EXPLOITING ANNOTATORS' TYPED DESCRIPTION OF EMOTION PERCEPTION TO MAXIMIZE UTILIZATION OF RATINGS FOR SPEECH EMOTION RECOGNITION

Huang-Cheng Chou[1,2], *Wei-Cheng Lin*[1], *Chi-Chun Lee*[2], *Carlos Busso*[1]

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

ABSTRACT

The decision of ground truth for *speech emotion recognition* (SER) is still a critical issue in affective computing tasks. Previous studies on emotion recognition often rely on consensus labels after aggregating the classes selected by multiple annotators. It is common for a perceptual evaluation conducted to annotate emotional corpora to include the class "other," allowing the annotators the opportunity to describe the emotion with their own words. This practice provides valuable emotional information, which, however, is ignored in most emotion recognition studies. This paper utilizes easy-accessed natural language processing toolkits to mine the sentiment of these typed descriptions, enriching and maximizing the information obtained from the annotators. The polarity information is combined with primary and secondary annotations provided by individual evaluators under a label distribution framework, creating a complete representation of the emotional content of the spoken sentences. Finally, we train multitask learning SER models with existing learning methods (soft-label, multi-label, and distribution-label) to show the performance of the novel ground truth in the MSP-Podcast corpus.

Index Terms— Emotion recognition, Distribution-label learning, Soft-label learning, Multi-label learning

1. INTRODUCTION

Emotion recognition is receiving more attention in human-centric computer interaction. With the proliferation of ubiquitous speech-based interfaces, *speech emotion recognition* (SER) is an appealing modality to estimate emotion [1]. While many studies have proposed solutions for SER, there are still critical challenges to address. One of them is the ground truth labels used to train SER models. The most common approach to annotate an emotional corpus is with perceptual evaluations, where evaluators provide their judgments by completing questionaries with fixed options after listening or watching a stimulus. However, emotion perception is a subjective process, leading to essential disagreements in the labels [2–4]. Most previous research on *emotion classification* (EC) regards the disagreement as noise, calculating a consensus label relying on the *wisdom of the crowd*. However, the consensus labels hide relevant information in the annotations, ignoring the natural subjectivity and variability in human emotion perception. While some studies had employed soft-label/multi-label learning to learn the subjectivity and uncertainty in the labels [5–11], exploring further this area can lead to more robust SER models where all the data is used, even ambiguous samples without clear consensus labels.

Conducting perceptual evaluations with a fixed number of options can lead to bias in the labels [12], where a class may be the best option available, even when it does not describe well the emotional content in the stimulus. A common solution to attenuate this problem is to allow annotators to select "other," or "none of the above," providing the opportunity to describe the emotions with their own words. This approach has been used in several emotional corpora [3, 4, 13, 14]. The typed descriptions in the labels have never been used to enrich the emotional labels used in emotion recognition tasks. Instead, most prior studies directly discard the class "other." These typed descriptors provide valuable information to characterize the emotional content in the stimulus, and they should be used.

This paper proposes a novel method to maximize the utilization of emotional annotations by exploiting the typed descriptions included in the database. Our approach mines these typed descriptions using *natural language processing* (NLP) tools, transforming them into a three-class polarity vector (positive, negative, and ambiguous). We rely on the *linguistic inquiry and word count* (LIWC) 2015 toolkit [15] to understand the sentiment of the data samples. We evaluate several training strategies for SER using soft-label, multi-label, and distribution-label leanings with the polarity information. By fully utilizing the emotional annotations from multiple evaluators, we demonstrate clear improvements in performance by using polarity labels. We also showed that using distribution-label learning is more suitable when the classification task includes more classes.

2. BACKGROUND

This study addresses how to leverage information from perceptual evaluations to create a more robust ground truth for EC. This section discusses relevant studies on soft-learning, multi-label learning, and distribution-label learning for EC.

2.1. Soft-label Learning for EC

Instead of using a one-hot vector for the consensus label, a soft-label is generated by considering all the evaluations provided by multiple annotators. It consists of a vector, where each dimension corresponds to an emotional class. The values assigned to the dimensions represent the proportion that each class was selected by the evaluators. For example, if we have a four-class emotion classification task (N: neural, A: anger, S: sadness, and H: happiness), and five evaluators annotated the stimulus as "H, H, H, N, N", then the soft-label is (0.4, 0.0, 0.0, 0.6). Soft-label learning can reduce the loss of data discarded when estimating consensus labels, exploring the uncertainty and variability among annotators' ratings during training [6]. Studies have shown that this approach often leads to better performance than consensus labels [6, 7, 9, 16].

A common practice when using soft labels is to evaluate the model performance with consensus labels [6, 9]. As a result, the ambiguous sentences without consensus labels in the test set are not used to evaluate the model performance. However, these ambiguous stimuli are often seen in daily interactions. Furthermore, these studies often formulate soft-label learning as a single-label task. How-

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

ever, a stimulus can be equally described by more than one emotional class (e.g., depressed and disappointed), especially for stimuli conveying emotions with ambiguous boundaries [17].

2.2. Multi-label Learning for EC

Multi-label learning is introduced to train EC models to deal with the coexistence of multiple emotions. The conventional multi-label learning approach is the "multiple-hot" format. For instance, when we use the example presented in Section 2.1, the multi-label vector is (1, 0, 0, 1). Kang et al. [18] formulated EC as a multi-label task (not a single-label task). They employed multi-label learning to train EC models and used various evaluation metrics for multi-label emotion classification (e.g., macro F1-score, accuracy metric, and Hamming Loss) by converting the prediction probabilities into binary predictions by setting the value 0.5 as a threshold. However, this approach does not distinguish if some emotions may be more prominent than another (e.g., in the example, the emotion "happiness" is more relevant than the class "neutral"). Therefore, it is questionable if the multi-label approach can correctly represent the detailed perceived emotional content. Therefore, studies have modified conventional "multiple-hot" vectors in the multi-label approach into a "soft" vector [10,19]. While this approach is similar to the soft-label in Section 2.1, the formulation during training followed a multi-label approach using the binary cross-entropy cost function instead of the cross-entropy used in the soft-label approach. However, these studies still utilized the highest probability category as the final prediction to assess performance on the test set. This setting loses the principle that multiple emotions can coexist. In contrast, our study uses "soft" multi-label as ground truth, but relies on multi-label classification metrics to estimate model performance.

2.3. Distribution-label Learning for EC

The distribution-label learning approach creates a distribution as the output vector, and the goal is to reduce its distance with the distribution for the ground truth [20]. This objective is often achieved with cost functions such as the *Kullback–Leibler divergence* (KLD). With KLD, the output probabilities during inference do not have to be discretized into binary classes as required by soft-label (cross-entropy) or multi-label (binary cross-entropy) approaches. This study uses the distribution-label learning formulation proposed by Ren et al. [21], which used this method for multimodal sentiment analysis.

3. RESOURCES

3.1. The MSP-Podcast corpus

We use release 1.9 of the MSP-Podcast corpus [3], which consists of 55,283 sentences in the train set, 9,546 sentences in the development set, and 16,570 sentences in the test set. All the audio recordings come from spontaneous conversations from podcasts available on audio-sharing websites. The collection of this corpus builds on the retrieval-based method proposed in Mariooryad et al. [22]. The annotation of the corpus includes primary emotions (i.e., the most dominant emotion) and secondary emotions (i.e., all the emotions perceived by the evaluators). The primary emotions contain nine options: anger, sadness, happiness, surprise, fear, disgust, contempt, neutral, and other. The evaluators have to select a single class. The secondary emotions include the primary emotions plus eight more classes: amusement, frustration, depression, concern, disappointment, excitement, confusion, and annoyance (17 options in total). The evaluators can select all the classes that they want. They are asked to include the emotion selected as the primary emotion. This study uses primary and secondary emotions. Notice that the primary and secondary emotions have the class "other." The evaluators are asked to describe the emotions in their own words. This typed description of their emotional perception allows us to mine additional emotional information. Every speaking turn is annotated by at least five evaluators with a crowdsourcing protocol adapted from Burmania et al. [23]. Table 1 shows the annotations for one of the sentences in the corpus for the primary (second column) and secondary (third column) emotions. Each row in the table represents the annotations from an evaluator. For the class *other*, we add in brackets the description typed by the evaluators. The study of Lotfian and Busso [3] provides more details about the corpus.

3.2. Acoustic Feature Extraction and Pre-processing

The analysis of Keesing et al. [24] revealed that the wav2vec feature set [25] is one of the most effective feature extraction approaches for SER tasks. Therefore, we rely on this feature set as the input for our SER system The wav2vec extracts feature representation for a 30 ms window of raw audio (consider a total of 210 ms reception field), producing a 512-dimensional vector. We use the z-normalization function to normalize all the features, where the parameters for the mean and standard deviation are estimated from the train set.

4. METHODOLOGY

4.1. Polarity Label Processing

Our study explores the benefits of using the typed descriptions provided by evaluators when they selected the class "other" in the primary or secondary emotions. We propose a three-dimensional polarity vector (positive, negative, and ambiguous words). We mine the emotional information from the typed descriptions using NLP tools, using the following steps. First, we transform the typed description into lowercase, employing the Google Spell Check API via SerpApi to correct spelling mistakes. Then, we determine if the typed words belong to the primary or secondary classes, since some of the typed descriptions were variations of the options provided in the questionnaire (e.g., "angry" instead of "anger") or variations were adjectives were added to describe the degree of the emotion (e.g., "slightly angry"). For these cases, we add these selections to the corresponding primary or secondary classes. Next, we use the LIWC 2015 toolkit [15] to classify the typed descriptions into positive and negative emotions. We get 6,852 (primary) and 5,605 (secondary) words grouped in the positive class and 9,131 (primary) and 2,893 (secondary) words in the negative class. Figures 1(a) and 1(c) visualize the typed words assigned to the positive and negative classes, respectively. If the words can not be categorized into positive or negative classes by LIWC, we assign these words to the ambiguous class. There are 4,129 (primary) and 7,536 (secondary) words assigned to the ambiguous emotions. Figure 1(b) visualizes the typed words assigned to the ambiguous class. The next step is to correct cases where an annotator omitted the primary emotion as part of the secondary emotions, as instructed during the perceptual evaluation. For the example in Table 1, the third annotator selected the class "surprise" as the primary emotion but not as the secondary emotions. We correct cases like this one by adding "surprise" as a

Table 1. One annotation example in the MSP-PODCAST. The words in "()" is the typed descriptions; ID means the unique raters.

ID	Primary Emotion	Secondary Emotion
1	Other(Inquisitive)	Excited,Other(Inquisitive)
2	Other(Energetic)	Other(Energetic)
3	Surprise	Amused,Other(interested)
4	Other(curiosity)	Amused,Concerned,Confused
5	Neutral	Neutral

secondary emotion. This correction also includes the class "other." For the example in Table 1, the fourth annotator selected the class "other," providing the typed description *curiosity*. This term is not used for the secondary emotions. Therefore, we add this term as part of the secondary emotions.

Some sentences do not have typed descriptions since the class "other" was not selected by any of the annotators. Furthermore, the primary and secondary emotions can also inform the polarity of the speaking turns. For these reasons, we include all the secondary emotions in the estimation of the polarity (notice that primary emotions are already included in the secondary emotions). Based on the LIWC 2015 toolkit, the terms anger, sadness, fear, disgust, contempt, frustration, depression, concern, disappointment, confusion, and annoyance are considered as negative emotions, the terms happiness, amusement, and excitement are considered as positive emotions, and the terms surprise and neutral are considered as ambiguous emotions. Finally, we follow the operation used in previous studies by estimating soft-labels [6, 8] by normalizing the polarity vectors, so their sum adds to 1. We also use the label smoothing strategy proposed by Szegedy et al. [26] to smooth the vector using the smoothing parameter 0.05, where a small probability is given to emotional classes with zero value. The same soft-label approach is also used for the primary (8D vector) and secondary (16D vector) emotional labels.

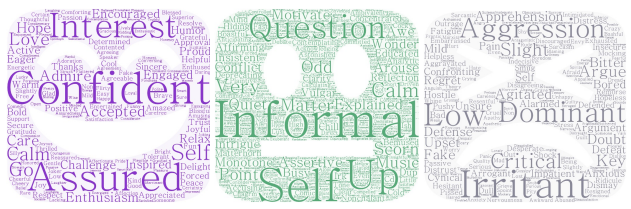
4.2. Learning Label Processing

We hypothesize that using detailed information from the annotators can lead to better ground truth models for SER. We aim to leverage annotations from primary emotions (**P**), secondary emotions (**S**), and our proposed polarity labels (**Po**). We implement multi-task experiments to show whether the SER performance can be improved by adding the polarity labels. Figure 2 shows the multi-task frameworks, which have one, two, or three tasks. The tasks are combinations of (**P**), (**S**) or (**Po**). Section 4.3 describes the core SER model.

We leverage three label learning methods to train the SER models: soft-label, multi-label (soft-label like instead of multiple-hot approach), and distribution-label (Section 2). The main differences between these three approaches are the cost function and the assumption about the coexistence of emotion labels. The soft-label learning method uses the *cross-entropy* (CE), which assumes that the classification task can only have a single class. The multi-label learning approach uses *binary cross-entropy* (BCE), which allows having multiple coexisting classes. The distribution-label learning approach uses KLD as a cost function, which also allows multiple classes.

4.3. Emotion Classification Model

We follow the chunk-level SER modeling methodology proposed by Lin and Busso [27] as our core model. The model hierarchically captures emotional-relevant information at the frame-level, chunk-level, and sentence-level. The approach splits the speaking turns into a fixed number of chunks, which have a fixed duration regardless of the duration of the recordings. This goal is achieved by using a dynamic chunking formula that changes the overlap between chunks. The chunk representations are combined with a chunk-level attention mechanism, which can be efficiently implemented since the number of chunks is fixed. The framework results in improved recognition performance and robustness regardless of the duration of the sentences. In this paper, we choose to use *long short-term memory* (LSTM) as the chuck-level feature encoder equipped with the *RNN-AttenVec* chunk-level attention model, which was one of the best combinations proposed by Lin and Busso [27]. For each chuck, we extracted the wav2vec feature maps, which are processed by the LSTM. The last time-step output of the LSTM feature encoder is extracted as the chunk-level representation vector of each input data chunk. These chunk-level representations are then fed into the *RNN-AttenVec* to perform the chunk-level attention, obtaining a weighted (via the trainable attention weights) combination vector as the final sentence-level representation to train with the desired emotion tasks. The detailed network architectures and parameters are the same as the original paper [27], with the exception of the number of hidden nodes, which are set to 512 for all the layers to match the dimension of the wav2vec input. Notice that the model structure is flexible, and it could be replaced by any other sentence-level SER models to train using our proposed polarity labels.



(a) Positive words (b) Ambiguous words (c) Negative words

Fig. 1. Example of sentiment words in the MSP-Podcast corpus.

5. EVALUATION

5.1. Experimental Setup

We evaluate different multi-task experiments by considering subsets of the output layers for the primary, secondary, and/or polarity labels. The ground truth and the data to train and test all the models are the same in all the experiments for fair comparisons. Unlike most prior studies on emotion classification, we did not discard data or labels in the training and testing stages. For simplicity, the weights for all the tasks in the cost function are equal in the multi-task experiments. The activation function of the output layer is the softmax for CE and KLD, and the sigmoid for BCE. We use the Adam optimizer with a learning rate set to 10^{-4}. We use batch sizes of 128, training the models for 100 epochs, and 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.

We consider multiple evaluation metrics to compare the predicted labels with the ground truth. First, we consider the cosine similarity (**Cosine**), which measures the similarity between the predicted and ground truth vectors. Second, we use the root mean-square-error (**RMSE**), which estimates the differences between the ground truth and the model prediction. The next three metrics follow the metrics proposed by Fei et al. [28] to evaluate multi-label classification performance: macro F1-score (**maF1**), ranking loss (**RL**), and hamming loss (**HL**). The multi-label learning method requires hard labels to estimate performance, so the values in the soft-label vector have to be binarized. The threshold used to convert the prediction probabilities into binary values depends on the dimensions of the output layer in the task. The thresholds for **P** is 1/8, for **S** is 1/16, and for **Po** is 1/3. We consider that an emotion is present if its probability value is larger than the corresponding threshold.

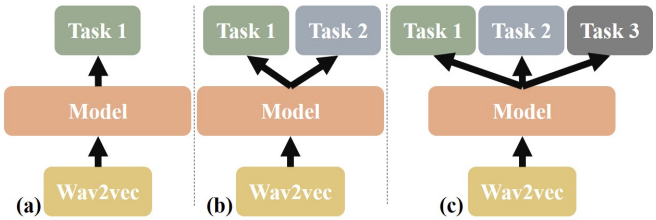


Fig. 2. The overview of multi-task experiments. (a) One single-task; (b) Two multi-tasks; (c) Three multi-tasks.

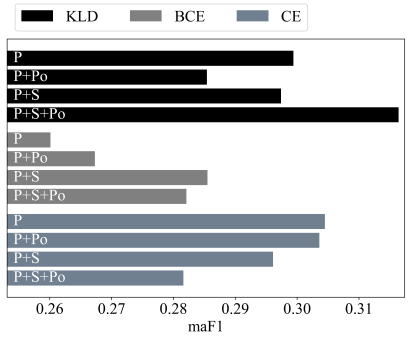
Table 2. Single-task and multi-task evaluation results using primary emotions (P), secondary emotions (S), and the polarity labels (Po). The symbol ↑ means that the performance increases with higher values of the metric. The symbol ↓ means that the performance increases with lower values of the metric. The bold numbers indicate the best performance for a given cost function. The underlined numbers represent the best results in the task across all cost functions.

Task	Loss	Target	RMSE↓	Cosine↑	HL↓	RL↓	maF1↑
P	KLD	P	0.194	0.597	0.287	0.510	0.299
		P+Po	**0.194**	**0.607**	**0.285**	**0.509**	0.285
		P+S	0.194	0.598	0.286	0.510	0.297
		P+S+Po	0.196	0.586	0.291	0.524	**0.316**
	BCE	P	0.201	**0.584**	0.279	0.527	0.260
		P+Po	0.203	0.574	**0.277**	0.534	0.267
		P+S	**0.199**	0.580	0.284	**0.523**	**0.286**
		P+S+Po	0.201	0.571	0.280	0.535	0.282
	CE	P	0.194	0.602	0.292	0.511	**0.305**
		P+Po	**0.193**	**0.602**	0.288	0.508	0.304
		P+S	0.195	0.598	0.287	0.513	0.296
		P+S+Po	0.194	0.601	**0.281**	**0.507**	0.282
S	KLD	S	0.089	0.613	**0.314**	0.579	0.326
		S+Po	0.088	0.610	0.321	**0.568**	**0.380**
		S+P	**0.088**	**0.615**	0.315	0.574	0.340
		S+P+Po	0.089	0.606	0.318	0.577	0.364
	BCE	S	0.090	0.605	**0.313**	**0.576**	**0.350**
		S+Po	**0.090**	**0.610**	0.319	0.582	0.349
		S+P	0.092	0.600	0.329	0.615	0.307
		S+P+Po	0.092	0.591	0.317	0.596	0.333
	CE	S	0.089	0.610	0.316	**0.568**	**0.359**
		S+Po	0.089	0.609	0.317	0.579	0.347
		S+P	0.090	0.610	0.323	0.595	0.323
		S+P+Po	**0.088**	**0.617**	**0.316**	0.571	0.339
Po	KLD	Po	0.240	0.799	0.369	0.502	0.452
		Po+P	**0.238**	**0.808**	0.376	0.510	0.442
		Po+S	0.241	0.793	**0.365**	**0.496**	0.460
		Po+P+S	0.241	0.791	0.373	0.505	**0.475**
	BCE	Po	0.266	0.760	0.371	0.507	0.441
		Po+P	**0.254**	**0.772**	0.366	0.498	0.465
		Po+S	0.259	0.771	**0.361**	**0.494**	0.435
		Po+P+S	0.259	0.762	0.366	0.498	**0.474**
	CE	Po	0.243	0.792	0.365	**0.497**	0.457
		Po+P	0.238	0.803	**0.365**	0.498	0.455
		Po+S	0.238	0.796	0.378	0.510	0.490
		Po+P+S	**0.235**	**0.805**	0.373	0.506	**0.492**

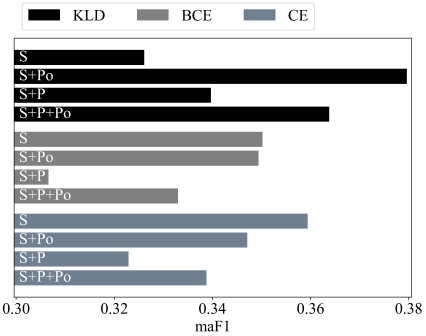
5.2. Experimental Results

Table 2 summarizes the overall results. For each metric, we add the symbol ↑ to indicate that the result is better when the value is higher. Otherwise, we add the symbol ↓. The table has three blocks for the prediction of primary (**P**), secondary (**S**) and polarity **Po** labels (column *Task*). Each of the blocks has different combinations of multi-task formulation trained using KLD (distribution-label learning), BCE (multi-label learning), or CE (soft-label learning), as indicated in column *Loss*. The column *Target* indicates the learning targets for the models. For instance, P means a single-task model that is only trained to predict primary emotions; P+Po means a multi-task model that is trained to estimate primary and polarity labels. The bold numbers indicate the best performance for a condition when trained with a given cost function. The bold and underlined numbers indicate the best result for a task for a given metric across all cost functions. Figure 3 visualizes the maF1 results for the prediction of primary and secondary emotions.

Is the polarity label useful? Table 2 shows that most of the best results correspond to cases when the polarity label is included in the multi-task model. For example, the model to predict primary emotions trained with KLD using the multi-task system P+S+Po achieves a 6.4% relative improvement in maF1 over the multi-task system P+S (see Fig. 3(a)). In the prediction of secondary emotion using KLD, the multi-task model S+Po gives a 16.56% performance gain over the single task learning model S (see Fig. 3(b)). The typed descriptions have valuable emotional information, showing that the polarity label is indeed helpful to improve the recognition performance for primary and secondary emotions.



(a) The macro-F1 scores for primary emotion recognition.



(b) The macro-F1 scores for secondary emotion recognition.

Fig. 3. maF1 results for primary and secondary emotion predictions.

Is there any difference between soft-label, multi-label, and distribution-label learning? The distribution-label learning with KLD achieves the best results in two out of five metrics on the classification of primary emotions and three out of five metrics on the classification of secondary emotions. The soft-label learning with CE achieves the best results in two out of five metrics for the classification of primary emotions, one out of five metrics for the classification of secondary emotions, and two out of five metrics on the classification of the polarity label. Table 2 and Figure 3 shows that multi-label learning with BCE often leads to worse performance across conditions. We conclude that distribution-label learning is more suitable when the classification task includes more classes than the other two learning methods.

6. CONCLUSION AND FUTURE WORK

This paper maximized the emotional information from the annotators' typed descriptions to generate a three-class polarity label. This approach fully exploits all the emotional information included in the perceptual evaluations during the training and evaluation of the EC model. The results demonstrated that the proposed three-class polarity label is helpful to increase the recognition accuracy of primary and secondary emotions. Moreover, our results showed that distribution-label learning with KLD as the cost function is more suitable when the dimension of the emotional descriptor is high. Our future work will explore other emotional attributes extracted from the typed descriptions that describe the perceived emotion beyond its sentiment (e.g., activation, dominance).

7. REFERENCES

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

[2] C. Busso and S.S. Narayanan, "Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, September 2008, pp. 1670–1673.

[3] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[4] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[5] Z. Zhang, J. Deng, E. Marchi, and B. Schuller, "Active learning by label uncertainty for acoustic emotion recognition," in *Interspeech 2013*, Lyon, France, August 2013, pp. 2856–2860.

[6] 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.

[7] 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.

[8] Atsushi Ando et al., "Soft-Target Training with Ambiguous Emotional Utterances for DNN-Based Speech Emotion Classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4964–4968.

[9] 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.

[10] A. Ando et al., "Speech emotion recognition based on multi-label emotion existence model," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2818–2822.

[11] 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.

[12] J. A. Russell, "Forced-choice response format in the study of facial expression," *Motivation and Emotion*, vol. 17, no. 1, pp. 41–51, March 1993.

[13] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[14] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, January-March 2019.

[15] J.W. Pennebaker, R. Booth, R. Boyd, and M. Francis, "Linguistic inquiry and word count: LIWC2015," Operator's manual, Pennebaker Conglomerates, Austin, TX, 2015.

[16] E. Mower, M.J. Mataric, and S.S. Narayanan, "A framework for automatic human emotion classification using emotional profiles," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1057–1070, May 2011.

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

[18] Xin Kang, Xuefeng Shi, Yunong Wu, and Fuji Ren, "Active learning with complementary sampling for instructing class-biased multi-label text emotion classification," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[19] Y. Kim and J. Kim, "Human-Like Emotion Recognition: Multi-Label Learning from Noisy Labeled Audio-Visual Expressive Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5104–5108.

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

[21] Y. Ren, N. Xu, M. Ling, and X. Geng, "Label distribution for multimodal machine learning," *Frontiers of Computer Science*, vol. 16, no. 1, pp. 1–11, February 2022.

[22] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[23] 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.

[24] A. Keesing, Y.S. 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.

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

[26] 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)*, Las Vegas, NV, USA, June 2016, pp. 2818–2826.

[27] 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.

[28] 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)*, New York, NY, USA, February 2020, vol. 34, pp. 7692–7699.