INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Learning to Recognize Per-rater's Emotion Perception Using Co-rater Training Strategy with Soft and Hard Labels

Huang-Cheng Chou, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University
MOST Joint Research Center for AI Technology and All Vista Healthcare
hc.chou@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

An individual's emotion perception plays a key role in affecting our decision-making and task performances. Previous speech emotion recognition research focuses mainly on recognizing the emotion label derived from the majority vote (hard label) of the speaker (i.e., producer) but not on recognizing per-rater's emotion perception. In this work, we propose a framework that integrates different viewpoints of emotion perception from other co-raters (exclude target rater) using soft and hard label learning to improve target rater's emotion perception recognition. Our methods achieve [3.97%, 1.48%] and [1.71%, 2.87%] improvement on average unweighted accuracy recall (UAR) on the three-class (low, middle, and high class) [valence, activation (arousal)] emotion recognition task for four different raters on the IEMOCAP and the NNIME databases, respectively. Further analyses show that learning from the soft label of co-raters provides the most robust accuracy even without obtaining the target rater's labels. By simply adding 50% of a target raters annotation, our framework performance mostly surpasses the model trained with 100% of raters annotations.

Index Terms: Speech Emotion Recognition (SER), rater perception, BLSTM-DNN, soft label learning

1. Introduction

The manner that we perceive and interpret other's emotional states affects how we interact and make decisions in daily life. For example, when students perceive that a teacher's actions as positive emotion toward their performances, this perception has a positive influence on a student's learning goal and enjoyment in English learning [1]. Also, negative emotion perception (e.g., anger and disgust) is one of the detrimental impact factors for police to make an erroneous decision (shot or not) [2]. In the field of sport, perceiving coaches' behaviors as demonstrating negative emotion harms elite child athletes performances [3]. In the buyer-seller interaction, the salesman, who can accurately appraise the emotions of others, is often better at using a strategy of customer-oriented selling and has a positive impact on sales performance [4].

While there exists a wealth of research in advancing speech emotion recognition (SER), most (if not all) of these works assume by having multiple annotators to rate a sample of behavior data, by taking the 'majority vote', it would correspond to the 'ground truth' of the expressive subject's (producer) emotion states. This assumption ignores that emotion itself is very *individualized* (both in terms of perceiving and expressing it), which is known to be related to one's own past experiences [5]. Only recently, few works have investigated this *individualized* aspect for SER. For example, Li et al. [6, 7] integrate each speaker's personal attributes through attention mechanism to improve *expressed* (producer) emotion recognition, and Chou et al. [8] recently models the subjectivity and differences across annotators to improve classification performance on 'majority vote' of expressed emotion. While there are works in music and social circle recommendation system that has targeted 'per-annotator' recognition [9, 10], limited if any of the works in conventional SER setting has contributed in per-rater emotion perception recognition.

In this work, we aim to improve per-raters emotion perception recognition system such that it could be further used to understand and potentially affect an individual's decision making across various application fields. However, in real life, it is unrealistic to be able to collect large enough *annotated* data from an individual rater, which hinders such a computational work to be carried out. To mitigate this issue, our idea is to integrate other rater's (co-raters) existing labels of both soft (distributional) and hard (majority-voted) labels, which ensures every rated annotation is fully utilized [8], in advancing per-rater's emotion perception recognition.

With rapid development in the field of speech emotion recognition, there are many types of methods to extract emotional information from each utterance, such as acoustic-prosodic feature extraction [11], spectrum [12, 13], and even raw data [14]. In addition, many researchers design the learning models vary with the characteristics of different feature inputs, e.g., Convolutional Neutral Networks (CNN) [15, 16, 17], Bidirectional Long Short-Term Memory (BLSTM) with attention mechanism [18], Generative Adversarial Networks (GAN) [19], or multi-head attention architecture [13, 20]. The recent state-of-the-art models are usually combinations of these models and inputs based on different setups, e.g., context information [21], or multiple attribute [20]. However, there is no study on per-rater emotion recognition in the field of SER to our knowledge; hence, in this work, we use the BLSTM-DNN model [18] as our main network building block due to its robust accuracy in single utterance speech modeling for emotion recognition across a variety of setups.

Specifically, we proposed a network architecture to perform per-rater emotion recognition by simultaneously leveraging the label uncertainty and the co-raters annotations on the IEMOCAP [22] and the NNIME [23] database. Due to its enhanced modeling capacity by including modeling of other raters and variability of annotations, our proposed model achieves [3.97%, 1.48%] and [1.71%, 2.87%] improvement on average UAR on the three-class [valence, activation (arousal)] task for four different raters individually on two databases, respectively. We observe that 1) integrating co-raters' labels indeed improves target rater's recognition rates, 2) when co-training with other raters, we only need 50% of the target rater's annotations to achieve the best performance, 3) soft label training strategy provides the most robust recognition results.

http://dx.doi.org/10.21437/Interspeech.2020-1714

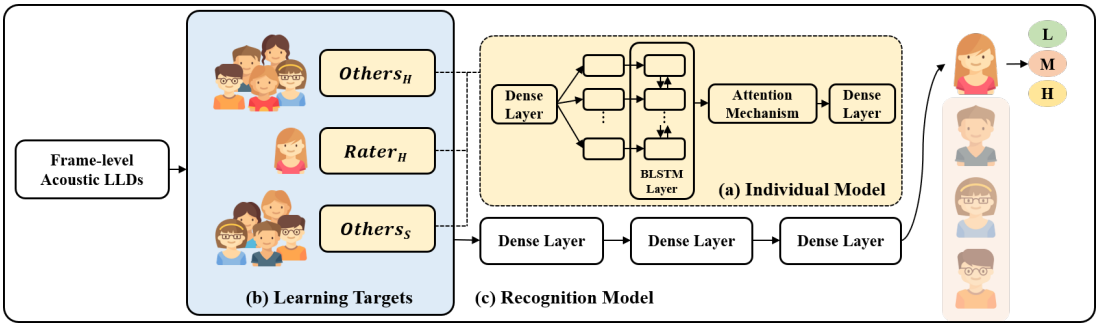

Figure 1: *An illustration of per-rater's emotion perception recognition model (L, H, and M means that low, middle, and high class).*

2. Research Methodology

Figure 1 illustrates our proposed framework used in this paper. We propose the model for joint training of three models, where each one learns from different viewpoints of emotion perception. In total, there are three basic core components showed in Figure 1: 1) learning from pooling all other raters' annotations (soft label), 2) learning from target rater-only annotations, and 3) learning from majority vote (hard label) from co-raters' annotations (Note that hard label and soft label excludes target raters' annotations). These models are then finally concatenated to learn the final per-rater emotion perception recognition (showed in Figure 1 (c)). The main framework used within each model is based on the structure proposed in [18], which contains an initial dense layer, then a bidirectional long short-term recurrent network with attention mechanism, and a final dense layer (BLSTM-DNN) (showed in Figure 1 (a)). In the fusion stage, we add one concatenate layer and three additional dense layers with ReLU as activation function [24] and batch normalization to perform fine-tuning.

2.1. Database and Soft Label Representation

Our framework is evaluated on the IEMOCAP [22] and the NNIME [23] database. We focus on valence and activation (arousal) classification since these attributes receive the most ratings in both databases and the boundaries between categories of emotion are fuzzy rather than discrete [25]. In addition, annotation on valence and activation in both databases are on a 1 to 5 scale (1 is "very negative" ("very active") and 5 is "very positive" ("very inactive") for valence (or activation)). Moreover, we follow recent works on the IEMOCAP database [19, 26] to transform raw annotations of each utterance of both databases into 3-class classification. The middle class equals to the original rating of 3, and the other two classes are annotations values that are smaller than 3 (low class) or higher than 3 (high class). A hard label is given by majority vote (conventional method), and a soft label is obtained by dividing each rated value by the total number of annotations such that the soft label of each sample sums to one, e.g. if three annotators give labels of 4, 2, and 1, the soft label would be [0.66, 0.00, 0.33] ([Low, Middle, High]). To be noticed, this work differs from [26] and follows [27] setting, which uses all the data even though not all of the utterances have majority vote (all of them, however, would have a soft label distribution).

2.1.1. IEMOCAP

IEMOCAP contains about 12 hours of audio-video recordings of dyadic interactions with 10 different actors split into pairs over 5 sessions in English. To be noticed, we use leave-one-session-out cross-validation when evaluating the results of IEMOCAP. There are 10039 utterances (4784 improvised turns, 5255 scripted turns) in the database that has been given emotion labels by 2 to 4 annotators. There are 18 unique raters including 6 of them who are actors themselves and 12 persons who are the observed naive raters. Half of the observed raters and all of the actors annotate emotional dimensional attributes, i.e., valence and arousal, on a 1 to 5 scale, and "the other" six observed naive raters and all actors label emotion categories. That is, there is no same person between two observed naive rater groups rated both the emotional dimensional attributes and the emotion categories. In this work, we finally focus on four "observed naive raters" of the IEMOCAP that annotated emotional dimensional attributes, i.e., E1, E2, E3, and E4, because E5 and E6 do not have enough samples (only 58). In Table 1 (Per-R), E1, E2, E3, and E4 in the IEMOCAP are Rater #1, #2, #3, #4.

2.1.2. NNIME

NNIME is a public database that includes approximately 11-hour worth of audio, video, electrocardiogram data, and transcripts. Audio data were manually segmented. The hypothesized interaction scenario was assumed to be in a real-life home setting, such as the living room, dormitory, or bedroom, and all subjects were native Mandarin Chinese speakers. There are in total of 22 dyadic pairs over six different atmospheres including happiness, anger, sadness, neutral, surprise, and frustration. We split these 22 pairs into five groups in the evaluation process. The original annotations were only given on the continuous-in-time level or whole-session level. In other words, the NNIME database did not have sentence-level annotations. Therefore, we recruited another 7 native Mandarin Chinese speakers as annotators (4 females, 3 males) with age ranged from 20 to 31 to label emotional dimensional attributes, such as valence and activation (arousal). In this work, we only used 4773 speech utterances (all are improvised turns) and chose four raters of the NNIME as target raters with the most annotated number of samples. To collect high-quality annotations, we allowed raters to change annotations until they were satisfied. In Table 1 (Per-R), E2, E3, E4, and E5 in the NNIME are Rater #1, #2, #3, #4, respectively.

2.2. Per-Rater Emotion Recognition

2.2.1. Acoustic Features

We extract acoustic features using the same setting as in [8, 27], which is based on the emobase.config in the OpenSmile toolbox [28]. This acoustic feature set extracts frame-level descriptors of loudness, fundamental frequency (F0), voice probability, zero-cross rate, 12-dimensional Mel-Frequency Cepstral Coefficients (MFCCs), the first derivatives of them, and the second

Table 1: *A summary of results on the three-class valence (Val.) and activation (Act.) for IEMOCAP and NNIME database in UAR (%).*

Database	Task	Per-R	$Rater_*$		$Others_H$			$Others_S$			$Others_{H-S}$	
Add % Rater			100%	50%	100%	50%	0%	100%	50%	0%	100%	50%
IEMOCAP	Val.	#1	51.67	50.69	53.02	52.28	47.52	53.63	52.46	47.83	**53.70**	52.86
		#2	49.98	47.41	51.67	49.48	47.21	**52.12**	50.51	48.61	52.01	50.48
		#3	51.25	46.48	52.47	51.59	48.34	54.22	53.76	53.45	**54.72**	53.20
		#4	42.37	41.53	50.93	49.85	47.91	**51.84**	49.82	49.57	50.70	51.22
	Act.	#1	58.95	58.33	59.85	59.49	56.38	**59.95**	58.94	56.03	59.83	59.31
		#2	58.20	55.44	58.69	56.74	50.67	**59.40**	58.27	57.87	59.04	58.17
		#3	66.12	62.24	67.10	65.66	55.13	**67.26**	65.12	57.36	67.17	65.15
		#4	54.75	53.79	56.77	55.26	49.96	**58.30**	57.20	52.31	57.87	56.59
NNIME	Val.	#1	43.91	41.84	45.12	43.93	44.55	45.41	43.66	43.38	**45.45**	44.17
		#2	40.06	39.31	41.60	39.87	38.33	**43.13**	40.75	41.86	42.75	40.81
		#3	44.00	41.06	44.88	43.31	43.01	**46.14**	45.46	44.77	45.68	45.01
		#4	44.24	42.73	43.16	44.35	43.59	44.52	45.31	46.20	**45.17**	45.88
	Act.	#1	53.61	51.19	57.58	55.42	53.49	58.06	55.73	43.46	**58.29**	56.47
		#2	51.97	48.36	53.50	52.33	51.62	54.89	54.59	42.96	**55.05**	54.56
		#3	53.54	51.60	53.96	52.68	50.09	54.58	53.64	41.41	**54.63**	53.33
		#4	51.73	48.07	53.28	51.18	44.60	**54.53**	52.04	39.59	54.35	52.07

derivatives of MFCCs and loudness. It contains 45-dimensional acoustic features per frame. All acoustic features are extracted at 60ms frame length size and 10ms frame step size, which are further normalized for each speaker using z-score normalization and downsampled by averaging every 3 frames.

2.2.2. Learning Strategies

The main components of our models are based on the BLSTM-DNN structure as previously proposed [8, 27]. In this work, our goal is to account for the label variability and inclusion of co-raters' views to perform target rater's perceived emotion recognition. We train a BLSTM-DNN with two different learning targets: hard labels and soft labels.

A hard label means that the ground truth is obtained using a majority of all ratings, e.g., if the ratings are [3, 3, 2] on one utterance, the ground truth of this data is [0, 1, 0]; however, this voting processing loses potential emotional information that naturally exists in the subjective emotion appraisal. Therefore, we also use soft labels (Section 2.1) to retain every original emotional rating as a target for learning per-rater perception.

2.2.3. Rater-specific and Co-raters Modeling

Emotion perception varies with person to person on the same utterance due to the nature of individual idiosyncrasy and subjectivity [29]. We jointly model the target rater's emotion perception data with co-raters' data within the IEMOCAP and the NNIME databases in our proposed recognition architecture. We define two types of models: $Others$ and $Rater_*$. In more detail, $Others$ takes all of the annotators' ratings excluding the target rater, for example, if we want to perform rater-specific emotion recognition on $Rater_1$, we exclude data samples from $Rater_1$s in the training of $Others$. We further would obtain two different models depending on whether the learning target is set to be a soft label or hard label (Section 2.2.2). Moreover, we use hard label training for all of our $Rater_*$ models since each rater gives one rating only on one sample, e.g., training a model with the utterances annotated by each rater (note the data amount will be different for each rater).

2.2.4. Final Prediction Layer

We freeze all of the sub-models (the structure is showed in Figure 1 (b)), i.e., two $Others$ models, and four $Rater$ models, and concatenate their last layer representation before the softmax to be fed into additional three dense layers with ReLU activation function. Three layers include batch normalization before the ReLU activation function and a dropout with 50% dropout rate. Finally, we add softmax layer to perform the final three-class valence and arousal emotion recognition task. Note that we only use the target rater's annotated data in this fine-tuning stage. The complete structure is illustrated in Figure 1.

3. Experimental Setup and Results

3.1. Experimental Setup

All of our models consist of a major BLSTM-DNN component which includes two dense layers with the ReLU activation function, one BLSTM with attention mechanism layer, and finally one dense layer with softmax function (classification layer). The number of hidden units in the IEMOCAP (in the NNIME) are [256, 128, 256, 3] ([128, 64, 128, 3]) in the first dense layer, BLSTM with attention mechanism layer, the last dense layer, and classification layer, respectively. In the late fusion stage, the number of hidden units in the IEMOCAP (in the NNIME) in the three-layer deep neural network and the final prediction layer are [256, 128, 256, 3] ([128, 64, 128, 3]), respectively. A dropout layer is added for all layers excluding prediction layers with 50% dropout rate.

All experiments are evaluated using leave-one-session-out cross-validation and five-fold cross-validation with the metric of unweighted average recall (UAR) on the IEMOCAP and the NNIME databases, respectively. The batch size and learning rate in the IEMOCAP (in the NNIME) are 64 and 5*1e-4 (5*1e-3), respectively, and the number of the epoch is 200 with early stopping criteria in all conditions with cross-entropy loss minimization. All of the hyper-parameters are selected based on the validation set. ADAMMAX is used as the optimizer.

3.1.1. Models Comparison

We compare different results obtained for each of components of our complete architectures:

$Rater_*$ model: Every $Rater_*$ model is trained with the rater-specific annotated utterances only using the target rater's annotation as the learning target.

$Others_S$ model: This model uses all annotated utterances excluding the target raters utterances with soft label as the learning target. (Note that the utterances used here must be labeled by at

Table 2: *A summary of average results (UAR) from $Rater_*$, $Others_H$, and $Others_H$ using 100% per-rater's annotations over four raters on the high (H), middle (M), low class (L) valence (Val.) and activation (Act.) for IEMOCAP and NNIME database in UAR (%).*

Database	Task	3-Class	$Rater_*$	$Others_H$	$Others_S$
IEMOCAP	Val.	H	69.48	68.41	68.30
		M	52.80	60.80	59.65
		L	45.78	45.20	46.12
	Act.	H	69.48	68.41	68.30
		M	42.70	46.64	49.93
		L	66.33	66.75	65.45
NNIME	Val.	H	41.62	58.71	54.38
		M	66.80	48.29	53.38
		L	20.74	24.06	26.64
	Act.	H	57.41	64.33	62.30
		M	28.85	27.16	33.64
		L	71.88	72.25	70.59

least two other raters excluding target rater, the same criterion is applied to the $Others_H$ model).

$Others_H$ model: This model uses all annotated utterances with co-raters' consensus voting.

Proposed model ($Others$ and $Rater_*$ fusion): This model is our proposed model that combines all co-rater's $Others$ model (training target can be either soft, hard or both), and then it is added with the $Rater_*$ target rater's own model trained with 0%, 50%, and 100% of the total target rater's annotated utterances, respectively.

3.2. Experimental Results and Analyses

Table 1 shows the complete recognition results. Our proposed framework obtains the best overall emotion recognition accuracy for $Rater_1$, $Rater_2$, $Rater_3$, and $Rater_4$ in the IEMOCAP on three-class [valence, activation (arousal)] with UAR of [53.7%, 59.95%], [52.12%, 59.40%], [54.72%, 67.26%], and [51.84%, 58.30%], respectively. Moreover, the best overall results for $Rater_1$, $Rater_2$, $Rater_3$, and $Rater_4$ in the NNIME obtains UAR of [45.45%, 58.29%], [43.13%, 55.05%], [46.14%, 54.63%], and [45.17%, 54.53%], respectively. The proposed method surpasses $Rater_*$ models (trained with 100% of target rater's annotated utterances) by [2.14%, 1.00%] ([1.54%, 4.68%]), [3.07%, 3.08%] ([2.14%, 1.00%]), [3.47%, 1.14%] ([2.14%, 1.09%]), and [9.48%, 3.55%] ([0.93%, 2.8%]) absolute in the IEMOCAP (in the NNIME), which indicates the importance in leveraging co-rater s' labeling information.

To be more specific, our methods achieve [3.97%, 1.48%] and [1.71%, 2.87%] improvement on average unweighted accuracy recall (UAR) on the three-class (low, middle, and high class) [valence, activation (arousal)] emotion recognition task for four different raters on the IEMOCAP and the NNIME, respectively. According to the results, our proposed method, especially $Others_S$, provides a larger boost on three-class valence and arousal tasks on both databases, and $Rater_*$, fusing co-rater's soft label information provides more stable recognition rates than other types of fusion combinations.

One key observation is that even when reducing the $Rater_*$ model's training data to a half, most of the proposed method can achieve higher performance than simply using $Rater_*$ with 100% of annotated training utterances. This insight demonstrates the potential application in real-world scenarios when it is difficult to collect a large number of target rater's emotion perceptual annotations, but the model can learn from co-raters' information.

Especially, in Table 2, $Others_S$ with soft label training obtains better recognition rates for low class and middle class on valence and activation task than $Other_H$ with hard label training. On the other hand, $Other_H$ with hard label training achieves a better recognition rate for low class on activation task than $Other_H$ with hard label training.

Table 3: *A summary of Root Mean Square Error (RMSE) between the annotations of $Rater_*$ and $Others_H$ (showed in $Others_H$ column), $Rater_*$ and $Others_S$ (showed in $Others_S$ column).*

Database	Per-R	Task	$Others_H$	$Others_S$	Task	$Others_H$	$Others_S$
IEMOCAP	#1	Val.	**0.421**	0.423	Act.	0.529	**0.506**
	#2		**0.379**	0.395		0.555	**0.520**
	#3		**0.325**	0.356		0.643	**0.573**
	#4		**0.358**	0.382		0.572	**0.520**
NNIME	#1	Val.	0.385	**0.358**	Act.	0.511	**0.445**
	#2		0.418	**0.377**		0.451	**0.397**
	#3		0.364	**0.340**		0.558	**0.469**
	#4		0.397	**0.363**		0.603	**0.515**

Furthermore, even if we do not have the target raters annotation at all, we compare two types of $Others$ models with the soft label and hard label training. The results show that using soft label training is overall better than using hard label training. We further compute the Root Mean Square Error (RMSE) as measures of concordance between different labeling used as a learning target in our work. According to Table 3, we show that the Root Mean Square Error (RMSE) between $Others_S$ and $Rater_*$ is larger than $Others_H$ and $Rater_*$, but the performance of $Others_S$ is still better than $Others_H$'s, which indicates the variability in the soft label learning strategy provides a more robust representation even when used in facilitate learning to recognize the target rater's perception. Interestingly, we also observe that learning to use the distributional label in the IEMOCAP database is indeed useful in predicting higher-level emotional dimension, valence. This finding is similar as the previous work [26] when using soft label training on cross-corpus valence tasks.

By including $Others$ model as an additional representation in late fusion, our proposed method can learn to integrate multiple complementary information from different perceptual viewpoints. Moreover, our experiments demonstrate that compared to the conventional method (hard label), the soft label is a better alternative way to model different emotion perception of each annotator.

4. Conclusion

Understanding how an individual perceives one another's emotional state is important as it often underlies our decision-making process and impacts task performances. In this work, we propose a framework that jointly models different viewpoints of emotion perception from co-raters labeling in improving target rater's emotion perception recognition. Our methods achieve [3.97%, 1.48%] and [1.71%, 2.87%] improvement on average UAR on the three-class [valence, activation] emotion recognition task for four different raters on the IEMOCAP and the NNIME databases, respectively. To the best of our knowledge, while there are recent works in studying *individualized* emotion recognition, this is one of the first works that have integrated co-raters' emotion labeling to focus a target rater's emotion *perception* to enable *individualized* emotion-sensing module. In the future, we would investigate more in detail specifically what are some of the key differences between individual annotators when they rate the same behavior data, and how it relates to the annotator's personal emotional experiences to obtain further insights on the variability of emotion perception.

5. References

[1] D. Urhahne, "Teacher behavior as a mediator of the relationship between teacher judgment and students' motivation and emotion," *Teaching and Teacher Education*, vol. 45, pp. 73–82, 2015.

[2] J. L. Harman, D. C. Zhang, and S. G. Greening, "Basic processes in dynamic decision making: How experimental findings about risk, uncertainty, and emotion can contribute to police decision making." *Frontiers in Psychology*, vol. 10, p. 2140, 2019.

[3] A. E. Stirling and G. A. Kerr, "The perceived effects of elite athletes' experiences of emotional abuse in the coach–athlete relationship," *International Journal of Sport and Exercise Psychology*, vol. 11, no. 1, pp. 87–100, 2013.

[4] B. Kidwell, R. G. McFarland, and R. A. Avila, "Perceiving emotion in the buyer–seller interchange: the moderated impact on performance," *Journal of Personal Selling & Sales Management*, vol. 27, no. 2, pp. 119–132, 2007.

[5] J.-A. BACHOROWSKI and M. J. Owren, "Sounds of emotion: Production and perception of affect-related vocal acoustics," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 244–265, 2003.

[6] J.-L. Li and C.-C. Lee, "Attentive to individual: A multimodal emotion recognition network with personalized attention profile." in *INTERSPEECH*, 2019, pp. 211–215.

[7] C.-C. L. Jeng-Lin Li, "Attention learning with retrievable acoustic embedding of personality for emotion recognition," in *2019 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019.

[8] H.-C. Chou and C.-C. Lee, "Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5886–5890.

[9] A. Nanopoulos, D. Rafailidis, P. Symeonidis, and Y. Manolopoulos, "Musicbox: Personalized music recommendation based on cubic analysis of social tags," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 407–412, 2009.

[10] X. Qian, H. Feng, G. Zhao, and T. Mei, "Personalized recommendation combining user interest and social circle," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 7, pp. 1763–1777, 2013.

[11] S. Yoon, S. Dey, H. Lee, and K. Jung, "Attentive modality hopping mechanism for speech emotion recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3362–3366.

[12] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7144–7148.

[13] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7179–7183.

[14] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[15] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *INTERSPEECH*, 2017, pp. 1089–1093.

[16] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2666–2670.

[17] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7174–7178.

[18] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[19] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2746–2750.

[20] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7189–7193.

[21] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6685–6689.

[22] 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*, vol. 42, no. 4, p. 335, 2008.

[23] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, and C.-C. Lee, "NNIME: The Nthu-Ntua Chinese Interactive Multimodal Emotion corpus," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 292–298.

[24] V. Nair and G. E. Hinton, "Rectified Linear Units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

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

[26] J. Gideon, M. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, 2019.

[27] A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, Y. Ijima, and Y. Aono, "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)*. IEEE, 2018, pp. 4964–4968.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[29] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–8.