Exploiting Co-occurrence Frequency of Emotions in Perceptual Evaluations To Train A Speech Emotion Classifier

Huang-Cheng Chou[1,2], *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
[2]Department of Electrical Engineering, National Tsing Hua University, Taiwan

Huang-Cheng.Chou@UTDallas.edu, cclee@ee.nthu.edu.tw, busso@UTDallas.edu

Abstract

Previous studies on *speech emotion recognition* (SER) with categorical emotions have often formulated the task as a single-label classification problem, where 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). Some studies have regarded SER problems as a multi-label task, predicting multiple emotional classes. However, this formulation does not leverage the relation between emotions during training, since emotions are assumed to be independent. This study explores the idea that emotional classes are not necessarily independent and its implications on training SER models. In particular, we calculate the frequency of co-occurring emotions from perceptual evaluations in the train set to generate a matrix with class-dependent penalties, punishing more mistakes between distant emotional classes. We integrate the penalization matrix into three existing label-learning approaches (hard-label, multi-label, and distribution-label learning) using the proposed modified loss. We train SER models using the penalty loss and commonly used cost functions for SER tasks. The evaluation of our proposed penalization matrix on the MSP-Podcast corpus shows important relative improvements in macro F1-score for hard-label learning (17.12%), multi-label learning (12.79%), and distribution-label learning (25.8%).

Index Terms: Speech emotion recognition, emotion co-occurrence, multi-class classification, multi-label classification.

1. Introduction

Speech emotion recognition (SER) plays an essential role in human-centered computer interaction. Speech is one of the most convenient modalities to recognize human emotions, given the ubiquity of speech-based interfaces. 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 common 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 [3–7]. This methodology ignores the chance of having co-occurring emotions, which is quite common with emotional behaviors (e.g., a sentence conveying a mixture of happiness and surprise). Recently, a new multi-label emotion classification formulation has allowed a system to predict multiple co-occurring emotions [8–11]. While multi-label learning allows ground truth with multiple valid classes, this formulation does not model the relation between emotions, assuming that they are independent. In this paper, we hypothesize that emotions are correlated and that exploiting the frequency of co-occurring emotions during perceptual evaluations during training can lead to more powerful SER systems

The observation of the co-occurrence of emotions in human interactions is common and natural [12]. Xu et al. [13] utilized the frequency of co-occurring emotions in perceptual evaluations to initialize the connection of emotions in their proposed graph-based *deep neural network* (DNN). Steidl et al. [14] use common confusions between emotions from evaluators to assess the performance of an SER system. Other studies have used soft labels to account for secondary emotions also conveyed in the speech [15–17]. These studies have shown the importance of considering all the evaluators included in the perceptual evaluation, even if they do not agree with the consensus labels.

This study explores the use of co-occurrence of emotions during the training process of an SER system. We calculate the co-occurrence statistic matrix to capture the relations between emotions selected during perceptual evaluations. We transform this matrix into a penalization matrix impacting the objective functions by penalizing predictions of infrequent co-occurring emotions. Our implementation integrates the penalization matrix into the existing cost functions as a "penalty loss" to produce a higher loss value when the model predicts infrequent co-occurring emotions. For example, anger and contempt do not often occur in perceptual evaluations so a joint prediction of these emotions will be penalized more than a prediction of common co-occurring emotions such as anger and sadness. This approach does not regard emotions as independent, leveraging instead the relationship between emotions.

To demonstrate the benefits of using the "penalty loss", we use the *RNN-AttenVec* chunk-level attention model proposed by Lin and Busso [18] as our SER model using three existing training strategies with and without the proposed "penalty loss." The experimental results on the MSP-podcast corpus [19] demonstrate consistent improvement in performance with the proposed "penalty loss," leading to important relative improvement in macro F1-score using hard-label learning (17.12%), multi-label learning (12.79%), and distribution-label learning (25.8%). Our main contributions are:

- Utilizing for the first time the prior knowledge of co-occurrence of emotions to train a SEC model.
- Proposing an elegant implementation to incorporate the "penalty loss" in the model, which is flexible and can be easily applied to any emotion classification framework using existing label learning approaches.

2. Background

This study investigates how to exploit the frequency of co-occurring emotions in perceptual evaluations to train a SER system. We discuss relevant literature on hard-label learning, multi-label learning, and distribution-label learning for emotion classification. As an illustration, we consider a four-class emotion classification task (N: neutral, A: anger, S: sadness, and H:

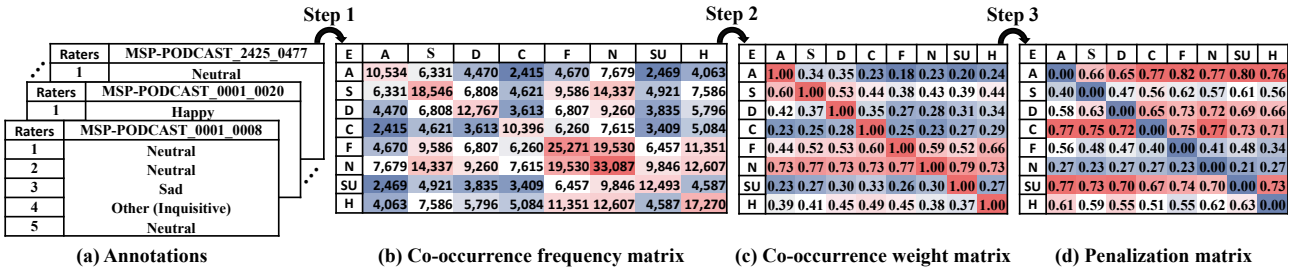

Figure 1: *The figure shows the procedure to create the penalization matrix. Primary emotion includes neutral (N), anger (A), sadness (S), happiness (H), surprise (SU), fear (F), disgust (D), and contempt (C). The process is explained in Section 3.1.*

happiness), where a sentence is annotated by five annotators, each contributing one label. An example of the labels for one sample is "S, S, S, N, N."

2.1. Hard-label Learning for Emotion Recognition

When emotional databases are annotated with perceptual evaluations, most SER studies rely on consensus labels as ground truth obtained by aggregating individual perceptual emotional annotations with rules such as the majority vote or the plurality rule. This strategy generates a one-hot vector with a single emotional class (hard-label), ignoring secondary emotions marked by annotators that disagreed with the consensus labels. For the example, the hard label is (0, 0, 1, 0). Some studies had proposed soft-labels [15–17], allowing a sample to have more than one emotion. Chou et al. [20] modeled each annotator to embrace the subjectivity in the perceptual evaluations. These approaches can use sentences in the train set, even if the evaluators do not reach an agreement. Hard-label learning regards SER as a single-label task, where emotions are considered independent, and co-occurrence of emotions is impossible. We still show improved results when using the proposed "penalty loss."

2.2. Multi-label Learning for Emotion Recognition

Multi-label learning can train emotion classifiers to recognize multiple emotions per data sample. The ground truth is a "multiple-hot" vector, where more than one emotion can be selected. For the example, the multi-label is (1, 0, 1, 0). Examples of multi-label learning approaches in emotion recognition include the work of Zhang et al. [9, 10], and Ju et al. [11]. While the multi-label learning approach is defined as a multi-label task where multiple emotions are possible, this formulation still assumes that emotions are independent. Furthermore, the conventional multi-label learning approach cannot determine if some emotions might be more dominant than others (e.g., primary versus secondary emotions). We follow the approach proposed by Kang et al. [8] to train models using multi-label learning to demonstrate the improvements of SER models using the proposed "penalty loss."

2.3. Distribution-label Learning for Emotion Recognition

Distribution-label learning [21] aims to predict a distribution as the output. It assumes that the label is distributional, and the sum of the probabilities across classes is 1. For the example, the distribution-label is (0.4, 0.0, 0.6, 0.0). The training process minimizes the distribution distance between the ground truth and the predicted distributions. Examples of distribution-label learning approaches in SER include the work of Chou et al. [22] and Wu et al. [23]. We follow the approach proposed by Chou et al. [22], which used distribution-label learning for SER to convert perceptual evaluations into the distributional label for training our SER model. The objective function of the conventional distribution-label approach is the *Kullback-Leibler divergence* (KLD). We investigate the classification results obtained when considering the proposed "penalty loss."

2.4. Co-occurrence of Emotions

While distribution-label learning and multi-label learning can predict more than one emotion per sample, these formulations do not directly learn the relation of co-occurring emotions (e.g., the probability of co-occurrence of anger and sadness is higher than the probability of co-occurrence of anger and happiness). Alhuzali et al. [24] proposed the *label-correlation aware* (LCA) loss to reduce the distance between co-occurring emotions and increase the distance between non-co-occurring emotions on a multi-label text emotion classification task. The LCA loss value should increase when the model predicts non co-occurring emotions. Deng et al. [25] proposed a multi-label focal loss to maximize the difference between positive emotions (e.g., "Love", "Joy") and negative emotions ("Anger", "Hate"). However, they did not consider the co-occurrence of positive and negative emotions, since it is possible that people feel at the same time negative and positive emotions [26]. The most relevant study to this paper is the work of Xu et al. [13], which leveraged prior knowledge about the co-occurrence of emotions to initialize the connections between emotional nodes in their graph-based neural network. Inspired by these studies, we utilize the co-occurrence of emotions to generate a penalty weight from the perceptual evaluations of the training set, which we use so that our model understands the relations between emotions.

3. Methodology

This sections describes our approach, which is implemented with as an eight-class problem, using the primary emotions from the MSP-Podcast corpus (Sec. 4.1).

3.1. Emotion Co-occurrence Penalty Weights

Figure 1 shows the three steps to generate the penalization matrix. In step 1, we follow the process proposed by Xu et al. [13] to calculate the co-occurrence matrix from the perceptual evaluations in the train set. This matrix has the frequency of co-occurring emotions, creating an 8×8 symmetric matrix. For instance, the matrix in Figure 1 (b) for ("A", "A") is 10,534, indicating that *anger* was selected in 10,534 speaking turns. In 7,679 of these sentences, an evaluator also assigned the label *neutral*. Therefore, the position ("A","N") and ("N","A") in the matrix is 7,679. In step 2, we obtain a probability-like co-occurrence matrix by dividing the numbers of each column by the number of samples annotated with the emotion associated with that column. For instance, consider the first column ("A") in Figure 1 (b). The frequency that *anger* co-occurred with other emotions are 10,534, 6,331, 4,470, 2,415, 4,670, 7,679, 2,469, and 4,063. These numbers are divided by the number of

sentences annotated with *anger* (10,534) to get the probabilities of co-occurrence of emotions: 1.00, 0.60, 0.42, 0.23, 0.44, 0.73, 0.23, 0.39. These numbers show, for example, that the co-occurrence probability of anger and sadness (0.60) is higher than anger and contempt (0.23). We refer to this matrix as the co-occurrence weight matrix (Fig. 1 (c)), which is no longer symmetric given the column-wise normalization. In step 3, we transform this matrix into a "penalization matrix" to punish the SER models if the models predict infrequent co-occurring emotions. The transformation is simple – we subtract each element in the co-occurrence weight matrix from one to get the penalization matrix (Fig. 1 (d)). This approach increases the loss if the model predicts the co-occurrence of distant emotions.

3.2. Learning Label Processing

We apply the penalization matrix to the three existing label learning approaches discussed in Section 2: hard-label learning, multi-label learning, and distribution-label learning. These three learning methods use different cost functions. We use *cross-entropy* (CE) for hard-label learning, *binary cross-entropy* (BCE) for multi-label learning and *Kullback–Leibler divergence* (KLD) for distribution-label learning. Besides, we use the label smoothing strategy proposed by Szegedy et al. [27] to smooth the ground truth vector of the hard-label and distribution-label using the smoothing parameter 0.05, where a small probability is given to emotional classes with zero value. For the multi-label ground truth vector, we add a small value where the value of the vector is zero (i.e., 10^{-6}).

3.3. Penalized Objective Functions

We introduce an elegant derivation to apply the penalization matrix in the objective functions. We define the $N \times K$ matrices for the ground truth ($\boldsymbol{Y^T}$) and model prediction ($\boldsymbol{Y^P}$), where N is the number of sentences in the train set, and the K is the number of emotional classes in the task. The values in each row of $\boldsymbol{Y^T}$ depend on whether we are using hard-label, multi-label, or distribution-label learning. We estimate the loss value matrix ($\boldsymbol{\mathcal{L}} \in \mathbb{R}^{N \times K}$) as follow:

$$\begin{aligned}\boldsymbol{\mathcal{L}} &= f_{loss}(\boldsymbol{Y^T}, \boldsymbol{Y^P}) \\ &= \begin{bmatrix} f_{loss}(Y_{11}^T, Y_{11}^P) & \cdots & f_{loss}(Y_{1j}^T, Y_{1j}^P) \\ \vdots & \ddots & \vdots \\ f_{loss}(Y_{i1}^T, Y_{i1}^P) & \cdots & f_{loss}(Y_{ij}^T, Y_{ij}^P) \end{bmatrix}_{N \times K},\end{aligned} \tag{1}$$

where f_{loss} is the objective function for the task (e.g., BCE, CE, or KLD). The entries of $\boldsymbol{\mathcal{L}}$ are $f_{loss}(Y_{ij}^T, Y_{ij}^P)$, with $i \in \{1 \dots N\}$ and $j \in \{1 \dots K\}$, which calculates the loss for each emotion and for each sentence. Then, we apply the penalization matrix to Equation 1. We denote the penalization matrix as $\boldsymbol{P} \in \mathbb{R}^{K \times K}$ (Fig. 1 (d)). We propose the penalty loss (*PL*) as follows:

$$\begin{aligned}PL &= \textstyle\sum_{i=1}^{N}(\boldsymbol{\mathcal{L}}_i \cdot \boldsymbol{P}) \\ &= \textstyle\sum_{i=1}^{N}(\sum_{j=1}^{K}\sum_{z=1}^{K} P_{jz} \cdot f_{loss}(Y_{ij}^T, Y_{ij}^P)).\end{aligned} \tag{2}$$

When we replace f_{loss} with the CE, BCE, and KLD loss functions, the equations become:

$$PL^{CE} = -\sum_{i=1}^{N}(\sum_{j=1}^{K}\sum_{z=1}^{K} P_{jz} \cdot Y_{ij}^T \cdot \log(Y_{ij}^P)), \tag{3}$$

$$\begin{aligned}PL^{BCE} = -\sum_{i=1}^{N}(\sum_{j=1}^{K}\sum_{z=1}^{K}(P_{jz} \cdot Y_{ij}^T \cdot \log(Y_{ij}^P) \\ + P_{jz} \cdot (1 - Y_{ij}^T) \cdot \log(1 - Y_{ij}^P))),\end{aligned} \tag{4}$$

$$PL^{KLD} = \sum_{i=1}^{N}(\sum_{j=1}^{K}\sum_{z=1}^{K} P_{jz} \cdot Y_{ij}^T \cdot \log(\frac{Y_{ij}^T}{Y_{ij}^P})). \tag{5}$$

Our penalty loss is added to the original loss for the task during training. If *L* represents the CE, BCE, or KLD loss, the total loss is defined in Equation 6, where $\alpha \in \mathbb{R}$ (e.g., 0.5, 1.0) and β is either 1 or 0.

$$\mathscr{L}^{Total} = \beta L + \alpha \cdot PL. \tag{6}$$

4. Experimental Settings

4.1. The MSP-Podcast Corpus

We use the MSP-Podcast corpus [19] to evaluate our proposed method. The collection of the corpus is an ongoing effort, where we use release 1.9. There are 55,283 speech segments in the train set, 9,546 speech segments in the development set, and 16,570 speech segments in the test set. The recordings come from spontaneous, real-world podcasts available on audio-sharing websites. The protocol of the data collection, including the pipeline to identify segments is described in detail in Lotfian and Busso [19]. The emotional annotations of the database are obtained with a crowdsourcing protocol that tracks the quality of the evaluations in real-time [28]. The perceptual evaluation includes primary emotions, which consist of eight classes: *neutral* (N), *anger* (A), *sadness* (S), *happiness* (H), *surprise* (SU), *fear* (F), *disgust* (D), and *contempt* (C). Annotators need to select one of the options. If none of these emotions is suitable for describing the speech segment's emotional content, annotators can select the "other" and provide the emotion. At least five annotators annotate each speaking turn and the consensus label is obtained with the plurality rule. This paper ignores annotations with the label "other," forming an eight-class classification task. We only use speaking turns with a consensus label for hard-label learning, discarding cases without agreement. The annotations also include secondary emotions and emotional attributes, which are not used in this study.

4.2. Acoustic Features

Keesing et al. [29] investigated the most effective features for SER by exploring many existing feature sets on various public emotional datasets. Their study concluded that the wav2vec feature set [30] is one of the most compelling feature extraction approaches. Therefore, we extract the 512-dimensional wav2vec feature as the input of our models. Before training the models, we apply the z-normalization function to normalize all the features by the mean and standard deviation, estimating these parameters on the train set.

4.3. Implementation Details

We used the chunk-level emotion recognition modeling methodology proposed by Lin and Busso [18] as our SER model. This framework deals in a principled way with sentences of different duration. It transforms a sentence with an arbitrary duration into a fixed number of chunks with a fixed duration by changing the overlap between the chunks. We follow the suggestion of the paper to use *long short-term memory* (LSTM) as the chunk-level feature encoder equipped with the *RNN-AttenVec* chunk-level attention model [18]. This combination learns emotional-relevant information at the frame-level, chunk-level, and sentence-level from the input features. We train and run the evaluation on Tensorflow 2.0 [31], using a NVIDIA GeForce RTX 3090. Lin and Busso [18] provides

more details about the network architecture. The details about the model parameters are the same as the ones used by Chou et al. [22]. Following insights from previous studies, we use the softmax function as the activation function of the output layer for CE [3, 5, 7] and KLD [21, 22], and the sigmoid function as the activation function for BCE [8, 32]. 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, we set the value of α in Equation 6 to either 0.5 or 1.0. We also run the experiments using only PL (without L; $\alpha > 0$; $\beta = 0$) or only L (without PL; $\alpha = 0$; $\beta = 1$)).

5. Evaluation

We follow the approach used in Lin and Busso [18] that randomly splits the original test set into 30 small subsets with similar size, reporting the average results in each evaluation metric. This approach allows us to conduct a two-sample, two-tailed t-test to evaluate the statistical significance of all results between the proposed method and baselines. If the p-value is less than 0.05, the result is considered statistically significant.

5.1. Evaluation Metric

We use multiple evaluation metrics to compare the predicted labels with the ground truth according to the different label learning methods. For the hard-label learning, we measure classification performance with *unweighted average recall* (UAR), *unweighted average precision* (UAP), *macro F1-score* (maF1), *micro F1-score* (miF1), and *weighted F1-score* (weF1). For multi-label learning and distribution-label learning, we use the metrics used in Fei et al. [32]: *hamming loss* (HL), *ranking loss* (RL), *coverage error* (COVE), and maF1. We also add miF1 and weF1 to evaluate multi-label classification performance. The predictions for multi-label and distribution-label learning methods have to be binarized for estimating performance. For multi-label learning, we use the threshold 0.5 to convert the prediction probabilities into "multiple-hot" binary vectors [8, 32]. For distribution-label learning, we adopt the value used by Chou et al. [22] setting the threshold to 1/8 to convert the prediction probabilities into the binary vectors.

5.2. Experimental Results

Table 1 shows the overall classification performance over different settings. For the metrics, we add the symbol $\uparrow$ to show that the result is better when the value is higher, and the symbol $\downarrow$ otherwise. The symbol $*$ shows that the results are statistically better than the baseline model trained without the proposed penalty loss (e.g., $\alpha = 0$ in Eq. 6). Table 1 has three main parts for the three-loss functions: CE (hard-label learning), BCE (multi-label learning), and KLD (distribution-label learning). The column for β indicates if the loss L is considered ($\beta = 1$) in the experiments or not ($\beta = 0$). The column α is the weight value for the PL loss. The bold numbers indicate the best performance for a given loss function.

Is the penalty loss (PL) useful for SER? Table 1 demonstrates that most of the best results are from models trained with the proposed loss (PL) over three different label learning approaches. For instance, for a single-label classification task, the model trained with PL when α equals 0.5 achieves the best performance in four out of the five evaluation metrics. This model achieves a 16.22% relative improvement in maF1 over the baseline. The model with PL ($\alpha = 1.0$) reaches the best results in four out of the six evaluation metrics for multi-label classification. This model also achieves 7.31% in relative improvement in maF1 over the baseline. While no model dominates the evaluation metrics on the distribution-label classification task, most of the best results are from the models using the proposed PL. If we focus on maF1, the best model using KLD achieves a 25.8% relative improvement over the baseline method. It also outperforms the state-of-the-art performance reported by Chou et al. [22] (31.6% in maF1). We conclude that the proposed penalty loss (PL) is indeed helpful to improve the model recognition performance on the primary emotion classification task. The table also shows that adding the main loss L to the proposed loss PL often improves performance.

Table 1: *Single-label and multi-label results for the eight-class SER task. The symbol $*$ indicates that the results of the models using the proposed penalization matrix are statistically significant over the baseline ($\beta = 1$;$\alpha = 0$).*

Single-label Classification							
Loss	β	α	**UAR** $\uparrow$	**UAP** $\uparrow$	**maF1** $\uparrow$	**miF1** $\uparrow$	**weF1** $\uparrow$
CE	**1**	**0**	0.144	0.133	0.111	0.424	0.318
	1	**0.5**	**0.156***	**0.137***	0.129*	**0.425**	**0.347**
	1	**1**	0.155*	0.136	**0.130***	0.408	0.346
	0	**1**	0.154*	0.136	0.128*	0.396	0.341
Muli-label Classification							
Loss	β	α	**HL** $\downarrow$	**RL** $\downarrow$	**COVE**$\downarrow$	**maF1** $\uparrow$	**miF1** $\uparrow$	**weF1** $\uparrow$
BCE	**1**	**0**	0.304	0.603	6.899	0.219	0.466	0.352
	1	**0.5**	0.303	0.608	6.928	0.215	0.462	0.348
	1	**1**	**0.303**	**0.587**	**6.837**	0.235	**0.482**	0.370
	0	**1**	0.305	0.597	6.871	**0.247***	0.477	**0.378**
Distribution-label Classification							
Loss	β	α	**HL** $\downarrow$	**RL** $\downarrow$	**COVE**$\downarrow$	**maF1** $\uparrow$	**miF1** $\uparrow$	**weF1** $\uparrow$
KLD	**1**	**0**	**0.294**	0.511	6.279	0.283	0.522	0.431
	1	**0.5**	0.308	**0.507**	6.220	0.322*	**0.533**	0.471*
	1	**1**	0.315	0.509	**6.214**	0.330*	0.532	0.475*
	0	**1**	0.337	0.530	6.284	**0.356***	0.526	**0.496***

We aim to understand whether the proposed method helps the model better capture the expected co-occurrence matrix. We use the Frobenius norm to evaluate the distance between the co-occurrence matrices obtained with the ground truth (train set) and the model predictions exclusively implemented with PL ($\alpha = 1$; $\beta = 0$) or L ($\alpha = 0$; $\beta = 1$). When using PL, the distance decreased from 4.27 to 4.00 for the multi-label approach and from 4.13 to 3.39 for the distribution-label approach. The co-occurrence matrix predicted with the proposed loss is closer to the target co-occurrence matrix.

6. Conclusions and Future Work

This study utilized the frequency of co-occurring emotions to generate a penalty weight to train an SER model. The proposed penalty loss considers the relationship between emotions, punishing more errors between distant emotions. We evaluated the proposed loss with three existing label learning approaches (hard-label learning, multi-label learning, and distribution-label learning). The results demonstrate that the proposed penalty loss improves performance for most evaluations on an eight-class primary emotion classification task. Our future work will explore other emotion corpora and framework (e.g., AuxFormer [33]) to validate the generalization of the proposed penalty loss.

7. Acknowledgements

This work was supported by the MOST under Grants 110-2917-I-007-016, 110-2221-E-007-067-MY3, and 110-2634-F-007-012, and the NSF under Grant CNS-2016719.

8. References

[1] A. 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, September 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] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, Brisbane, QLD, Australia, April 2015, pp. 4749–4753.

[4] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[5] Z. Aldeneh and E. Mower Provost, "Using regional saliency for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2741–2745.

[6] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*. Calgary, AB, Canada: IEEE, April 2018, pp. 5084–5088.

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

[8] X. Kang, X. Shi, Y. Wu, and F. Ren, "Active learning with complementary sampling for instructing class-biased multi-label text emotion classification," *IEEE Transactions on Affective Computing*, vol. Early Access, 2022.

[9] D. Zhang, X. Ju, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal multi-label emotion detection with modality and label dependence," in *Empirical Methods in Natural Language Processing (EMNLP 2020)*, Virtual Conference, November 2020, pp. 3584–3593.

[10] D. Zhang, X. Ju, W. Zhang, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing," in *AAAI Conference on Artificial Intelligence (AAAI 2021)*, vol. 35, Virtual Conference, February 2021, pp. 14 338–14 346.

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

[12] 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.

[13] P. Xu, Z. Liu, G. I. Winata, Z. Lin, and P. Fung, "EmoGraph: Capturing emotion correlations using graph networks," *ArXiv e-prints (arXiv:2008.09378)*, pp. 1–7, August 2020.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

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

[24] H. Alhuzali and S. Ananiadou, "SpanEmo: Casting multi-label emotion classification as span-prediction," in *European Chapter of the Association for Computational Linguistics (EACL 2021)*, vol. 1, Virtual conference, April 2021, p. 1573?1584.

[25] J. Deng and F. Ren, "Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning," *IEEE Transactions on Affective Computing*, vol. Early Access, 2022.

[26] G. Asmundson and J. Katz, "Understanding the co-occurrence of anxiety disorders and chronic pain: state-of-the-art," *Depression and Anxiety*, vol. 26, no. 10, pp. 888–901, October 2009.

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

[28] 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.

[29] 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.

[30] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 3465–3469.

[31] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Symposium on Operating Systems Design and Implementation (OSDI 2016)*, Savannah, GA, USA, November 2016, pp. 265–283.

[32] 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.

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