Analysis of Emotion Annotation Strength Improves Generalization in Speech Emotion Recognition Models

Joao Palotti[*]
Earkick
joao@earkick.com

Gagan Narula[*]
Earkick
gagan@earkick.com

Lekan Raheem
Earkick
lekanwraheem@gmail.com

Herbert Bay[†]
Earkick
herbert@earkick.com

Abstract

Recent advances in speech emotion recognition (SER) have relied on a mix of acted and in-the-wild research datasets. It is unclear whether annotations in these datasets are of similar strength or quality, can reliably be detected by other human annotators, and to what extent emotion classification knowledge can be transferred between acted and in-the-wild data. A well known, large in-the-wild dataset for emotion classification and sentiment analysis is the CMU-MOSEI video dataset. The raw annotations of CMU-MOSEI are "soft labels" on a Likert scale. Usually, experiments are performed with a simple binarization of these fine-grained labels. In this work, we re-annotated 1% of the data from two acted and two in-the-wild datasets to analyze the strength of emotion annotation per label, compare annotation accuracy between acted and in-the-wild data, and identify an appropriate threshold for CMU-MOSEI label binarization. We report a significant improvement (7% increase on weighted average F_1) using the same model architecture in emotion classification by simply identifying a better threshold for CMU-MOSEI. Further, we show that emotion annotation strength of acted and in-the-wild data is similar, and that the same model architecture generalizes to the same extent when trained on acted and tested on in-the-wild data, and vice-versa.

1. Introduction

Speech Emotion Recognition (SER) is a subset of the general Automated Emotion Recognition (AER) problem, where human speech in acoustic signals is used to perform tasks such as identify emotions, sentiment, predict valence and arousal etc. [1,2]. The advent of deep learning has led to an important increase in the accuracy of SER models working directly on mel-spectrogram representations [3] or the raw audio waveform [4–6].

Since deep networks typically thrive on rich datasets, the SER research community has open-sourced large multimodal video datasets annotated with discrete emotion categories [7, 8] and valence-arousal dimensional values [9]. In this work, we study one of the key features that qualitatively distinguishes datasets – i.e., whether a dataset was built from *acted emotions* (e.g., IEMOCAP [10], CREMA-D [11], RAVDESS [12], EMODB [13]) or *emotions in-the-wild* (CMU-MOSEI [14], Aff-Wild2 [15], MSP-Podcast [16]).

Research algorithms for SER often train or evaluate models on a mix of acted and in-the-wild datasets. However, in-the-wild natural emotion may well be out-of-distribution for acted emotion. In-the-wild datasets possess the advantage that they contain many different speakers, ethnicities, and accents, as well as environmental noise, recording quality, background chatter, music, etc. Therefore, they require deep network models to learn more robust representations of emotional speech by becoming insensitive to confounding stimuli. With over 65 hours of video data and more than 1000 speakers, CMU-MOSEI [14] is the largest in-the-wild dataset. A unique aspect of CMU-MOSEI is that emotions are annotated on a $[0, 3]$ Likert scale, which effectively creates *soft labels*. Previous works train and evaluate on CMU-MOSEI with *hard labels* derived by applying a threshold ($t > 0.$) to each emotion annotation [17,18]. It is unclear whether this choice of threshold is justified.

The first contribution of this paper is investigating the degree of knowledge transfer between acted and in-the-wild datasets and whether inter-annotator agreement varies across these two types of emotional audio data. The second contribution of this paper is performing an in-depth analysis of the effect of different thresholds on inter-annotator agreement and SER model accuracy. For both contributions, three authors in our work re-annotated a random subset (1%) of the combination of four datasets: two acted, namely CREMA-D and RAVDESS, and two in-the-wild, namely Affwild2 and CMU-MOSEI, to analyze inter-annotator agreement disaggregated by emotion and dataset

[*] These authors contributed equally to this work
[†] Correspondent Author

type. Further, we varied the threshold applied to CMU-MOSEI labels and report the effect on SER model scores.

The above-mentioned two contributions of this paper led us to define and shed light on the following research questions:

- **RQ1:** Are there emotions that are easier to annotate than others?
- **RQ2:** Is it easier to annotate acted data or in-the-wild data?
- **RQ3:** What is the appropriate threshold for the CMU-MOSEI dataset?
- **RQ4:** How does changing the threshold for the CMU-MOSEI dataset affect SER model performance?
- **RQ5:** Is there a transfer between acted and in-the-wild datasets?

1.1. Related Work

Automated speech emotion recognition is typically studied using two tasks: (1) Valence-Arousal Estimation [15,19] or (2) discrete emotion classification [2]. We focus our efforts on the latter (2) in this study. Recently, transformer-based [20] deep neural network models have reported state-of-the-art results on benchmark datasets for SER such as for MSP-Podcast [21], IEMOCAP [22] and CMU-MOSI [6]. The conclusion from these studies is that large pre-trained models such as HuBERT [4] and wav2vec2 [5] are excellent initializations for fine-tuning a transformer-based network on a downstream SER task. Furthermore, they conclude that optimizing the entire network except the initial convolutional layers is necessary for good performance. We utilize these findings for our architecture in this study, choosing HuBERT base model as a pretrained backbone. Current weighted average F_1 scores on CMU-MOSEI are on the order of 0.8. Multimodal transformer networks show greater accuracy, presumably because of the complementary information present in text and visual domain such as facial expressions, gestures and body language [17, 23] reaching weighted average F_1 scores of 0.875.

Although it has been shown that people from different cultural backgrounds can recognize over 20 emotions in human speech using cues such as prosody [24, 25], and non-linguistic vocal bursts [26, 27] the understanding of emotional states observed in others also shows cultural differences [28–31]. Annotators on services like Amazon Mechanical Turk often work within specific contexts and cultures, for example, the CMU-MOSEI annotators were based in India. This may influence the certainty with which emotions are interpreted, regardless of prior experience and instruction because there can be subtle differences across cultures . Empirically, this is reflected in the inter-annotator correlation and the average rating of emotion intensity in CMU-MOSEI, with stronger emotions receiving higher average scores on the $[0, 3]$ Likert scale. Thus, the threshold applied to each sample's rating to map *soft labels* to *hard labels* is of considerable significance because it reflects the *certainty* of the label. We investigate the effect of different threshold values on subjective human annotation and knowledge transfer from CMU-MOSEI.

It has been observed previously that acted emotion may not reflect *real* emotion in human study participants [32,33]. Nonetheless, acted emotional speech remains of significant importance in SER research [6, 25] with IEMOCAP [10] being regularly used as a benchmark SER dataset. Acted data is recorded in controlled lab environments. It can often be exaggerated in comparison to in-the-wild data, which (1) has a more natural context, (2) is often evoked by stimuli or self-produced, and (3) is recorded in variable settings. We investigate whether it is easier to annotate acted data than in-the-wild data, and whether there is knowledge transfer between these two types of datasets, and if in-the-wild data is superior in generalization to acted data because of these characteristics.

2. Methods

In order to answer our five research questions, we perform two main tasks in this paper. The first task consists of re-annotating samples from four different datasets. Insights gained from the first task are used in the second task, in which we conduct ML experiments with the same four datasets.

2.1. Datasets

We use the following four datasets in our experiments (a summary of their characteristics is shown in Table 1):

- **CMU-MOSEI** [14, 18]: the CMU Multimodal Opinion Sentiment and Emotion Intensity dataset is a large-scale multimodal dataset for sentiment and emotion analysis in videos. The dataset consists of 1,000 videos extracted from YouTube of speakers with a diverse range of demographics and topics. Each video is annotated with continuous and categorical sentiment and emotion labels. The annotations are obtained using 3 annotators. Their agreement was calculated using Krippendorff's alpha [34] and shown in the first numeric column of Table 4. The dataset includes annotations for the following six emotions: anger, disgust, fear, happiness, sadness, and surprise.
- **Aff-Wild2** [15, 19, 35–41]: the Aff-Wild2 dataset is an extension of the Aff-Wild database [42]. The dataset contains video recordings of 539 participants (361 for training and 178 for validation) displaying

Dataset	Label Type	Acquisition	Orig. Size	Relabeled
CMU-Mosei [14]	Multi-label, multi-scale	In-the-Wild	22,777	228
Aff-Wild2 [15]	Multi-class	In-the-Wild	3,180	32
CREMA-D [11]	Multi-class	Acted	7,442	75
RAVDESS [12]	Multi-class	Acted	1,104	12

Table 1. We randomly sampled 1% of the data from four commonly used datasets in the SER area and re-labeled the sample using the same approach used by crowdsourcing workers that annotated CMU-MOSEI dataset.

spontaneous and posed facial expressions in natural settings. The videos have been captured under uncontrolled conditions with diverse variations in facial poses, head motions, and illumination. The dataset includes the same labels as CMU-MOSEI plus ”neutral”.

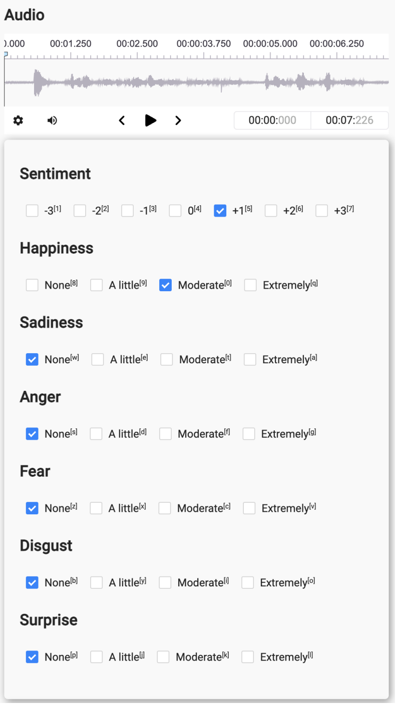
- **CREMA-D** [11]: Crowd-sourced Emotional Multimodal Actors Dataset consists of audio and video recordings of 91 professional actors (48 men and 43 women) performing 12 sentences and 12 words, each with different emotions including anger, disgust, fear, happiness, neutral, sadness, and surprise.

- **RAVDESS** [12]: The Ryerson Audio-Visual Database of Emotional Speech and Song is a multimodal database of emotional speech and song. The dataset consists of audio and video recordings of 24 professional actors, who were asked to perform a variety of emotional expressions, including calm, happy, sad, angry, fearful, surprise, and disgust. The audio recordings include speech, as well as sung performances, while the video recordings include both frontal and profile views of the actors' faces. The dataset includes annotations for the following six emotions: anger, calm, disgust, fear, happiness, sadness and surprise. We ignored the samples for “calm” because none of the other datasets had this label. We also only use speech samples and not song samples.

2.2. Data Labeling

A cornerstone of this work involves relabeling previously annotated data from four datasets discussed in Section 2.1. To better utilize the varying levels of annotations provided by CMU-MOSEI, we conducted a smaller-scale experiment. For this, we employed LabelStudio v1.7.1 [†] to relabel a random sample of 1% of the audio files. We followed the same labeling procedure as CMU-MOSEI [14] and used the same six basic emotions: anger, disgust, fear, happiness, sadness, and surprise.

Additionally, we annotated 1% of the other datasets used in this study, namely Aff-Wild2, CREAM-D, and RAVDESS. We also annotated the estimated age group in 9 categories (teens, 20s, 30s, 40s, 50s, 60s, 70s, 80s, 90s) and the estimated gender expression of the speaker in the audio in two options (male/female). These additional tasks were conducted to compare human performance on similar annotations.



Sentiment

□ -3[1] □ -2[2] □ -1[3] □ 0[4] ☑ +1[5] □ +2[6] □ +3[7]

Happiness

□ None[8] □ A little[9] ☑ Moderate[0] □ Extremely[q]

Sadiness

☑ None[w] □ A little[e] □ Moderate[t] □ Extremely[a]

Anger

☑ None[s] □ A little[d] □ Moderate[f] □ Extremely[g]

Fear

☑ None[z] □ A little[x] □ Moderate[c] □ Extremely[v]

Disgust

☑ None[b] □ A little[y] □ Moderate[i] □ Extremely[o]

Surprise

☑ None[p] □ A little[j] □ Moderate[k] □ Extremely[l]

Figure 1. Label studio screenshot of the system used to collect labels from the audio files. The same categories used in the original CMU-MOSEI dataset were used here.

To create our annotations, three authors of this paper independently annotated each sampled audio file without time limits or other constraints. Note that all three annotators hailed from different cultures, countries and native languages. The labels were then mapped to an integer value as follows: *None* → 0, *A little* → 1, *Moderate* → 2, *Extremely* → 3, as done in the original CMU-MOSEI dataset [14]. The final label for a given sample is given by the average of the 3 annotations and rescaled from 0 to 1. The possible labels for any sample in both our annotations and the original ones range from 0 to 1 in steps of 0.11.

2.3. Evaluation Metrics

Following the CMU-MOSEI's original data collection [18], we calculated the inter-rating agreement among our 3 independent assessors using Krippendorff's α coefficient using the *interval data function* as the difference function, except for age, in which we used the nominal function [34]. Krippendorff's $\alpha = 1$ indicates a perfect agreement, $\alpha = 0$ indicates the complete absence of agreement

[†] https://labelstud.io/

and $\alpha < 0$ indicates that disagreements are systematic and exceed what can be expected by chance [34].

After averaging the scores of each assessor, each sample has a float value from 0 to 1, which we could interpret as the intensity of that given emotion, as assessed by the annotators. We calculated the Pearson correlation (r) between our average and those from the original CMU-MOSEI dataset for each emotion separately. Pearson's r ranges from $r = -1$ (total negative correalation) to $r = +1$ (total positive correlation).

For the emotion classification task, we needed a threshold to binarize the annotation for each emotion. While the usual choice is $t = 0.01$ (e.g., [18]), we measured the agreement among the two groups of annotators for different thresholds using the Cohen's κ coefficient [43]. In particular, we examined 3 thresholds: low ($t = 0.01$), best ($t = 0.23$), and high ($t = 40$) and visualize their results with a multi-label confusion matrix (MLCM) [44]. The Cohen's κ coefficient can also range from -1 to +1, and its usual interpretation is that values ≤ 0 indicates no agreement, $.01 - .20$ for none to slight, $.21–.40$ for fair, $.41–.60$ for moderate, $.61–.80$ for substantial, and $\geq .81$ for almost perfect agreement [43].

We report weighted and unweighted (macro) average F_1 scores as well as the Matthews Correlation Coefficient (MCC), which is a suitable metric for unbalanced datasets, such as the ones studied in this work [45]. MCC's interpretation is the same as Perason's r interpretation. Finally, whenever possible, we report the average and the 95% confidence interval around the mean (i.e., value $\pm$ .95 ci).

2.4. Machine Learning Experiments

We conducted our experiments with state-of-the-art techniques based on foundation models for ASR [46]. Similar to the work of Wagner et al. [6], we used a straightforward multi-head architecture on top of a HuBERT-base transform-based backbone [4]. We freeze the convolutional layers and only fine-tune the weights in the transformer part of the backbone during training. We added a two layer residual feed-forward network to map the encoding generated by HuBERT-base to emotion logits, after averaging the encodings over the sequence dimension. Our preliminary experiments showed that this additional capacity helped improve performance over simply mapping the sequence averaged encoding directly to emotion logits. Our architecture is shown in Figure 2.

We binarized the labels per emotion using a pre-defined threshold th and removed samples that did not have any true emotion after applying the binary threshold. For the first part of our experiments, we only fine-tuned our models on CMU-MOSEI, using the sub-sampled training and validation sets. We then evaluate each model on the CMU-MOSEI test set, and complete CREMA-D, RAVDESS and Aff-Wild2 datasets.

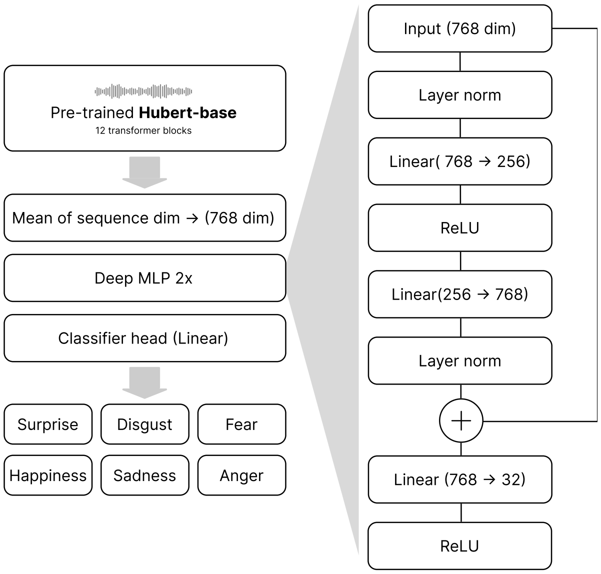


Figure 2. Model architecture used for all experiments. After encoding a waveform sampled at 16 KHz with a pretrained HuBERT-base backbone, we mapped the encoded data to emotion logits through a two layer, residual feed-forward network.

For the second part of our experiments, we study the degree of knowledge transfer between acted and in-the-wild datasets by first training on acted data and testing on in-the-wild, followed by training on in-the-wild data and testing on acted data. We used the label threshold for CMU-MOSEI identified from our analysis of emotion annotation correlation, as previously stated.

In our experiments, we used a batch size of 60 samples, a learning rate of $1e-4$, and trained the models for 15 epochs, with early stopping criteria using binary F_1 and patience set to 5. We used the AdamW optimizer with Adam $\beta = (0.9, 0.99)$ and weight decay of $1e-5$. We did not use a learning rate scheduler. Finally, we used binary cross entropy as the loss function for each label and averaged the loss to get a scalar value.

3. Results

From the initial 347 audio samples annotated, we removed a total of 30 samples. These samples came from CMU-MOSEI (16 – 7% of the annotated files) and Aff-Wild2 (14 – 43%). We dropped these files from our analysis for the following reasons: multiple speakers talking simultaneously (7), audio containing a large portion of songs, music or video-game background (16), or synthetic voice (2). These samples may have enough visual information to allow for labeling from the original annotators but were not suitable for audio-only experiments. We used the remaining 317 samples to analyze agreement between our annotators, the annotation accuracy across labels by comparing to the original annotations, and between acted and in-the-wild datasets. This attempts to answer *RQ1* and *RQ2* in Section 3.1. We also used the annotations to answer *RQ-3* in Section 3.2 where we compare our annotaitons with the original set from CMU-MOSEI and derive a threshold for binarizing label values. Finally, we take our study of thresholds to answer *RQ4* and *RQ5* in Section 3.3 and assess the impact of our analysis on ML experiments.

3.1. Studying our inter-rater agreement

RQ1: Are there emotions that are easier to annotate than others?

To address this research question, we measure the level of agreement or correlation among the annotators. We look into Pearson's r correlation and Krippendorff'sα coefficient for that.

Tables 2, 3 and 4 illustrate our findings. Table 2 shows Pearson's r of the new annotations made by the authors of this paper. While Happiness ($r = .45$) and Anger ($r = .42$) have the highest correlations, Disgust ($r = .27$), with the lowest correlation, is not much lower, showing a generally good level of agreement among our annotators.

Comparing the average of our annotations for each label with the original annotations for CMU-MOSEI in Table 3, we find that Pearson's correlation can go as low as only $r = .05$ for Fear. Sadness ($r = .27$) and Surprise ($r = .23$) are in the middle range, while Disgust ($r = .46$), Happiness ($r = .50$), and Anger ($r = .51$) are the highest values. The average correlation across emotions is 0.34.

Table 4 shows a deeper picture of our annotations with Krippendorff's α disaggregated by emotions and dataset acquisition method (here, CMU-MOSEI, combined acted data, combined in-the-wild data and all datasets). With multiple factors in place, it is difficult to assert that some emotions are easier to annotate than others. For example, Fear has the highest agreement for the *acted* datasets ($\alpha = .41$) but the lowest for the *in-the-wild* ones ($\alpha = .09$). Similarly, Happiness seems easier to annotate in-the-wild ($\alpha = .39$) than when acted ($\alpha = .21$). Finally, the last column of Table 4 shows that annotating audio for emotions ($\alpha = .31$) is a comparatively much harder task than annotating it for gender ($\alpha = .96$) or age group ($\alpha = .52$).

RQ2: Is it easier to annotate acted data or in-the-wild data?

We measure how easy it is to annotate a dataset by how much agreement the annotators have. In Table 4, we show the combinations of datasets with respect to their acquisition mode (in-the-wild or acted). According to Krippendorff's α, the *acted* data ($\alpha = .30$) is slightly easier to annotated than *in-the-wild* data ($\alpha = .25$), but the difference is small.

	Happiness	Sadness	Anger	Surprise	Disgust	Fear	Average
Annots 1 and 2	.46	.35	.36	.27	.21	.30	.33 ± .09
Annots 1 and 3	.51	.33	.61	.27	.41	.34	.41 ± .13
Annots 2 and 3	.37	.25	.30	.36	.20	.43	.32 ± .09
Average	.45 ± .18	.31 ± .13	.42 ± .41	.30 ± .13	.27 ± .29	.36 ± .17	.35 ± .12

Table 2. Pairwise Pearson correlation (r) for the 317 audio samples annotated by our 3 annotators.

	Happiness	Sadness	Anger	Surprise	Disgust	Fear	Average
Ours Vs Original	.50	.27	.51	.23	.46	.05	.34 ± .19

Table 3. Pairwise Pearson correlation (r) for the 212 CMU-MOSEI annotations intersection between our annotations and the original ones from CMU-MOSEI [18]. Fear presented the lowest correlation when comparing our annotations with the original ones.

RQ1 and *RQ2* are also answered in Table 4. For example, Happiness has a higher agreement for in-the-wild ($\alpha = .39$) datasets than acted ($\alpha = .21$), while Fear is a higher agreement for acted datasets ($\alpha = .41$) than in-the-wild ($\alpha = .09$).

3.2. Identifying a better threshold for CMU-MOSEI using agreement between new and original annotations

RQ3: What is the appropriate threshold for the CMU-MOSEI dataset?

We make use of Cohen's κ agreement between the original CMU-MOSEI annotations and the new annotations created by us to find the optimal threshold for the CMU-MOSEI dataset. Figure 3 shows Cohen's κ agreement per emotion as the threshold th varies. The left panel shows th vs Cohen's κ annotations for CMU-MOSEI, and the right panel for all four datasets combined. In both cases, the highest κ averaged over emotions was found at $th = .23$.

We picked 3 thresholds from low to high, and to better understand what each choice of threshold implies, we plotted their multi-label confusion matrix in Figure 4 [44]. The trade-off between a lower ($th = .01$) and a higher (e.g., $th = .40$) threshold is the number of examples for each emotion that will be available versus the likelihood of these samples being correctly labeled. The left panel, showing results for ($th = .01$) in Figure 4, for instance, shows that Happiness is often misunderstood as Surprise by the different sets of assessors (first row). On the other extreme, the right most panel showing results for ($th = .4$) in Figure 4 indicates that many of the emotions are not detected by the our assessors (last column). The diagonal elements for ($th = .23$) are consistently higher than either of the other thresholds, without a reduction in the overall number of samples detected indicating that it is the best choice for threshold.

Emotions	CMU-MOSEI [18] (Original, n=20477)	CMU-MOSEI (n=212)	Acted Comb (n=87)	In-the-Wild Comb (n=230)	All Comb (n=317)
Happiness	0.41	0.36	0.21	0.39	0.39
Sadness	0.12	0.26	0.24	0.27	0.27
Anger	0.18	0.26	0.40	0.32	0.38
Surprise	0.09	0.12	0.33	0.22	0.29
Disgust	0.21	0.17	0.20	0.20	0.22
Fear	0.02	0.08	0.41	0.09	0.33
Average	0.17 ± 0.14	0.21 ± 0.11	0.30 ± 0.10	0.25 ± 0.11	0.31 ± 0.07
Gender	-	0.96	0.95	0.96	0.96
Age Group	-	0.57	0.33	0.58	0.52

Table 4. Categorical agreement per emotion state calculated using Krippendorff's α reliability coefficient [34]. The first column shows the values of the original CMU-MOSEI dataset [18], while the other columns show various slices of the data re-assessed in this work. The *Acted Comb.* column shows the combination samples from CREMA-D and RAVESS datasets, while *In-the-Wild Comb.* merges CMU-MOSEI and Aff-Wild2 datasets. We also compute α for gender (2 classes) and age group (9 classes).

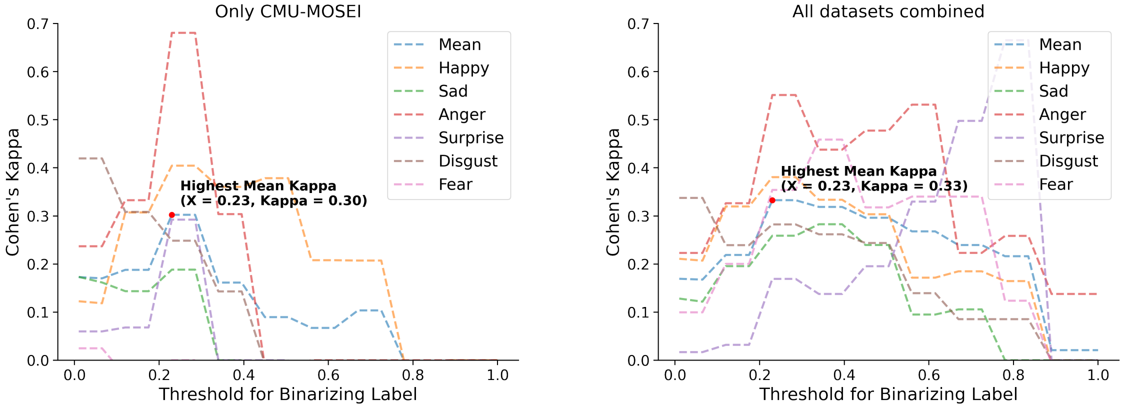


Figure 3. Cohen's κ agreement for all six emotions between the original annotations and our labels. The highest average Cohen's κ of 0.31, occurred at a CMU label threshold of 0.23. This threshold means that audio files that were labeled as weak by two of the three assessors (or medium by only one assessor) should receive the final label as FALSE for that emotion.

3.3. Impact on Deep Learning Experiments

RQ4: How does changing the threshold for the CMU-MOSEI dataset affect SER model performance?

To measure how different thresholds affect the performance of SER models, we investigate the same 3 thresholds studied in Section 3.2. In Table 5, we present the weighted average F_1, unweighted average (Macro) F_1 and the Matthew's correlation coefficient (MCC) averaged (± .95 confidence interval) over emotions for models trained using different thresholds and tested on datasets. With only two exceptions, the use of $th = .23$ reached the highest scores across the board. In particular, it seems to be a good threshold that generalizes well for other datasets, being the best threshold for all test cases including Aff-Wild2, CREMA-D and RAVDESS. The average increase in weighted F_1 was 7% across test datasets.

RQ5: Is there a transfer between acted and in-the-wild datasets?

Our final experiments consist of evaluating the transfer between acted and in-the-wild datasets. For that, we trained on the training sets of in-the-wild datasets (Aff-Wild2 and CMU-MOSEI with $th = .23$) and evaluate it on both acted (CREMA-D and RAVDESS combined) and the test sets of in-the-wild data with two different thresholds ($th = .01$ and $th = .23$). Table 6 shows the results of our experiments. The differences between each pair of experiments was surprisingly low, showing similar levels of information transfer between acted and in-the-wild datasets.

4. Discussion

In this work, we analyzed the strength, (or overall quality) of emotion annotations in acted and in-the-wild SER datasets. We manually re-annotated 1% of the data. However, one limitation of this is that our annotators only heard the audio, while the original annotators used video with audio and transcribed text which could have increased the

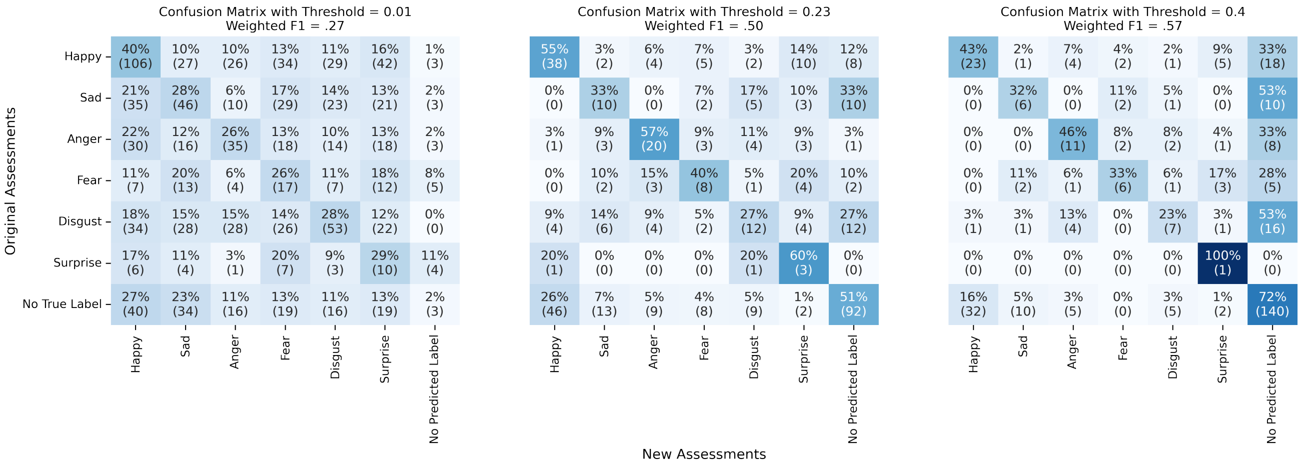


Figure 4. Multi label confusion matrix for the 3 thresholds studied in this work. Left panel, low $th = .01$; middle panel, best $th = .23$; right panel, high $th = .40$. In each cell, we have both the frequency (top) and the actual counts (bottom) of elements for that cell. The caption on each panel shows the weighted $F1$ score obtained from each confusion matrix.

Training on	Testing on					
CMU-MOSEI with...	CMU-MOSEI $Th = 0.01$	CMU-MOSEI $Th = 0.23$	CMU-MOSEI $Th = 0.40$	Whole Aff-Wild2	Whole CREMA-D	Whole RAVDESS
Evaluation Metric = Weighted F_1						
$Th = 0.01$	75.47 ± 8.99	81.31 ± 13.54	82.55 ± 14.25	80.76 ± 17.95	69.43 ± 8.45	72.73 ± 9.95
$Th = 0.23$	$\mathbf{76.36 \pm 9.48}$	$\mathbf{89.93 \pm 12.34}$	$\mathbf{92.23 \pm 13.03}$	$\mathbf{89.23 \pm 14.04}$	$\mathbf{78.79 \pm 5.28}$	$\mathbf{78.03 \pm 4.46}$
$Th = 0.40$	72.86 ± 14.18	82.20 ± 28.08	84.23 ± 30.89	82.13 ± 27.14	65.32 ± 24.80	63.78 ± 32.70
Evaluation Metric = Macro F_1						
$Th = 0.01$	$\mathbf{59.67 \pm 4.98}$	51.34 ± 3.35	47.11 ± 2.44	46.24 ± 6.49	53.02 ± 5.48	55.29 ± 8.26
$Th = 0.23$	53.37 ± 7.07	$\mathbf{56.19 \pm 5.56}$	$\mathbf{53.95 \pm 3.82}$	$\mathbf{50.71 \pm 2.61}$	$\mathbf{58.45 \pm 15.13}$	$\mathbf{57.41 \pm 13.59}$
$Th = 0.40$	49.88 ± 6.12	48.08 ± 8.68	46.30 ± 10.61	45.72 ± 7.44	43.49 ± 9.27	41.60 ± 13.81
Evaluation Metric = MCC						
$Th = 0.01$	$\mathbf{19.92 \pm 10.00}$	13.87 ± 9.50	10.34 ± 8.03	1.27 ± 2.84	12.92 ± 12.79	18.22 ± 15.11
$Th = 0.23$	13.54 ± 12.99	$\mathbf{14.38 \pm 13.14}$	$\mathbf{11.98 \pm 11.37}$	$\mathbf{3.69 \pm 6.01}$	$\mathbf{20.10 \pm 25.48}$	$\mathbf{18.62 \pm 22.00}$
$Th = 0.40$	7.69 ± 10.65	5.92 ± 7.73	4.95 ± 5.30	0.27 ± 1.71	2.37 ± 6.55	1.79 ± 4.40

Table 5. Results of training models using only CMU-MOSEI with different threshold and testing it with different datasets. Values are shown as average ± 95% confidence. Higher values per set are highlighted in bold.

quality of their annotations. Surprisingly, our results show that the average agreement between our annotators is higher than the average agreement among the original annotators of the CMU-MOSEI dataset. Second, certain labels are more easily annotated in one type of dataset versus the other. For example, Fear has the lowest inter-annotator agreement among both the original annotators of CMU-MOSEI and our annotations on in-the-wild data. Conversely, Happiness had lower agreement values for us on acted data than in-the-wild. Happiness and Anger are generally easier to identify and annotate. The average agreement across labels is similar for both acted and in-the-wild data. We believe that the number of samples we re-annotated is a limitation of our approach because minority emotions like Surprise and certain datasets like RAVDESS ($n = 12$) were poorly represented. We encourage further annotation of such benchmark datasets by the wider community to obtain more reliable, validated samples. Third, correlation analysis of the new annotations revealed a better value for thresholding the soft labels in CMU-MOSEI. The new threshold improved the performance of a speech emotion classification model both on the CMU-MOSEI test set, as well generalization on the other three datasets. Note that, simply increasing the threshold actually reduced performance. Increasing the threshold may be akin to dataset pruning [47], which has been shown to improve model performance. We hypothesize that there may be more optimal threshold values that can be estimated with greater number of annotated samples and more annotators, or simply by running several experiments with more fine-grained thresholds. Finally, we report

Training	Testing		
	Acted	In-the-Wild With $Th = 0.01$	In-the-Wild With $Th = 0.23$
Evaluation Metric = Weighted F_1			
In-the-Wild	76.06 ± 3.45	$\mathbf{79.80 \pm 9.73}$	89.65 ± 11.15
Acted	$\mathbf{77.73 \pm 3.08}$	79.56 ± 10.44	$\mathbf{89.93 \pm 12.77}$
Evaluation Metric = Macro F_1			
In-the-Wild	54.55 ± 11.29	$\mathbf{54.90 \pm 7.57}$	54.58 ± 4.73
Acted	$\mathbf{56.87 \pm 11.80}$	53.70 ± 6.41	$\mathbf{54.88 \pm 4.42}$
Evaluation Metric = MCC			
In-the-Wild	15.22 ± 21.31	$\mathbf{12.48 \pm 13.92}$	10.58 ± 10.70
Acted	$\mathbf{17.13 \pm 18.86}$	12.07 ± 11.63	$\mathbf{11.42 \pm 10.75}$

Table 6. Study on learning transfer from in-the-wild and acted datasets. In-the-wild datasets are Aff-Wild2 and CMU-MOSEI with $th = .23$. Acted datasets are CREMA-D and RAVDESS. Values are shown as average ± 95% confidence. Higher values per set are highlighted in bold.

that acted and in-the-wild datasets show similar levels of information transfer for this discrete emotion classification task. This comes as a surprise because the hypothesis is that in-the-wild datasets have a greater assortment of recording qualities, ethnicities and speaker identities. However, this result does agree with the fact that the average agreement of our annotators was similar for acted and in-the-wild data.

Note that, the fine grained annotations of CMU-MOSEI may be used in a regression setting, using the concordance correlation coefficient as a loss (or mean squared error or similar). However, we believe our argument regarding emotion label strength would matter nonetheless because it leads to a smaller, cleaner dataset.

Finally, our annotations, together with the architecture used in this paper, are publicly available at https://github.com/earkick/abaw2023.

5. Conclusion

Our analysis of emotion annotation strength and its impact on subjective agreement and SER model performance is an important step toward a better understanding of how we can utilize both acted and in-the-wild SER data. The community needs even larger SER and multimodal emotion datasets that can be validated in an open manner by the public, for example, as done by the CommonVoice project by the Mozilla Foundation [48]. We believe that closely examining research datasets for label uncertainty as well as how datasets relate to each other are fruitful areas for further work aligning with data-centric AI approaches.

References

[1] J. de Lope and M. Graña, "An ongoing review of speech emotion recognition," *Neurocomputing*, 2023. 1

[2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, 2018. 1, 2

[3] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, 2019. 1

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," 2021. 1, 2, 4

[5] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. 1, 2

[6] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," 2022. 1, 2, 4

[7] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, 1992. 1

[8] D. Keltner, J. L. Tracy, D. Sauter, and A. Cowen, "What basic emotion theory really says for the twenty-first century study of emotion," *Journal of nonverbal behavior*, 2019. 1

[9] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, 1977. 1

[10] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, 2008. 1, 2

[11] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, 2014. 1, 3

[12] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, 2018. 1, 3

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, *et al.*, "A database of german emotional speech.," in *Interspeech*, 2005. 1

[14] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *ACL*, 2018. 1, 2, 3

[15] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," in *British Machine Vision Conference BMVC*, 2019. 1, 2, 3

[16] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, 2019. 1

[17] Z. Tang, J. Cho, Y. Nie, and M. Bansal, "Tvlt: Textless vision-language transformer," in *NeurIPS*, 2022. 1, 2

[18] P. P. Liang, R. Salakhutdinov, and L.-P. Morency, "Computational modeling of human multimodal language: The MOSEI dataset and interpretable dynamic fusion," in *First Workshop and Grand Challenge on Computational Modeling of Human Multimodal Language*, 2018. 1, 2, 3, 4, 5, 6

[19] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," 2021. 2

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017. 2

[21] S. Srinivasan, Z. Huang, and K. Kirchhoff, "Representation learning through cross-modal conditional teacher-student training for speech emotion recognition," 2022. 2

[22] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 finetuning for improved speech emotion recognition," 2023. 2

[23] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *EMNLP*, 2022. 2

[24] R. L. C. Mitchell and E. D. Ross, "Attitudinal prosody: what we know and directions for future study," *Neurosci. Biobehav. Rev.*, 2013. 2

[25] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner, "The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures," *Nat. Hum. Behav.*, 2019. 2

[26] D. T. Cordaro, D. Keltner, S. Tshering, D. Wangchuk, and L. M. Flynn, "The voice conveys emotion in ten globalized cultures and one remote village in bhutan," *Emotion*, 2016. 2

[27] A. S. Cowen, H. A. Elfenbein, P. Laukka, and D. Keltner, "Mapping 24 emotions conveyed by brief human vocalization," *Am. Psychol.*, 2019. 2

[28] J. C. Jackson, J. Watts, T. R. Henry, J.-M. List, R. Forkel, P. J. Mucha, S. J. Greenhill, R. D. Gray, and K. A. Lindquist, "Emotion semantics show both cultural variation and universal structure," *Science*, 2019. 2

[29] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," *J. Pers. Soc. Psychol.*, 1994. 2

[30] P. Laukka, H. A. Elfenbein, N. S. Thingujam, T. Rockstuhl, F. K. Iraki, W. Chui, and J. Althoff, "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features," *J. Pers. Soc. Psychol.*, 2016. 2

[31] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proceedings of the National Academy of Sciences*, 2010. 2

[32] R. Cowie, "Perceiving emotion: towards a realistic understanding of the task," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 2009. 2

[33] J. Wilting, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech.," in *Interspeech*, vol. 2006, p. 9th, 2006. 2

[34] K. Krippendorff, "Computing Krippendorff's alpha-reliability," *University of Pennsylvania*, 2011. 2, 3, 4, 6

[35] D. Kollias, P. Tzirakis, A. Baird, A. Cowen, and S. Zafeiriou, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges," *arXiv preprint arXiv:2303.01498*, 2023. 2

[36] D. Kollias, "Abaw: learning from synthetic data & multi-task learning challenges," in *European Conference on Computer Vision*, Springer, 2023. 2

[37] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Distribution matching for heterogeneous multi-task learning: a large-scale face study," *arXiv preprint arXiv:2105.03790*, 2021. 2

[38] D. Kollias and S. Zafeiriou, "Analysing affective behavior in the second abaw2 competition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[39] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, "Analysing affective behavior in the first abaw 2020 competition," in *15th IEEE International Conference on Automatic Face and Gesture Recognition*, 2020. 2

[40] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Face behavior a la carte: Expressions, affect and action units in a single network," *arXiv preprint arXiv:1910.11111*, 2019. 2

[41] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: Valence and arousal 'in-the-wild'challenge," in *Computer Vision and Pattern Recognition Workshops*, IEEE, 2017. 2

[42] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *IJCV*, 2019. 2

[43] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, 2012. 4

[44] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-label confusion matrix," *IEEE Access*, 2022. 4, 5

[45] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, 2020. 4

[46] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021. 4

[47] B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. S. Morcos, "Beyond neural scaling laws: beating power law scaling via data pruning," 2022. 7

[48] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020. 8