Integration of Annotator-wise Estimations for Emotion Recognition by Using Group Softmax

Yuuki Tachioka*
* Denso IT Laboratory, Tokyo, Japan
E-mail: tachioka.yuki@core.d-itlab.co.jp Tel: +81-3-6419-2315

***Abstract*—In emotion recognition, a major modeling difficulty arises from the different perceptions of emotion from annotator to annotator. Generally, it is common to use a one-hot (dominant) emotion label, which is obtained using majority voting by annotator-wise (minor) emotion labels. Previous studies show that the introduction of soft-target labels, which consider the frequency of annotator-wise labels, improves emotion recognition performance. However, these studies did not use minor emotion labels directly. Another study used multi-task learning to handle dominant and minor emotions independently, but this independent modeling is inappropriate because the two are closely related. We propose a sequential model composed of multiple annotator-wise classifiers and their majority voting to estimate dominant emotion. When using multiple classifiers, classifier imbalance, where the difficulty of classification is different from classifier to classifier, causes performance degradation. To address this classifier imbalance problem, we assign a group softmax to multiple annotator-wise classifiers. Experiments show that majority voting by estimated annotator-wise emotions improves the estimation performance for dominant emotions when compared with conventional methods that estimate dominant emotion directly. In addition, the proposed method is effective not only for speech emotion recognition but also for speech and text emotion recognition.**

I. Introduction

Automatic speech emotion recognition, which estimates speakers' emotions via voice, is important for many applications such as speaker state estimation [1] or appropriate dialog generation [2], [3]. Conventional methods have used handcrafted utterance-level features [4], [5] to classify emotion types, but the estimation accuracy is relatively low. The utterance-level feature does not provide enough information to estimate emotion because emotion is dynamic. Deep neural networks (DNNs) have improved performance [6] using frame-level features that capture a time sequence of speech features in detail. Essentially, speech emotion recognition utilizes a similar framework of automatic speech recognition, but they are different in two perspectives. First, the amount of available training data is much smaller than that for speech recognition because it costs more to collect and annotate training data containing various emotions. Second, different annotators perceive different emotions in the same speech, and one utterance can communicate multiple emotions, which is an important property of emotion recognition [7].

The latter is more serious than the former because additional expenditures can solve the first, but the latter is an essential problem of emotion perception and is difficult to solve. To address this variety of annotator-wise labels (minor emotion labels), it is common to use majority voting, which averages minor emotion labels to obtain a single dominant emotion label [8]. The terms 'dominant' and 'minor' are defined in [9]. Conventional studies have used this dominant emotion label as a one-hot target label [6], [10], whereas, to exploit minor emotion labels, some studies use soft-target labels that have non-zero values for multiple emotions instead of one-hot labels [11], [12]. Soft-target labels are obtained using the frequency of the annotations by all annotators. This considers mixed minor emotions, but this label ambiguity makes model training difficult. To address this problem, Ando et al. proposed adding a presence/absence binary decision for the target emotion before classification [9]. Other approaches, instead of estimating soft-target emotion labels, prepared multiple annotator-wise models to estimate each annotator's minor emotion label directly [13]. This model improved the estimation accuracy for dominant emotion labels by averaging annotator-wise estimated emotions instead of estimating dominant emotions directly; however, this method performs multiple inferences per annotator since each annotator requires its own model.

Another approach is to model dominant and minor emotions within the framework of multi-task learning [14]. This paper demonstrates the effectiveness of considering minor emotions that differ from dominant emotion. However, the dominant emotion is closely related to minor emotions because the dominant emotion label is derived from majority voting based on minor emotion labels; thus, it is inappropriate to model dominant and minor emotions independently.

To address these problems (multiple inference and independent modeling), we propose sequential modeling of annotator-wise minor emotions and dominant emotions. Our model can estimate annotator-wise emotions and dominant emotion sequentially using a one-time inference and does not need multiple inferences per annotator [13].

Generally, in the case of using multiple classifiers such as [13], [14], there is a classifier imbalance problem in which the difficulty of classification is different from classifier to classifier because of data distribution imbalance [15]. Our emotion recognition model also faces this classifier imbalance problem, as the difficulty of estimating annotator-wise labels is different from annotator to annotator. For example, some minor labels, which are similar to the dominant label, are easy to be recognized but other minor labels, which are very different from the dominant label, are difficult. In the field of image processing, to address this classifier imbalance problem, Li et

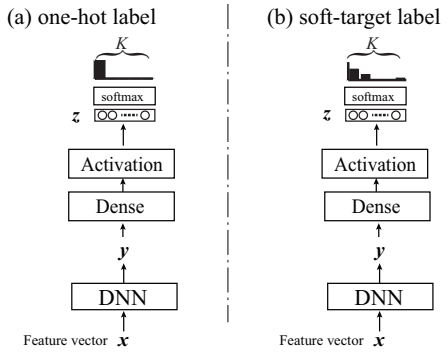

Fig. 1. Conventional emotion recognition model (one-hot label and soft-target label). 'Dense' is a single fully-connected layer.

al. [16] proposed a group softmax. For training, to reduce the influence of classifier imbalance, this method assigns different types of classification labels to different classifiers according to the difficulty of the classification. For testing, to obtain an estimation result for the original classification problem, the estimated results from different classifiers are reordered and rescaled. For this classifier imbalance in the annotator labels, we assign a group softmax to multiple annotator-wise classifiers. The dominant emotions are then obtained using a simple and weighted majority voting by the annotator-wise emotions.

This paper is organized as follows. Sec. II describes the conventional method that aims to estimate dominant emotion. Sec. III describes the proposed method, which models annotator-wise minor emotion and dominant emotion sequentially. In Sec. IV, speech emotion recognition experiments using an open dataset confirmed the effectiveness of the proposed method.

II. Conventional Emotion Recognition Model

Fig. 1 shows the baseline model for speech emotion recognition. There are K-types of dominant emotions to be classified. The input feature is $\boldsymbol{x}$, and the output from the DNN is $\boldsymbol{y}$, which is input into the dense layer for K-class classification. Fig. 1(a) depicts the most basic model, which estimates a one-hot target emotion [6]. A one-hot target label is obtained using majority voting by the annotator-wise minor emotion labels. Fig. 1(b) shows the sort-target label [9], [11], [12], which is based on the frequency of all annotator-wise labels. To convert a frequency to a probability, the total sum of the labels is normalized to unity.

III. Integration of Annotator-wise Estimation Results

A. Annotator-wise estimation using group softmax

Fig. 2 shows the proposed model structure. The DNN output $\boldsymbol{y}$ is branched into N annotator-wise classifiers, where the number of annotators is N. For these N annotator groups, the softmax operation is applied separately. There are N loss functions corresponding to each label to be optimized. For

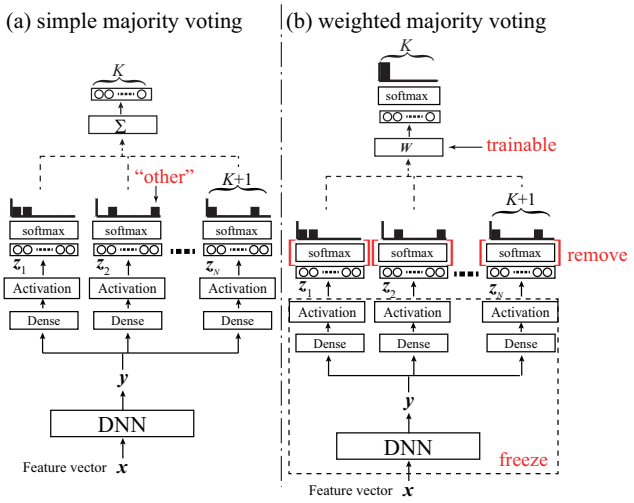

Fig. 2. Proposed sequential emotion recognition model employing annotator-wise classification and majority voting.

annotator-wise estimation, we need to process $(K+1)$-types of emotion labels because it is necessary to add "other" emotion type to the target K emotions. For training on dominant emotion labels, training data labeled with other emotions can be removed, but annotator-wise labels inevitably include emotions other than target emotions even if the dominant emotion is in the target. Thus, each annotator-wise output, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, ..., and $\boldsymbol{z}_N$, is associated with the $(K+1)$-dimensional labels. After group softmax activation [16], the model parameters are optimized using the cross-entropy criterion for each annotator.

B. All-data training

For conventional methods trained on the dominant labels, it is difficult to use speech with non-target emotions for training. In some cases, speech annotated as other emotions in the dominant label can be perceived as a target emotion by some annotators. These types of data cannot be used by the conventional method, whereas the proposed method can use all of the training data because it can manage other emotions as the "other" class, which increases the number of training data with the target emotion. This is an advantage of the proposed method.

C. Reduced loss for other emotions

In emotion recognition, other emotions, as introduced in Sec. III-D1, are less important than the target emotions. In many applications, it is effective to reduce loss values for unimportant classes [17], [18]. Here, the loss for $(K+1)$ emotions is $\boldsymbol{l} \in \mathbb{R}^{K+1}$. We multiply the loss by a $(K+1)$-dimensional weight vector $\boldsymbol{w}_l = [1, ..., 1, \kappa]^\top$ that has a reduced weight κ for "other" emotions, where $\top$ is a transpose. This reduces the influence of the annotator-wise emotions that deviate from the dominant emotion and reduce the overfitting during training, especially in the case when all the data is used during training in Sec. III-B.

D. Integration of annotator-wise estimation results

1) Majority voting: For evaluation, to estimate the dominant emotion by integrating annotator-wise estimation results, after the probabilities associated with the other emotions are discarded, the emotion with the highest posterior probability is selected among the K emotion types, as in Eq. (1).

$$\arg\max_{1\leq k\leq K} \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{z}_n[k] \tag{1}$$

This is a simple majority voting by the annotator-wise minor emotions as shown in Fig. 2 (a).

2) Weighted majority voting: In the previous section, simple majority voting was used, but the label reliability can be different from annotator to annotator. To place more weight on a more reliable annotator's estimate, we simply extend simple majority voting to weighted majority voting by the annotator-wise classification results to estimate the dominant emotion, as shown in Fig. 2(b). After the output from each classifier $\boldsymbol{z}_n$ ($\in \mathbb{R}^{K+1}$) is concatenated as $\boldsymbol{z} = [\boldsymbol{z}_1^\top, \boldsymbol{z}_2^\top, ..., \boldsymbol{z}_N^\top]^\top$ and the weight matrix $\boldsymbol{W} \in \mathbb{R}^{K\times((K+1)N)}$ is multiplied by $\boldsymbol{z}$, we identify the emotion with the highest posterior probabilities as:

$$\arg\max_{1\leq k\leq K}(\boldsymbol{W}\boldsymbol{z})[k] \tag{2}$$

When we connect a weighted voting component with the annotator-wise classifiers, we have two connection options. First, there are two types of connections: direct connections that directly connect them and connections after softmax removal, which remove softmax activation from the annotator-wise estimations $\boldsymbol{z}$ before their connection ($s = \{0, 1\}$). Second, there are two types of training: on in which the entire model is optimized and one in which only weight $\boldsymbol{W}$ is optimized after freezing the annotator-wise classifiers ($f = \{0, 1\}$). We train $\boldsymbol{W}$ for weighted majority voting in the four possible cases about (s, f).

- $(0, 0)$: After direct connection, the entire model is optimized.
- $(1, 0)$: After connection with softmax removal, the entire model is optimized.
- $(0, 1)$: After direct connection, only weight $\boldsymbol{W}$ is optimized with freezing classifiers.
- $(1, 1)$: After connection with softmax removal, only weight $\boldsymbol{W}$ is optimized with freezing classifiers.

To make the result consistent to simple majority voting, the initial value of $\boldsymbol{W}$ in Eq. (2) was constructed using N-times repetition of the $K \times K$ diagonal matrix and the K-dimensional zero vectors. Using this initial matrix, the estimation result was the same as that of simple majority voting[1] using Eq. (1).

$$\boldsymbol{W} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & \dots & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & & 0 & 0 & \vdots & 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots & \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & \dots & 0 & 0 & \dots & 1 & 0 \end{bmatrix}^\top$$

TABLE I
THE NUMBER OF LABELS IN TERMS OF THE ANNOTATORS.

annotator	# of labels	ratio
C-E1	**4,376**	0.297
C-E2	**4,141**	0.281
C-E3	191	0.013
C-E4	**4,000**	0.272
C-E5	268	0.018
C-E6	488	0.033
C-F1	216	0.015
C-F2	184	0.012
C-F3	263	0.018
C-M1	204	0.014
C-M3	273	0.019
C-M5	124	0.008

IV. EXPERIMENT

A. Dataset

We used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database [8], an open dataset that is frequently used for speech emotion recognition experiments. This dataset contains approximately twelve-hour dialog speech data annotated with nine types of emotion labels. The speakers are five males and five females. In this experiment, as previous studies did [9], [10], we focused on the four (=K) types of emotions with sufficient data (happy, neutral, sad, and angry). There are twelve annotators as shown in Table I.

IEMOCAP contains utterances with scripts and improvisation. For training, both data were used, whereas for evaluation, only improvisation data were used. To clarify the effectiveness of annotator-wise modeling, we did not use data extension or change the class weights for the loss function to adjust the emotion type data imbalance. For the ten speakers, using the leave-one-out method, ten evaluations were performed. The performance was evaluated in terms of a weighted accuracy that was averaged over the ten evaluations.

B. Model setup

We employed the long short-term memory (LSTM [19])-based model[2] proposed in [10] as the baseline of multimodal emotion recognition[3]. For speech emotion recognition, we used 35-dimensional acoustic features such as energy flux, zero-cross rate, and spectrum envelope. We also performed text emotion recognition considering the content of the utterances.

[1] If there are no constraints for weight matrix, the weighted majority voting is the same as the score fusion using logistic regression.

[2] https://github.com/bagustris/Apsipa2019_SpeechText

[3] Our baseline performance was different from the one reported in [10], because for this experiment, we modified the provided code. For example, the provided code did not use cross validation.

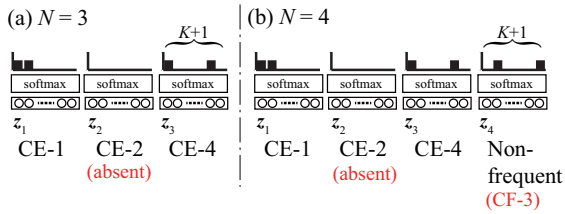

Fig. 3. Example of nodes associated with each annotator in the cases of $N = 3$ and $N = 4$. Annotators are CE-1, CE-4, and CF-3.

We used 300-dimensional text features, which are continuous-value vectors from GloVe-based word embedding [20]. For speech and text emotion recognition, we input speech and text features into different LSTM layers and then into concatenated hidden layer outputs after the LSTM layers.

IEMOCAP contains emotion labels annotated by three or four annotators. Table I displays the number of emotion labels from each annotator in IEMOCAP, which shows that annotators C-E{1,2,4} are the most frequent among the twelve annotators. We set the number of annotators N to three or four, which is the maximum number of annotators. The former ($N = 3$) focuses only on the most frequent annotators, and the latter ($N = 4$) dealt with one non-frequent annotator rather than the three most frequent annotators. For the three most frequent annotators, the same node n corresponded to each annotator. Annotators C-E{1,2,4} were associated with nodes {1,2,3}, respectively. Fig. 3 shows an example of a certain utterance annotated by three annotators (C-E1, C-E4, and C-F3). In the case of $N = 3$, only nodes $\boldsymbol{z}_1$ and $\boldsymbol{z}_3$ have target labels, and node $\boldsymbol{z}_2$ has (K+1)-dimensional zero vectors. In the case of $N = 4$, nodes $\boldsymbol{z}_1$, $\boldsymbol{z}_3$, and $\boldsymbol{z}_4$ have target labels, and node $\boldsymbol{z}_4$ has a target label from annotator C-F3. When annotators produced multiple emotion labels for one utterance, a target vector was averaged. For example, when one annotator identified the emotions "neutral" and "happy", the target vector was $[0.5,0.5,0,0,0]^\top$. The structure of the proposed model is shown in Fig. 4. Batch normalization [21] was applied after the LSTM layers, and the Adam optimizer [22] was used. For loss computation, the reduced weight κ for the other emotion labels was set to 0.5 and 0.1.

C. Baseline result

The baseline result is shown in the first line of Table II. Its accuracy for speech emotion recognition was 56.3%. Its accuracy for text emotion recognition was 54.1%, which was lower than that of speech emotion recognition. However, using both modalities, accuracy was improved by 8.6 points in terms of absolute value because usage of both modalities improves the cases in which emotions cannot be recognized solely from the content of the utterance or acoustic features. This result demonstrates the effectiveness of using multiple modalities [10].

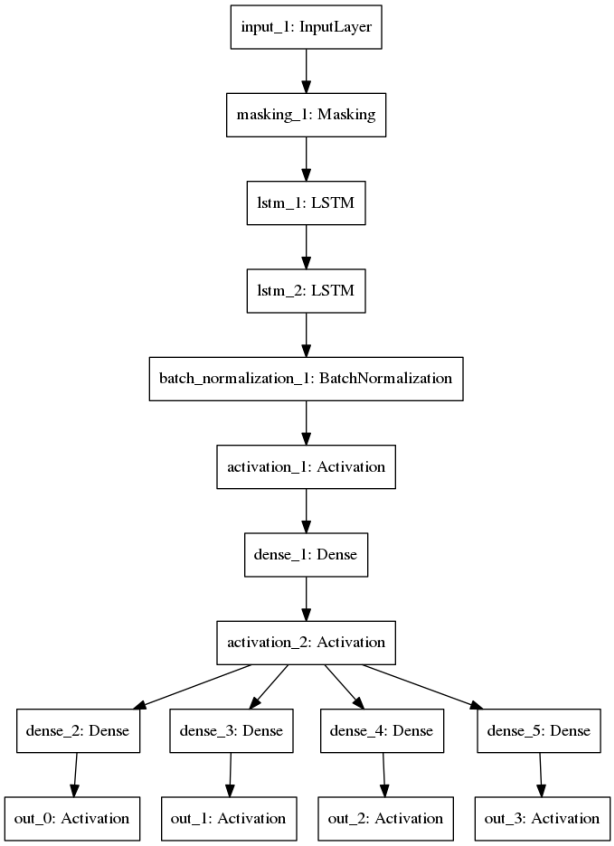

Fig. 4. Structure of the annotator-wise classification model.

TABLE II
EMOTION RECOGNITION ACCURACY [%] (BASELINE AND THE PROPOSED ANNOTATOR-WISE CLASSIFICATION WITH SIMPLE MAJORITY VOTING).

	speech	text	speech & text
baseline	56.3	54.1	64.9
$N = 3$	58.8	**54.8**	64.3
$N = 4$	**59.4**	54.5	**64.9**

D. Simple majority voting by annotator-wise estimation results

The second and third rows of Table II show the results of the proposed annotator-wise estimation using simple majority voting. In most cases, the performance was improved. The proposed method was especially effective in speech emotion recognition. The best configuration was $N = 4$, which considered an additional non-frequent annotator.

E. All-data training and reduced loss for other emotions

Table III shows the speech recognition accuracy when training utilized all of the data with reduced loss κ ($= 0.5$ and 0.1), as discussed in Sec. III-C. The $\kappa = 1$ result from Table III corresponds to the second column of Table II, which shows the effectiveness of using all of the data for training, which is an advantage of the proposed method. Except for case $\kappa = 0.1$, training with all of the data improved the performance.

TABLE III
EMOTION RECOGNITION ACCURACY (SPEECH) OF ALL-DATA TRAINING USING REDUCED LOSS FOR OTHER EMOTIONS. WE COMPARED TRAINING USING ONLY TARGET EMOTION DATA WITH TRAINING USING ALL EMOTION DATA.

	target only			all data		
κ	1	0.5	0.1	1	0.5	0.1
$N=3$	58.8	59.8	59.8	59.4	59.1	55.8
$N=4$	59.4	57.9	60.5	60.5	**60.7**	58.4

TABLE IV
EMOTION RECOGNITION ACCURACY [%] FOR THE FOUR TYPES OF WEIGHTED MAJORITY VOTING.

	target only				
type (s,f)	-	(0,0)	(1,0)	(0,1)	(1,1)
$N=3$	**58.8**	56.7	56.3	58.6	57.7
$N=4$	59.4	58.5	55.3	**59.8**	59.3
	all data				
type (s,f)	-	(0,0)	(1,0)	(0,1)	(1,1)
$N=3$	**59.4**	57.1	55.1	58.4	56.4
$N=4$	**60.5**	58.2	56.8	59.5	59.3

F. Weighted majority voting

Table IV shows the results for the four types of weighted majority voting. "Type" in the table corresponds to the four cases of (s, f) discussed in Sec. III-D2. Retraining entire systems without freezing annotator-wise classifiers ($f = 0$) degraded the performance. Freezing the classifier is necessary. In some cases in which $s = \{0, 1\}$ and $f = 1$, the performance improved slightly. Type (0,1) was the best among the weighted majority voting methods, but unfortunately, no significant improvement is observed when employing simple majority voting ('-' in Table).

G. Best configurations

The experiments above show that when $N = 4$, the accuracy was the best and that reduced loss ($\kappa = 0.5$) was effective for some cases. Among the weighted voting methods, the (0,1)-type was the best. Using these configurations, we compared the performance of speech, text, and speech and text emotion recognition with the baseline.

Table V shows the results. For all cases of speech emotion recognition, the proposed method improved the performance. The best configuration was all-data training with reduced loss ($\kappa = 0.5$), which achieved 60.6% accuracy (a 4.3 point improvement). For text emotion recognition, the proposed method improved the performance by 0.7 points. The all-data training was ineffective for text emotion recognition. For speech and text emotion recognition, the best accuracy of 66.3% was achieved.

TABLE V
EMOTION RECOGNITION ACCURACY [%] USING THE BEST CONFIGURATIONS ($N = 4$).

emo	κ	type (s,f)	speech	text	speech & text
baseline			56.3	54.1	64.9
target	1	(0,1)	59.8	**54.8**	**66.3**
target	0.5	(0,1)	58.8	53.9	64.5
all	1	(0,1)	59.5	53.8	65.0
all	0.5	(0,1)	**60.6**	53.5	65.1

V. CONCLUSION

For speech emotion recognition, we consider annotator-wise minor emotion labels in addition to dominant emotion labels and propose a sequential model of annotator-wise classifiers that output minor labels and use majority voting to output a dominant label. Classifiers are trained on each annotator's estimate using group softmax to reduce the influence of classifier imbalance. Experiments demonstrate that the proposed majority voting by annotator-wise estimations improved the emotion recognition performance of the conventional model, which estimates dominant emotions directly. Our proposed model's computational cost is almost the same as that of the conventional models because dominant emotions can be estimated by a one-time inference, which is an advantage of the proposed method. In addition, the proposed method is effective not only for speech emotion recognition, but also for text and speech and text emotion recognition.

REFERENCES

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

[2] C. Huang, O. R. Zaïane, A. Trabelsi, and N. Dziri, "Automatic dialogue generation with expressed emotions," in *Proc. NAACL-HLT*, 6 2018, pp. 49–54.

[3] Y. Zhang and M. Huang, "Overview of the NTCIR-14 short text generation subtask: Emotion generation challenge," in *Proc. NTCIR Conference on Evaluation of Information Access Technologies*, 6 2019, pp. 316–327.

[4] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM)*. ACM, 10 2010, pp. 1459–1462.

[5] 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.

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

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

[8] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 12 2008.

[9] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Speech emotion recognition based on multi-label emotion existence model," in *Proc. INTERSPEECH*, 2019, pp. 2818–2822. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2524

[10] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech emotion recognition using speech feature and word embedding," in *Proc. APSIPA*, 11 2019, pp. 519–523.

[11] 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.

[12] 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. ICASSP*, 2018, pp. 4964–4968.

[13] A. Ando, T. Mori, S. Kobashikawa, and T. Toda, "Speech emotion recognition based on listener-wise perception model," in *Proc. Acoustical Society of Japan*, 9 2020, pp. 777–778.

[14] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Proc. INTERSPEECH*, 2018, pp. 951–955.
[15] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *Proc. International conference on intelligent computing*. Springer, 2005, pp. 878–887.
[16] Y. Li, T. Wang, B. Kang, S. Tang, C. Wang, J. Li, and J. Feng, "Overcoming classifier imbalance for long-tail object detection with balanced group softmax," in *Proc. CVPR*, 2020, pp. 10 991–11 000.
[17] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017, pp. 2999–3007.
[18] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. CVPR*, 2019, pp. 9268–9277.
[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.
[20] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162
[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICLR*, 2015, pp. 448–456.
[22] D. Kingma and L. Ba, "Adam: A method for stochastic optimization," in *Proceedings of ICLR*, 2015. [Online]. Available: https://arxiv.org/pdf/1412.6980.pdf