Model Smoothing using Virtual Adversarial Training for Speech Emotion Estimation using Spontaneity

Toyoaki Kuwahara, Ryohei Orihara[a], Yuichi Sei[b], Yasuyuki Tahara[c] and Akihiko Ohsuga
Graduate School of Information and Engineering, University of Electro-Communications, Tokyo, Japan

Keywords: Deep Learning, Cross Corpus, Virtual Adversarial Training, Emotion Recognition, Speech Processing, Spontaneity.

Abstract: Speech-based emotion estimation increases accuracy through the development of deep learning. However, most emotion estimation using deep learning requires supervised learning, and it is difficult to obtain large datasets used for training. In addition, if the training data environment and the actual data environment are significantly different, the problem is that the accuracy of emotion estimation is reduced. Therefore, in this study, to solve these problems, we propose a emotion estimation model using virtual adversarial training (VAT), a semi-supervised learning method that improves the robustness of the model. Furthermore, research on the spontaneity of speech has progressed year by year, and recent studies have shown that the accuracy of emotion classification is improved when spontaneity is taken into account. We would like to investigate the effect of the spontaneity in a cross-language situation. First, VAT hyperparameters were first set by a preliminary experiment using a single corpus. Next, the robustness of the model generated by the evaluation experiment by the cross corpus was shown. Finally, we evaluate the accuracy of emotion estimation by considering spontaneity and showed improvement in the accuracy of the model using VAT by considering spontaneity.

1 INTRODUCTION

In recent years, with the development of deep learning, research on speech-based emotion recognition is progressing day by day. Emotion recognition is gaining attention as a task in communication between humans and between humans and machines, and it is thought that it can be used to improve the dialogue quality of dialogue agents communicating with humans. Emotion recognition problems can be classified into emotion classification problems and emotion estimation problems. Emotion classification problems include classification of specific emotions such as joy and sadness. Emotion estimation problems include estimating emotions as a degree of valence, activation, and superiority. In addition, various studies have been conducted to form appropriate consensus to express human perception(Laukka, 2005),(Laukka et al., 2005).

Emotion recognition based on speech basically extracts features (Mel-Frequency Cepstrum Coefficients, loudness, Voice Probability, etc.) from speech and trains an estimator or classifier using them. Mel-Frequency Cepstrum Coefficients(MFCC) is a feature value that represents the characteristics of the vocal tract taking into account the features of human speech perception. However, the problem with speech-based emotion recognition is that if the voice recording environment used to collect the training data is different from the actual voice recording environment, the emotion recognition accuracy decreases. This is because the extracted features may vary depending on the ambient conditions during recording, the attributes of the speakers, and the equipment used for recording. Assuming actual use of a spoken dialogue agent, changes in feature values are likely to occur easily due to differences in the environment in which the agent exists, the attributes of the conversation partner, and the recording device of each agent. In addition, there is a lack of supervised speech corpus available for training, and biases in the language, culture, speakers, etc. used in the corpus as a problem of speech-based emotion recognition. Continued use of such a corpus is thought to affect recognition results(Kim et al., 2010).

[a] https://orcid.org/0000-0002-9039-7704
[b] https://orcid.org/0000-0002-2552-6717
[c] https://orcid.org/0000-0002-1939-4455

Kuwahara, T., Orihara, R., Sei, Y., Tahara, Y. and Ohsuga, A.
Model Smoothing using Virtual Adversarial Training for Speech Emotion Estimation using Spontaneity.
DOI: 10.5220/0008958405700577
In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020) - Volume 2*, pages 570-577
ISBN: 978-989-758-395-7; ISSN: 2184-433X

Many researches studies have been conducted on emotion recognition(such as(Han et al., 2014),(Li et al., 2013)), and various approaches(Kim et al., 2017)(Sahu et al., 2018) have been taken to solve these problems. In addition, to make up for the shortage of corpora, several studies have created corpora specific to different languages and cultures(Parada-Cabaleiro et al., 2018)(Kamaruddin et al., 2012). Some studies are also working to increase training data using semi-supervised learning(Sahu et al., 2018), (Kuwahara et al., 2019).

In recent years, research on the spontaneity of utterances has been widely studied(Dufour et al., 2009),(Dufour et al., 2014),(Tian et al., 2015), and there is also research that improves accuracy by considering spontaneity in emotion classification based on speech(Mangalam and Guha, 2017); however, this work did not use spontaneity in emotion estimation task. These studies simply show that spontaneity is somehow related to speech and does not go any further into quantitative analysis. In our study, spontaneity in utterance refers to whether or not humans naturally expressed utterances. For example, when a speaker speaks naturally with his / her own consciousness, there is spontaneity, but when a speaker reads scripts and utters it, there is no spontaneity. However, there are not many studies that consider the spontaneousness of utterance in current emotion recognition.

Therefore, in this study, we created an emotion estimation model using Virtual Adversarial Training (VAT), which is semi-supervised learning, and evaluated the improvement of the robustness of the model among other languages. In addition, we propose consideration of the spontaneity and evaluate the accuracy of emotion estimation using the spontaneity classification.

The structure of this paper is as follows. In section 2, we explain the method used in this research and the related method. In section 3 we show the difference between the proposed method and the regression problem of VAT and emotion estimation method considering spontaneity. In section 4 we present the experiment and results followed by a summary in section 5. Future prospects are described in section 6.

2 RELATED WORKS

In this section, we explain the emotion classification method related to the method used in this study. Given the set of N labeled data points $\boldsymbol{x_i}, y_i, i = 1,\ldots,N$, we represent the DNN output for the point $\boldsymbol{x_i}$ as $\theta(\boldsymbol{x_i})$. $\theta(\boldsymbol{x_i})$ is a vector of probabilities that the DNN assigns to each class in the label space spanned by y. We define a loss function based on the DNN outputs and the one-hot vectors $\boldsymbol{y_i}$ corresponding to labels y_i, as shown below. $V(\theta(\boldsymbol{x_i}),\boldsymbol{y_i})$ is the loss of $\boldsymbol{x_i}$ that can be defined by cross entropy, mean square error, and so on.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} V(\theta(\boldsymbol{x_i}),\boldsymbol{y_i}) \tag{1}$$

Afterward, we describe adversarial training and virtual adversarial training, which are methods of adding a normalization term adverse to model output to this loss function.

2.1 Adversarial Training (AT)

AT, as proposed by Goodfellow(Goodfellow et al.,), is a training method that adds errors in model output when perturbation is added to training data $\boldsymbol{x_i}$ as adversarial losses. D is a non-negative function for quantifying the distance between the prediction $\theta(\boldsymbol{x_i}+\boldsymbol{r_i^a})$, and the target $\boldsymbol{y_i}$. γ is a tunable hyperparameter that determines the tradeoff between the original loss function and adversarial loss.

$$\mathcal{L}_{adv} = \mathcal{L} + \lambda * \frac{1}{N}\sum_{i=1}^{N} D(\boldsymbol{y_i}, \theta(\boldsymbol{x_i}+\boldsymbol{r_i^a})) \tag{2}$$

The perturbation $\boldsymbol{r_i^a}$ is determined by the equation below. ε is a hyperparameter that determines the search neighborhood for $\boldsymbol{r_i^a}$.

$$\boldsymbol{r_i^a} = \underset{\boldsymbol{r}:\|\boldsymbol{r}\|\leq\varepsilon}{\arg\max} D(\boldsymbol{y_i}, \theta(\boldsymbol{x_i}+\boldsymbol{r})) \tag{3}$$

Considering $\|\boldsymbol{r}\|$ to be the Euclidean norm, $\boldsymbol{r_i^a}$ in Equation 3 can be approximated as following:

$$r_i^a \approx \varepsilon \frac{g}{\|g\|_2}, where\ g = \nabla x_i D(y_i, \theta(x_i)) \tag{4}$$

By using the above calculations, it is the purpose of AT to smooth the generated model by adding adversarial loss to the normal loss function and improve robustness.

2.2 Virtual Adversarial Training (VAT)

Similar to AT, Virtual Adversarial Training (VAT) is a training method that adds the adversarial loss obtained by applying a perturbation to the input of the original loss function. VAT uses output of DNN instead of original label to derive adversarial loss unlike AT. The loss function is defined as follows.

$$\mathcal{L}_{vadv} = \mathcal{L} + \lambda * \frac{1}{N}\sum_{i=1}^{N} D(\theta(\boldsymbol{x_i}), \theta(\boldsymbol{x_i} + \boldsymbol{r_i^v})) \quad (5)$$

where the adversarial perturbation $\boldsymbol{r_i^a}$ for the training example $\boldsymbol{x_i}$ is defined as following:

$$\boldsymbol{r_i^v} = \underset{\boldsymbol{r}:\|\boldsymbol{r}\|\leq\varepsilon}{\arg\max} D(\theta(\boldsymbol{x_i}), \theta(\boldsymbol{x_i} + \boldsymbol{r})) \quad (6)$$

Then, define the following equation. KL is the Kullbac-Leibler divergence (KL divergence), which represents a distance measure between the probability distributions. $\boldsymbol{\theta}$ is the parameters of a model.

$$\Delta_{KL}(\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{\theta}) \equiv KL[p(y|\boldsymbol{x}, \boldsymbol{\theta}) || p(y|\boldsymbol{x} + \boldsymbol{r}, \boldsymbol{\theta})|] \quad (7)$$

Let this be a nonnegative function D and convert Equation (6) as following:

$$\boldsymbol{r_i^v} = \underset{\boldsymbol{r}:\|\boldsymbol{r}\|_2\leq\varepsilon}{\arg\max} \Delta_{KL}(\boldsymbol{r}, \boldsymbol{x_i}, \boldsymbol{\theta}) \quad (8)$$

Then, local distributional smoothing (LDS) is defined as following:

$$LDS(\boldsymbol{x_i}, \boldsymbol{\theta}) = \Delta_{KL}(\boldsymbol{r_i^v}, \boldsymbol{x_i}, \boldsymbol{\theta}) \quad (9)$$

Then, Equation(5) can be shown as following:

$$\mathcal{L}_{vadv} = \mathcal{L} + \lambda * \frac{1}{N}\sum_{i=1}^{N} LDS(\boldsymbol{x_i}, \boldsymbol{\theta}) \quad (10)$$

The computing algorithm of $\boldsymbol{r}_{v-adv}$ is described in (Miyato et al., 2015). As can be seen from Equation(5), the adversarial loss of VAT does not depend on the correct labels. Thus, it can be used in semi-supervised training scenarios where the first term $\mathcal{L}$ is computed using labeled data, and the second term $\mathcal{L}_{vadv}$ is computed using both labeled and unlabeled data.

3 GENERATION OF EMOTION ESTIMATION MODEL USING VAT CONSIDERING SPONTANEITY

3.1 Generation of Emotion Estimation Model using VAT

In this study, we apply VAT to emotion estimation which is a regression problem. The overview of the system used in this study is shown in Fig. 1. As we cannot access large unlabeled data for this experiment, we calculated only using the available data set.

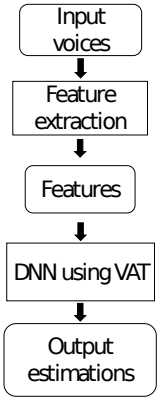

Figure 1: The overview of the emotion estimation system using VAT.

3.2 Difference in VAT in Estimation and Classification

Unlike the emotion classification problem, the output of emotion estimation problem is a numerical value. In VAT, when deriving the KL divergence, we use the probability vector allocated to each class that outputted $\boldsymbol{x_i}$ and $\boldsymbol{x_i} + \boldsymbol{r_i^v}$ by DNN. However, in the case of a regression problem, as the output is a continuous value, the KL divergence cannot be derived. According (Miyato et al., 2018), in AT, the normal distribution which centered at y_i with constant variance can be used as $\boldsymbol{y_i}$ to calculate $\mathcal{L}_{adv}$ for regression tasks. Therefore, to solve this problem when deriving the KL divergence, we defined a normal distribution with an average value of $\theta(\boldsymbol{x})$ and a variance of 1 as $p(\theta(\boldsymbol{x}))$. We then calculated the KL divergence between $p(\theta(\boldsymbol{x_i}))$ and $p(\theta(\boldsymbol{x_i} + \boldsymbol{r_i^v}))$.

3.3 Emotion Estimation System Considering Spontaneity

In this study, we propose an emotion estimation system considering spontaneity in order to improve the accuracy of emotion estimation based on the spontaneity of speech. The overview of the system used in this study is shown in Fig. 2.

The DNN that determines spontaneity is trained using all utterances, and the DNN that performs emotion estimation is trained using only utterances with and without spontaneity. The feature used for discrimination of spontaneity and the feature used for emotion estimation are the same. When speech data is fed to the system, features are extracted from the speech, and the spontaneity is first determined using that feature. After that, if there is spontaneity, the features are given to the estimator that trained only by

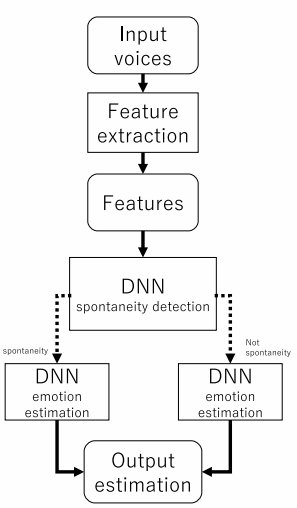

Figure 2: The overview of the estimation system using spontaneity.

utterances with spontaneousness, and if there is no spontaneity, the features are given to the estimator trained only by utterances without spontaneity. Finally, the estimator outputs an estimate of emotion.

4 EXPERIMENT

In this study, first we performed a single corpus experiment for setting hyperparameters, and then performed a cross-corpus experiment using the empirical hyperparameters to evaluate the accuracy of emotion estimation using different corpora to show robustness of the model.

In addition, in order to investigate the contribution to emotion estimation by considering the spontaneity of utterances, we first generated an emotion estimation model using VAT for utterances with spontaneity and utterances without spontaneity. After that, we compared the accuracy and investigated the effect of spontaneity on emotional expression.Then, we compared the estimation accuracy of the emotion estimation system considering the spontaneity of the utterance and the normal emotion estimation system, and evaluated the contribution to the emotion estimation by considering the spontaneity.

4.1 Features

In this experiment, we used the features that were used in INTERSPEECH 2009 Emotion Challenge(Schuller et al., 2009). The list of the features is shown in the Table 1.

We use features that are total 32 dimensions including the average power, pitch frequency, zero crossing rate, voice probability (proportion of harmonic components to the total power), 12-dimensional MFCC for each frame, and their Δ. 12 kinds of statistics (average, standard error, maximum value and its position, minimum value and its position, range of value, kurtosis, skewness, first order regression coefficient and intercept, regression error) are computed from 32 features. We calculated 12 kinds of statistics (amean, standard deviation, max, maxPos, min, minPos, range, skewness, kurtosis, linregc1, linregc2, linregerrQ) using a total of 32 dimensions: RMS enregy, F0, ZCR, voice probability, MFCC (12 dimentions), and *Delta* for each of them. In the end, we used 32 * 12 = 384 dimensions to use as input to the model in total.

Table 1: Features.

Feature	Δ	Statistics
RMS energy	Δ RMS energy	amean
F0	Δ F0	standard deviation
ZCR	Δ ZCR	max, maxPos, min, minPos
voice Prob	Δ voice Prob	range, skewness, kurtosis
MFCC 1-12	Δ MFCC 1-12	linregc 1, linregc2, linregerrQ

To extract features, we used the open-source software OpenSMILE(Eyben et al., 2010) developed for research, such as speech recognition, music recognition, para-language recognition as developed by Munich Technical University.

4.2 Experimental Setup

We use a DNN as our regression model, such that the output layer consisted of one node (corresponding to an emotion value), with ReLU activation function. The DNN had two hidden layers with the number of neurons in each layer set to 1200 and 600. The objective function $V(\theta(\boldsymbol{x_i}), \boldsymbol{y_i})$ was chosen to be the mean squared error loss in our experiments. We implemented the VAT model training in pytorch(Paszke et al., 2017) and performed optimization using Adam. Our evaluation metric was mean absolute error (MAE).

4.3 Single Corpus Experiment

In this experiment, we aimed to optimize the model by using one corpus and evaluating the change in accuracy of emotion estimation by changing hyperparameters. Initially, we fixed ε and changed λ to examine the influence of the model due to the change of λ. Then, we examined the influence of ε on the model by fixing λ to the value that was most accurate in the previous process and changing ε. In this experiment, we used the Interactive Emotional Dyadic Motion Cap-

ture (IEMOCAP)(Busso et al., 2008) dataset, which is a dialogue corpus containing five groups of one-on-one dialogues. In this experiment, 1/7 of 10,039 utterances of all subjects were used as test data, and the rest were used as training data and valence annotated with estimated numerical values 1 to 5.

4.4 Cross-corpus Experiment

In this experiment, to investigate the robustness of the model generated by VAT, a corpus other than one used to train the model was used as input, and its accuracy was compared to ordinary DNN. For this input, we used the Utsunomiya University (UU) Spoken Dialogue Database for Paralinguistic Information Studies(Mori et al., 2011). In this database, 6 pairs of utterance contents with intimate relationships are recorded in Japanese, and a total of 4,840 utterances are recorded. The emotional states are annotated with 6 abstract dimensions whose value range are 1 to 7: they are pleasant-unpleasant, aroused-sleepy, dominant-submissive, credible-doubtful, interested-indifferent, and positive-negative. In this experiment, we converted pleasant-unpleasant, aroused-sleepy, and dominant-submissive labels to a scale of 1 to 5, and entered it into the model used in all the experiments. The pleasant-unpleasant label corresponds to val label, aroused-sleepy corresponds to the act label, and dominant-submissive label corresponds to the dom label in IEMOCAP respectively.

4.5 Evaluation of Contribution to Emotion Estimation by Considering Spontaneity

This experiment consists of two stages.

In the first step, an emotion estimation model using VAT was generated for both spontaneous and non-spontaneous utterances, in order to investigate the effect of the utterance spontaneity on emotion estimation. Then, we compared the accuracy and investigated the effect of spontaneity on emotional expression. In this experiment, in IEMOCAP, we treated the utterances labeled with *improvisation*(47.8%) as utterances with spontaneity, and the utterances labeled *script* labels (52.2%) as utterances without spontaneity.

In the next step, based on the results, for the emotion that were determined to significantly contribute for estimation by consideration of spontaneity, we compared the accuracy in cases where spontaneity is considered and where it is not. In the system that considered utterance spontaneity, we used a DNN that determined spontaneity as our classification model, such that the output layer consisted of one node (corresponding to spontaneity), with sigmoid activation function. The DNN had three hidden layers with the number of neurons in each layer set to 1200, 600 and 300. We used a cross-entropy error for the loss function. The two DNNs, which are divided by the presence or absence of spontaneity, were trained in advance using hyperparameters similar to those in the cross-corpus experiment using only spontaneous utterances and non-spontaneous utterances in IEMOCAP.

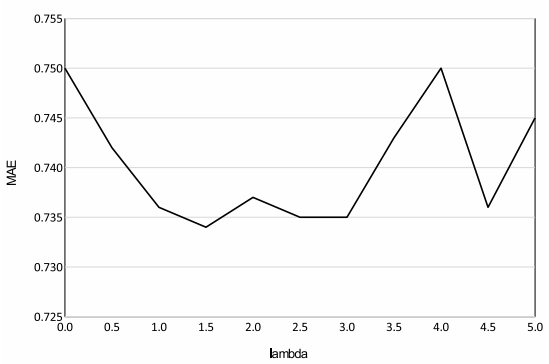

Figure 3: Mean absolute error of emotion estimation changing λ.

In a system that did not consider spontaneity, DNN using VAT was trained under the same conditions as the previous experiment using all utterances of IEMOCAP. We used 90 % of all utterances for training to discriminate spontaneity, and the remaining 10 % was used as test data to compare the accuracy of both systems.

4.6 Results and Discussions

4.6.1 Single Corpus Experiment

Fig. 3 shows the result when λ is changed when ε is fixed to 1.0.

When $\lambda = 1.5$, MAE became the smallest: MAE = 0.734. The cause of the decrease in precision when λ exceeds 1.5 is that the effect of hostile loss is larger because λ is a variable that determines the overall proportion of adversarial loss to total loss. It is conceivable that a great influence appears in the original loss function. From this results, we next experimented with $\lambda = 1.5$.

Fig. 4 shows the result when ε is changed and when λ is fixed to 1.5.

When $\varepsilon = 1.5$, MAE became the smallest: MAE = 0.726. When ε exceeded 1.5, the noise became too large, and excessive smoothing was performed, so it can be estimated that the accuracy was reduced. From

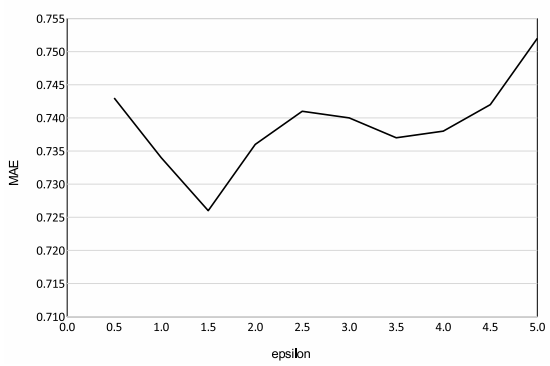

Figure 4: Mean absolute error of emotion estimation changing ε.

Table 2: Results of MAE of the emotion estimation in cross-corpus experiment.

	val	act	dom
DNN with VAT	**0.6964**	0.6633	**0.7180**
DNN without VAT	0.7011	**0.6631**	0.7243

these results of the single corpus experiment, we experimented with $\varepsilon = 1.5$ and $\lambda = 1.5$ in the cross-corpus experiment.

4.6.2 Cross-corpus Experiment

Table 2 shows the resultant MAE of the emotion estimation of DNN and DNN with VAT.

MAE of the emotion estimation of DNN with VAT of the val and dom labels are less than DNN without VAT. On the other hand, in the act label, the MAE of the estimation of DNN without VAT is less than its of DNN with VAT.

In the val and dom labels, there are differences of MAE between DNN with VAT and DNN without VAT. From these results, we consider that there are differences between Japanese and English in the val and dom labels(val: p>0.0005, dom:p>0.001), and VAT improves the robustness of the model against the gap of language.

On the other hand, the accuracy of DNN using VAT in the act label was lower than the accuracy of normal DNN. From this result, it is considered that the difference in expression of act between Japanese and English can not be compensated by smoothing using VAT, and it did not lead to improvement of estimation accuracy.

In addition, as a future prospect based on these, it is possible to find differences in emotional expression between languages and cultures by examining the change in the accuracy of emotion estimation due to the use of VAT between more languages and between cultures.

Table 3: Results of MAE of the emotion estimation which has spontaneity and which has not.

	val	act	dom
Spontaneity	0.814	0.580	0.657
Not spontaneity	0.690	0.594	0.685

Table 4: Results of MAE of the emotion estimation which using spontaneity and which not.

	Using spontaneity	Not using spontaneity
val	**0.749**	0.761

4.7 Evaluation of Contribution to Emotion Estimation by Considering Spontaneity

First, the table 3 shows the accuracy of emotion estimation for which only spontaneity utterances was used in training and only non-spontaneity was used.

From the table 3, it can be seen that the estimation accuracy is significantly different between utterances with and without spontaneity in the val label. Therefore, in the next experiment, we compared the accuracy of the val label between the system considering spontaneity and the normal system.

Table 4 shows the results of accuracy comparison between the system considering the spontaneity in the val label and the normal system.

From the table 4, it can be seen that the system which is considering spontaneity is more accurate than the system which is not considering spontaneity. As a result, it is thought that the consideration of spontaneity of utterance contributes to the improvement of the accuracy of emotion estimation.

5 CONCLUSION

In this study, in order to improve the degradation of recognition accuracy due to the difference between training data and actual data, we aimed to improve the robustness by smoothing the speech-based emotion estimation model using VAT. In addition, we aimed to improve the accuracy of speech-based emotion estimation by estimating emotions considering utterance spontaneity. As the results of a single corpus experiment, $\varepsilon = 1.5$, and $\lambda = 1.5$, were judged to be optimal in the DNN used in the experiment. Results of the cross-corpus experiment of DNN using VAT can estimate the val and dom labels better than DNN without VAT. Therefore the robustness against the change of the language of input data was improved by using VAT. Furthermore, in the verification experiment of accuracy change considering spontaneity, the MAE of

the val label without considering spontaneity is 0.761, but when considering MAE, it is 0.749. Therefore, by considering spontaneity, we were able to show a contribution to emotion estimation.

6 FUTURE WORKS

In this study, we experimented with three simple layers of DNN. However, if the network was changed, the optimal hyperparameter settings could also change. Moreover, we simply decided hyperparameters by experiments, hence it will be a future work to use known optimization algorithms to decide them. Further, we used a normal distribution with variance 1 as the distribution for measuring KL divergence, but it may be possible to improve the accuracy of emotion estimation by changing the variance and verifying the change in VAT accuracy. Furthermore, in this study, we conducted a cross-corpus experiment between Japanese and English, however as a future task, we will investigate the improvement of robustness by VAT by conducting a cross-corpus experiment in other languages and culture areas. In addition, in this study, we compared the estimation accuracy using IEMOCAP. IEMOCAP was used for both training and evaluating. Therefore it is considered future works to evaluate the contribution of estimation accuracy considering spontaneity using other language corpora. Finally, we need to conduct subjective assessment experiments to understand how the estimation error affects the human perception.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP17H04705, JP18H03229, JP18H03340, 18K19835, JP19H04113, JP19K12107.

REFERENCES

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Dufour, R., Estève, Y., and Deléglise, P. (2014). Characterizing and detecting spontaneous speech: Application to speaker role recognition. *Speech communication*, 56:1–18.

Dufour, R., Jousse, V., Estève, Y., Béchet, F., and Linarès, G. (2009). Spontaneous speech characterization and detection in large audio database. *SPECOM, St. Petersburg*.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. corr (2015).

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*.

Kamaruddin, N., Wahab, A., and Quek, C. (2012). Cultural dependency analysis for understanding speech emotion. *Expert Systems with Applications*, 39(5):5115–5133.

Kim, J., Englebienne, G., Truong, K. P., and Evers, V. (2017). Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning. *arXiv preprint arXiv:1708.03920*.

Kim, J., Lee, S., and Narayanan, S. S. (2010). An exploratory study of manifolds of emotional speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5142–5145. IEEE.

Kuwahara, T., Sei, Y., Tahara, Y., Orihara, R., and Ohsuga, A. (2019). Model smoothing using virtual adversarial training for speech emotion estimation. In *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, pages 60–64. IEEE.

Laukka, P. (2005). Categorical perception of vocal emotion expressions. *Emotion*, 5(3):277.

Laukka, P., Juslin, P., and Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, 19(5):633–653.

Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I., Valentin, E., and Sahli, H. (2013). Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 312–317. IEEE.

Mangalam, K. and Guha, T. (2017). Learning spontaneity to improve emotion recognition in speech. *arXiv preprint arXiv:1712.04753*.

Miyato, T., Maeda, S.-i., Ishii, S., and Koyama, M. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*.

Miyato, T., Maeda, S.-i., Koyama, M., Nakae, K., and Ishii, S. (2015). Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*.

Mori, H., Satake, T., Nakamura, M., and Kasuya, H. (2011). Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, 53(1):36–50.

Parada-Cabaleiro, E., Costantini, G., Batliner, A., Baird, A., and Schuller, B. (2018). Categorical vs dimensional perception of italian emotional speech. *Proc. Interspeech 2018*, pages 3638–3642.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

Sahu, S., Gupta, R., Sivaraman, G., and Espy-Wilson, C. (2018). Smoothing model predictions using adversarial training procedures for speech based emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4934–4938. IEEE.

Schuller, B., Steidl, S., and Batliner, A. (2009). The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.

Tian, L., Moore, J. D., and Lai, C. (2015). Emotion recognition in spontaneous and acted dialogues. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 698–704. IEEE.