\mathcal{L}_{BCE}$ (Eq. 1)	-	-	0.352	0.539	0.489	0.335	0.276	0.176
$\mathcal{L}_{CBL}$ (Eq. 2)	-	✓	0.367	0.548	0.504	**0.311***	**0.263***	0.154
$\mathcal{L}_{P}$ (Eq. 4)	0.2	-	0.320	0.535	0.464	0.352	0.292	0.170
	0.5		0.360	0.531	0.490	0.365	0.292	0.200
	0.8		0.331	0.539	0.473	0.345	0.286	0.171
	1.0		0.329	0.535	0.470	0.351	0.289	0.177
$\mathcal{L}_{P+CB}$(Eq. 5)	0.2	✓	**0.401***	0.560*	0.532*	0.328	0.270	0.177
	0.5	✓	0.385*	0.555	0.518	0.339	0.277	0.183
	0.8	✓	0.400*	**0.573**	**0.537***	0.329	0.273	0.170
	1.0	✓	0.371	0.555	0.507	0.316	0.266	0.158

6. Conclusions

This paper investigated the relationships between the confidence and the accuracy of predictions of multi-label SER systems. We surprisingly observed that multi-label SER systems' predictions are under-confident instead of over-confident. We combined two strategies to simultaneously improve performance and calibration: the emotion co-occurrence weight penalty function and the class-balanced objective function. The results revealed that the class-balanced loss can not only calibrate the predictions (7.16 % improvement gain in ECE), but also improve classification performance (4.26% improvement gain in macro-F1 score) of SER systems compared to the baseline model. The emotion co-occurrence weight penalty function leads to further improvements in classification performance. Finally, we proposed the multi-label TS calibration method to separately calibrate the prediction of individual emotions. Using this strategy, the ECE metric improves for all models with gains between 15.4% and 20%. Using the developed multi-label TS calibration method with the two loss function strategies, we obtain the best recognition accuracy (0.401 in macro-F1 score), achieving an ECE equal to 0.27. Compared to the baseline model, this model has a 13.92% performance gain in macro-F1 score and a 2.22% improvement gain in ECE. This study showed the importance of paying attention to the calibration of a system to improve confidence in SER tasks.

7. Acknowledgements

This research was supported by NOVATEK Fellowship and the NSF under Grant CNS-2016719. We also thank Tz-Ying Wu and Te-Cheng Hsu for their valuable comments.

8. References

[1] L. Goncalves and C. Busso, "Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks," in *Interspeech 2022*, Incheon, South Korea, September 2022, pp. 1168–1172.

[2] 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.

[3] 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, Dec. 2008.

[4] 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.

[5] 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, Dec. 2020, pp. 12 449–12 460.

[6] S.-W. Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 1194–1198.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning (ICML 2017)*, Sydney, NSW, Australia, August 2017, pp. 1321–1330.

[8] X. Wang, H. Liu, C. Shi, and C. Yang, "Be confident! towards trustworthy graph neural networks via confidence calibration," in *Conference on Neural Information Processing Systems (NeurIPS 2021)*, vol. 34, Virtual, December 2021, pp. 23 768–23 779.

[9] S. Zhao, M. Kim, R. Sahoo, T. Ma, and S. Ermon, "Calibrating predictions to decisions: A novel approach to multi-class calibration," in *Conference on Neural Information Processing Systems (NeurIPS 2021)*, vol. 34, Virtual, Dec. 2021, pp. 22 313–22 324.

[10] K. Sridhar and C. Busso, "Speech emotion recognition with a reject option," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 3272–3276.

[11] ——, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *IEEE international conference on acoustics, speech and signal processing (ICASSP 2020)*, Barcelona, Spain, May 2020, pp. 8384–8388.

[12] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, Long Beach, CA, USA, June 2019, pp. 9260–9269.

[13] Z. Zhong, J. Cui, S. Liu, and J. Jia, "Improving calibration for long-tailed recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, Nashville, TN, USA, June 2021, pp. 16 484–16 493.

[14] 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.

[15] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *International Journal of Computer Vision*, vol. 127, pp. 884–906, November 2019.

[16] H. Fei, Y. Zhang, Y. Ren, and D. Ji, "Latent emotion memory for multi-label emotion classification," in *AAAI Conference on Artificial Intelligence (AAAI 2020)*, vol. 34, New York, NY, USA, February 2020, pp. 7692–7699.

[17] 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.

[18] 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.

[19] 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, Nov. 2021.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] A. M. Davani, M. Díaz, and V. Prabhakaran, "Dealing with disagreements: Looking beyond the majority vote in subjective annotations," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 92–110, January 2022.

[25] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 951–955.

[26] ——, "Over-sampling emotional speech data based on subjective evaluations provided by multiple individuals," *IEEE Transactions on Affective Computing*, vol. 4, no. 12, pp. 870–882, October-December 2021.

[27] 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.

[28] P. Swarup, R. Maas, S. Garimella, S. H. Mallidi, and B. Hoffmeister, "Improving ASR confidence scores for Alexa using acoustic and hypothesis embeddings," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2175–2179.

[29] S. Wu, A. Hadachi, D. Vivet, and Y. Prabhakar, "This is the way: Sensors auto-calibration approach based on deep learning for self-driving cars," *IEEE Sensors Journal*, vol. 21, no. 24, pp. 27 779–27 788, December 2021.

[30] A. Karandikar, N. Cain, D. Tran, B. Lakshminarayanan, J. Shlens, M. C. Mozer, and B. Roelofs, "Soft calibration objectives for neural networks," in *Conference on Neural Information Processing Systems (NeurIPS 2021)*, vol. 34, Virtual, December 2021, pp. 29 768–29 779.

[31] 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.

[32] 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.

[33] T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.