ENTRAINMENT ANALYSIS FOR ASSESSMENT OF AUTISTIC SPEECH PROSODY USING BOTTLENECK FEATURES OF DEEP NEURAL NETWORK

Keiko Ochi[1], *Nobutaka Ono*[2], *Keiho Owada*[3], *Miho Kuroda*[3], *Shigeki Sagayama*[4], *Hidenori Yamasue*[3,5]

[1] Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2] Department of Computer Science, Tokyo Metropolitan University, Hino, Japan
[3] Faculty of Medicine Tokyo University, Tokyo, Japan,
[4] Professor Emeritus, University of Tokyo, Tokyo, Japan,
[5] Department of Psychiatry, Hamamatsu University School of Medicine, Hamamatsu, Japan

ABSTRACT

In the present study, we quantify entrainment characteristics of conversation with the aim of automatic assessment of the severity of autism spectrum disorder (ASD). We focus on pairs of utterances immediately before and after turn-takings, which have prosodic/acoustic similarities.

The clinical severity of ASD is estimated by the bottleneck features obtained by an hourglass-shaped deep neural network (DNN) in the neural entrainment distance (NED) method used to measure the degree of entrainment. The DNN is firstly pre-trained using a large conversation corpus in various daily situations and then fine-tuned with conversations during the Autism Diagnostic Observation Schedule (ADOS) assessment. Absolute difference vectors are calculated from the bottleneck feature vectors between a pair of utterances. Centroid and variance of the absolute difference vectors are combined with the speech features discovered in our previous study in order to estimate the scores of ASD severity.

Consequently, the estimated scores significantly correlate with the actual observed ADOS 'Reciprocity' scores with a coefficient of 0.70. This result shows the effective use of fine-tuning technique with data of typically developed individuals and, furthermore, reveals social communication deficits in ASD individuals represented by utterances adjacent to turn-takings.

Index Terms— entrainment, fine-tuning, spoken dialogue, turn-taking, autism spectral disorder

1. INTRODUCTION

Autism spectrum disorder (ASD) is a widely prevalent neurodevelopmental disorder: one out of 54 children have ASD [1]. The core symptoms of ASD are characterized by deficits in social communication and interactions, including nonverbal communicative behaviors such as prosody. Currently, the diagnosis and assessment of ASD rely on the subjective rating by trained experts. For instance, the Autism Diagnostic Observation Schedule (ADOS), one of the gold-standard diagnosis tools used worldwide, requires a qualified rater to assess the video recording of the interaction between a testee and a qualified administrator. The administration and rating are also time-consuming, thus it is difficult to use repeatedly. This limitation hinders its frequent use, which is necessary to detect time-course changes in clinical severity for the development of novel treatment or medication.

To overcome these problems, the development of an objective and quantitative assessment is desirable. In recent years, many attempts to automatically diagnose and assess ASD have been successfully made utilizing easy-to-obtain biomarkers such as audio or video signals because they reflect the characteristics of communicative interactions [2]-[13]. In addition to the pathognomonic characteristics of prosody that are particularly frequently studied [2]-[8], some studies pointed out that turn-taking features such as turn-taking gaps are also essential descriptors of core symptoms [8]-[10]. From these observations, turn-taking is considered as one of the critical factors in the interaction in people with ASD.

Entrainment is a research topic that attracts attention in conversational analysis fields. In general, in a conversation, the prosodic or linguistic features between two speakers correlate when the conversation or the ongoing task is successfully reciprocal [11]-[13]. Recently, it has been revealed that entrainment in prosody in a conversation is less in the speakers with ASD than those with typical development (TD) [14][15]. Wynn *et al.* found that adults with TD show entrainment in speech rate while it is not observed in those with ASD [14]. Lehnert-LeHouiller and her colleagues showed that children with ASD have deficits of prosodic entrainment, whereas neurotypical peers make their prosody similar to their interlocutor in the course of conversation [15]. Our previous research also showed that the correlation of prosodic features in a conversation could estimate the ADOS score over a time course [10].

In the analyses of entrainment, the prosodic similarity is observed between the utterances just before and after turn-taking [16][17]. Prosodic features such as the fundamental frequency (F_0) range and intensity range of the last utterance of a turn correlate with that of another speaker's subsequent turn. From the observation that turn-taking manner is affected by ASD, the adjoint utterances around turn-taking could be a good descriptor of conversational characteristics of ASD; however, the correlation between the utterances just before and after turn-taking is not studied well for the conversation of people with ASD.

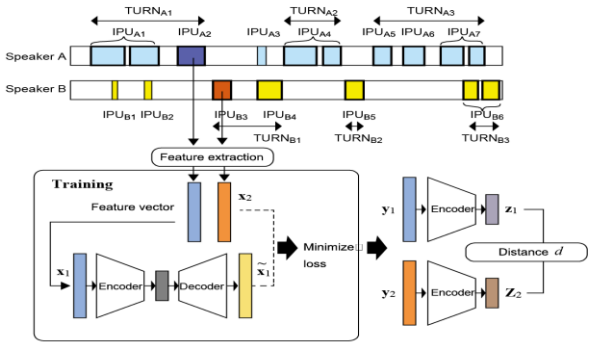

Fig. 1 Overview of NED-based approach

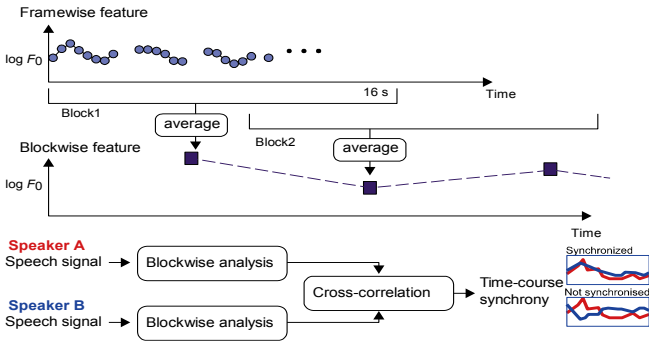

Fig. 2 Blockwise analysis of prosodic features

Some studies have proposed to model the degree of entrainment during conversation [18][19]. Nasir *et al.* developed a novel approach of unsupervised machine learning using an hourglass-shaped DNN, which produces the features of the next turn from the input of the features of the preceding turn [19]. It can be considered that the bottleneck feature vector represents the essential information concerning the features of the next turn. They showed that the distance between the two bottleneck features (called 'Neural Entrainment Distance (NED)') represents the degree of entrainment in a conversation.

In our study, we adopt the NED approach to the clinical application by modifying some input/output features, and we measure the entrainment using the bottleneck feature of the DNN. We hypothesize that the degree of entrainment of the utterances just before and after the turn-taking is lower in adults with ASD. In some related studies, deep autoencoders are embraced in the detection of specific sounds which rarely arise, such as anomaly sound monitoring [20]-[22]. Because the speech data of individuals with ASD is limited, we attempt to detect them as rare samples in a semi-supervised manner. We consider that the autoencoder-like DNN trained by TD data can detect turn-taking manner specific for speakers with ASD.

As in other speech evaluation areas, DNN based approaches are actively utilized for automatic diagnoses/assessment [3]-[5]. Acoustic features and linguistic features effectively estimated the severity of ASD using the BERT-based approach with a correlation between estimated and observed scores as high as 0.60 [4]. However, collecting a large corpus of autistic speakers is not easy because of privacy problems and clinical diagnosis requirements. Thus, we conduct pre-training using task-nonlimited corpus without considering ASD. In the subsequent fine-tuning, the semi-structural conversation of ADOS activity is used to fit the DNN to the task.

2. RELATED WORKS

2.1. NED-based approach

The concept of the NED-based approach [19] is dimensional compression using an autoencoder-like hourglass-shaped DNN. The input and output acoustic/prosodic features are extracted from pairs of utterances, the final inter-pausal unit (IPU) of a turn (IPU$_{A2}$), and the initial IPU of the next turn (IPU$_{B3}$). The DNN is trained to minimize the loss between the output and the feature vector of IPU$_{B3}$ (See. Figure 1). After training, the degree of entrainment of an arbitrary pair of IPUs can be obtained by measuring the distance of the output vectors of the encoder ($\mathbf{z}_1$, $\mathbf{z}_2$).

2.2. Automatic assessment using synchrony

In our previous work, we used time-course synchrony within a whole session, together with other speech features [10]. The speech features consist of prosodic features and voicing- or turn-related features. Fig. 2 shows the flow of prosodic feature extraction. Firstly, framewise prosodic features (F_0 and intensity) are extracted at the 10 ms frame step. Blockwise prosodic features were calculated by averaging within each block of 16 s. We obtained the within-session statistics of the framewise and blockwise features. The degree of prosodic synchrony was calculated as the cross-correlation between the blockwise prosodic features of two speakers. Voicing- or turn-related features consist of speech rate, the within-session mean and SD of turn-taking gap, time ratio of pause to turn. The best combination listed in Table 1 provided 0.59 of correlation between the estimated and observed ADOS score with regard to the 'Reciprocity' score.

3. DATASETS

3.1. Corpus of Everyday Japanese Conversation (CEJC)

CEJC includes everyday multiparty conversations in various settings [23]. We used the Monitor Version of CEJC, which consists of 118 conversational situations among three to nine people. 44.4% of the participants were male speakers. The duration of conversations is about 4-5 min on average. In the recording of CEJC, a total of about 50 hours of audio data was used, recorded by an IC recorder hanging from speakers' necks. In this corpus, utterances are annotated in long utterance units (LUUs) [24] for each speaker.

3.2. ADOS administration data

We used the semistructured dialog between a clinician and a participant to fine-tune the model (for details, see our earlier

Table 1. Selected 9 features used for the estimation of Reciprocity

Category	Feature
Framewise	Mean of log F_0
Blockwise	SD of blockwise mean of log F_0
Framewise	SD of intensity
Blockwise	SD of blockwise mean of intensity
Blockwise	Corr. of blockwise mean of intensity
Timing	Speech rate [mora/s]
Voicing-related	Mean of log speaking time
Turn-taking-related	Mean of log turn-taking gap

paper [10]). The data was collected as a part of a clinical trial of medicine (Oxytocin nasal spray) in the University of Tokyo. The clinical trial was approved by the institutional review board of University of Tokyo Hospital and registered (UMIN000015264). The participants were informed beforehand through written consent.

A total of 65 male adults with ASD aged 18-48 years and 17 age-/intelligence-matched male controls aged 21-34 years received Activity 7 in ADOS Module 4. This activity includes questions and answers about emotions. The duration of a session ranged from 101 to 389s. One participant withdrew their agreement, and two participants' recordings were incomplete and so were excluded. In addition, one participant, whose numbers of turn-takings and utterances were not enough for analysis, was also excluded in this study.

As for the participants with ASD, this dataset includes the ADOS score: a psychiatrist (HY) administrated all activities ADOS on the participants, and a qualified psychiatrist (KM) rated the score by watching the video recordings of ADOS administration. The IPU segments are manually added for each speaker.

4. PROPOSED APPROACH

4.1. Turn-taking detection

In this study, to automatically extract turn-taking and its preceding/subsequent utterances, we heuristically regarded the point the speaker changed with another speaker's utterance longer than 500 ms duration as turn-taking.

4.2. Feature extraction

While adhering fundamentally to the NED-based approach, we introduced the speech features considered to be related to the characteristics of ASD. Prior to dimension compression in the bottleneck architecture, we selected input features that embody the dynamic and static characteristics of prosodic and acoustic features. Additionally, we used the modified DNN with an asymmetric shape by adding a supplemental output feature (turn-taking gap).

For each frame in utterances (e.g., a pair of IPU_{A2} and IPU_{B3} in Fig. 1), F_0, intensity, and mel-frequency cepstral coefficients (MFCCs) were extracted. The number of dimensions of the MFCCs was 13 in this study. These features are converted to Z-score by using the following normalization:

$$\overline{x_k} = \frac{x_k - m}{s} \quad (1),$$

where x_k, m, and s are the features at kth frame, and the mean and SD values throughout a session, respectively.

After that, the delta- and delta-deltas of them were calculated as dynamic features. We also conducted line-fitting and curve-fitting of the second-order polynomial to the framewise features. Specifically, we obtained the coefficients of a_1, a_2, b_1, b_2, and b_3 to minimize $\sum_k (y_k - a_1 k + a_2)^2$ and $\sum_k (y_k - b_1 k^2 + b_2 k + b_3)^2$ for each IPU, where y_k denotes the framewise features (x_k, its delta or delta-delta). Finally, we used seven statistics, the mean, SD, the five coefficients mentioned above, as representative values. Thus, a total of 315 (15×3×7) dimensions were obtained for each IPU.

In addition, the SD values of unnormalized F_0 and intensity were added to the input and output features of the DNN. Z-score normalization was carried out to eliminate individual differences and capture the relative change of the prosodic features. However, the unnormalized SD is considered to be necessary to distinguish whispering or monotonous small voice characteristic of autistic people [10]. We used Praat [25] for the extraction of F_0 and intensity values. Therefore, the input and output vectors of the DNN had 316 and 317 dimensions, respectively.

4.3. Pre-training

We first conducted pre-training of the DNN using the CEJC for brief training of the model with the general conversational settings. A 5-layer fully connected DNN, with hidden layers consisting of $64-16-64$ hidden neurons was implemented using Keras. Although CEJC is annotated in an LUU unit, the fine-tuning in the subsequent step could compensate for the difference in the utterance unit of the two datasets.

The mean-squared-logarithmic error (MSLE) was used for the loss function, and Adam optimizer was applied. The number of epochs was set to 100. The ReLU and linear activation functions were used for the hidden layers and output layer, respectively. For training, 88.3% of 39152 utterance samples were used for training and the rest was used for validation. Although three or more people speak in each session of CEJC, turn-taking occurs between the two of them. Therefore, the input feature vector can be obtained from the multiparty conversation.

4.4. Fine-tuning

We conducted fine-tuning using the dialog between a participant and an administrator of Activity 7 in ADOS to modify the pre-trained DNN to fit the semistructured interview setting. Only the data of participants with TD were used to capture the tendency of entrainment in typical cases. In this process, the weights of the first hidden layer were fixed. We used 4438 training samples and 178 validation samples.

4.5. Bottleneck features and statistics of their features

We obtained the bottleneck features of the DNN by inputting the speech features of Activity 7 of the participant with ASD. To calculate the relationship between the utterances just before and after turn-taking, we input both of the utterances. The output of the encoder has 16 dimensions; therefore, for each participant, we obtain the same number of pairs of bottleneck feature vectors as the turn-takings. Firstly, let us input feature vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ to obtain the output bottleneck feature vectors, $\mathbf{z}_1$ and $\mathbf{z}_2$. The absolute difference between the pair feature vectors is given as follows:

$$w_k = |z_{1k} - z_{2k}| \quad (4)$$

where w_k denotes the kth component of the output of the encoder, z_{1k} and z_{2k}, respectively. After that, the feature vectors within a session were aggregated into two 16-dimensional vectors, the mean and SD vectors, whose components represent the centroid and variance.

4.6. Regression analysis

We performed a regression analysis using Support Vector Regression (SVR) to investigate whether the proposed bottleneck features can estimate the severity (ADOS score); SVR is considered suitable to fit the limited data size. We used the 'e170' package in the statistical environment R [26]. A linear kernel and a leave-one-out cross-validation were used to estimate three categories of ADOS scores, 'Reciprocity', 'Communication', and 'Repetition'. Their maximum scores are 14, 8, and 4, respectively.

We used the proposed bottleneck features of 32 dimensions in addition to the speech features selected in the previous study (e.g., for Reciprocity score estimation shown in Table 1. For Communication and Repetition, see [10]). We selected the optimal combination of the dimensions of the bottleneck features using Forward Feature Selection (FFS), where we select the best feature adding dimensions one-by-one.

5. RESULTS

Table 2 shows the correlation coefficients of the estimated and observed ADOS scores and the mean absolute errors (MAEs) of the estimated values. For comparison, we show the results of our previous work (baseline) and the setting when the fine-tuning was not used (w/o fine-tuning). The proposed method provided the best correlation and MAE for all three score categories. As for the Reciprocity score, four components from the mean vector and nine components from the SD vector were selected by FFS.To observe the bottleneck feature when utterances of the ASD and TD groups are used as inputs, we show examples of their components with the rated ADOS score in Fig. 3. The separated distribution of the two groups can be interpreted as these components representing some aspects of the difference between the groups.

6. DISCUSSIONS AND CONCLUSIONS

The results showed that the 'Proposed' method achieved the best accuracies in the regression experiment. This indicates

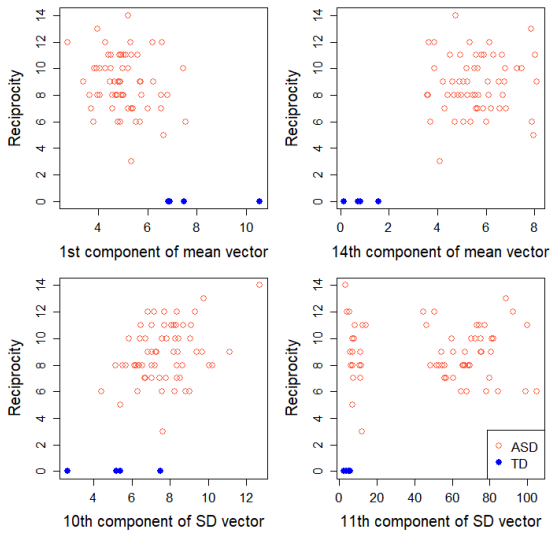

Fig. 3 Examples of scatterplots between the statistics of absolute difference of a component of bottleneck features and Reciprocity scores. Note that TD individuals were not rated in ADOS, although they are plotted at zero. In drawing this figure, we omitted four randomly selected TD participants from the fine-tuning to observe the bottleneck features when.using them as inputs.

Table 2. Correlations and MAE (MAE shown in parentheses) between the estimated and the observed ADOS score.

Method	Reciprocity	Communication	Repetition
Baseline	0.59 (1.23)	0.49 (0.96)	0.18 (1.23)
w/o fine-tuning	0.60 (1.23)	0.59 (0.90)	0.49 (0.61)
Proposed	**0.70** (1.18)	0.62 (0.84)	0.49 (0.61)

the effectiveness of fine-tuning using the TD participants' data. In addition, adjacent utterances just before and after turn-taking can be characterized by deficits in the social communication of people with ASD. In the leave-one-out cross-validation, the autistic participant with the lowest reciprocity score (3) had an incorrectly higher estimated score (7.99), whereas the participant with the most severe Reciprocity score (14) was successfully estimated (14.19). The lowest-score participant repeated the question word with a restrained voice, regardless of the speech volume of the administrator. His ADOS reciprocity sub-score regarding non-verbal communication with speech prosody was rated as high (i.e. 2 among score range 0−2) whereas other sub-item scores related to other reciprocal features such as facial expression were low. This may lead to an incorrectly estimated high score. Future works include applying the automatic evaluation of turn-taking and prosodic/acoustic entrainment to therapies such as computer-assisted social skill training.

6. ACKNOWLEDGEMENTS

This research was supported by Japan Agency for Medical Research and Development (AMED) under Grant Number JP16dm0107134 and a JSPS KAKENHI Grant-in-Aid for Scientific Research (A) (Grant Number: 20H00613).

7. REFERENCES

[1] M. J. Maenner, K. A. Shaw, J. Baio, A. Washington, M. Patrick, M. DiRienzo, et al., "Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years ─ Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016," *Morbidity and mortality weekly report Surveillance summaries*, vol.69, no. 4, pp.1─12J, 2020.

[2] Y. Nakai, R. Takashima, T. Takiguchi, S. Takada, "Speech Intonation in Children with Autism Spectrum Disorder," *Brain and Development*, vol. 36, no. 6, pp. 516–522, 2014.

[3] M. Eni, I. Dinstein, M. Ilan, I. Menashe, G. Meiri, and Y. Zigel, "Estimating Autism Severity in Young Children from Speech Signals using a Deep Neural Network," *IEEE Access*, vol. 8, pp. 139489–139500, 2020.

[4] T. Saga, H. Tanaka, H. Iwasaka, and S. Nakamura, "Objective Prediction of Social Skills Level for Automated Social Skills Training Using Audio and Text Information," In Companion Publication of the 2020 International Conference on Multimodal Interaction, pp. 467–471, 2020.

[5] S. Levy, M. Duda, N. Haber, and D. P. Wall, "Sparsifying Machine Learning Models Identify Stable Subsets of Predictive Features for Behavioral Detection of Autism," *Molecular Autism*, vol. 8, no.1, pp. 1–17, 2017.

[6] M. Kumar, P. Papadopoulos, R. Travadi, D. Bone, and S. Narayanan, "Improving Semi-Supervised Classification for Low-Resource Speech Interaction Applications," *Proceedings of ICASSP 2018*, pp. 5149–5153, pp. 2018.

[7] D. Bone, M. P. Black, A. Ramakrishna, R. Grossman, and S. S. Narayanan, "Acoustic-Prosodic Correlates of 'Awkward' Prosody in Story Retellings from Adolescents with Autism," *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] D. Bone, S. Bishop, R. Gupta, S. Lee, and S. S. Narayanan, "Acoustic-Prosodic and Turn-Taking Features in Interactions with Children with Neurodevelopmental Disorders," *Proceedings of Interspeech* 2016, pp. 1185–1189, 2016.

[9] P. Heeman, R. Lunsford, E. Selfridge, L. Black, L. van Santen "Autism and Interactional Aspects of Dialogue," *Proceedings of 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue.*pp. 249–252, 2010.

[10] K. Ochi, N. Ono, K. Owada, M. Kojima, M. Kuroda, S. Sagayama, S., and H. Yamasue, "Quantification of Speech and Synchrony in the Conversation of Adults with Autism Spectrum Disorder," *PLoS ONE*, vol. 14, no. 12, e0225377, 2019.

[11] J. Thomason, H. V. Nguyen, and D. Litman, "Prosodic entrainment and tutoring dialogue success," *International Coference on Artificial Intelligence in Education*, pp. 750−753. 2013.

[12] M. Mizukami, K. Yoshino, G. Neubig, D. Traum, and S. Nakamura, "Analyzing the effect of entrainment on dialogue acts," *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 310−318, 2016.

[13] N. Lubold and H. Pon-Barry, "Acoustic-prosodic Entrainment and Rapport in Collaborative Learning Dialogues.," *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge* pp. 5−12, 2014.

[14] C. J. Wynn, S. A. Borrie, and T. P. Sellers, "Speech Rate Entrainment in Children and Adults with and Without Autism Spectrum Disorder," *American Journal of Speech-Language Pathology,* vol. 27, no. 3, pp. 965−974, 2018.

[15] H. Lehnert-LeHouillier, S. Terrazas, and S. Sandoval, "Prosodic Entrainment in Conversations of Verbal Children and Teens on the Autism Spectrum," *Frontiers in Psychology*, 11, 2718, 2020.

[16] R. Levitan, S. Benus, A. Gravano, and J. Hirschber, "Entrainment and Turn-taking in Human-human Dialogue," *2015 AAAI Spring Symposium Series*, 2015.

[17] V. Cabarrao, F. Batista, H. Moniz, I. Trancosoa nd A. Mata, "Acoustic-prosodic Entrainment in Structural Metadata Events," *Proceedings of Interspeech 2018*, pp. 2176−2180, 2018.

[18] Z. Rahimi and D. Litman, "Entrainment2vec: Embedding entrainment for multi-party dialogues," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8681−8688, 2020..

[19] M. Nasir, B. Baucom, S. Narayanan,and P. Georgiou, "Towards an Unsupervised Entrainment Distance in Conversational Speech Using Deep Neural Networks," *Proceedings of Interspeech 2018*, pp 3423−3427, 2018.

[20] O. I. Provotar, Y. M. Linder and M. M. Veres,. "Unsupervised Anomaly Detection in Time Series Using Lstm-based Autoencoders," *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)* pp. 513−517, 2019..

[21] R. Giri, F. Cheng, K. Helwani, S. V. Tenneti, U. Isik, amd A. Krishnaswamy, "Group Masked Autoencoder Based Density estimator for audio anomaly detection. In DCASE 2020 Workshop.

[22] Suefusa, K., Nishida, T., Purohit, H., Tanabe, R., Endo, T., & Kawaguchi, Y. (2020, May). Anomalous sound detection based on Interpolation Deep Neural Network," `*Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 271−275, 2020.

[23] H. Koiso, Y. Den, Y. Iseki, W. Kashino, Y. Kawabata, K. Nishikawa, et al., "Construction of the Corpus of Everyday Japanese Conversation: An Interim Report," *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* 2018

[24] T. Maruyama, Y. Den and H. Koiso "Design and Annotation of Two-level Utterance Units in Japanese," *Search of Basic Units of Spoken Language*, pp. 155-180, John Benjamins , 2020.

[25] P. Boersma, "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound. Proceedings of the Institute of Phonetic Sciences.," vol. 17, no. 1193, pp. 97–110, 1993..

[26] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, "e1071: Misc Functions of the Department of Statistics (e1071)," TU Wien. R package version 1. pp. 6–7. 2011.