Automatic Naturalness Recognition from Acted Speech Using Neural Networks

Bagus Tris Atmaja* Akira Sasou* and Masato Akagi†
* National Institute of Advanced Industrial Science and Technology, Japan
E-mail: b-atmaja@aist.go.jp, a-sasou@aist.go.jp
† Japan Advanced Institute of Science and Technology, Japan
E-mail: akagi@jaist.ac.jp

***Abstract*—This study proposes an automatic naturalness recognition from an acted dialogue. The problem can be stated that: given speech utterances with their naturalness labels, is it possible to recognize these labels automatically? By what methods? And how to evaluate these methods? We evaluated two supervised classifiers to investigate the possibility of recognizing naturalness automatically in acted speech: long short-term memory and multilayer perceptron neural networks. These classifiers accept inputs in the form of acoustic features from a speech dataset. Two kinds of acoustic features were evaluated: low-level and high-level features. This initial study on automatic naturalness recognition of speech resulted in a moderate performance of the assessed systems. We measured the performance in concordance correlation coefficients, Pearson correlation coefficients, and root mean square errors. This study opens a potential application of speech processing techniques for measuring naturalness in acted dialogue, which benefits for drama- or movie-making in the future.**

***Index Terms*—speech naturalness recognition, acted dialogue, paralinguistic information, speech processing, speech analysis**

I. Introduction

The measure of success of a movie or drama is usually taken after the production or after the movie is being premiered (post-production). A common measure for a movie's success is the performance on the box office sales (e.g., in [1]). If the measure is taken after the post-production, as reported in that paper, no further improvements can be made except on the next release. A method should be proposed to evaluate the performance of actors while in the production stage; this paper proposes a method to evaluate the naturalness of acted speech via machine learning methods.

In speech processing, it is common to study the effects of particular acoustic features on the related phenomena. For instance, the study conducted by Mairano et al. [2] has found the correlation between acoustics features and sentiment analysis: pitch parameters show a modest correlation to valence, while rhythmic and spectral parameters showed a correlation to arousal. In more general emotion recognition, Blanton [3] stated that "the effect of emotions upon the voice is recognized by all people." Following this idea, there should be a similarity in naturalness recognition: the effect of naturalness in speech is recognized by all people.

Speech naturalness, according to [4], is "perception of the degree to which speech meets the typical patterns in terms of intonation, voice quality, rate, rhythm, and intensity, with respect to the syntactic structure of the utterance." In acted dialogue, like in a drama or a movie making, the naturalness of speech influences the quality of actors' performance. For instance, in an angry scene, the actors should be able to speak as naturally as possible as the real angry people. The failure to perform natural angry speech will make the quality of the movie degraded since the act is not natural. The source of unnatural speech is the ambiguity of the attention of the actors, focusing on both acting and the naturalness of acting. As in emotional speech, the perception of naturalness in speech can be estimated by such acoustic and prosodic features.

Instead of investigating acoustic and prosodic features that correlate with naturalness in speech, this pilot study fed and evaluated sets of acoustic features in supervised manners. pyAudioAnalysis, developed by Giannokopoulos [5], were utilized as acoustic features either by feeding low-level or high-level features to the classifiers. The models are then trained to match these acoustic features to the given labels. This training process involved training and development partitions of the dataset is conducted in the training phase. In the test phase, the models predict the labels given merely the acoustic features. Comparing the predicted labels from the model with the ground-truth labels from the dataset by such metrics can be used to measure the performance of the speech naturalness recognition system.

Naturalness in speech can be applied in many areas. In [6], the authors evaluated speech naturalness in different genders, including transmasculine and transfeminine speakers. The finding showed that transgenders were rated less natural than cisgenders. In [7], the authors developed a 9-point scale to evaluate speech naturalness for stutterers. The authors found significant differences between the speech naturalness of stutterers and non-stutters. In contrast to the research in psychophysics by subjective test, this study aims at recognizing naturalness of speech automatically from an acted dialogue, an artificial dialogue that is intended to be as natural as possible.

The contribution of this paper is the first study of automatic naturalness recognition of speech in acted dialogue. One of the potential applications in the future is to evaluate the naturalness performance of actors. Most studies of speech naturalness recognition have been carried out in either evaluating quality of speech synthesis or perceptual evaluation by human annotation. To the best of our knowledge, no

study has been found on evaluating speech naturalness on acted dialogue. Given the benefits of such future applications resulted from this study, e.g., actors evaluation, it is worth investigating the effectiveness of supervised machine learning methods to build the automatic speech naturalness recognition by machine. The rest of this paper describes the problems, the dataset, the acoustic features, the classifiers, metrics to measure the performances, the experiment results and their discussions, and, finally, the conclusions.

II. Problem Statement

The problem on this study can be stated as follows:

1) Given speech utterances (provided in .wav files) from acted dialogues with naturalness recognition labels measured at 5-point scales, is it possible to recognize these labels automatically using neural networks?
2) How to perform automatic naturalness recognition and evaluate the methods?

III. Methods

A. Dataset

This research employed MSP-IMPROV dataset [8]. The dataset is a corpus of acted dialogue to study emotion perception via multimodal information (audio and visual modalities). There are 8438 utterances divided into four scenarios: Target-improvised, Other-improvised, Target-read, and Natural interaction. The number of speakers is 12 in six sessions, with two speakers for each session. For splitting into training and test partitions, the first five sessions were allocated for training (6816 utterances), while the rest of the sixth session was for the test (1622 utterances).

The MSP-IMPROV dataset aims to promote naturalness in affective speech corpus. Hence, this dataset is suitable for testing a method for recognizing the naturalness of speech in acted dialogue. The average number of annotators in the MSP-IMPROV dataset was five raters for scenarios except Target - improvised. In this scenario, the average number of annotators was 28. Given this high number of annotators, the naturalness labels are more reliable than the smaller number of annotators. The naturalness labels were rated on a 5-point Likert-like scale from 1 (most acted) to 5 (most natural). These labels were converted into a floating-point scale [-1, 1] when we fed them into the model.

Table I shows an excerpt of utterances in the MSP-IMPROV dataset along with their naturalness ratings (labels). This naturalness information in speech is close to the category of paralinguistic information defined in [9]. The speakers attempt to perform natural speech (cf. acted speech). The degree of naturalness rating is an average value of at least five annotators. Although the reliability of the annotation in the dataset is not high (Cronbach's $\alpha = 0.44$) [10], this is the only publicly available speech dataset that includes a naturalness rating in the annotations. This low reliability of the naturalness rating (cf. dimensional emotion labels) also highlights the difficulty of the perceptual speech emotion recognition task. Similar results may apply to automatic recognition by computers.

TABLE I
Sample of utterances and their corresponding naturalness ratings from MSP-IMPROV dataset

Utterance	Rating
How can I not?	3.2
I'm so tired, I'm not looking for the class this morning	4.5
I can skip class. What do you talking about?	4.2
No, it is a really big deal	3.4
Yeah.	2.2
I'm quite sure that we will find some way or another	1.6

B. Acoustic Features

pyAudioAnalysis [5] is an open-source acoustic feature extraction tool for a wide range of audio analysis applications. Based on the previous research on automatic speech emotion recognition, it has been found that acoustic features extracted by this tool are effective to predict dimensional and categorical emotions [11], [12]. The extracted acoustic features by this tool even obtained better results than a specially-designed acoustic feature extractor [11].

We used both low-level descriptors (LLDs, extracted per frame) and high-level statistical functions (HSFs, extracted per utterance) to evaluate speech naturalness recognition. LLD represents frame information, while HSF generalizes information from all frames within an utterance. Either LLD or HSF may be informative for recognizing naturalness. Table II shows both types of features, which were used in this study.

TABLE II
List of acoustic features used for input; LLD: low-level descriptor; HSF: high-level statistical function

LLD	Zero crossing rate (ZCR), energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, 13 MFCCs, 12 chroma vectors, chroma deviation
HSF	Mean, Std

C. Classifiers

Two classifiers were evaluated in this study, LSTM and MLP. The architectures of these classifiers are briefly explained below. The choices of these classifiers are based on the previous findings in the speech emotion recognition task [13], [14].

An LSTM model consists of four layers with units of (512, 256, 128, 64), shown in Fig. 1. The number of layers and their unit was optimized via brute-force experiments. The choice and discussion about this number of layers are given in the next section. A batch normalization layer is added before four LSTM layers to speed up the computation process. All four LSTM layers return their all final output sequence; hence, a flatten layer is needed after the last LSTM layer. The last layer is a one-unit dense layer to predict the output of the naturalness score, ranging from -1 (very unnatural) to 1 (very natural). This classifier is implemented in the TensorFlow toolkit with Keras framework.

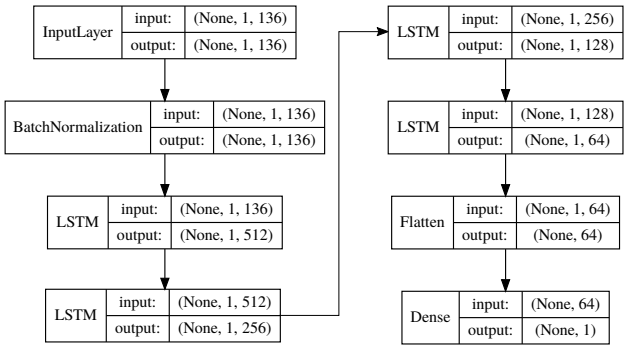

Fig. 1. Architecture of LSTM model

An alternative classifier is a classical MLP, which consists of three layers with units of (64, 32, 16). This small network showed better results than larger networks in some experiments. The comparison and discussion related to this evaluation are given in the "Results and Discussion." This network is adopted from [14], which performed well on the dimensional emotion recognition task. The network uses a logistic activation function, Adam optimization, ten patiences of stopping criteria, and an initial learning rate of 0.001. This classifier is implemented in scikit-learn toolkit [15].

D. Evaluation Metrics

We adopted metrics from speech emotion recognition for this dimensional speech naturalness recognition task. The motivation of this choice is the similarity of the annotation method, by a five-point scale, among both tasks. Three metrics are measured between predictions (x) and gold-standard labels (y), following the work in dimensional speech emotion recognition [16]. These metrics are CCC (concordance correlation coefficient), PCC (Pearson correlation coefficient), and RMSE (Root Mean Square Error). As in dimensional emotion recognition, CCC is chosen as the primary metric because of its robustness over PCC and RMSE.

The first metric CCC is formulated as follows:

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

where σ is the standard deviation, σ^2 is the variance, and μ is a mean value. ρ is the Person correlation coefficient (PCC) between two variables formulated as follows,

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

As the third metric is RMSE which measure the discrepancy between two continuous errors and is formulated as follows,

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}((x_i - \mu_x)^2 + (y_i - \mu_y)^2)}. \quad (3)$$

Finally, a public repository was made to reproduce the research reported in this study. The repository includes all codes and HSF input features and excludes the original speech dataset. The repository is available at `https://github.com/bagustris/snr`.

IV. Results and Discussion

We present results of the speech naturalness recognition study in four tables and discuss these results in three topics. We evaluate both LLD and HSF features on both MLP and LSTM classifiers on this speech naturalness recognition study. For each condition, we varied the number of layers and units in classifiers with 11 variations. Hence, there are 44 conditions in total (2 feature types $\times$ 2 classifiers $\times$ 11 variations). Tables III – VI show the results from MLP with LLD, LSTM with LLD, MLP with HSF, and LSTM with HSF, respectively.

As a baseline system, we chose MLP with LLD since this architecture combines a common approach of both classifier and features used in speech processing techniques. The results show poor CCC results (interpreted based on Altman [17]). In 11 variations, the best CCC score is 0.16 from 2 layers MLP with 512 and 256 units (Table III).

TABLE III
PERFORMANCE OF NATURALNESS RECOGNITION USING **MLP** CLASSIFIERS WITH **LLD** FEATURES IN DIFFERENT LAYERS AND UNITS

# layers	# units	CCC	PCC	RMSE
1	16	- 0.005	-0.008	0.334
2	32, 16	- 0.003	-0.025	0.313
3	64, 32, 16	0.106	0.197	0.307
4	128, 64, 32, 16	0.147	0.247	0.305
5	256, 128, 64, 32, 16	0.124	0.184	0.314
6	512, 256, 128, 64, 32, 16	0.115	0.237	0.304
5	512, 256, 128, 64, 32	0.112	0.268	0.301
4	512, 256, 128, 64	0.124	0.200	0.309
3	512, 256, 128	0.139	0.234	0.308
2	**512, 256**	**0.160**	**0.250**	**0.304**
1	512	0.126	0.200	0.310

On the LSTM networks, the highest CCC score improved to 0.269 using three layers of LSTM with 512, 256, and 128 units (Table IV). This result shows the benefit of utilizing deep neural networks (DNN) over a neural network (NN) approach. On the same input features, the performance improved from poor to moderate.

TABLE IV
PERFORMANCE OF NATURALNESS RECOGNITION USING **LSTM** CLASSIFIERS WITH **LLD** FEATURES IN DIFFERENT LAYERS AND UNITS

# layers	# units	CCC	PCC	RMSE
1	16	0.127	0.149	0.438
2	32, 16	0.138	0.169	0.476
3	64, 32, 16	0.211	0.224	0.374
4	128, 64, 32, 16	0.225	0.241	0.356
5	256, 128, 64, 32, 16	0.255	0.260	0.357
6	512, 256, 128, 64, 32, 16	0.115	0.237	0.304
5	512, 256, 128, 64, 32	0.230	0.260	0.367
4	512, 256, 128, 64	0.242	0.247	0.360
3	**512, 256, 128**	**0.269**	**0.274**	**0.357**
2	512, 256	0.143	0.134	0.431
1	512	0.131	0.161	0.343

On the third and four conditions, we adopted global feature sets based on [18] and on the basis of [19]. Similar to the dimensional speech emotion recognition task, we observe

significant improvements in HSF application over LLD. On the MLP approach, the improvement of the CCC score is from 0.16 to 0.228. This moderate result was obtained using a three-layer network with 64, 32, and 16 units or nodes (Table V).

TABLE V
PERFORMANCE OF NATURALNESS RECOGNITION USING **MLP** CLASSIFIERS WITH **HSF** FEATURES IN DIFFERENT LAYERS AND UNITS

# layers	# units	CCC	PCC	RMSE
1	16	0.215	0.302	0.299
2	32, 16	0.220	0.294	0.302
3	**64, 32, 16**	**0.228**	**0.311**	**0.299**
4	128, 64, 32, 16	0.210	0.286	0.302
5	256, 128, 64, 32, 16	0.203	0.287	0.302
6	512, 256, 128, 64, 32, 16	0.219	0.294	0.302
5	512, 256, 128, 64, 32	0.214	0.293	0.301
4	512, 256, 128, 64	0.213	0.305	0.298
3	512, 256, 128	0.196	0.279	0.302
2	512, 256	0.199	0.284	0.302
1	512	0.197	0.288	0.300

Next is the last condition, LSTM layers with HSF features. This architecture, the proposed method in this research, achieves the highest CCC scores among other conditions. A number of 68 statistics (34 means and 34 stds) are fed into LSTM network per utterance. Although LSTM usually performed on different time steps, we use a single time step since it proved to be useful on the previous SER research [11]. In this case, the network will perform three forward passes instead of a single pass. Using a more extensive network – 4 LSTM layers with 512, 256, 128, 64 units – obtained a CCC score of 0.302 (Table VI). The trends obtained by CCC scores are similar to those of PCC and RMSE, meaning that if CCC improved, PCC also improved while RMSE decreased (the smaller error, the better). Among these three metrics, we choose CCC as the primary metric due to its superiority among others [20].

TABLE VI
PERFORMANCE OF NATURALNESS RECOGNITION USING **LSTM** CLASSIFIERS WITH **HSF** FEATURES IN DIFFERENT LAYERS AND UNITS

# layers	# units	CCC	PCC	RMSE
1	16	0.258	0.273	0.363
2	32, 16	0.245	0.259	0.363
3	64, 32,16	0.268	0.290	0.347
4	128, 64,32,16	0.245	0.272	0.346
5	256, 128, 64, 32, 16	0.280	0.299	0.360
6	512, 256, 128, 64, 32, 16	0.300	0.314	0.357
5	512, 256, 128, 64, 32	0.284	0.313	0.330
4	**512, 256, 128, 64**	**0.302**	**0.327**	**0.339**
3	512, 256, 128	0.267	0.286	0.355
2	512, 256	0.273	0.299	0.345
1	512	0.274	0.280	0.353

In addition of those results, the following discusses an interpretation of the general results, some important issues regarding the different input features and classifiers, the different number of layers and its nodes, and naturalness cf. dimensional emotion recognition.

A. Interpretation of Automatic Naturalness Recognition Results

This research investigates the feasibility of building an automatic naturalness recognition system from an acted dialogue. We show that it is feasible to build a speech naturalness recognition system using a common speech processing technique: acoustic feature extraction and classification. The results show poor to moderate CCC score performances, ranging from 0.160 to 0.302 for the best of each pair of a feature set and a classifier. This interpretation is based on [17] where CCC less than 0.2 is categorized as poor, higher than 0.8 is categorized as excellent, and in between these scores (0.2 – 0.8) is categorized as moderate.

It will be beneficial to compare the results obtained here by automatic recognition with human annotation to gain insight into the current performance of naturalness recognition by machines vs. by humans. For comparison, in [21], the authors reported that the performance of their automatic speech emotion recognition system is close to human performance. However, since the labels in this naturalness recognition study are from human annotation, that comparison can't be made. The only metric showing the performance of human annotation is Cronbach's alpha, which measures the internal consistency among annotators. In this case, the naturalness score is 0.44, which is categorized as bad according to [22]. Note that in Table 9 in the reference [8], the reported scores are the classification performance of automatic emotion recognition among different naturalness scores, not the performance of naturalness recognition.

B. Comparing Input Features and Classifiers

This paper aims to study the possibility of recognizing naturalness in a speech via a neural network and deep neural networks as classifiers. The input to the classifiers is acoustic features. We evaluated two different input features and two different classifiers. The evaluation is carried out mainly by measuring the CCC scores.

On comparing the different input features on the same classifier, we observe that HSF performed better than LLD. It means that Std+Mean are also better at representing information related to the naturalness of speech than LLD (in addition to the speech emotion recognition task). Aside from the small feature size compared to LLD, the processing time needed by HSF features are also small. HSF takes seconds to minutes to finish, while LLD takes minutes to hours. Mean+Std is intended to capture both commonalities of all features within an utterance and their dynamics as in the speech emotion recognition task. The result shows that this HSF representation is more informative than LLD to capture naturalness information in speech.

On comparing two classifiers on the same input features, we found that LSTM performed better than MLP. This result, in contrast, is different from [14], in which MLP performed better than LSTM in various experiments for the SER task. The LSTM network could model the input-output connection between HSF features and the naturalness labels. The past

information used in the unimodal LSTM may contribute to the naturalness recognition performance. Since we did not explore other classifiers, the obtained moderate results show a need to go beyond these standard classifiers, aside from the acoustic features that highly correlate with naturalness in speech.

Between acoustic feature and classifier, the contribution of the latter is higher than the first. Changing MLP to LSTM improves CCC from 0.16 to 0.269 on LLD features and from 0.228 to 0.302 on HSF features. Nevertheless, both research directions – acoustic features and classifiers – are suggested to be explored to continue this initial study on automatic speech naturalness recognition.

C. Varying the number of layers and units

A deep neural network (DNN) is an evolution of the conventional neural network (NN) by applying deeper layers to learn representation at the input and map them to the labels. We compared MLP to LSTM with various numbers of layer and unit (node). The number of layers is evaluated from one to six, with 11 variations as shown in Table III – VI.

A baseline for all pairs of a feature set and a classifier is one MLP or LSTM layer with 16 nodes. The choice of the single layer and small nodes is to investigate whether speech naturalness recognition can be modeled in the simplest form. As shown in Table III with negative scores, this single layer NN cannot model the speech naturalness recognition with LLD features and MLP classifiers. The obtained results show poor to moderate CCC scores in the other three pairs of features and classifiers, meaning that we can build the model within the system with a single layer NN with a proper combination of feature and classifier.

Using DNN – NN with more than a layer – improves the recognition rate by CCC score. We increased the number of layers by doubling the number of nodes in the previous layer. For instance, if the first single layer is 16 nodes, then in the next variation with two layers are 32 and 16 nodes, and so on. After six layers, we reduced the last layer one by one. For instance, (512, 256, 128, 64, 32) is reduced to (512, 256, 128, 64) in the next variation. The last variation will only have a single layer with 512 nodes/units. We found that the optimum performance for each pair of a feature and a classifier is within the boundary of one layer to six layers. It means that the search boundary is sufficient for the number of layers and nodes evaluated in this study. The optimum number of layers is different in each pair of input features and classifier, showing a dependency of this variable to the input features and the classifier.

Among four best results for each pair of a classifier and an acoustic feature set, the highest is achieved with the deepest and widest network, i.e., HSF with LSTM on four layers with (512, 256, 128, 64) units. However, it is of interest to discover that thinner network with (64, 32, 16) units performed the best result on a pair of MLP and HSF. While the authors cannot explain this phenomenon except that dependency between the features and the classifiers occurs, this topic could be investigated for future research. Aside from this topic, data augmentation and regularization are also to worth study in this new application of speech processing techniques.

D. Naturalness and dimensional emotion recognition

Although we did not perform dimensional recognition in this study, one can make a direct connection between results obtained in this naturalness recognition study and the previous dimensional emotion recognition study [14]. This topic aims to find the difficulties of recognizing naturalness and emotion in a speech by automatic recognition systems. We treated the same speaker-independent scenario and the same CCC scores to judge the performance as previous emotion recognition study. The highest CCC score obtained on this research for naturalness recognition is higher than valence recognition (0.290) but lower than arousal and dominance recognition (0.556 and 0.402). These results, in contrast, are different from the agreement of the evaluators of the dataset reported in [10]. In the Cronbach's alpha (α) statistics, the order of the agreement is valence, arousal (activation), dominance, and naturalness (0.89, 0.73, 0.54, and 0.44). Our combined studies (this naturalness study and previous emotion recognition study) reveal the order of difficulty from the hardest to the easiest is arousal, dominance, naturalness, and valence. Note that the annotators voted the emotion and naturalness labels through audio and visual information (video), while our naturalness recognition system only processed audio information. The multimodal process may contribute to the different order of difficulty levels to recognize naturalness and emotion. Our results suggest that for automatic recognition, recognizing naturalness in speech is easier than recognizing valence but is harder than recognizing arousal and dominance.

V. Conclusions

This pilot study demonstrated an automatic naturalness recognition from speech through neural network and deep neural network techniques as classifiers. The first is composed of a single-layer network, while the latter is composed of multiple layers. We addressed two issues, the possibility to recognize naturalness in a speech by such methods and metrics to evaluate these methods. The answer to the first issue is that it is possible to recognize naturalness in speech automatically by adopting the classical building blocks in speech processing. The main building blocks are dataset, feature extraction, and classification. The acoustic features are extracted from the speech dataset on both frame and utterance bases. The classifiers, MLP and LSTM, acquire acoustic features and map them to the correspondence labels. Three metrics were evaluated to investigate the second issue on measuring the performance of speech naturalness recognition methods: CCC, PCC, and RMSE. The CCC was chosen as the primary metric since it is more challenging (lower score) and is used in speech emotion recognition tasks. The combination of high-level acoustic features with the LSTM classifier achieves the highest performance in this study.

We plan to study the acoustic features that correlate with the naturalness of speech and concurrent naturalness and emotion

recognition in speech for future studies. The appropriateness of features can vary from unnatural speech to natural speech, as revealed in other domains [23]. Combining naturalness and emotion recognition in speech is an integrated way to build an intelligent system for several tasks, concurrently.

VI. Acknowledgment

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

References

[1] E. A. Antipov and E. B. Pokryshevskaya, "How to Measure the Power of Actors and Film Directors?" *Empir. Stud. Arts*, vol. 34, no. 2, pp. 147–159, 2016.

[2] P. Mairano, E. Zovato, and V. Quinci, "Do sentiment analysis scores correlate with acoustic features of emotional speech?" in *AISV Conf.*, 2019.

[3] S. Blanton, "The voice and the emotions," *Q. J. Speech*, vol. 1, no. 2, pp. 154–172, jul 1915. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00335631509360475

[4] J. S. Damico and M. J. Ball, "The SAGE Encyclopedia of Human Communication Sciences and Disorders," *SAGE Encycl. Hum. Commun. Sci. Disord.*, 2019.

[5] T. Giannakopoulos, "pyAudioAnalysis: An open-source python library for audio signal analysis," *PLoS One*, vol. 10, no. 12, pp. 1–17, 2015.

[6] B. Merritt and T. Bent, "Perceptual Evaluation of Speech Naturalness in Speakers of Varying Gender Identities," *J. Speech, Lang. Hear. Res.*, vol. 63, no. 7, pp. 2054–2069, jul 2020. [Online]. Available: http://pubs.asha.org/doi/10.1044/2020_JSLHR-19-00337

[7] R. R. Martin, S. K. Haroldson, and K. A. Triden, "Stuttering and speech naturalness," *J. Speech Hear. Disord.*, vol. 49, no. 1, pp. 53–58, feb 1984. [Online]. Available: http://pubs.asha.org/doi/10.1044/jshd.4901.53

[8] C. Busso, S. Parthasarathy, A. Burmania, M. Abdelwahab, N. Sadoughi, E. M. Provost, S. S. Member, S. Parthasarathy, S. S. Member, A. Burmania, M. Abdelwahab, N. Sadoughi, and E. Mower Provost Member, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, jan 2017. [Online]. Available: http://dx.doi.org/10.1109/TAFFC.2016.2515617

[9] H. Fujisaki, "Prosody, Information, and Modeling with Emphasis on Tonal Features of Speech," in *Work. Spok. Lang. Process.*, 2003.

[10] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 374–388, 2016.

[11] B. T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning," *APSIPA Trans. Signal Inf. Process.*, vol. 9, no. May, p. e17, may 2020.

[12] S. Sahu, V. Mitra, N. Seneviratne, and C. Espy-Wilson, "Multi-Modal Learning for Speech Emotion Recognition: An Analysis and Comparison of ASR Outputs with Ground Truth Transcription," in *Interspeech 2019*. ISCA: ISCA, sep 2019, pp. 3302–3306.

[13] M. Macary, M. Lebourdais, M. Tahon, Y. Estève, and A. Rousseau, "Multi-corpus Experiment on Continuous Speech Emotion Recognition: Convolution or Recurrence?" in *Int. Conf. Speecn Comput.*, 2020, pp. 304–314.

[14] B. T. Atmaja and M. Akagi, "Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition," in *2020 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2020 - Proc.*, Auckland, 2020, pp. 325–331.

[15] 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, É. Duchesnay, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825—-2830, 2011.

[16] 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 *AVEC 2016 - Proc. 6th Int. Work. Audio/Visual Emot. Challenge, co-located with ACM Multimed. 2016*, 2016, pp. 3–10. [Online]. Available: http://dx.doi.org/10.1145/2988257.2988258

[17] D. G. Altman, *Practical statistics for medical research*. CRC press, 1990.

[18] M. Schmitt, N. Cummins, and B. W. Schuller, "Continuous Emotion Recognition in Speech Do We Need Recurrence?" in *Proc. Interspeech 2019*. ISCA: ISCA, sep 2019, pp. 2808–2812.

[19] 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.

[20] V. Pandit and B. Schuller, "The many-to-many mapping between concordance correlation coefficient and mean square error," *arXiv*, pp. 1–32, 2019.

[21] Y. Chiba, T. Nose, and A. Ito, "Multi-stream attention-based BLSTM with feature segmentation for speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, 2020, pp. 3301–3305. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1199

[22] M. Wati, S. Mahtari, S. Hartini, and H. Amalia, "A Rasch model analysis on junior high school students' scientific reasoning ability," *Int. J. Interact. Mob. Technol.*, vol. 13, no. 7, pp. 141–149, 2019.

[23] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *2015 Int. Conf. Affect. Comput. Intell. Interact. ACII 2015*, 2015, pp. 698–704.