Predicting Inter-annotator Agreements to Improve Calibration and Performance of Speech Emotion Classifiers

Huang-Cheng Chou
Department of Electrical Engineering, National Tsing Hua University, Taiwan
hc.chou@gapp.nthu.edu.tw

Abstract

Emotion recognition is receiving more attention in human-centric computer interaction. With the proliferation of ubiquitous speech-based interfaces, *speech emotion recognition* (SER) is an appealing modality to estimate emotions. However, the disagreement on emotional annotations between annotators is still a critical issue in affective computing tasks. Previous studies on SER with categorical emotions often rely on consensus labels after aggregating the classes selected by multiple annotators and formulating the task as a single-label classification problem. The emotions are considered orthogonal to each other. However, previous studies have indicated that emotions can co-occur, especially for more ambiguous emotional sentences (e.g., a mixture of happiness and surprise). Therefore, the SER task might be defined as a multi-label task. Furthermore, most of the existing state-of-the-art models are based on deep neural networks. Previous studies discovered that modern neural networks are poorly calibrated and the predictions of the models are often over-confident, under-confident, or both. This Ph.D. work hypothesizes that the disagreement between annotators on the provided annotations might cause unreliable and uncertainty in the predictions of speech emotion classifiers. I aim to investigate whether predicting the agreements among annotators on sentence-level annotations can improve the calibration of speech emotion classifiers. Then, I plan to modify the existing state-of-the-art calibration method to jointly train the SER systems to observe whether they are getting better-calibrated speech emotion classifiers.

Index Terms: Speech emotion recognition, model calibration, multi-label classification, distribution-label learning

1. Motivation

Speech emotion recognition (SER) plays an essential role in human-centered computer interaction. Given the ubiquity of speech-based interfaces, speech is one of the most convenient modalities for recognizing human emotions. Emotional labels used to train SER systems are often derived from perceptual evaluations. However, emotion perception is subjective and evaluators often have different emotional perceptions when listening to the same speech [1, 2]. The standard approach in SER studies is to regard the disagreement of emotional annotations as noise and use a major vote or plurality rule to generate a "*clear*" consensus label as the ground truth [3, 4]. This methodology only allows each speech sentence to have only one emotion label and defines the SER task as a "*single-label*" task. Additionally, this methodology ignores valuable emotional information and the chance of having co-occurring emotions, which is quite familiar with emotional behaviors (e.g., a sentence conveying a mixture of happiness and surprise). This methodology discards the data without a consensus label and the data can not be used to evaluate the SER systems. Recently, in our previous works [5, 6], we regard the SER task as multi-label tasks and use the "soft" multi-label format as the ground truth. However, the disagreement of ground truth for SER is still a critical issue in affective computing tasks. I hypothesize that the disagreement between annotators on the provided annotations might cause unreliable and uncertainty in the predictions of speech emotion classifiers. Therefore, the predictions of the SER systems might be over-confident, under-confident, or both. In this Ph.D. research, I will follow our aforementioned works and use the "soft" multi-label format as the ground truth and investigate whether predicting the agreements among annotators can improve the calibration of speech emotion classifiers.

2. Key Research Question

I aim to investigate whether the SER models need to be calibrated and explore whether modeling ambiguity in perceptual emotional evaluations can calibrate SER models. The ambiguity (the level of inter-annotator agreement) in perceptual evaluation is challenging for emotion recognition and the disagreement between annotators directly affects the performance of emotion classifiers. Most studies use majority vote or plurality rule to make labels for emotion recognition. However, these operations ignore valuable emotional information. In this research, I aim to use an entropy-based method proposed by Steidl et al. [7] to measure the ambiguity of each data sample between annotators. The entropy will be zero if all annotators provide the same emotion class on the same data. Otherwise, the more the annotators disagree, the higher the entropy. Besides, one of the well-known evaluation metrics to measure model calibration is *Expected Calibration Error* (ECE) [8]. The ECE is calculated by the difference between a weighted average accuracy and the confidence for a given bin represents the calibration gap. The ECE value of a perfectly calibrated model is zero. Inspired by this work [8], I modify the original ECE for a single-label task into the one for a multi-label task as the calibration metric. Ideally, the predictions of calibrated SER systems might have lower uncertainty when the confidence of predictions is higher.

3. Methodology

3.1. Emotion Classification Model

We use the release version 1.10 MSP-Podcast corpus [9] to evaluate our proposed method because the data source comes from real world settings instead acted emotional states. We focus on 8-class primary emotion recognition. An analysis by Keesing et al. [10] revealed that the wav2vec feature set [11] is one of the most effective features extraction approaches for SER tasks. Therefore, I rely on this feature set as the input for the SER system. I use the z-normalization function to normalize all the features, where the parameters for the mean and standard deviation are estimated from the train set. I follow the chunk-level SER modeling methodology proposed by Lin and Busso [12] as the core model. I choose to use *long short-term memory* (LSTM) as the chuck-level feature encoder equipped with the *RNN-AttenVec* chunk-level attention model, which was one of the best combinations proposed by Lin and Busso [12].

3.2. Inter-annotator Agreement Estimation

I use the entropy-based measure proposed by Steid et al. [7] to estimate inter-annotator agreements. Their method leaves

Table 1: *Results for the eight-class SER task and ambiguity recognition (the column,* ***Ambiguity****). The symbol ↑ means that the performance increases with higher values of the metric. The symbol ↓ means that the performance increases with lower values of the metric. The bold numbers indicate the best performance for a given evaluation metric.*

Weights		**Calibration**	**Distribution Similarity**						**Multi-label Classification**						**Ambiguity**	
α	β	**ECE ↓**	**Chebyshev↓**	**Clark↓**	**Canberra↓**	**KLD↓**	**Cosine↑**	**RMSE↓**	**HL↓**	**RL↓**	**COVE↓**	**maF1↑**	**miF1↑**	**weF1↑**	**Cosine↑**	**RMSE↓**
1	0	0.108	0.394	1.955	5.199	0.815	0.658	0.172	0.286	0.511	6.327	0.244	0.522	0.395	-	-
0.9	0.1	0.111	**0.392**	1.951	5.182	0.815	0.662	0.172	0.286	0.510	6.314	0.251	0.523	0.402	0.748	0.742
0.8	0.2	0.109	0.393	1.961	5.206	**0.810**	0.663	**0.172**	0.287	0.508	6.306	0.253	0.525	0.404	0.749	0.742
0.7	0.3	0.120	0.393	1.958	5.195	0.812	**0.663**	0.172	0.289	0.510	6.306	**0.267**	0.525	**0.417**	**0.752**	0.741
0.6	0.4	0.117	0.392	**1.939**	**5.157**	0.820	0.659	0.172	**0.284**	0.511	6.314	0.257	0.524	0.409	0.750	0.742
0.5	0.5	0.116	0.392	1.954	5.188	0.812	0.662	0.172	0.288	0.509	6.308	0.260	0.525	0.410	0.751	**0.741**
0.4	0.6	0.113	0.393	1.941	5.158	0.821	0.658	0.173	0.285	0.514	6.346	0.241	0.520	0.393	0.749	0.741
0.3	0.7	0.115	0.394	1.944	5.165	0.824	0.658	0.173	0.287	0.514	6.346	0.247	0.520	0.398	0.749	0.742
0.2	0.8	0.101	0.397	1.967	5.216	0.819	0.660	0.173	0.289	0.512	6.338	0.246	0.522	0.395	0.749	0.742
0.1	0.9	**0.084**	0.397	1.976	5.238	0.813	0.661	0.172	0.286	**0.506**	**6.303**	0.239	**0.526**	0.391	0.748	0.742

each annotator in succession and averages the computed entropy value (the level of inter-annotator agreement) for each annotator over the left-out annotators. I follow the denotation in [7] and define the value as the ambiguity recognition task. The overall inter-annotator agreement entropy mean of each data sample can be estimated by:

$$H(s) = \frac{1}{N}\sum_{n=1}^{N}(H(\bar{n}, s)). \tag{1}$$

3.3. Multi-task Learning and Objective Functions

3.3.1. Distribution-label Emotion Learning

In the emotion recognition task, given the ground truth ($\boldsymbol{Y^T}$) and model prediction ($\boldsymbol{Y^P}$), we use Kullback–Leibler divergence (KLD) as the objective function.

$$L^E = KLD(Y^T, Y^P). \tag{2}$$

3.3.2. Inter-annotator Agreement Learning

In the ambiguity (the level of inter-annotator agreements) estimation task, given the ground truth ($\boldsymbol{Y^T}$) and model prediction ($\boldsymbol{Y^P}$), we use Mean squared error (MSE) as the loss function.

$$L^A = MSE(Y^T, Y^P). \tag{3}$$

3.3.3. Multi-task Learning

I use different weights by setting the α or β to range from 0.1 to 0.9. The final objective function is as follows:

$$\mathscr{L}^{Total} = \alpha \cdot L^E + \beta \cdot L^A. \tag{4}$$

4. Results and Future Plan

4.1. Experimental Settings

The details about the model structure and its parameters are the same as the *two multi-task* ones used by Chou et al. [5]. Following insights from previous studies, I use the softmax function as the activation function of the output layer for KLD [5,6,13]. We use the Adam optimizer with a learning rate set to 0.0001, and with a batch sizes of 128. We train the models for 25 epochs selecting the best model based on the lowest loss on the development set. The best model is used to assess the system on the test set. To observe the effect of the proposed loss on the model performance, I set the value of α in Equation 4 to the range from 0.1 to 0.9 and the sum of α and β equals 1.

4.2. Evaluation Metrics

I use multiple evaluation metrics to compare the predicted labels with the ground truth. For calibration measure, I modified the *Expected Calibration Error* (ECE) [8] into the multi-label ECE. For distribution similarity measure, I use the metrics used in Fan et al. [14]: *chebyshev distance* (*Chebyshev*), *clark distance* (*Clark*), *canberra distance* (*Canberra*), *Kullback–Leibler divergence* (*KLD*), *cosine similarity* (*Cosine*). I also add root mean square error (*RMSE*) to evaluate the differences of values between predictions and labels. For multi-label classification performance, I use the metrics used in Fei et al. [15]: *hamming loss* (HL), *ranking loss* (RL), *coverage error* (COVE), and *macro F1-score* (maF1). I also add *micro F1-score* (miF1) and *weighted F1-score* (weF1) as evaluation metrics. I adopt the value used by Chou et al. [5, 6] setting the threshold to 1/8 to convert the prediction probabilities into the binary vectors.

4.3. Experimental Results To Date and Future Plan

Table 1 shows the overall classification performance over different settings. The column α and β are the weight values for the L^E and L^A losses, respectively. Most of the best results are from models trained with the L^A loss. While no model dominates the evaluation metrics, most of the best results are from the models predicting the ambiguity task. If we focus on maF1 and weF1, the best model trained with L^A when α equals 0.7 and β equals 0.3 achieves a 9.47% and 5.51% relative improvement over the baseline method, respectively. If we focus on ECE, the model trained with L^A when α equals 0.1 and β equals 0.9 achieves the best result. However, the calibration of model is not significantly improved in our experiments. I aim to follow the suggestion presented in [16] to use the additional dataset for calibrating *deep neural network*-based systems and include the robustness calibration metrics proposed by the paper [17, 18] and use the new start-of-the-are SER model proposed by Wagner et al. [19]. My future plan is to propose a novel calibration method based on [16] for the multi-label emotion classification task and to explore other ambiguity measure method, such as Wong et al. [20]. We also want to investigate the performance of the calibration method proposed by Tellamekala et al. [21] on multimodal emotion recognition task. We want the predictions of calibrated SER systems to have lower uncertainty when the confidence of predictions is higher.

5. Challenges and Expected Contributions

Based on the results in Table 1, the ambiguity recognition task is not significantly helpful on calibration of emotion predictions. Hopefully I can have some suggestion after the doctoral consortium from a panel of experts. I expect the following contributions of this Ph.D. work: (1) demonstrate the relationship between the calibration of SER systems and inter-annotator agreements; (2) propose a novel calibration method for SER systems.

6. References

[1] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[2] 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.

[3] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2822–2826.

[4] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2697–2709, September 2020.

[5] H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. Busso, "Exploiting Annotators' Typed Description of Emotion Perception to Maximize Utilization of Ratings for Speech Emotion Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7717–7721.

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

[7] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ""Of all things the measure is man" automatic classification of emotions and inter-labeler consistency," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 317–320.

[8] M. Pakdaman Naeini, G. Cooper, and M. Hauskrecht, "Obtaining Well Calibrated Probabilities Using Bayesian Binning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, Feb. 2015.

[9] 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.

[10] A. Keesing, Y. Koh, and M. Witbrock, "Acoustic features and neural representations for categorical emotion recognition from speech," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 3415–3419.

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

[12] W.-C. Lin and C. Busso, "Chunk-Level Speech Emotion Recognition: A General Framework of Sequence-to-One Dynamic Temporal Modeling," *IEEE Transactions on Affective Computing*, vol. Early Access, 2022.

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

[14] Y. Fan, H. Yang, Z. Li, and S. Liu, "Predicting Image Emotion Distribution by Learning Labels' Correlation," *IEEE Access*, vol. 7, pp. 129 997–130 007, 2019.

[15] 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.

[16] A. Karandikar, N. Cain, D. Tran, B. Lakshminarayanan, J. Shlens, M. C. Mozer, and B. Roelofs, "Soft Calibration Objectives for Neural Networks," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 29 768–29 779.

[17] A. Kumar, P. S. Liang, and T. Ma, "Verified Uncertainty Calibration," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[18] R. Roelofs, N. Cain, J. Shlens, and M. C. Mozer, "Mitigating bias in calibration error estimation," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Eds., vol. 151. PMLR, 28–30 Mar 2022, pp. 4036–4054.

[19] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," 2022.

[20] K. Wong and P. Paritosh, "k-Rater Reliability: The correct unit of reliability for aggregated human annotations," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 378–384.

[21] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar, "COLD Fusion: Calibrated and Ordinal Latent Distribution Fusion for Uncertainty-Aware Multimodal Emotion Recognition," 2022.