INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speech Emotion Recognition based on Multi-Label Emotion Existence Model

Atsushi Ando, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, Yushi Aono

NTT Media Intelligence Laboratories, NTT Corporation, Japan

atsushi.ando.hd@hco.ntt.co.jp

Abstract

This paper presents a novel speech emotion recognition method that addresses the ambiguous nature of emotions in speech. Most conventional methods assume there is only a single ground truth, the dominant emotion, though utterances can contain multiple emotions. In order to solve this problem, several methods that consider ambiguous emotions (e.g. soft-target training) have been proposed. Unfortunately, training them is difficult since they work by estimating the proportions of all emotions. The proposed method improves both frameworks by evaluating the presence or absence of each emotion. We expect that it is much easier to estimate just presence/absence of emotions rather than trying to determine proportions of each, and the deliberate assessment of emotion existence information will help to estimate the proportion of each or dominant class more precisely. The proposed method employs two-step training. Multi-Label Emotion Existence (MLEE) model is trained first to estimate whether each emotion is present or absent. Then, the dominant emotion recognition model with hard- or soft-target labels is trained by means of the intermediate outputs of the MLEE model so as to utilize cues of emotion existence for inferring the dominant. Experiments demonstrate that the proposed method outperforms both hard- or soft-target based conventional emotion recognition schemes.

Index Terms: emotion recognition, multi-label classification, convolutional neural network (CNN), spectrograms

1. Introduction

Automatic emotion recognition (AER) is an important technology for better understanding of human communication. There will be many applications such as voice-of-customer analysis in contact center calls [1], human-like responses in spoken dialog systems [2] and driver state monitoring [3]. The key requirement in realizing AER-based applications is to identify the dominant emotion of the speaker in an utterance. The aim of this paper is categorical emotion recognition from speech.

A large number of emotion recognition methods have been investigated. The traditional approaches are based on utterance-level heuristic features including the statistics of frame-level acoustic features such as pitch, power and Mel-Frequency Cepstral Coefficients (MFCC) [4, 5]. However, it is difficult to create truly effective features because emotional cues exhibit great diversity. In contrast with these approaches, recent studies employ Deep Neural Networks (DNNs) to learn emotion-related features automatically. Their works include DNN-based classifiers and frame-level inputs [6, 7]. Recurrent Neural Networks (RNNs) have been employed as the classifier to allow use of local characteristic for recognition [8, 9]. Attention mechanism has also been utilized to focus on specific regions of an utterance [9, 10]. It has been reported that low-level signals such as raw waveforms or frequency-domain spectrograms are suitable as the input because they have rich information [11, 12]. Therefore, some of the latest works process spectrograms in Convolutional Neural Networks (CNNs) [10, 11, 13–15].

One of the problems existing research is that the ambiguous nature of emotion is seldom considered. Most conventional methods employ hard-target labels, i.e. the dominant emotion label is taken as the one ground truth. However, some studies posit that several minor emotions may be present in some utterances [16, 17]. This ambiguity will hinder the training of the dominant emotion recognition model.

Several studies have considered emotion ambiguity in dominant emotion recognition. The typical approach is based on soft-target labels as they reflect the frequency of the annotations [18, 19]. However, soft-target training forces the model to estimate the proportion of each emotion in an utterance, which will greatly complicate learning since all the intensities of emotions must be evaluated properly. Another is specifying secondary emotions in addition to the dominant emotion through the multi-task learning framework [20]. The secondary emotions are those that more than half of the annotators perceived as minor emotions. Though it offers better performance than attempting just dominant emotion recognition, it is unclear how well the ambiguity of emotions are utilized because multi-task learning indirectly transfers knowledge of secondary emotions.

In this paper, we present a novel dominant emotion recognition method that improves on conventional hard-/soft-target based methods by directly handling the ambiguity of emotions. The key idea of the proposed method is to evaluate the presence or absence of individual emotions in dominant emotion recognition. We hypothesize that it is much easier to estimate the presence/absence of emotions than their proportions, and determination of emotion existence will help to evaluate their proportions. The proposed method introduces the new task of estimating the presence/absence of emotions, called Multi-Label Emotion Existence (MLEE); it employs two-step training. MLEE model is trained first to estimate whether each emotion is present or absent. Then a dominant emotion recognition model with hard-/soft-target labels is trained by the intermediate outputs of MLEE model so as to utilize knowledge of the presence/absence of each emotion. The main strength of the proposed method is that it can directly utilize knowledge of multiple emotions in dominant emotion recognition. Another advantage is that utterances that have no dominant target emotions can be used for MLEE training since they are regarded as negative samples in the MLEE task. Experiments demonstrate that the proposed method improves both hard- and soft-target based dominant emotion recognition performance.

This paper is organized as follows. Section 2 briefly introduces conventional emotion recognition methods. The proposed method, based on modeling multi-label emotion existence, is shown in Section 3. Evaluation experiments are reported in Section 4 and the conclusion is given in Section 5.

http://dx.doi.org/10.21437/Interspeech.2019-2524

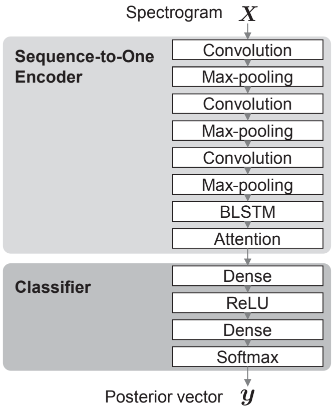

Figure 1: *An example of the structure of the conventional model.*

2. Conventional Methods

2.1. Emotion Recognition based on Spectrograms

This section describes a dominant emotion recognition method based on spectrograms with CNN-based model [10,13–15]. The advantage of this method is that the model can utilize the rich information contained in spectrograms.

Let $\boldsymbol{X}$ be the spectrogram of an utterance and $c \in \{c_1, \cdots c_K\}$ be the dominant emotion of the utterance, where K is the total number of target emotion classes. Estimated emotion $\hat{c}$ is obtained when the model evaluates posterior probabilities $p(c \mid \boldsymbol{X}, \boldsymbol{\theta})$,

$$\hat{c} = \arg\max_{c} p(c \mid \boldsymbol{X}, \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ is a set of model parameters.

An example of the model structure is shown in Figure 1. The model consists of CNN, Bidirectional Long Short-Term Memory (BLSTM), attention and dense (fully-connected) layers. CNNs with pooling layers are stacked to aggregate the local characteristics of the input spectrograms. BLSTMs handle CNN output sequences to learn forward/backward temporal behaviors from arbitrary length inputs, and the attention layer focuses on particular regions. These structures are regarded as an encoder that generates fixed-length emotion-related features from variable-length inputs. The output of the encoder is used to evaluate posterior probability vector $\boldsymbol{y} = [p(c_1 \mid \boldsymbol{X}, \boldsymbol{\theta}), \cdots, p(c_K \mid \boldsymbol{X}, \boldsymbol{\theta})]^\top$ by dense layers. Note that some proposals did not employ attention [14] or LSTM layers [10], but we regard them as being conventional because they are reported to be effective in handling variable-length inputs [9, 15].

In the training step, the model parameters are updated by the loss function $\mathcal{L}$ based on softmax cross entropy,

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{k=1}^{K} q(c_k) \log p(c_k \mid \boldsymbol{X}, \boldsymbol{\theta}), \tag{2}$$

where $q(c_k)$ is the reference class distribution. Hard targets, which means that one annotated emotion class is dominant and the rest are zero, are widely used:

$$q(c_k) = \begin{cases} 1 & \text{if } k = \arg\max\limits_{k} \dfrac{\sum_n h_k^n}{\sum_{k'} \sum_n h_{k'}^n} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where h_k^n is binary label-existence which is 1 if the n-th annotator gives class label c_k, otherwise 0. Utterances that have no dominant target emotion are excluded from the training data.

2.2. Soft-target Training

One of the problems of the conventional model with hard targets is that it ignores minority emotions. To solve this problem, soft-target labels which represent annotation frequency are used as the reference ground truth. This paper uses smoothed soft-target labels [18] as follows:

$$q(c_k) = \frac{\alpha + \sum_n h_k^n}{\alpha K + \sum_{k'} \sum_n h_{k'}^n} \tag{4}$$

where α is the smoothing coefficient of the label. Soft-targets enable the knowledge of minority emotions to be used in training. Another advantage is that they allow the use of utterances that have no dominant target emotion but at least one of the target emotions is present, which increases the amount of training data available. However, soft-target assumes the existence of class distributions, i.e. intensities of all target emotions, which will complicates training.

3. Proposed Method

This section introduces a new emotion recognition method based on Multi-Label Emotion Existence (MLEE). MLEE estimates the existence probabilities of individual emotions, which significantly enhances the determination of the dominant emotion. Three types of dominant emotion recognition methods are presented in Figure 2.

3.1. Multi-Label Emotion Existence (MLEE)

The proposed method, MLEE, introduces a new target variable e which represents emotion existence. $e \in \{e_0, e_1\}$ and e_0 means that the emotion does not surely appear in the utterance, while e_1 means the emotion does exist and should be assessed so as to decide the dominant emotion. The reference distribution of existence of particular emotion $q(e_1|c_k)$ is automatically defined by multiple annotations,

$$q(e_1 \mid c_k) = \begin{cases} 1 & \text{if } \exists n, h_k^n = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

MLEE uses this model and a loss function for multi-label classification. This paper employs the same structure as the conventional model except that the final layer is replaced with a sigmoid activation function as the multi-label model. The loss function is class-wise binary cross entropy, which is common in multi-label classification,

$$\begin{aligned} \mathcal{L}_e(\boldsymbol{\theta}_e) = -\sum_{k=1}^{K} \{ & q(e_1 \mid c_k) \log p(e_1 \mid c_k, \boldsymbol{X}, \boldsymbol{\theta}_e) \\ & + (1 - q(e_1 \mid c_k)) \log(1 - p(e_1 \mid c_k, \boldsymbol{X}, \boldsymbol{\theta}_e)) \}, \end{aligned} \tag{6}$$

where $p(e_1 \mid c_k, \boldsymbol{X}, \boldsymbol{\theta}_e)$ is the existence probability of class c_k in the input utterance, and is an element of MLEE output, $\boldsymbol{y}_e = [p(e_1 \mid c_1, \boldsymbol{X}, \boldsymbol{\theta}_e), \cdots, p(e_1 \mid c_K, \boldsymbol{X}, \boldsymbol{\theta}_e)]^\top$. $\boldsymbol{\theta}_e$ is a set of MLEE parameters.

Main strength of MLEE is that it learns the feature extractor of existences of individual emotions, which can directly enhance dominant emotion recognition performance. Another advantage is that all the utterances in an annotated corpus can be

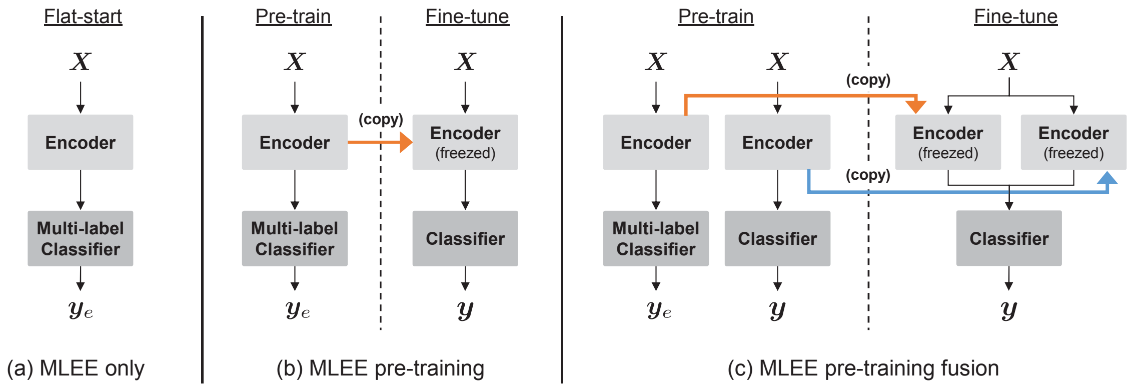

Figure 2: *Overviews of the proposed methods based on Multi-Label Emotion Existence (MLEE) model.*

used for training. The conventional method (hard/soft-target) can only use the utterances that contain at least one target emotion annotation. However, MLEE can learn from even those utterances that contains no target emotions because they work as negative samples in emotion existence learning. This increase of training data yields better estimation model.

3.2. Dominant Emotion Recognition based on MLEE

3.2.1. MLEE only

MLEE outputs the existence probabilities of individual classes. Thus it can be directly used for dominant emotion recognition by selecting the highest probability,

$$\hat{c} = \arg\max_{c_k} p(e_1 \,|\, c_k, \boldsymbol{X}, \boldsymbol{\theta}_e). \tag{7}$$

Note that there is no guarantee of the correct dominant emotion class being selected by this criteria because $p(e_1 \,|\, c_k, \boldsymbol{X}, \boldsymbol{\theta}_e)$ just represents the appearance of each class, not its strength relative to others.

3.2.2. MLEE pre-training

MLEE determines the dominant emotion in two steps; estimating the existence of individual emotions in the utterance, then the emotions present are compared to select the dominant one.

Two-step training with MLEE is employed for this strategy as shown in Figure 2(b). First, MLEE model is trained from scratch to obtain the encoder of emotion-existence-related features, in a pre-training operation. The dominant emotion recognition model is the fine-tuned with MLEE encoder outputs to learn the definition criteria of dominant emotion decision. In this step, MLEE encoder parameters are frozen and only the classifier part (dense layers and activation functions) of the model is updated to retain the knowledge provided by the emotion-existence feature extractor.

3.2.3. MLEE pre-training fusion

The third takes the same approach for estimating the dominant emotion. The difference is that this method employs not only the pre-trained MLEE encoder but also the encoder of dominant emotion estimation. This is because these two encoders provide different bits of information about emotional cues, and both are useful for identifying the dominant emotion.

Figure 2(c) overviews MLEE pre-training fusion. Both dominant emotion estimation and MLEE models are trained independently in pre-training. The encoders of these models are used as parallel encoders for the dominant emotion recognition model. The outputs of the two encoders are concatenated to yield the single vectors that are input to the classifier part. As in MLEE pre-training, the two encoders are frozen in the fine-tuning step.

Table 1: *Ratio of emotion existence in each dataset.*

	neu	*hap*	*sad*	*ang*
dominant	58.2%	15.4%	30.5%	15.9%
all	45.2%	24.0%	19.8%	17.8%

4. Experiments

4.1. Setup

The experiments used the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [21]. It has approximately 12 hours of dialog speech performed by 10 speakers (5 male and 5 female). The recording format was 16 kHz, 16 bit linear-PCM. All utterances were divided and labeled by three annotators into 10 emotional categories. The dataset contains scripted and improvised sessions and we used only the improvised speech for all experiments because the scripted data may depend on linguistic cues or contextual information. Four target emotions were considered: neutral (*neu*), happy (*hap*), sad (*sad*) and angry (*ang*). This configuration follows previous work [10,14,15]. This process yielded 4784 speech utterances. Numbers of dominant emotions in the utterances were 1099 *neu*, 284 *hap*, 608 *sad*, 289 *ang* and 2504 other emotions (surprise, excited, no-dominant emotion, etc.).

Leave-one-speaker-out cross validation was conducted for all evaluations. In each validation run, 8 speakers were used for training, one speaker was used for development, and the remaining speaker was used in the test. To increase the amount of training data, we performed data augmentation by means of speed perturbation with speed factors of 0.9, 1.1 [12]. Furthermore, we prepared two training dataset: *dominant* and *all*. *dominant* data were utterances that contained one dominant emotion. *all* included the utterances whose dominant emotion was others of the four targets, in addition to *dominant* set. This was because MLEE permits the use of all utterances for training, as mentioned in Section 3.1. The ratios of emotion existence in each dataset are shown in Table 1.

Table 2: *Network architectures of emotion recognition model.*

	Layer-type	Parameters
Encoder	CNN	16 ch, [12×16] kernel, [2×2] stride
	CNN	24 ch, [8×12] kernel, [1×1] stride
	CNN	32 ch, [5×7] kernel, [1×1] stride
	BLSTM	1 layer, 128 dim.
	attention	self-attention [22], 1 head
Classifier	Dense	1 layer, 64 dim.
	Dense	1 layer, 4 dim.

The conditions used in extracting spectrograms followed those of conventional studies [14, 15]. Frame length and frame shift length were 40 ms and 10 ms, respectively. The window type was Hamming window. DFT length was 1600 (10Hz grid resolution) and we used 0-4 kHz frequency range, which yielded 400 dimensional log power spectrograms. All the spectrograms were z normalized using the mean and variance of the training dataset.

The baseline was a dominant emotion recognition model [14] with hard-target or soft-target labels [18]. In *all* dataset training, the utterances with no dominant target emotion were eliminated in the hard-target case, while those with no target emotions were also discarded in soft-target. The structure of the baseline was that shown in Figure 1; the parameters are shown in Table 2. Each CNN layer was followed by batch normalization [23], rectified linear activation function and 2×2 max pooling layers. Early stopping was performed using development set loss as the trigger. Optimization method was Adam [24] with learning ratio of 0.0005. In the training step, inverse values of the class frequencies were used as class weights to mitigate the class imbalance problem [25]. Minibatch size was 16. Soft-target smoothing coefficient was 0.75.

The proposed methods were MLEE only, MLEE pre-training and MLEE pre-training fusion. The model structure and parameters of MLEE were same as the baseline except for replacing output activation function with the sigmoid function. In the fine-tuning step of MLEE pre-training and MLEE pre-training fusion, teacher labels of the dominant emotion estimation model were either hard or soft-target. The learning ratio was 0.0002 in flat-start or pre-training of MLEE, but 0.0001 in fine-tuning and 0.00005 in fine-tuning fusion. The other training conditions were same as for the baseline. The inverse values of existent/non-existent label frequencies in each emotion class were used as class weights for multi-label loss. Both the baseline and the proposed methods were implemented by PyTorch [26].

Two evaluation metrics common in emotion recognition research were employed; weighted accuracy (WA) and unweighted accuracy (UA). WA is the classification accuracy of all utterances and UA is the average of individual emotion class accuracies.

4.2. Results and Discussions

Recognition accuracies are shown in Table 3. Note that for the dominant emotion estimation model with hard-target labels, results of *dominant* and *all* training set were the same because the valid utterances were fully matched.

Table 3: *Comparison of weighted accuracy (WA) and unweighted accuracy (UA). 'label' is the target of dominant emotion recognition model.*

		Train set			
		dominant		*all*	
	label	WA	UA	WA	UA
Dominant recog.	hard	59.5	61.6	59.5	61.6
	soft	62.5	62.9	64.4	63.9
MLEE only	-	**66.3**	60.9	64.4	64.0
MLEE pre-train	hard	62.7	62.6	65.6	63.0
	soft	64.3	62.3	64.6	64.4
MLEE pre-train fus.	hard	63.4	62.9	65.1	63.9
	soft	64.4	**63.7**	**66.1**	**65.4**

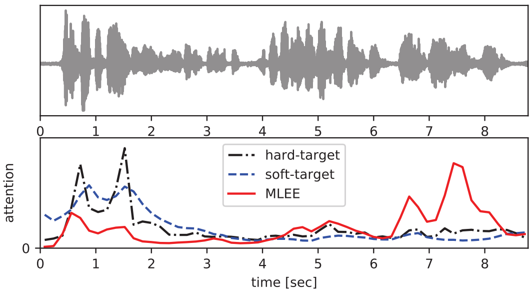

Figure 3: *An example of attention-layer outputs.*

Comparing the two label types in the dominant emotion estimation model, soft-target showed better accuracy in both *dominant* and *all* training dataset. It is considered that soft-targets can utilize the utterances that have minor target emotions for training, which improves overall performance. On the other hand, the MLEE yielded comparable performances to soft-target training in both dataset. This indicates that the MLEE model has the ability to handle the ambiguity of emotions properly as soft-targets. MLEE pre-training outperformed the dominant emotion recognition model with both hard and soft target labels. It means that the information of emotion existence provided by MLEE is effective for dominant emotion recognition. Furthermore, MLEE pre-training fusion achieved the highest performance. The dominant emotion estimation model and emotion existence estimation model will extract different features, and both are useful in identifying the dominant emotion. To confirm this we focused on the attention weights of the encoders because they reflect the regions of interests in each model. An example of the attentions by the dominant model with hard/soft-target labels and MLEE is shown in Figure 3. Some peaks, e.g. those from 0.5 to 2.0 second, were similar in all methods, while MLEE also focused on different areas. Though all methods consider local characteristics of emotions, MLEE can detect a wider variety of emotional cues.

5. Conclusions

This paper presented a new emotion recognition method that can handle the ambiguity of emotions properly. The key idea of the proposed method is introducing the estimation of multi-label emotion existence (MLEE) as an auxiliary task to support dominant emotion recognition. Three MLEE variants were examined: MLEE only, MLEE pre-training, and MLEE pre-training fusion. There were two strengths in MLEE; it can directly utilize knowledge of ambiguity in dominant emotion recognition, and even the utterances that have no dominant target emotions can be used. Experiments showed that the proposed methods outperformed a conventional dominant emotion recognition method. Future works include further evaluations by other corpora.

6. References

[1] L. Devillers, C. Vaudable, and C. Chastagnol, “Real-life emotion-related states detection in call centers: a cross-corpora study,” in *Proc. of INTERSPEECH*, 2010, pp. 2350–2353.

[2] J. C. Acosta, “Using emotion to gain rapport in a spoken dialog system,” in *Proc. of NAACL HLT Student Research Workshop and Doctoral Consorium*, 2009, pp. 49–54.

[3] A. Tawari and M. Trivedi, “Speech based emotion classification framework for driver assistance system,” in *Proc. of IEEE IVS*, 2010, pp. 174–178.

[4] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Communication*, vol. 53, pp. 1062–1087, 2011.

[5] I. Luengo, E. Navas, I. Hernàez, and J. Sànchez, “Automatic emotion recognition using prosodic parameters,” in *Proc. of INTERSPEECH*, 2005, pp. 493–496.

[6] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in *Proc. of INTERSPEECH*, 2014, pp. 223–227.

[7] J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in *Proc. of INTERSPEECH*, 2015.

[8] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in *Proc. of ICASSP*, 2017, pp. 2227–2231.

[9] C. W. Huang and S. S. Narayanan, “Attention assisted discovery of sub-utterance structure in speech emotion recognition,” in *Proc. of INTERSPEECH*, 2016, pp. 1387–1391.

[10] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, “An attention pooling based representation learning method for speech emotion recognition,” in *Proc. of INTERSPEECH*, 2018, pp. 3087–3091.

[11] P. Tzirakis, J. Zhang, and B. Schuller, “End-to-end speech emotion recognition using deep neural networks,” in *Proc. of ICASSP*, 2018, pp. 5089–5093.

[12] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, “Emotion identification from raw speech signals using DNNs,” in *Proc. of INTERSPEECH*, 2018, pp. 3097–3101.

[13] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, “Speech emotion recognition using spectrogram & phoneme embeddings,” in *Proc. of INTERSPEECH*, 2018, pp. 3688–3692.

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

[15] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, “Emotion recognition from variable-length speech segments using deep learning on spectrograms,” in *Proc. of INTERSPEECH*, 2018, pp. 3683–3687.

[16] J. Tao, Y. Li, and S. Pan, “A multiple perception model on emotional speech,” in *Proc. of ACII*, 2009, pp. 1–6.

[17] A. Ortony, G. L. Clore, and A. Colins, *The cognitive structure of emotions*. Cambridge University Press, 1988.

[18] 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 *Proc. of ICASSP*, 2018, pp. 4964–4968.

[19] H. M. Fayek, M. Lech, and L. Cavedon, “Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels,” in *Proc. of IJCNN*, 2016, pp. 566–570.

[20] R. Lotfian and C. Busso, “Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning,” in *Proc. of INTERSPEECH*, 2018, pp. 951–955.

[21] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[22] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” in *Proc. of ICLR*, 2017.

[23] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. of ICLR*, 2015, pp. 448–456.

[24] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. of ICLR*, 2015.

[25] C. Elkan, “The foundations of cost-sensitive learning,” in *Proc. of IJCAI*, 2001, pp. 973–978.

[26] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” in *Advances in NIPS*, 2017.