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

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

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, University of Texas at Dallas (UTD), USA

[2] Behavioral Informatics & Interaction Computation (BIIC) lab,
Department of Electrical Engineering, National Tsing Hua University (NTHU), Taiwan

Model Calibration

What is model calibration?

- **It is a method to modify the predictions of models to improve the consistence between model accuracies and predictions' probabilities**
  - If the classifier is well-calibrated, given 100 predictions, each with a confidence of 0.8, we expect that 80% of them should be correctly classified

Why does model calibration matter?

- Guo et al. [1] discovered that predictions of modern neural networks are often over-confident in computer vision and document classification tasks
  - e.g., high confidence for predictions with low accuracies

When does model calibration matter?

- Access probabilities of predictions for a richer interpretation
  - e.g., analyze the model shortcomings or provide the uncertainty to the end-users

[1] Guo et al.. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

Research Questions & Hypotheses

Do *modern Speech Emotion Classifiers (SECs) need model calibrations*?:

- Yes. Our preliminary results show that ***SECs*** using one [1] of the state-of-the-art frameworks are under-confident
- Reliability Scenario: the most intuitive way to improve the reliability of predictions is to reject predictions [2,3] based on probabilities
- Issue: we may reject too many “low confident” predictions for samples that are actually correctly predicted

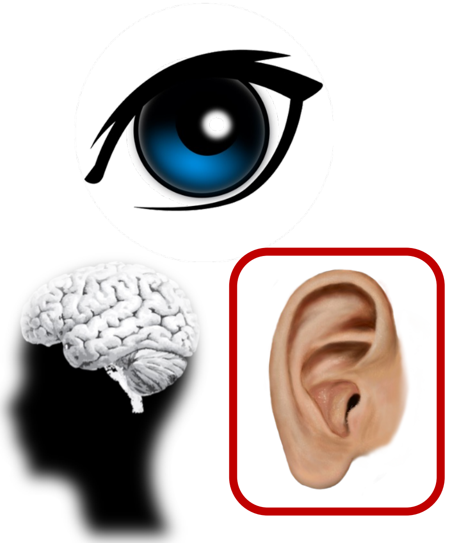
Hypotheses: the following factors could improve SECs’ calibration and classification performances

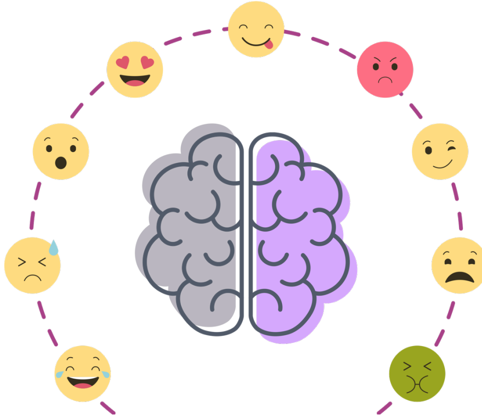
- Consider emotion co-occurrence
- Deal with the imbalance of emotional classes
- A multi-label post-calibration *temperature scaling* (TS) method

[1] Wagner et al. (2023). Dawn of the transformer era in speech emotion recognition: closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence.
[2] Sridhar, K., & Busso, C. (2019, September). Speech Emotion Recognition with a Reject Option. In *Interspeech* (pp. 3272-3276).
[3] Sridhar, K., & Busso, C. (2020, May). Modeling uncertainty in predicting emotional attributes from spontaneous speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8384-8388). IEEE.

Emotion Co-occurrence - Subjectivity of Emotion Perception

Emotional stimulus | **Emotion perception** | **Emotion decoding** | **Perceptual evaluations**





Different emotional experiences!

Speech Emotion Recognition (SER)

Emotion Co-occurrence - Example in the MSP-Podcast Corpus

Task: 8-class Primary Emotion:
File name: 0004_0073.wav
Annotations:

- **Rater 1: Neutral**
- **Rater 2: Neutral**
- **Rater 3: Sad**
- **Rater 4: Angry**
- **Rater 5: Disgust**

Quantity of classes on one sample:

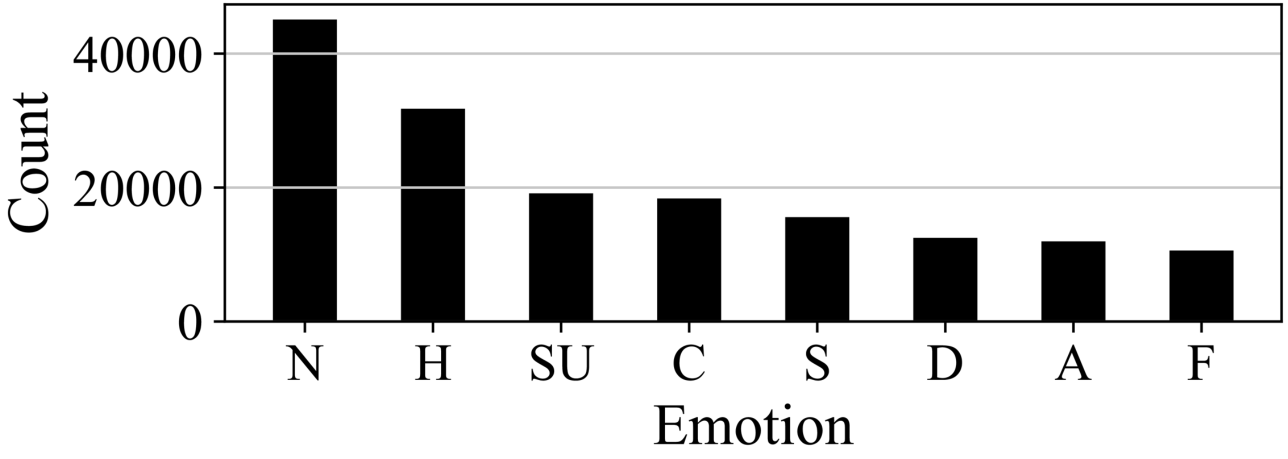
- **Neutral: 2**
- **Sad: 1**
- **Angry: 1**
- **Disgust: 1**

- **Emotion co-occurrence**
- **Disagreements among raters**
- **One sample has multiple annotations**

Imbalance of Emotional Annotations: MSP-Podcast Corpus

Imbalance annotation distribution:

- **Zhong et al. [1] found that the class imbalance makes a model more miscalibrated**
- **The right-hand side figure** shows the **imbalance** in the emotional distribution of the MSP-Podcast corpus based on the multi-label setting
  - One sentence might have more than one emotional label
- **Most previous studies on SER have ignored imbalanced class distribution**



N: neutral
H: happiness
SU: surprise
C: contempt
S: sadness
D: disgust
A: anger
F: fear

[1] Zhong, Z., Cui, J., Liu, S., & Jia, J. (2021). Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16489-16498)

Calibration Method

- **Temperature scaling calibration [1] is the most common post-calibration way to calibrate models**
- It was originally used for single-label tasks
- We adapt it for multi-label tasks, and we will introduce in detail in the next sessions

[1] Guo et al.. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

Methodology

Purpose:

- Explore whether **considering emotion co-occurrence, imbalance of emotional annotations, and a calibration method can make models better calibrated and improve performance**

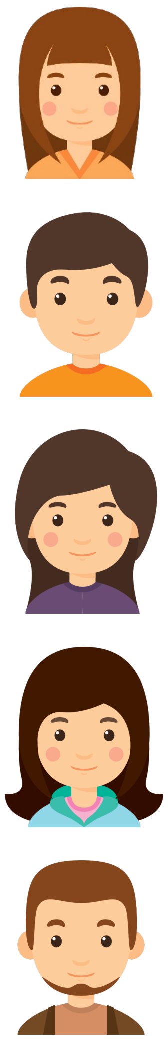
Method:

- **Emotion co-occurrence:** jointly training with emotion co-occurrence weight penalty loss function [1]
- **Imbalance of emotional annotations:** jointly training with class-balanced loss function [2]
- **Post-Calibration**: modifying the existing *temperature scaling* (TS) calibration [3] **for multi-label classification tasks**

[1] Chou, H. C., Lee, C. C., & Busso, C. (2022). Exploiting co-occurrence frequency of emotions in perceptual evaluations to train a speech emotion classifier. In *Proc. Interspeech* (Vol. 2022)..
[2] Cui, et al. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9268-9277).
[3] Guo et al.. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

Decision of Labels for Speech Emotion Recognition (SER)

UT Dallas MSP Multimodal Signal Processing Laboratory



Neutral

Neutral

Sad

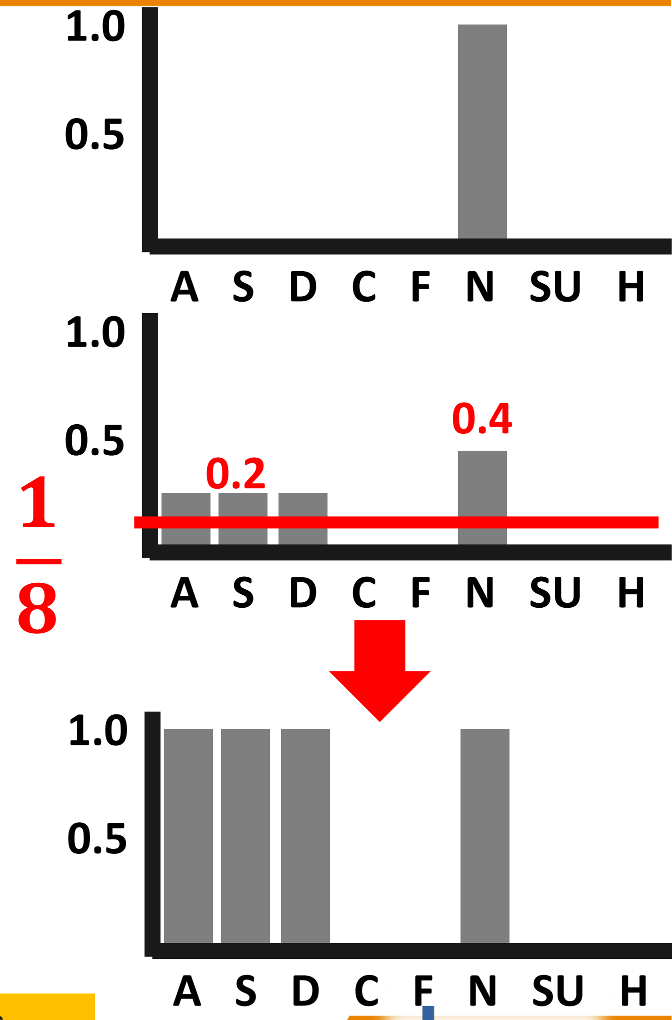
Angry

Disgust

Single-label Task → Hard-label — Neutral

Multi-label Task [1,2] → Distribution-label

Multi-label Task [1,2] → Multi-label — Neutral, Sad, Angry, Disgust



[1] Riera, P. et al. (2019). No sample left behind: Towards a comprehensive evaluation of speech emotion recognition system. In *Proc. Workshop on Speech, Music and Mind 2019*.

[2] Chou, H. C. et al. (2022, May). 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)* (pp. 7717-7721). IEEE.

Emotion Co-occurrence Weight Penalty Loss function

Step 1 → Step 2 → Step 3

Raters	MSP-PODCAST_2425_0477
1	Neutral

Raters	MSP-PODCAST_0001_0020
1	Happy

Raters	MSP-PODCAST_0001_0008
1	Neutral
2	Neutral
3	Sad
4	Other (Inquisitive)
5	Neutral

(a) Annotations

E	A	S	D	C	F	N	SU	H
A	10,534	6,331	4,470	2,415	4,670	7,679	2,469	4,063
S	6,331	18,546	6,808	4,621	9,586	14,337	4,921	7,586
D	4,470	6,808	12,767	3,613	6,807	9,260	3,835	5,796
C	2,415	4,621	3,613	10,396	6,260	7,615	3,409	5,084
F	4,670	9,586	6,807	6,260	25,271	19,530	6,457	11,351
N	7,679	14,337	9,260	7,615	19,530	33,087	9,846	12,607
SU	2,469	4,921	3,835	3,409	6,457	9,846	12,493	4,587
H	4,063	7,586	5,796	5,084	11,351	12,607	4,587	17,270

(b) Co-occurrence frequency matrix

E	A	S	D	C	F	N	SU	H
A	1.00	0.34	0.35	0.23	0.18	0.23	0.20	0.24
S	0.60	1.00	0.53	0.44	0.38	0.43	0.39	0.44
D	0.42	0.37	1.00	0.35	0.27	0.28	0.31	0.34
C	0.23	0.25	0.28	1.00	0.25	0.23	0.27	0.29
F	0.44	0.52	0.53	0.60	1.00	0.59	0.52	0.66
N	0.73	0.77	0.73	0.73	0.77	1.00	0.79	0.73
SU	0.23	0.27	0.30	0.33	0.26	0.30	1.00	0.27
H	0.39	0.41	0.45	0.49	0.45	0.38	0.37	1.00

(c) Co-occurrence weight matrix

E	A	S	D	C	F	N	SU	H
A	0.00	0.66	0.65	0.77	0.82	0.77	0.80	0.76
S	0.40	0.00	0.47	0.56	0.62	0.57	0.61	0.56
D	0.58	0.63	0.00	0.65	0.73	0.72	0.69	0.66
C	0.77	0.75	0.72	0.00	0.75	0.77	0.73	0.71
F	0.56	0.48	0.47	0.40	0.00	0.41	0.48	0.34
N	0.27	0.23	0.27	0.27	0.23	0.00	0.21	0.27
SU	0.77	0.73	0.70	0.67	0.74	0.70	0.00	0.73
H	0.61	0.59	0.55	0.51	0.55	0.62	0.63	0.00

(d) Penalization matrix

We denote the penalization matrix as P

This loss function [1] can *penalize* more infrequent emotions that do not co-occur

$$\mathcal{PL} = \sum_{i=1}^{N} (\mathcal{L}_i \cdot P)$$

$$= \sum_{i=1}^{N} \left(\sum_{j=1}^{K} \sum_{z=1}^{K} P_{jz} \cdot f_{loss}(Y_{ij}^{T}, Y_{ij}^{P}) \right)$$

f_{loss} = Binary Cross-Entropy (BCE) in the work

[1] Chou, H. C., Lee, C. C., & Busso, C. (2022). Exploiting co-occurrence frequency of emotions in perceptual evaluations to train a speech emotion classifier. In *Proc. Interspeech* (Vol. 2022).

Imbalanced Annotations: Class-balanced Loss Function

Class-balanced Sigmoid Cross-entropy Loss:

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

where

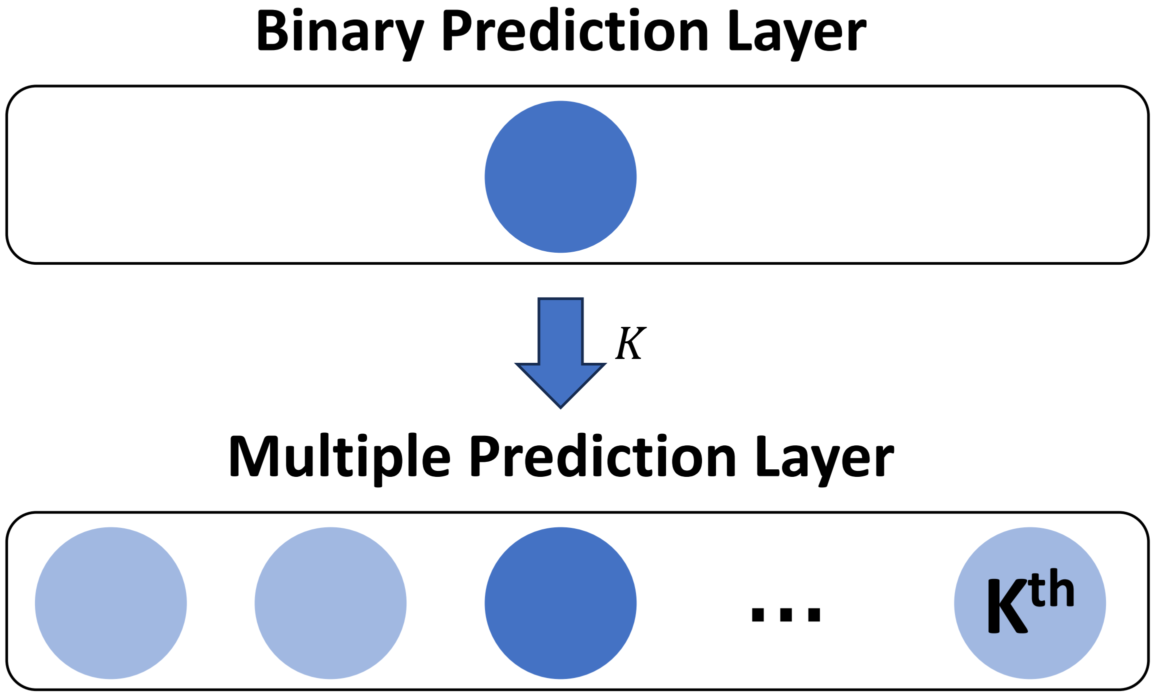
- K: number of emotional classes;
- $\beta \in (0, 1]$ hyperparameter ;
- n_j: number of positive samples in j^{th} emotion classes in the "train set"

Add a weighting factor to adjust the values of the used loss function based on the inverses of the class frequency

[1] Cui, et al. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9268-9277).

Multi-Label Temperature Scaling Calibration (1/3)

Let's consider multi-label tasks as multiple (K) binary classification tasks

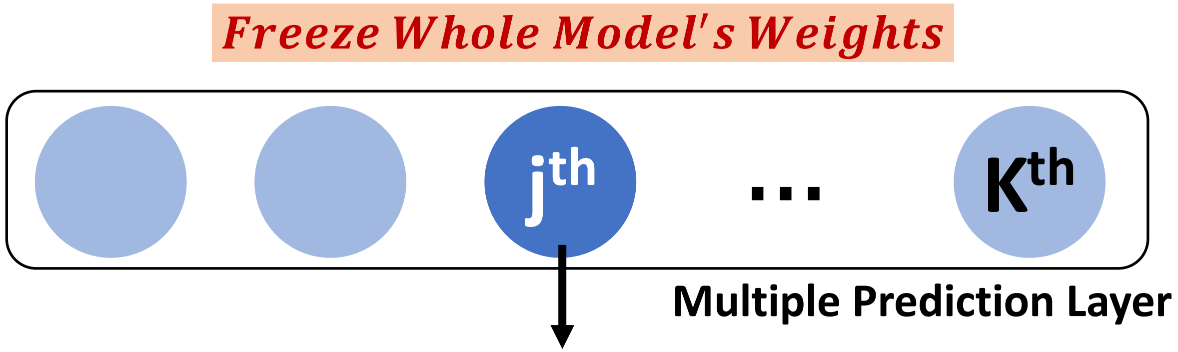


[1] Guo et al.. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

Multi-Label Temperature Scaling Calibration (1/2)

Step 1: Access probabilities of samples in the development set (N is numbers of samples)

- Extract the prediction probabilities from the pre-trained models by freezing the models' weights
- Converting vectors for the emotion *j* into $N \times 2$, ***denoted as*** $Z^P(j)$
- Confidence value (c) of each prediction is the maximum probability of the predictions



$$Z_0^P(j)$$

$$Z_1^P(j) = 1 - Z_0^P(j)$$

$$Z^P(j) = \left[Z_0^P(j)\ Z_1^P(j)\right]_{\mathrm{Nx2}}$$

$$c(j) = max(Z_0^P(j), Z_1^P(j))$$

Multi-Label Temperature Scaling Calibration (2/2)

Step 2: Get the logits matrix by using the LogSoftmax activation function

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

Step 3: Divide the logits vector by a learnable single scalar temperature (T)

- The model learns the optimal value for T by **minimizing the negative log-likelihood (NLL) loss on the development set**

$$\frac{a(j)}{\color{red}{T}(j)} \longrightarrow \textbf{Learnable T}$$

Step 4: Calculate the calibrated confidences:

$$T(j)\begin{cases} = 1;\ \text{maintain the original confidences} \\ > 1;\ \text{“smooth” the confidences} \end{cases}$$

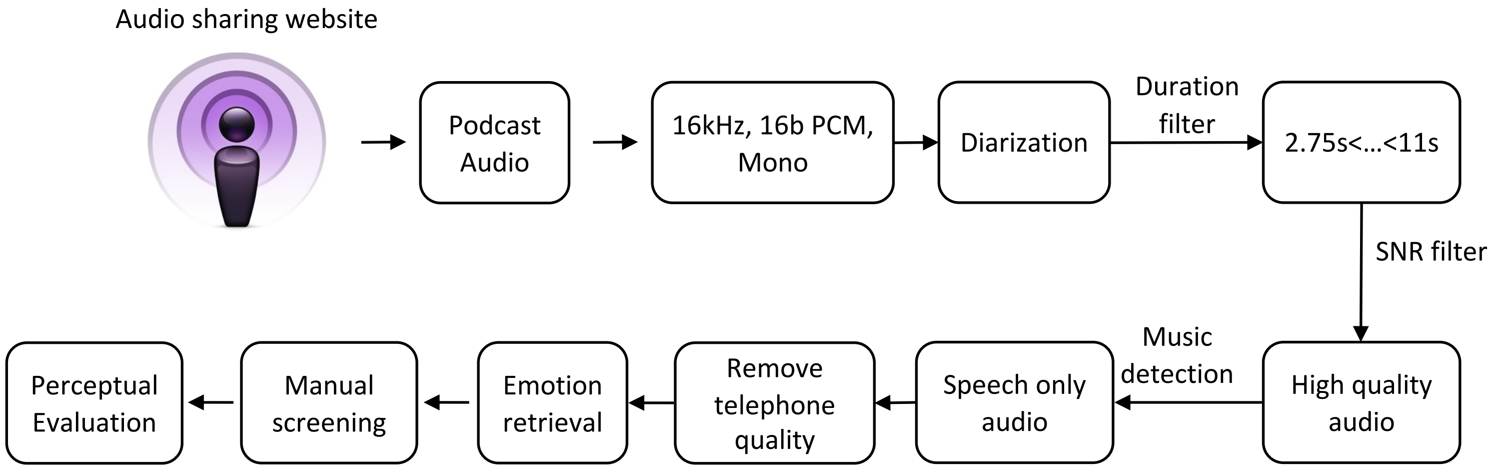
$$\mathrm{c(j)} = \boldsymbol{Softmax}\left(\frac{a(j)}{T(j)}\right) = \frac{\boldsymbol{exp}\left({}^{a(j)}/_{T(j)}\right)}{\sum_q^2 \boldsymbol{exp}\left({}^{a_q(j)}/_{T(j)}\right)}$$

Experiment Setup (Corpus): The MSP-Podcast Corpus (Version 1.10)

UT Dallas MSP Multimodal Signal Processing Laboratory

Audio sentences:

- Train set: 63,076
- Validation set: 10,999
- Test set: 16,903



Emotional Annotations:

- Crowdsource-based protocol
- Every sentence has **more than** 5 annotators
- Primary emotion (P) (Single-choice):
  - anger, sadness, happiness, surprise, fear, disgust, contempt, neutral, and ~~other (excluded)~~

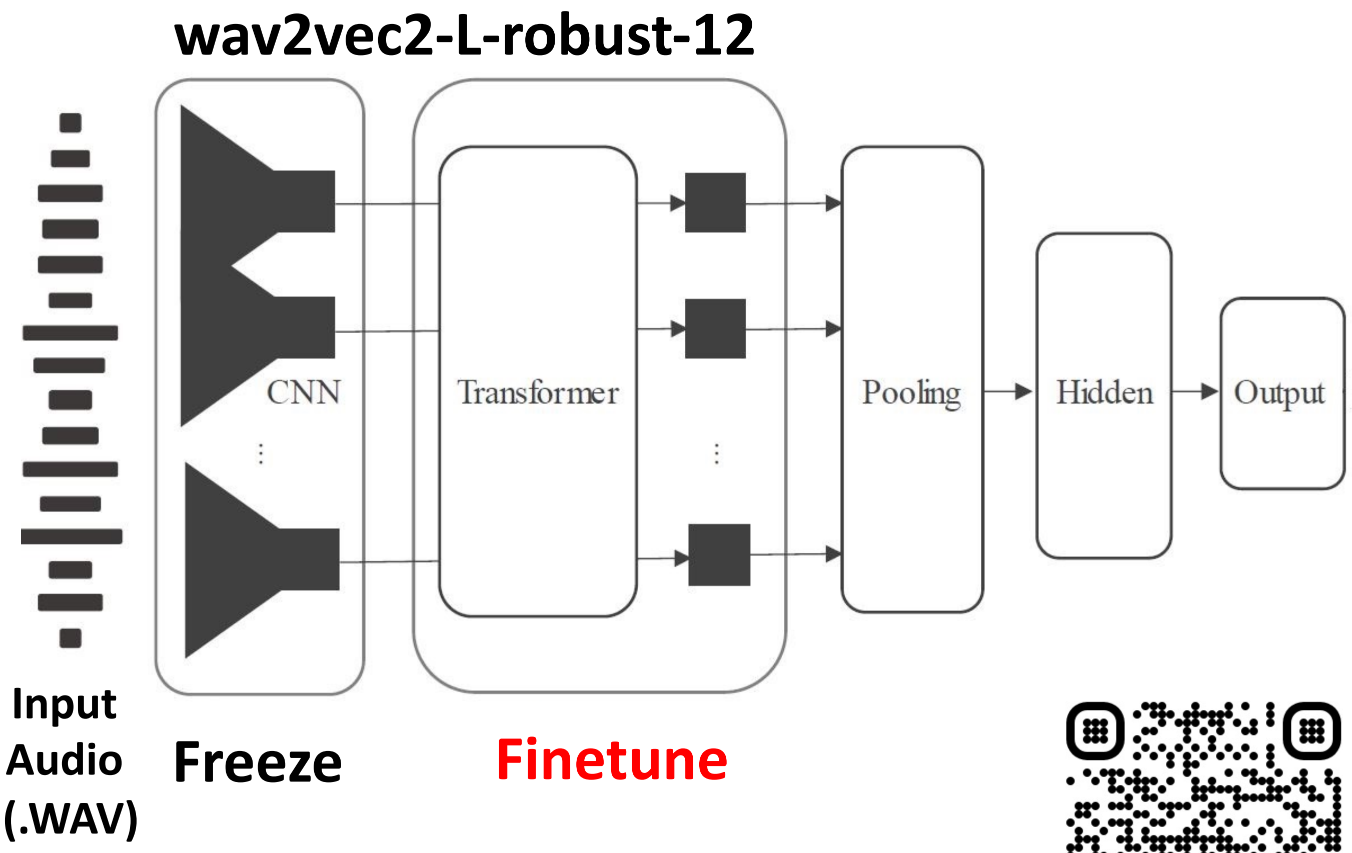
Access the MSP-Podcast

[1] Lotfian, R., & Busso, C. (2017). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, *10*(4), 471-483.

Experiment Setup (SER Framework)

We use the "wav2vec2-L-robust-12" model in [1] and follow the finetuning process:

- Freeze the weights of the *convolutional neural network* (CNN) layers
- Finetune the weights of the 12 transformer layers on the emotion corpus
- Use Adam optimizer with a learning rate of 0.0001 and a batch size of 32



Code of the Original paper[1]

[1] Wagner et al. (2023). Dawn of the transformer era in speech emotion recognition: closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Experiment Setup (Ablation Study)

Model evaluation of different systems on primary emotion recognition

1. $\mathcal{L}_{BCE}$: Binary cross-entropy loss (as Baseline)
2. $\mathcal{L}_{CBL}$: Consider class-balanced loss
3. $\mathcal{L}_{P}$: $(1\text{-}\alpha)\cdot \mathcal{L}_{BCE} + \alpha \cdot P\mathcal{L}$
4. $\mathcal{L}_{P+CB} = (1\text{-}\alpha)\cdot \mathcal{L}_{CBL} + \alpha \cdot \mathcal{L}_{P+CBL}$
5. **Apply the multilabel temperature scaling calibration method**
   - Can be used for the above No. 1, No.2, No.3, and No. 4 systems

$P\mathcal{L}$ = **Emotion co-occurrence weight penalty loss**

α = **0.0, 0.2, 0.5, or 0.8 (we did not optimize α in this work)**

[1] Chou, H. C., Lee, C. C., & Busso, C. (2022). Exploiting co-occurrence frequency of emotions in perceptual evaluations to train a speech emotion classifier. In *Proc. Interspeech* (Vol. 2022).

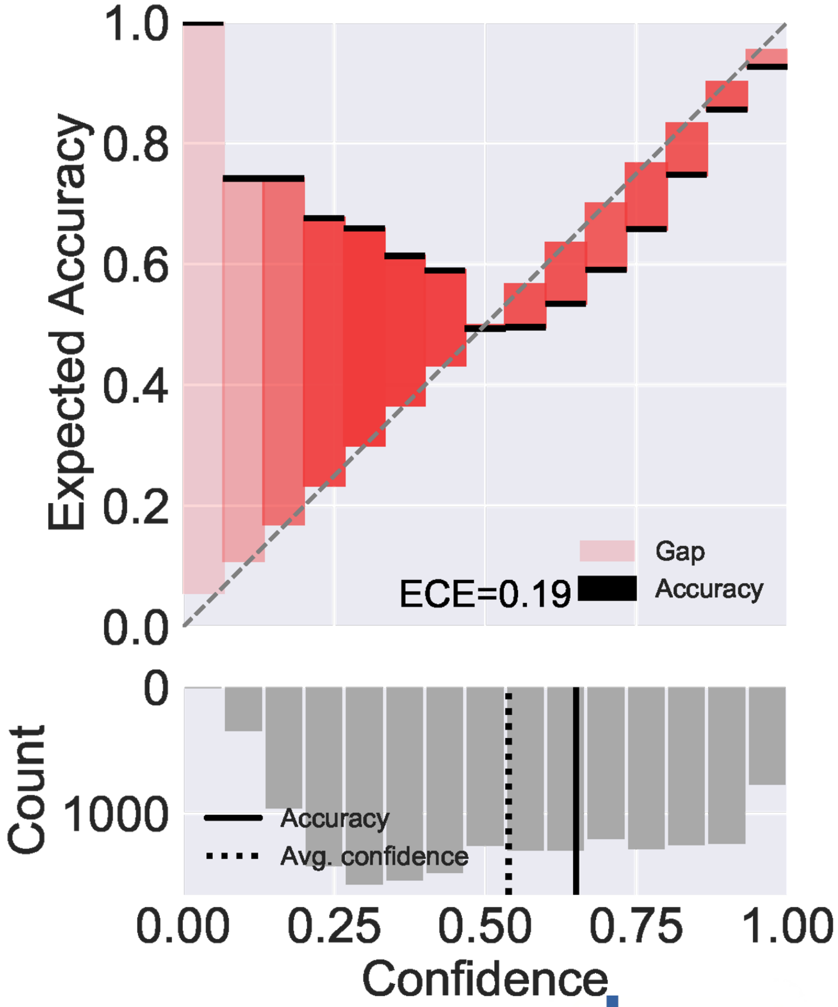
Evaluation Metric (Calibration and Classification Performance)

Calibration: Expected Calibration Error (ECE) [1]:

$$\boldsymbol{ECE} = \sum_{b=1}^{B} \frac{N^b}{N} \left| accuracy^b - confidence^b \right|,$$

where

- B = # of bins ($B = 15$, follow [1])
- N^b = # of samples in the b^{th} bin
  - We calculate the ECE scores for 8 emotions and show the averaged ECE as the final scores



Classification Performance:

- Macro-F1 (maF1) - Other metrics in the paper

[1] Guo et al.. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

Experiment Results & Statistical Analysis

Average of results:

- Split the original test set into 40 small subsets
- Report the average results in evaluation metrics

Tests of Significance:

- Perform a **two-tailed test**
- Evaluate the statistical significance of all results between the proposed method and baselines
  - Use the symbol * to indicate that the results of the models are statistically significant over the baseline

Research Question (1)

Can CBL and PL improve the performance and calibration of an SER system?

- $\mathcal{L}_{CBL}$ can improve both the classification performance and calibration of SER systems
- PL only can improve the classification performance based on $\mathcal{L}_P$ results
- Use CBL and PL simultaneously can improve the performance and calibration of SER

CBL: class-balanced loss

PL: emotion co-occurrence weight penalty loss

Loss	α	CB	maF1 ↑	ECE ↓
$\mathcal{L}_{BCE}$			0.352	0.335
$\mathcal{L}_{CBL}$		✓	0.367	0.311*
$\mathcal{L}_{P}$	0.2	-	0.320	0.352
	0.5		0.360	0.365
	0.8		0.331	0.345
	1.0		0.329	0.351
$\mathcal{L}_{P+CB}$	0.2	✓	0.401*	0.328
	0.5		0.385*	0.339
	0.8		0.400*	0.329
	1.0		0.371	0.316

Research Question (2)

Can we improve the confidence of the predictions of SER systems without performance drops?

- **CB✓** means the model is calibrated by the proposed multi-label TS calibration method
- CB sustains classification performances
- Improved ECE values with gains between 15.4% and 20%
- **Best performance and calibration achieved with CBL + PL + multi-label TS calibration**

CBL: class-balanced loss

PL: emotion co-occurrence weight penalty loss

Loss	α	CB	maF1 ↑	ECE ↓	ECE (CB✓) ↓	ECE Gain
$\mathcal{L}_{BCE}$			**0.352**	**0.335**	0.276	17.6%
$\mathcal{L}_{CBL}$:		✓	0.367	**0.311***	0.263	15.4%
$\mathcal{L}_{P}$	0.2	-	0.320	0.352	0.292	17.0%
	0.5		0.360	0.365	0.292	**20.0%**
	0.8		0.331	0.345	0.286	17.1%
	1.0		0.329	0.351	0.289	17.7%
$\mathcal{L}_{P+CB}$	0.2	✓	**0.401***	**0.328**	**0.270**	**17.7%**
	0.5		0.385*	0.339	0.277	18.3%
	0.8		0.400*	0.329	0.273	17.0%
	1.0		0.371	0.316	0.266	15.8%

Conclusions

Contribution:

- Multi-label speech emotion classifiers are under-confident
- **Emotion co-occurrence weight penalty function + Class-balanced objective function, + Multi-label calibration** simultaneously can improve performance and calibration of SER models

Results (8-class Primary emotion classification)

- **+2.22%** improvement gain in ECE
- **+13.92%** performance gain in macro- F1 score

Take-Home Message:

- **It is important to calibrate the SER system to improve its confidence and performance**

Thank You

Scan Me for Paper Full Text

傑出人才發展基金會
Foundation For The Advancement of Outstanding Scholarship

NSF

CNS-2016719

Contact **Huang-Cheng Chou**:
Email: hc.chou@gapp.nthu.edu.tw
LinkedIn: https://www.linkedin.com/in/huangchougchou/