EVERY RATING MATTERS: JOINT LEARNING OF SUBJECTIVE LABELS AND INDIVIDUAL ANNOTATORS FOR SPEECH EMOTION CLASSIFICATION

Huang-Cheng Chou, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

ABSTRACT

Emotion perception is subjective and vary with respect to each individual due to the natural bias of human, such as gender, culture, and age. Conventionally, emotion recognition relies on the consensus, e.g., majority of annotations (hard label) or the distribution of annotations (soft label), and do not include rater-specific model. In this paper, we propose a joint learning methodology that simultaneously considers the label uncertainty and annotator idiosyncrasy using hard and soft emotion label annotation accompanying with individual and crowd annotator modeling. Our proposed model achieves unweighted average recall (UAR) 61.48% on the benchmark emotion corpus. Further analyses reveal that emotion perception is indeed rater-dependent, using the hard label and soft emotion distribution provides complementary affect modeling information, and finally joint learning of subjective emotion perception and individual rater model provides the best discriminative power.

Index Terms— speech emotion recognition, BLSTM, annotator modeling, soft label learning

1. INTRODUCTION

Human behaviors are affected by many factors, e.g., society types [1] and emotional states [2]. Emotional states are one of the most influential factors on human behaviors and affect our decision-making [3] and even social circle [4]. Recently, the rapid evolution in artificial intelligence (AI) algorithms have enabled automated technologies reaching human-level performances; examples can be found in face verification [5] and fetal cardiac analysis [6]. Algorithmic development has also been observed in the field of affective computing, i.e., designing robust automated emotion recognition system. The use of emotion sensing technologies has been key to advance multiple modern human-centered applications in our life, e.g., healthcare [7], and commercial applications [8].

Conventional emotion recognition systems rely using the majority vote from multiple annotators as the ground truth to train the emotion recognizer. This ground truth is also called the *hard* emotion label. However, many factors would affect emotion perception, such as culture, gender, and age. Emotion annotation can naturally have disagreement and be ambiguous [9]. Hence, the hard label used in learning the classifier loses not only the variability of annotations but also the subjectivity in the emotion perceptual process. Recently, researchers have proposed to use soft label (distributional representation instead of single hard assignment) to characterize the *blended* emotion perception [10]. Further, Kobashikawa et al. devised a DNN-based model using soft emotion label as ground truth, which obtains a significantly improved performances on the benchmark IEMOCAP database [11].

While soft labeling approach provides better flexibility in characterizing the variability of emotion perception, using soft label as ground truth still ignores the idiosyncrasies of individual annotator due to its generation of label distribution by pooling over all annotators. However, individual difference in emotion perception has been linked to neuro-perceptual mechanism, e.g., the magnitude of amygdala activity [12]. It adds to the subjectivity in the rating, in fact, annotator modeling has recently also received attention, e.g., Kim et al. used agreement and disagreement annotations of each speaker in order to weight the training instances to learn speaker-dependent model [13]. Han et al. proposed a model to estimate perception uncertainty using the inter-rater disagreement level to improve the performance of continuous dimensional emotion tracking [14]. Guan et al. proposed crowd layers in their network architecture to model experts individually, where each individual model weight is averaged to perform ensemble recognition [15].

Motivated by these research, we propose a network architecture to perform speech emotion classification by simultaneously leveraging the label uncertainty and annotator idiosyncrasy through joint learning from hard assignment and soft emotion label distribution accompanying with individual and crowd annotator modeling - *making every rating counts*. Our proposed model obtains an improved four emotion label classification accuracy to 61.48% due to its enhanced modeling capacity by including multi-views modeling of annotators and variability of annotations. The rest of paper is organized as follows: section 2 describes our database and methodology, section 3 includes experimental setup and results, and we finally conclude with future work.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

Table 1. The number of hard label (single annotation) and soft label (two or more annotations) utterance for each model and the annotation distribution (ratio)

The number of soft and hard label utterance for each model				Annotation distribution (ratio)			
Model	Total	Soft label	Hard label	Neutral	Anger	Happiness	Sadness
$Crowd_H$	5531	0	5531	30.88%	19.94%	29.58%	19.60%
$Crowd_S$	7774	3185	4589	29.33%	17.77%	35.79%	17.10%
E_1	5954	44	5910	8.49%	21.21%	49.67%	20.64%
E_2	7845	38	7807	22.45%	26.58%	31.35%	19.62%
E_4	6429	212	6217	52.88%	12.41%	23.76%	10.95%
E_5	422	3	419	69.88%	15.29%	8.94%	5.88%
E_6	773	20	753	26.73%	15.76%	43.38%	14.22%

2. RESEARCH METHODOLOGY

Figure 1. depicts our overall framework used in this work. The core idea is to joint training with multiple models, where each one learns from a different emotion view point. In total, there are three basic core components: 1) learning from majority vote hard labels, 2) learning from soft label derived by pooling all annotators, 3) learning from individual annotator separately. These models are then finally concatenated to learn the final emotion recognition. The building block used within each model is based on the structure proposed in [16], which consists of initial dense layer, then a bidirectional long short term memory network with attention mechanism, and a final dense layer (BLSTM-DNN).

2.1. IEMOCAP Database

In this work, we use the IEMOCAP database [17]. It contains 12 hours of audio-video recordings of dyadic interactions with 10 different speakers split in pair over 5 sessions in English. There are 10039 utterances in the database that has been given emotion labels by 3 or 4 annotators. 12 unique raters annotate the database by choosing labels out of the 9 possible emotional labels per utterance (not restricted to single choice) - an example of annotation is given in Table 2. Emotional labels are happiness, anger, sadness, neutral, fear, surprise, frustration, excitement, and others.

To compare with other state of the arts, we choose the same evaluation data, which means each data is labeled with a single emotion state using the majority votes of 3 or 4 annotators. In this work, we concentrate on performing emotion recognition on four emotion classes (happiness, anger, sadness, and neutral). We merge the happiness and excitement categories as happiness. This includes a total of 5531 number of data samples to use in our recognition evaluation, and this particular setup is similar to past works in utilizing IEMOCAP as a benchmark dataset. The emotion classes distributions of utterances are: happiness: 29.58%, anger: 19.94%, sadness: 19.60%, and neutral: 30.88%. Out of the 12 annotators, only 5 annotators (indexed as E_1, E_2, E_4, E_5, and E_6) annotate enough utterances over all 5 sessions. Hence, for the *individual* annotator component, we only build models for these five annotators.

Table 2. Exemplary complete annotations of an utterance

Utterance name	Ses03M impro06 M027	
Annotator	$Annotation_1$	$Annotation_2$
E_1	Sadness	
E_2	Sadness	
E_4	Sadness	Anger
M_3	Sadness	Anger

2.2. Emotion Classification Framework

2.2.1. Acoustic Features

We extract frame-level utterance acoustic features using the openSMILE toolbox [18]. Emobase.config is used to extract 45 dimensional acoustic features including 12 dimensional Mel-Frequency Cepstral Coefficients (MFCCs), loudness, fundamental frequency (F0), voice probability, zero cross rate, the first derivatives of them, and the second derivatives of MFCCs and loudness. All features are extracted at 60ms fame length size and 10ms frame step size. They are further normalized to each speaker using z-score normalization, and then downsample by taking the average values of every 5 frames.

2.2.2. Learning Targets

All of our models are built based on the BLSTM-DNN structure as proposed in the previous work [16]. In this work, our aim is to account for both the variability in emotion perception and the subjective nature of individual annotator. To handle the variability in the emotion perception, we train BLSTM-DNN with two different learning targets: hard labels and soft labels.

Hard label, i.e., giving a single ground truth, as the learning target is the most natural and conventional way of train-

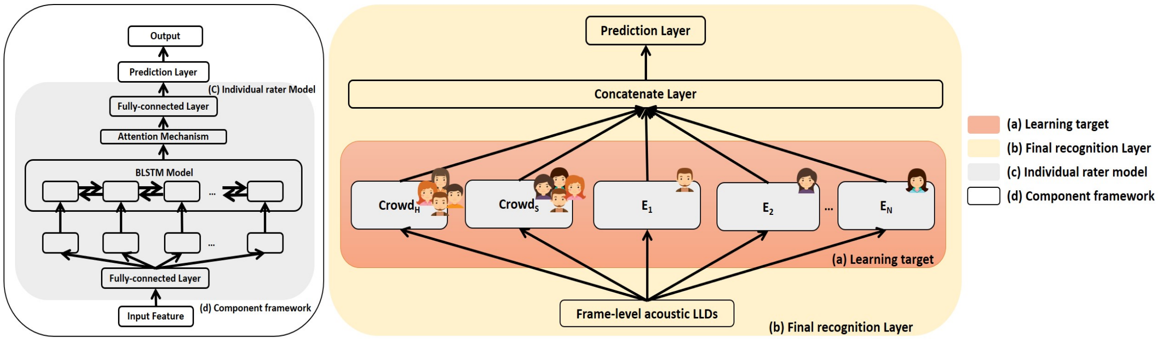

Fig. 1. (a) Learning target (b) Final recognition Layer (c) Individual rater model (d) Component framework

ing an emotion recognizer. To account for the fact that an utterance could be a *blended* emotion, Fayek et al. is the first to propose the use of soft labels to model the subjectivity of emotional annotations by integrating inter-annotator variability [10]. In addition, Kobashikawa et al. further presents an improved soft-target approach [11] defined as follows:

$$q(c_k) = \frac{\alpha + \sum_n h_k^{(n)}}{\alpha K + \sum_{k'} \sum_n h_{k'}^{(n)}} \quad (1)$$

where $q(c_k)$ is the reference class distribution, $h_k^{(n)}$ is the binary label-existence which is 1 if the n-th annotator gives class label c_k , and α is the smoothing coefficient. The modified soft label is equal to [10] if α=0.

In this paper, we set α=0.75 in this modified soft label form to be the same as used in [11].

2.2.3. *Annotator Modeling*

Everyone has different emotion perception toward the same utterance due to the nature of subjectivity and also individual idiosyncrasy. We further incorporate annotator modeling into our recognition architecture. We define two types of annotators: $Crowd$ (note crowd means the annotators in the used database) and E. $Crowd$ pools all of the annotator's annotation for each utterance, and E only learns from each individual rater. For $Crowd$, we will obtain two different models depending on whether the learning target is set to be hard label or soft label (Section 2.2.2). Moreover, we use soft label for all of our E, i.e., training a model for each individual annotator on the data they have annotated (note the data amount will be different for each annotator). Table 1 summarizes the total number of data points available for each model training and also shows the relative distribution of each emotion classes within that particular data cohort.

2.2.4. *Final Recognition Layer*

Once we obtain all of the model components, i.e., two $Crowd$ models and five E models, we save and freeze weights of each model. Finally, we concatenate the last layer representation before softmax of every BLSTM-DNN model and add an additional softmax layer to perform the final four class emotion prediction. The complete structure is illustrated in Figure 1.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

Our basic building block is BLSTM-DNN with attention. This model contains two dense layers (fully-connected layer) with Rectified Linear Unit (ReLU) activation function, one BLSTMs with attention layer and finally one dense layer with softmax function. The number of hidden units is 256 in the first dense layer, 128 in BLSTMs with attention layer, 256 in the last dense layer, and 4 in the last layer with softmax function. We include a dropout layer for all layers with 50% drop out rate.

We conduct our experiments using leave-one-session-out cross validation using the metric of unweighted average recall (UAR). The other hyperparameter, i.e., batch size, learning rate, and epochs, for each model is grid searched within the range of [8, 16, 32, 64, 128], [1e-6, 1e-5, 1e-4, 1e-3, 1e-2], [10, 20, 30, 40, 50, 100, 200], respectively. These parameters are chosen with early stopping criteria in all conditions to minimize cross entropy on the validation set. The optimizer used in this work is ADAMMAX [19].

3.1.1. *Models Comparison*

We further compare different performances obtained for each of components of our complete architecture:

- $Crowd_H$ model baseline:
  This model is essentially the same as the one proposed in that uses BLSTM-DNN learns from hard label [16].
- $Crowd_S$ model baseline:
  This model uses soft label training, which was proposed by Ando et al. in order to effectively leverage all annotated utterances [11].
- $Crowd_H$ and $Crowd_S$ fusion model baseline:
  This model leverages all $Crowd$ information by con-

catenating representation of both $Crowd_H$ and $Crowd_S$ to be fed into final softmax.

- E_* model:
  Each of the E_* model trains from using soft label learning on individual annotator.

- Proposed Model:
  This model is our final proposed model that leverages all $Crowd$ and E information by concatenating representation of both $Crowd_H$ and $Crowd_S$ to be fed into final softmax.

3.2. Experimental Results and Analyses

Table 3 shows a summary of the complete recognition performance over all comparison models. Our proposed framework model obtains the best overall emotion classification accuracies (61.48% UAR). This method surpasses previously best method of $Crowd_H$ [16] and $Crowd_S$ [11] by 3.18% and 4.36% absolute, respectively. Our results demonstrate the importance in modeling the subjective annotators' emotional information to further improves emotion classification results over the state-of-the-art.

One important observation is that when comparing between hard label and soft label learning, $Crowd_S$ obtains a better recognition rate for happiness compared to $Crowd_H$, where $Crowd_H$ works better for neutral and sadness. The complementary nature of $Crowd_H$ and $Crowd_S$ makes the integration critical in achieving further improved emotion recognition results. Happiness has been a challenge emotion class to recognize. Learning from soft label help improve this particular class may further indicate that happiness is a more distributed manifestation in the acoustic space as compared to other emotion state, e.g., anger and sadness.

Furthermore, while the individual model by itself does not obtain high recognition rates, most likely due to the biased view of an individual rater and the unevenly-distributed emotion class data for each annotator. For example, Table 3 shows that E_1 model has low recognition on neutral category but has good accuracy on happiness emotion, but the phenomenon is reversed for E_5 model. When examining Table 1's (left) emotion distribution for each model, it seems to be related to amount on the type of emotional data that each annotator has annotated. By explicitly including each annotator's model directly at the representation-level, our proposed method can learn to integrate multiple complementary information from each distinct individual view point of the emotion perception. Another point to raise is that, due to the use of individual rater model, we are capable of expanding the set of data used in training the recognizer as compared to conventional hard label approach, where the training data only comes from the set where there is consensus.

Table 3. Results on the IEMOCAP database

Model	Overall	Neutral	Anger	Happiness	Sadness
$Crowd_H$[16]	57.45%	55.71%	63.29%	45.02%	65.77%
$Crowd_S$[11]	57.12%	49.70%	62.98%	62.85%	53.14%
E_1	50.98%	8.04%	61.31%	77.24%	57.34%
E_2	59.68%	38.78%	64.35%	64.25%	62.61%
E_4	48.59%	81.29%	45.42%	38.20%	29.44%
E_5	37.62%	86.89%	47.62%	11.21%	4.75%
E_6	45.82%	36.85%	40.10%	60.39%	45.95%
$Crowd_{HS}$	58.58%	59.66%	59.31%	53.63%	61.71%
$Proposed$	**61.48%**	54.55%	64.51%	60.32%	66.56%

4. CONCLUSIONS AND FUTURE WORKS

The subjectivity and variability exist in the human emotion perception differs from person to person. In this work, we propose a framework that models the majority of emotion annotation integrated with modeling of subjectivity in improving emotion categorization performances. Our method achieves a promising accuracy of 61.48% on a four-class emotion recognition task. To the best of our knowledge, while there are many works in studying annotator subjectivity, this is one of the first works that have explicitly modeled jointly the *consensus* with *individuality* in emotion perception to demonstrate its improvement in classifying emotion in a benchmark corpus.

In our immediate future work, we will evaluate the proposed framework on other public large-scaled emotional database with multiple annotators, e.g., NNIME [20], to further justify its robustness. We also plan to extend our framework to includ other behavior attributes, e.g., lexical content and body movements. Furthermore, the subjective nature of emotion perception has been shown to be related to the rater personality [21], a joint modeling of rater's characteristics with his/her subjectivity in emotion perception may lead to further advancement in robust emotion recognition.

5. REFERENCES

[1] P Wesley Schultz and Lynnette C Zelezny, "Values and proenvironmental behavior: A five-country survey," *Journal of cross-cultural psychology*, vol. 29, no. 4, pp. 540–558, 1998.

[2] Sigal G Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," *Administrative Science Quarterly*, vol. 47, no. 4, pp. 644–675, 2002.

[3] Antoine Bechara, "The role of emotion in decision-making: evidence from neurological patients with orbitofrontal damage," *Brain and cognition*, vol. 55, no. 1, pp. 30–40, 2004.

[4] Susan Shott, “Emotion and social life: A symbolic interactionist analysis,” *American journal of Sociology*, vol. 84, no. 6, pp. 1317–1334, 1979.

[5] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[6] Weilin Huang, Christopher P Bridge, J Alison Noble, and Andrew Zisserman, “Temporal heartnet: towards human-level automatic analysis of fetal cardiac screening video,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 341–349.

[7] Min Chen, Yin Zhang, Yong Li, Mohammad Mehedi Hassan, and Atif Alamri, “Aiwac: Affective interaction through wearable computing and cloud technology,” *IEEE Wireless Communications*, vol. 22, no. 1, pp. 20–27, 2015.

[8] Songfan Yang, Le An, Mehran Kafai, and Bir Bhanu, “To skip or not to skip? a dataset of spontaneous affective response of online advertising (sara) for audience behavior analysis,” in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. IEEE, 2015, vol. 1, pp. 1–8.

[9] Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, “Interpreting ambiguous emotional expressions,” in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–8.

[10] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon, “Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels,” in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 566–570.

[11] Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima, and Yushi 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.

[12] Heather C Abercrombie, Stacey M Schaefer, Christine L Larson, Terrence R Oakes, Kristen A Lindgren, James E Holden, Scott B Perlman, Patrick A Turski, Dean D Krahn, Ruth M Benca, et al., “Metabolic rate in the right amygdala predicts negative affect in depressed patients,” *Neuroreport*, vol. 9, no. 14, pp. 3301–3307, 1998.

[13] Yelin Kim and Emily Mower Provost, “Leveraging inter-rater agreement for audio-visual emotion recognition,” in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 553–559.

[14] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller, “From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty,” in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 890–897.

[15] Melody Y Guan, Varun Gulshan, Andrew M Dai, and Geoffrey E Hinton, “Who said what: Modeling individual labelers improves classification,” *AAAI’18. AAAI Press, 2018.*, 2018.

[16] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[17] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[18] Florian Eyben, Martin Wöllmer, and Björn 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.

[19] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations*, 2015.

[20] Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee, “Nnime: The nthu-ntua chinese interactive multimodal emotion corpus,” in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 292–298.

[21] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe, “Ascertain: Emotion and personality recognition using commercial sensors,” *IEEE Transactions on Affective Computing*, , no. 2, pp. 147–160, 2018.