IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 June 2022, accepted 3 July 2022, date of publication 7 July 2022, date of current version 14 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3189481

RESEARCH ARTICLE

Speech Emotion and Naturalness Recognitions With Multitask and Single-Task Learnings

BAGUS TRIS ATMAJA[1,3], AKIRA SASOU[1], (Member, IEEE), AND MASATO AKAGI[2], (Life Member, IEEE)

[1]National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8560, Japan
[2]Japan Advanced Institute of Science and Technology, Nomi 923-1211, Japan
[3]Department of Engineering Physics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia (on leave)

Corresponding author: Bagus Tris Atmaja (b-atmaja@aist.go.jp)

This work was supported by the New Energy and Industrial Technology Development Organization (NEDO), Japan, under Project JPNP20006.

ABSTRACT This paper evaluates speech emotion and naturalness recognitions by utilizing deep learning models with multitask learning and single-task learning approaches. The emotion model accommodates valence, arousal, and dominance attributes known as dimensional emotion. The naturalness ratings are labeled on a five-point scale as dimensional emotion. Multitask learning predicts both dimensional emotion (as the main task) and naturalness scores (as an auxiliary task) simultaneously. The single-task learning predicts either dimensional emotion (valence, arousal, and dominance) or naturalness score independently. The results with multitask learning show improvement from previous studies on single-task learning for both dimensional emotion recognition and naturalness predictions. Within this study, single-task learning still shows superiority over multitask learning for naturalness recognition. The scatter plots of emotion and naturalness prediction scores against the true labels in multitask learning exhibit the lack of the model; it fails to predict the low and extremely high scores. The low score of naturalness prediction in this study is possibly due to a low number of samples of unnatural speech samples since the MSP-IMPROV dataset promotes the naturalness of speech. The finding that jointly predicting naturalness with emotion helps improve the performance of emotion recognition may be embodied in the emotion recognition model in future work.

INDEX TERMS Speech emotion recognition, speech naturalness recognition, multitask learning, affective computing, speech processing.

I. INTRODUCTION

Speech emotion recognition (SER) is an emerging field of study in the field of speech processing. The goal of SER is to predict an affective state within speech, either in terms of emotion categories, emotion dimensions/attributes, or both. In this article, the terms 'attribute' and 'dimension' will henceforth be used interchangeably for representing valence, arousal, and dominance (VAD) as measures of an affective state. While the research of categorical SER is well established, there is a debate on how many attributes of emotion should be investigated [1].

Among many, three-dimensional emotions with valence, arousal, and dominance is the most common attributes to represent an affective state or emotion. In this view, the affective state is measured from pole to pole in these three attributes. Categorical emotions such as anger and sadness can be represented in a valence-arousal space [2]. In addition, dominance shows the degree of control of the speaker for that emotion (e.g., degree of control of anger). Using these three attributes completes the representation of emotion recognition by a computer/robot that mimics the human response to emotion (as proposed in [3]).

Naturalness recognition from the speech is a new application of speech processing technique to predict the degree of naturalness score from unnatural to very natural scores for an utterance [4]. This technique can be applied for such

The associate editor coordinating the review of this manuscript and approving it for publication was Ananya Sen Gupta.

applications as filmmaking, theatrical show, and emergency call center. In the latter application, a naturalness score could help the call center staff determine the authenticity of the call.

While the previous study on speech naturalness recognition is only intended to predict naturalness scores within speech datasets [5]–[7], it is worth studying the performance of the method on two tasks (emotion and naturalness) since the dataset used in the previous method also provided dimensional emotion scores. The concurrent speech emotion and naturalness recognition could be approached by utilizing multitask learning: one is to predict valence, arousal, and dominance of the speech signal, and the other is to predict the naturalness score of the same utterance. Indeed, predicting valence, arousal, and dominance simultaneously also can be regarded as a multitask learning problem [8].

The study of multitask learning to tackle several problems together is not new. Parthasarathy and Busso [9] proposed multitask learning for jointly predicting valence, arousal, and dominance from cross datasets. Lee [10] proposed to predict language labels in addition to emotion categories by utilizing multitask learning approach for multilingual speech emotion recognition. Both kinds of research show the superiority of multitask learning compared to single-task learning. We adopt that multitask learning approach to tackle the problem of concurrent speech emotion and naturalness recognition.

This study contributes to the previous studies in two aspects. First, we show the ability to multitask learning dimensional emotions and naturalness scores simultaneously with a small loss in naturalness recognition performance scores (while improving dimensional emotion recognition scores). Second, we evaluated our models in a 6-fold cross-validation evaluation to fill the gap in the previous studies, which only evaluated the performance of the models on a single fold. This cross-validation evaluation enables us to infer conclusions from the results that are more reliable and accurate than in previous studies.

II. METHODS

A. DATASET

This study employed MSP-IMPROV dataset, a mix of an acted and natural-interaction corpus, to study emotion perception while promoting naturalness in the recording [11]. There are four scenarios recorded in the dataset: Target - improvised, Other - improvised, and Natural interaction, and Target - read sentences. Note that not all the speech data is acted; the natural interaction scenario is fully natural speech recorded during the breaks of recording.

The total number of utterances in the MSP-IMPROV dataset is 8438 samples, and all samples are evaluated in this study. The recording in the dataset is split into six sessions; each session contains an interaction of two speakers (male and female). At least five evaluators annotated emotion and naturalness labels. The emotional labels are provided in both categorical and dimensional emotions; we adopted dimensional emotion since it has the same scale as the naturalness score. The scores were on a five-point Likert-like scale; we normalized the scores to the range of −1 to 1 in the deep neural network (DNN) learning process.

TABLE 1. Number of utterances for training/test split for each fold in 6-fold cross-validation in the MSP-IMPROV dataset.

Fold	Training (%)	Test (%)
1	7449 (88%)	989 (12%)
2	7006 (83%)	1432 (17%)
3	6661 (79%)	1777 (21%)
4	7325 (87%)	1113 (13%)
5	6933 (82%)	1505 (18%)
6	6816 (81%)	1622 (19%)

The dimensional emotion contains three attributes; the naturalness score contains a single attribute. The attributes of dimensional emotion are valence, arousal, and dominance. Valence is the degree of positive or negative emotion, arousal refers to the level of activation from sleepiness (low) to awakeness (high), and dominance is the degree of control over the emotion [12]. The naturalness labels represent the most unnatural speech (score 1) to the most natural speech (score 5).

We split the dataset into two parts: training and test. A portion of 20% of the total training data is used for evaluation or development within the training phase (to adjust the weights of layers in the neural network models). We adopted leave-one-session-out (LOSO) cross-validation to evaluate the performance of the method. In this evaluation, one session is allocated for test data while the other five sessions are allocated for training data. Since each session is recorded by different speakers, this LOSO split is also a speaker-independent evaluation. The number of utterances in each session is different; Table 1 shows the number of utterances in each session/fold for training and test. The number of folds in that Table 1 also represents the session for the test set. For instance, Session 1 is used for the test set in fold 1 (while sessions 2-6 are for training). All test data is unseen (held-out) data except for the calculation of performances. The reported performances were the average of the performances of the six LOSO folds.

B. ACOUSTIC FEATURES

We evaluated four different acoustic features below. For each acoustic feature set, we only extracted high-level statistical functions (HSF) from all frames in an utterance. This feature type is also known as the global feature.

1) pyAudioAnalysis (pAA)

We extracted 136-dimensional features from the audio signal using the pyAudioAnalysis library [13]. The features include 34 acoustic features and their deltas (68 in total). For these 68 features, we calculated the mean and standard deviation of the feature values from all frames in each utterance. The frame size (window size) is 0.025 seconds with a frameshift length (hop size) of 0.01 seconds.

2) ComParE

We extracted 6373-dimensional features from the audio signal using the Python-version of openSMILE library [14]. The configuration for the feature set is ''ComParE_2016'' [15] with ''functionals'' feature level.

3) eGeMAPS

We extracted 88-dimensional features from the previous openSMILE library with the ''egemaps'' feature set. The configuration for the feature set is ''eGeMAPSv02'' [16] with also ''functionals'' feature level.

4) EMOBASE

We extracted 988-dimensional functional features from the previous openSMILE library with the ''emobase'' feature set. The configuration for the feature set is ''emobase', which is provided at the INTERSPEECH 2010 Paralinguistic Challenge [17].

All acoustic feature sets above are normalized to zero mean and unit variance over all data globally.

C. CLASSIFIERS

We evaluated the following two classifiers: multilayer perceptron and long short-term memory networks. The choice of these classifiers is based on the performance of the previous studies [4], [5]. The best values for hyperparameters are searched with a brute-force search mechanism. These values, shown in Table 2, are optimized for each classifier independently. For instance, the number of layers is searched in ranges [1..6] with a number of units/nodes in variations of the following: 16, 32, 64, 128, 256, and 512.

1) MULTILAYER PERCEPTRON (MLP)

We built a three-layer MLP with hidden shared layers of size 512, 256, and 128 units (Fig. 1) and logistic activation function. This simple MLP is trained on 200 maximum iterations (epoch) with ten patiences. The observation showed that the training epoch never reaches the maximum number of iterations for obtaining the best performance. The training is batched in size of 'auto', which chooses a minimum number between 200 or a number of samples. The model is implemented with MLPRegressor in scikit-learn toolkit [18].

2) LONG-SHORT TERM MEMORY (LSTM)

We built an LSTM network with three shared layers and an independent layer of size 128, 64, 32, and 16 units (Fig. 2). We used the default ''tanh'' activation for LSTM layers but ''ReLU'' for the Dense layer. The network is trained on 100 epochs with ten patiences of early stop criteria. The observation also showed that the training epoch never reaches the maximum number of epochs for obtaining the best performance. The training is batched in size 8. The model is implemented with LSTM and Dense layers in Tensorflow toolkit [19].

TABLE 2. Values of the hyperparameters for the MLP and LSTM networks.

Parameters	MLP	LSTM
Shared layers (SL)	3	3
Independent layers (IL)	-	1
Nodes (SL + IL)	(512, 256, 128)	(128, 64, 32, 16)
Optimizer	adam	RMSprop
Learning rate	0.001	0.001
Batch size	'auto'	8
Epoch	200	100
Early stop	Yes	Yes
Tolerance	10	10

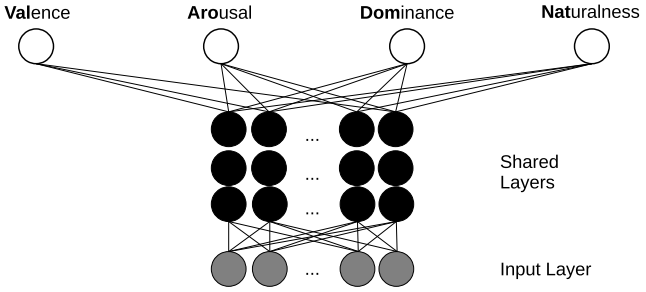

FIGURE 1. DNN architecture for MLP networks with MTL. The shared layers contain three MLP layers with 512, 256, and 128 nodes. For STL, the network predicts either valence-arousal-dominance scores or naturalness scores.

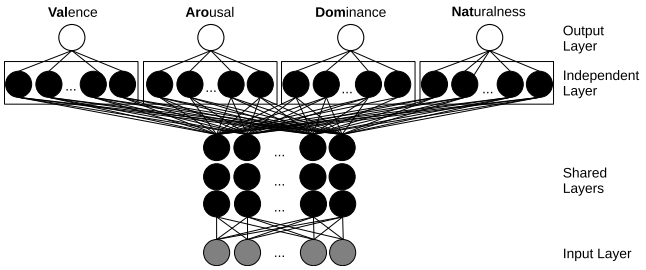

FIGURE 2. DNN architecture for LSTM networks with MTL. The shared layers contain three LSTM layers with 128, 64, and 32 nodes. The independent layer is four single dense layers with 16 nodes each. For STL, the output layer is either three single-node dense layers (emotion) or a single-node dense layer (naturalness).

D. EVALUATION METRIC AND LOSS FUNCTIONS

We evaluated the performance of the classifiers using the concordance correlation coefficient (CCC) between predictions and labels. CCC is claimed to be better than Pearson correlation since it penalizes deviation in scale (e.g., the prediction's scale is shifted from the original labels) [20]. It also replaces mean squared error in many modern multivariate regression analyses [21]. The CCC is formulated as follows:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{1}$$

where σ is the standard deviation, σ^2 is the variance, and μ is a mean value of the variable (prediction for x or label for y). ρ is the Person correlation coefficient (PCC) between two variables formulated as follows,

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}}. \tag{2}$$

A direct approach to maximize CCC is by minimizing the CCC loss function (CCCL) which is formulated as follows,

$$CCCL = 1 - CCC. \quad (3)$$

In single-task learning (STL), the loss function is either sum of three loss functions from valence ($CCCL_V$), arousal ($CCCL_A$), and dominance ($CCCL_D$) or naturalness ($CCCL_N$) (notice that jointly predicting VAD is called multitask learning in the previous study [9] but now is called STL in this study). In multitask learning (MTL), when CCC loss is used as a single metric for all arousal, valence, dominance, and naturalness, the $CCCL_{total}$ is a combination of those four CCC loss functions defined as follows,

$$CCCL_{tot} = CCCL_V + CCCL_A + CCCL_D + CCCL_N. \quad (4)$$

The equation 4 above applies to LSTM network. For the MLP network, we applied the similar mean squared error (MSE) loss function as follows,

$$MSE_{tot} = MSE_V + MSE_A + MSE_D + MSE_N. \quad (5)$$

While this research did not evaluate the CCC loss for MLP due to its difficulties in implementation with Scikit-learn [18], future research could tackle this limitation for a fair comparison with LSTM network.

It should be noted that for both CCC and MSE loss functions in MTL, we treated each attribute as important as the other (a factor of "1" for each attribute). For SER, with three attributes, the total importance is 0.75 compared to all tasks. For naturalness recognition with a single attribute, the importance factor is 0.25 compared to all tasks.

The pre-trained model generated by the research methods above will be hosted at https://github.com/bagustris/sner and free to use for academic and non-commercial purposes. The pre-trained model is trained on the whole dataset, i.e., all data is used for training without a test phase.

III. RESULTS AND DISCUSSION

We conducted experiments on the different features and classifiers. When experimenting with different features, we hold the classifier; when experimenting with different classifiers, we hold the features. The evaluations are performed in 6-fold CV (Table 3). The additional results on a single fold (with session six as a test set) are shown for a benchmark to the previous studies (Table 5).

A. EXPERIMENT RESULTS

Table 3 shows the results of the experiments of concurrent multitask learning for predicting valence, arousal, dominance, and naturalness. Using CCC as the evaluation metric, the results show moderate performance on all four tasks. All models achieved the highest performance in predicting arousal, strengthening the previous findings [5]. As it has been found in that paper, we also found that MLP achieved better performances than LSTM in predicting naturalness in addition to valence, arousal, and dominance.

TABLE 3. CCC scores (higher is better) for the 6-fold cross-validation (CV) speech and naturalness recognitions on the MSP-IMPROV dataset with MTL.

Feature	Val	Aro	Dom	Nat
MLP				
pAA	0.329	0.577	0.440	0.237
ComParE	**0.364**	**0.596**	**0.444**	0.273
eGeMAPS	0.352	0.574	0.436	0.239
emobase	0.361	0.559	0.421	0.264
LSTM				
pAA	0.276	0.542	0.416	0.279
ComParE	0.290	0.505	0.385	0.278
eGeMAPS	0.352	0.574	0.436	0.239
emobase	0.289	0.508	0.399	**0.288**

Among four different global acoustic feature sets and two classifiers, ComParE feature set achieved the highest performance among all variations. ComParE achieved the top performance on MLP networks (in terms of average CCC from all tasks with CCC average = 0.419) with small differences from the emobase feature set (CCC average = 0.401) and eGeMAPS (CCC average = 0.400). Feature set emobase achieved the top performance for predicting the naturalness score with CCC = 0.288 with LSTM networks, while ComParE attained CCC = 0.278 on the use of the same LSTM networks. Nevertheless, the performance of ComParE using MLP on this naturalness task was also competitive (CCC = 0.273) to that score by emobase using LSTM. In MLP models, this CCC score of ComParE for naturalness is the highest among the same MLP models.

B. VISUALIZATION BY SCATTER PLOTS

Following the previous research [5], we visualize the predictions of MLP model against true labels for dimensional emotions (Fig. 3) and naturalness score (Fig. 4). Both figures are obtained from the last fold, i.e., the sixth fold. It can be inferred from the scatter plots that the model fails in predicting very low and very high scores, particularly on naturalness recognition. The original value (in [1, 5]) was scaled to [−1, 1] for the deep learning model, while the obtained naturalness predictions are in [−0.14, 0.74] for naturalness. For the dimensional emotion (valence, arousal, and dominance) the predictions are in the range of [−0.66, 0.88]. This result is in line with the obtained CCC scores in Table 3 where dimensional emotions obtain higher CCC scores than naturalness.

One possible explanation for the low score of naturalness is that the model is trained on a low number of unnatural samples. The number of samples in MSP-IMPROV dataset based on scenarios are 620, 652, 2758, and 4381 for Target-read, Target-improvised, Natural-interaction, and Other-improvised. While we did not take into account this unbalance condition in the model, future research may consider balancing the dataset before training the model or developing a model which can handle unbalanced data.

Another potential solution to tackle the limitation of the current model is by shifting the continuous score to the

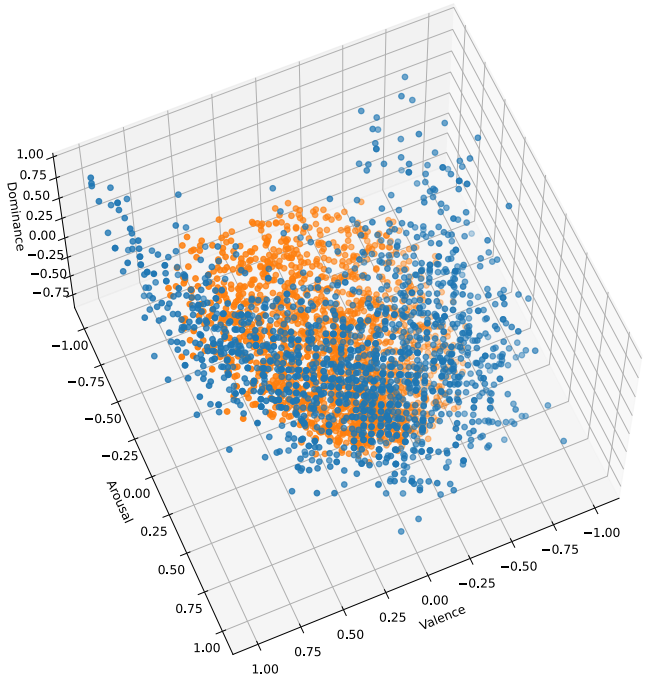

FIGURE 3. Scatter plot of the true labels (blue) and predictions (orange) for valence, arousal, and dominance (dimensional emotions) on the sixth fold as the test data by the MLP model and ComParE feature set. For numerical value on how close/far predictions vs. true labels, see CCC scores in Table 5.

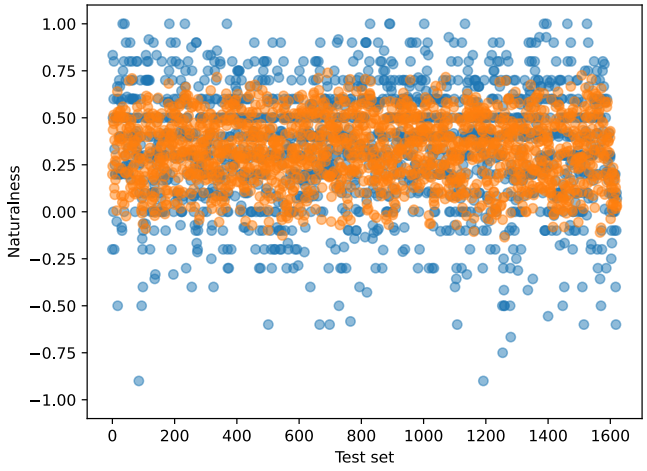

FIGURE 4. Scatter plot of the true labels (blue) and predictions (orange) for naturalness scores on the sixth fold; the MLP model with ComParE feature set fails to predict low and very high scores of naturalness.

ordinal categorical label. It has been argued in [22] that emotion is ordinal by nature. Instead of predicting the continuous score of valence, arousal, and dominance, the authors of [22] have shown that labeling emotion in ordinal rating has been found to be more accurate and beneficial. A similar approach (ordinal label) may also be applied to the naturalness score. The original score in [1, 5] may be mapped into the ordinal label of low, medium, and high categories.

C. MTL VS. STL

The previous results were obtained using MTL as proposed in this research. While the basic idea in this research is to combine speech emotion and naturalness recognition simultaneously, it is necessary to evaluate these tasks independently using single task learning (STL). Two STLs could be evaluated. One STL is to predict scores of valence, arousal, and dominance. Another STL is to predict naturalness score only.

TABLE 4. CCC scores (higher is better) for the 6-fold cross-validation (CV) speech and naturalness recognitions on the MSP-IMPROV dataset with STL; note that scores of Naturalness is obtained independently (in a single task learning) from scores of Valence, Arousal, and Dominance.

Feature	Val	Aro	Dom	Nat
MLP				
pAA	0.334	0.572	0.432	0.237
ComParE	0.352	**0.574**	**0.434**	**0.411**
eGeMAPS	**0.354**	0.544	0.419	0.410
emobase	0.352	0.551	0.416	0.399
LSTM				
pAA	0.272	0.536	0.415	0.273
ComParE	0.266	0.503	0.384	0.283
eGeMAPS	0.283	0.501	0.391	0.270
emobase	0.301	0.508	0.379	0.289

Table 4 shows our STL result for both emotion recognition and naturalness recognition. Note that the scores of valence, arousal, dominance, and naturalness for each row are obtained independently using a single model (a model that predicts three attributes in the case of VAD), either using MLP or LSTM. To our surprise, while the scores of VAD in STL evaluations are less than MTL, the scores of naturalness are higher than MTL in this research and the similar STL in the previous research [4]. The best result in previous research was obtained using statistical features of pAA with four-layer LSTM (512, 256, 128, 64). In this research, evaluating MLP with three layers (512, 256, 128) leads to better performances. Although the higher CCC scores obtained by SER could be explained by their importance factors, the discrepancy in naturalness performances between MTL and STL left room for improvement for future research.

From this evaluation of MTL vs. STL, it has been found that recognizing naturalness in speech helps improve the prediction of dimensional emotion but not vice versa. This information that naturalness improves recognition of emotion is in line with human perception of emotion, in which the more natural the speech is, the more likely the emotion to be recognized. Instead of multitask learning, the naturalness information (in the form of features) could be embedded into the speech features to improve the performance of the speech emotion recognition model in the future.

D. CROSS COMPARISONS

To evaluate the performance of the models in this study, we conducted cross-comparisons of the best models in this study to the previous results published on the MSP-IMPROV dataset. The results are shown in Table 5. Our system in this study is the only one that predicted both dimensional emotions and naturalness scores (MTL) in addition to STL. In this case, STL is either predicting emotion attributes:

TABLE 5. Summary of benchmarking results (CCC scores) in this study (MTL and STL) against previous studies (STL). Reference [8] used subsets of MSP-IMPROV with improvised and natural interaction parts only (MPSIN) using both acoustic and linguistic features.

Reference	Test setting	Val	Aro	Dom	Nat
This study MTL	Session 6	0.361	0.595	**0.452**	0.323
This study MTL	6-fold CV	**0.364**	**0.596**	0.444	0.273
This study STL	6-fold CV	0.352	0.574	0.434	**0.411**
[5]	Session 6	0.204	0.525	0.361	-
[4]	Session 6	-	-	-	0.302
[8]	Session 6 (MSPIN)	0.291	0.570	0.405	-

valence, arousal, and dominance, or predicting naturalness only. In addition, we also performed cross-validation for a more confident evaluation, while the previous studies only evaluated their models on the sixth session as the test set.

Table 5 shows that our results on both single-fold test and six-fold cross-validation are better than the previous studies for all tasks. We improved recognition of valence on Session 6 as the test set from 0.291 to 0.361 by employing ComParE feature set with MLP. Similarly, we improved arousal and dominance recognitions from 0.57 to 0.595 and from 0.405 to 0.452. For naturalness, we obtained a small improvement from 0.302 to 0.323 (MTL). Note that in [8], the authors only evaluated parts of the MSP-IMPROV dataset with improvised and Natural-interaction scenarios. The performance for the complete MSP-IMPROV dataset for that model (acoustic-linguistic fusion) may be lower since the remaining scenario contains Target-read, which sounds difficult if linguistic information is utilized.

Finally, a cross-validation evaluation is more reliable than a single-test evaluation. Our results on cross-validation evaluation show that our models are more accurate than the previous studies for dimensional emotion recognition. For naturalness recognition, which is new in the field of speech processing, there is a need for more accurate models. Although the multitask learning model is not optimized for emotion, the model predicts emotion better than naturalness. The weight of nodes in the deep learning model may tend to learn the emotion dimensions (three attributes) more than naturalness labels (a single attribute). The suggested studies in the previous subsections can be considered for future work.

IV. CONCLUSION

In this paper, we simultaneously evaluate speech emotion and naturalness recognitions by utilizing deep learning models with multitask learning and single-task learning approaches. The emotion model accommodates valence, arousal, and dominance attributes known as dimensional emotion. The naturalness ratings are labeled on a five-point scale as dimensional emotion. The results with multitask learning show improvement from previous studies on single task learning for both dimensional emotion recognition and naturalness recognition. Within this study, the performance of naturalness recognition with multitask learning is lower than that of single-task learning, whereas the performance of dimensional recognition is improved. The scatter plots of emotion and naturalness scores exhibit the lack of the model; it fails to predict the low and extremely high scores. The disadvantage of this study on naturalness prediction is possibly due to a low number of samples on unnatural speech. This presumption is based on the fact that MSP-IMPROV dataset is intended to promote the naturalness of speech.

Future research can be directed to studying the acoustic features that correlate to the naturalness of speech. The appropriateness of features can vary from unnatural speech to natural speech, as revealed in other domains [23]. A gap between the naturalness recognition performance of MTL and STL needs to be improved in the future. Future work may also include a balancing strategy to improve the model performance, as well as mapping continuous scores to ordinal labels. While the best results in this study are obtained with MLP with MSE losses, we believe a pair of MLP with CCC losses could improve the current results in terms of CCC scores.

REFERENCES

[1] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychol. Sci.*, vol. 18, no. 12, pp. 1050–1057, 2017.
[2] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Develop. Psychopathol.*, vol. 17, no. 3, pp. 715–734, Sep. 2005.
[3] I. Bakker, T. van der Voordt, P. Vink, and J. de Boon, "Pleasure, arousal, dominance: Mehrabian and Russell revisited," *Current Psychol.*, vol. 33, no. 3, pp. 405–421, 2014.
[4] B. T. Atmaja, A. Sasou, and M. Akagi, "Automatic naturalness recognition from acted speech using neural networks," in *Proc. APSIPA Annu. Summit Conf.*, Dec. 2021, pp. 731–736.
[5] B. T. Atmaja and M. Akagi, "Evaluation of error- and correlation-based loss functions for multitask learning dimensional speech emotion recognition," *J. Phys., Conf.*, vol. 1896, no. 1, Apr. 2021, Art. no. 012004.
[6] B. Merritt and T. Bent, "Perceptual evaluation of speech naturalness in speakers of varying gender identities," *J. Speech, Lang., Hearing Res.*, vol. 63, no. 7, pp. 2054–2069, Jul. 2020.
[7] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," in *Proc. Interspeech*, Oct. 2020, pp. 1748–1752.
[8] B. T. Atmaja and M. Akagi, "Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM," *Speech Commun.*, vol. 126, pp. 9–21, Feb. 2021.
[9] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. Interspeech*, Aug. 2017, pp. 1103–1107.
[10] S.-W. Lee, "The generalization effect for multilingual speech emotion recognition across heterogeneous languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5881–5885.
[11] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan. 2017.
[12] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emotions*, vol. 1, no. 1, pp. 68–99, Apr. 2010.
[13] T. Giannakopoulos, "PyAudioAnalysis: An open-source Python library for audio signal analysis," *PLoS ONE*, vol. 10, no. 12, pp. 1–17, Jan. 2015.
[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2013, pp. 835–838.
[15] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, vols. 8–12, 2016, pp. 2001–2005.

[16] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andrè, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, and S. S. Narayanan, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.
[17] S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, M. Christian, S. Language, P. Group, and D. Telekom, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, 2010, pp. 2794–2797.
[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[19] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and Others, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Des. Implement. (OSDI)*, 2016, pp. 265–283.
[20] M. Valstar, J. Gratch, B. Schuller, F. Ringevaly, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Panticz, "AVEC 2016—Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Work. Audio/Visual Emot. Challenge, Co-Located ACM Multimedia (AVEC)*, 2016, pp. 3–10.
[21] V. Pandit and B. Schuller, "The many-to-many mapping between concordance correlation coefficient and mean square error," Jul. 2020, *arXiv:1902.05180v6*.
[22] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 16–35, Jan. 2021.
[23] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, 2015, pp. 698–704.

BAGUS TRIS ATMAJA received the B.E. and M.E. degrees from the Sepuluh Nopember Institute of Technology, in 2009 and 2012, respectively, and the Ph.D. degree in information science with a focus on speech emotion recognition from the Japan Advanced Institute of Science and Technology, in 2021. Then, he was employed as a Docent with the Vibrastic Laboratory, Sepuluh Nopember Institute of Technology. He is currently a Postdoctoral Researcher with the Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST). His research interests include speech processing, including speech enhancement, source separation, and speech (emotion) recognition.

AKIRA SASOU (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in electrical engineering from Tokyo Denki University, in 1994, 1996, and 1999, respectively. He currently leads the Signal Processing Research Group, Department of Information Technology and Human Factor, National Institute of Advanced Industrial Science and Technology (AIST).

MASATO AKAGI (Life Member, IEEE) received the B.E. degree from the Nagoya Institute of Technology, in 1979, and the M.E. and Ph.D. (Eng.) degrees from the Tokyo Institute of Technology, in 1981 and 1984, respectively. He joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT), in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been a Faculty Member of the School of Information Science, JAIST, and is currently a Professor Emeritus. His research interests include speech perception, modeling of speech perception mechanisms in humans, and the signal processing of speech.