Interspeech 2024
1-5 September 2024, Kos, Greece

Entrainment Analysis and Prosody Prediction of Subsequent Interlocutor's Backchannels in Dialogue

Keiko Ochi[1], *Koji Inoue*[1], *Divesh Lala*[1], *Tatsuya Kawahara*[1]

[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
{ochi.keiko.5f,inoue.koji.3x}@kyoto-u.ac.jp,kawahara@i.kyoto-u.ac.jp

Abstract

This study investigates the characteristics of backchannels showing the entrainment to the interlocutor's speech. The prosodic features of the dialogues of attentive listening are analyzed to describe how the prosody of Japanese backchannels is affected by the preceding interlocutor's utterance. We adopt a support vector regression (SVR) to model the relationships between the prosodic features of backchannels and those of the preceding utterances. As a result, we found an interrelationship between the different types of features; in particular, the F0 of backchannels is highly correlated with the power of the preceding utterance. The regression analyses show that the combination of prosodic features of the preceding utterances achieves good prediction of both the F0 and power of backchannels. The findings of this study can be applied to the automatic generation of backchannels for spoken dialogue systems to show empathy and facilitate user's speech.

Index Terms: dialogue, backchannel, entrainment, spoken dialogue system, prosody

1. Introduction

Backchannels in conversations play critical roles in encouraging human-to-human and human-to-robot interactions [1]–[9]. Verbal backchannel is characterized by its function of continuing the interlocutor's speech. It is defined by brief utterances made near the end of the speaker's turn, responding without disrupting the continuity of the speaker's turn. Listeners can facilitate the speaker's speech by placing backchannels, which show understanding and empathy. Fluent speech is accompanied by appropriate backchannels [1]–[5]; therefore, the generation of effective backchannels is required to develop autonomous spoken dialogue systems such as virtual agents and humanoid robots.

Backchannels have been widely investigated in a variety of studies, some of which aim to reveal the characteristics of human conversations, whereas others are applied to spoken dialogue systems [6]-[14]. Most studies on the automatic prediction of backchannels have been focused on predicting the timing or form of backchannels [14]-[16]. Other studies used prosodic features to detect backchannels of human speakers in conversational data [17][18].

Entrainment is known as a phenomenon reflecting the success of a conversation or an ongoing collaborative task. When entrainment occurs, speakers and their interlocutors speak similarly in terms of acoustic, prosodic, and lexical features.

Backchannels and their preceding utterances exhibit entrainment similarly to other dialogue behaviors [22]-[24]. These studies revealed correlations between prosodic features between backchannels and interlocutors' preceding utterances. Backchannels correlate with prior utterances in terms of pitch, power, and vocal quality (e.g., harmonic-to-noise ratio (HNR), jitter, or shimmer) [23][24]. On the basis of these findings, spoken dialogue system could effectively facilitate a conversation by providing entrained backchannels to users.

The above studies investigated the relationship of the same kinds of prosodic/acoustic features of backchannels and an interlocutor's speech. However, a few studies on entrainment investigated the correlation between the different types of prosodic features. Moreover, only a few studies made prediction of following backchannels with satisfactory performance. Therefore, in this study, we investigate the inter- and intra-relationship between the fundamental frequency (F_0) and power. It is expected that the correlation analyses of these prosodic features can contribute to the development of spoken dialogue systems that produce naturally empathetic and synchronized backchannels.

In this paper, firstly, we describe the dataset used for the analyses and its recording configurations. In the next section, the extraction of prosodic features and adopted statistical models are presented. Then, we present the results of the correlation analyses and prediction performance of prosodic features of backchannels. Finally, we discuss and conclude the observed phenomena related to entrainment and the feasibility of future application to spoken dialogue systems.

2. Dataset

We used a Japanese conversational speech corpus provided by the JST ERATO project, which consists of spoken dialogue via Wizard of Oz using android ERICA [25]. The conversation task was one-on-one attentive listening, in which the participant in the role of a speaker told his/her experiences and impressions to a listener. In this corpus, 31 of 59 dialogue sessions were conducted by female speakers. The talk topics were the memories of food or travel he/she enjoyed and each session lasted approximately eight minutes.

Four human operators, who are actors and are not among the 59 participants, played the role of a listener while teleoperating the android robot, and were instructed to respond using "attentive listening," which entailed receptively listening to the participants' stories. They were required to provide backchannels and ask probing questions about. We analyzed only the pairs of a listener's backchannel and the user's preceding utterance. All speech data were recorded at a sampling rate of 16 kHz, and the quantization bit was 16.

10.21437/Interspeech.2024-628

Table 1: *Statistics of prosodic features.*

Statistic	F_0	power
Mean	mean_f	mean_p
Median	med_f	med_p
Standard deviation (SD)	SD_f	SD_p
Interquartile range (IQR)	IQR_f	IQR_p
Maximum value	max_f	max_p
Minimum value	min_f	min_p
Slope of the first-order regression line	slope_f	slope_p

3. Methods

3.1. Backchannels

In this study, we analyzed three types of response backchannels, namely, *un* (yeah), its repetition *un-un*, and *hai* (yes), which most frequently occur in Japanese daily conversations. *Hai* provides a more polite impression to listeners than *un*. The repetition of *hai* (*hai-hai*) was excluded from the analyses because of its impoliteness and low frequency of use. The total number of observed *un*, *un-un*, and *hai* were 2906, 566, and 1612, respectively in all sessions.

The speech data were transcribed manually and split into long utterance units (LUUs) [26] by an expert labeler. An LUU is a unit of utterances and approximately corresponds to a single sentence in written texts. The labeled information includes whether each LUU is a backchannel or a filler and the type of dialogue acts (DA). We differentiate labeled backchannels from non-backchannel utterances, such as *un* to show agreement by using criteria that exclude responses to questions.

3.2. Analyzed Speech Window

We examined four different window lengths to analyze the user's preceding utterances: one, two, four, and eight seconds. Referring to a study on global synchrony using 16-sec sliding windows [27], we chose a shorter duration to capture local prosody entrainment.

3.3. Prosodic Feature Extraction

We extracted F_0 and power of each frame using SVTool [28]. The width and step of the sliding frames were 32 ms and 10 ms, respectively. The F_0 extraction algorithm was based on the peak search of the autocorrelation of linear predictive coding (LPC) residual error signals. Power was calculated from the root-mean-square (RMS) values in dB of each frame. The logarithm of F_0 values was calculated for each voiced frame.

We normalized the log F_0 and power by subtracting the within-session mean value to compensate individual variations of vocal pitch and the difference of the distances to the microphone. Thus, the analyzed log F_0 and power can be considered relative amounts throughout each speaker's session.

3.4. Statistics of Prosodic Features

We calculated seven types of statistics of log F_0 and power as representative values as shown in Table 1. We used median values in addition to mean values because the former is sensitive to outliers. Mean values can be affected by F_0 extraction error and power during silent intervals. When the relationship between observed prosodic features (F_0 and power) and time frame index is regressed by Equation (2), the slope of the first-order regression is defined as a by Equation (1).

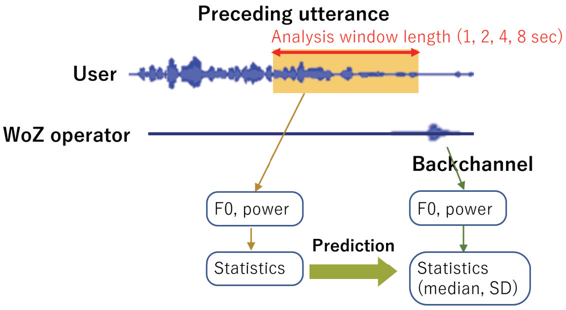

Figure 1: *Flow of feature extraction*

$$x_i = at_i + b + \epsilon_i, \quad (1)$$

$$a = \frac{\Sigma_{i=1}^n (x_i - \bar{x})(t_i - \bar{t})}{\Sigma_{i=1}^n (t_i - \bar{t})^2}, \quad (2)$$

$$b = \bar{x} - a\bar{t}, \quad (3)$$

where y_i, t_i, and ϵ_i represent a log F_0 or power value, time, and residual at i-th frame, respectively. $\bar{x}$ and $\bar{t}$ represent a mean prosodic features, (log F_0 or power) and mean time-frame index, respectively. In the speech segment of backchannels, the median and SD of log F_0 and power were also calculated to investigate the correlation with the preceding utterances. These statistics were chosen for the future application to automatic control of backchannels of a dialogue system. Figure 1 shows an outlook of the prosodic feature analyses.

3.5. Analysis Methods

We conducted a correlation analysis between each single prosodic feature of backchannels and that of the preceding utterance. Then, we used support vector regression (SVR) to predict the prosodic features of backchannels. Suppose that a set of observations and response variables x_n and y_n are given as training data, and a task to obtain the following linear model equation.

$$f(x') = \beta x + b \quad (4)$$

The task to find the $f(x)$ for which the value of the norm $(\beta'\beta)$ is minimized is formalized as a convex optimization problem that minimizes:

$$J(\beta) = \beta'\beta \quad (5)$$

This optimization problem is approximated by the following Lagrangian dual formalization, which introduces non-negative multipliers α_n, α_n^* for each observation x_n.

$$\begin{aligned} &\min_{\alpha,\alpha^*} \frac{1}{2}(\alpha-\alpha^*)^{\mathsf{T}}\mathbf{Q}(\alpha-\alpha^*) \\ &+\epsilon'\Sigma_{i=1}^l(\alpha+\alpha^*) + \Sigma_{i=1}^l y_i(\alpha-\alpha^*) \\ &\text{s.t.}: 0 \le \alpha_i, \alpha_i^* \le C, (i = 1, \ldots, l), \\ &\qquad \Sigma_{i=1}^l(\alpha_i - \alpha_i^*) = 0, \end{aligned} \quad (6)$$

where C represents the upper bound, $\mathbf{Q}$ represents a $l \times l$-dimensional semi-definite matrix. The parameter ϵ' was set to 0.1 in this study. The components of $\mathbf{Q}$ is defined as

$$Q_{i,j} \equiv K(x_i, x_j), \quad (7)$$

where $K(x_i, x_j) \equiv \phi(x_i)^{\mathsf{T}}\phi(x_j)$ represents a kernel. We used a radial basis function kernel defined in Equation (8), where $\mathbf{u}$

Table 2: *Spearman's rank sum correlation coefficient between prosodic features of backchannels and preceding interlocutor's utterances. The analysis window is set to eight seconds here. Only the coefficients of significant correlation are shown.*

Type	Feature	med_f	max_f	med_p	max_p
un	F_0 med.	-0.12	n.s.	0.26	0.32
	F_0 SD	-0.08	n.s.	0.14	0.16
	Power med.	n.s.	0.14	0.24	0.31
	Power SD	n.s.	0.11	0.19	0.25
unun	F_0 med.	-0.30	n.s.	0.27	0.35
	F_0 SD	-0.22	n.s.	0.24	0.30
	Power med.	-0.25	0.13	0.25	0.33
	Power SD	-0.13	0.12	n.s.	0.16
hai	F_0 med.	n.s.	0.14	0.11	0.11
	F_0 SD	n.s.	n.s.	0.09	0.09
	Power med.	-0.08	0.16	0.64	0.64
	Power SD	-0.13	0.09	0.56	0.55

and $\mathbf{v}$ are given as a sampled vectors in the input space:

$$K(\mathbf{u}, \mathbf{v}) = \exp\left(-\gamma \left|\mathbf{u} - \mathbf{v}\right|\right). \tag{8}$$

The parameter γ was set to 1.0 in this study.

The prediction accuracy was evaluated in terms of mean absolute error (MAE) between the predicted and observed values of the prosodic features. Starting from one feature, we gradually increased the number of features and compared the MAEs of all possible combinations of prosodic feature statistics of preceding utterances to find the best set of features. The MAE was calculated by averaging the results of five-fold cross-validation. To prevent overfitting, we stopped under the criteria that the improvement of the predicted residual error sum of squares (PRESS) was below a threshold.

4. Results

4.1. Correlation analysis

Table 2 shows the correlation coefficients between the prosodic features of backchannels and those of the user's preceding utterance. The correlation tests were conducted ($p < 0.05$) with p-values adjusted by the Bonferroni method controlling the false discovery rate (FDR). It is observed that many prosodic features of backchannels had a larger correlation with power-related features (med_p, max_p) of the preceding utterances than with F_0-related features (med_f, max_f).

4.2. Support Vector Regression

Tables 3–5 show the selected features and their estimation accuracy in terms of the correlation coefficient between the observed and predicted values. The MAE between the observed and predicted values are also presented. Note that the MAEs were calculated based on the normalized value with within-session mean and SDs. Table 3 gives a comparison of analysis window lengths of 1, 2, 4, and 8 seconds for "*un*", and it is shown that 8 seconds yielded the best accuracy.

In Table 4, we also compared three settings for feature selection to evaluate the contribution of F_0 and power for prediction for "*un*": selecting from all features (Setting 1), selecting from the $F0$-related features (Setting 2), and selecting from the power-related features (Setting 3). We used the analysis window length, the eight-second because it performed best among the four settings. We observed that both F_0 and power were selected from candidates in Setting 1. As a result, prediction of Setting 1 obtained better performance than Setting 2 and 3. Moreover, Setting 3 showed higher accuracies and lower MAEs than Setting 2.

Table 5 shows the results for all types of backchannels. It is confirmed that both F_0 and power features are used to predict most of the features of the backchannels. The prediction accuracy is improved from the case using the same type of features only.

5. Discussions

The correlation between the median of log F_0 of backchannels and the median of log F_0 of the preceding utterance was low. This result is not consistent with the similarity of the mean of F_0 between backchannels and its preceding utterances shown by Heldner et al. conducted on the analyses of English speech corpus [23]; however, a similar result was observed in a previous study conducted on Japanese datasets [29].

The difference in the accentual systems may cause inconsistent results between English and Japanese. In English, F_0 correlates with power [30] because of the influence of the stress-accent system. On the other hand, because Japanese has a pitch-accent system, linguistic information can affect the mean/median of log F_0. Comparing the normalized mean within the whole session, we found that the power of backchannels is distributed in a significantly narrow range than that of log F_0 ($p < 0.05$).

The F_0-related features of backchannels were weakly correlated with the SD of log F_0 of the preceding utterances. However, a higher correlation was observed between the F_0-related features of backchannels and the power-related features of the prior utterance. The power-related feature improved the prediction performance of prediction of F_0-related features of backchannels from the case using only F_0-related features.

On the other hand, the power-related features of backchannels were correlated more highly with the power-related features of the preceding utterances than those of F_0-related features. The results of SVR indicate that the use of F_0-related features does not highly contribute to the performance with regard to *un* and *un-un*. This may be because the listener produces backchannels according to the power of the speaker's voice. In addition, the listener may align with hot spots or the climax of the story behind the change of power. We also note that the maximum value of the power of the preceding utterances particularly has an important influence because it contributes to the performance of the SVR.

Using these results, we have implemented the control module of backchannel prosody for our spoken dialogue system to encourage the user's talk in real settings such as installation in a public space, elderly care facilities, and rehabilitation facilities.

6. Conclusions

In this study, we here investigated the relationship between the prosodic features of backchannels and their preceding utterances. The interrelationship between different prosodic features, F_0 and power, was observed beyond the local synchrony of the same prosodic feature. We found that the power-related features of the interlocutors' preceding utterance remarkably affect the control of both F_0 and power of the backchannels. Future work will include the evaluation of the effectiveness of

Table 3: *Pearson's correlation coefficient between the predicted and observed prosodic features of backchannels. The analysis window length was set to 1, 2, 4, and 8 sec.*

		Analysis window length							
		1 sec		2 sec		4 sec		8 sec	
	Predictee	MAE	r	MAE	r	MAE	r	MAE	r
un	F_0 med.	0.70	0.38	0.66	0.48	0.63	0.53	0.61	**0.56**
	F_0 SD	0.20	0.16	0.20	0.18	0.20	0.22	0.20	**0.24**
	Power med.	0.29	0.35	0.28	0.44	0.27	0.51	0.23	**0.55**
	Power SD	0.12	0.23	0.11	0.38	0.11	0.40	0.11	**0.44**

Table 4: *Pearson's correlation coefficient between the predicted and observed prosodic features of backchannels. The feature selection was conducted under the three settings: (1) Setting 1: selecting from all of 14 features, (2) Setting 2: selecting from seven F_0-related features, (3) Setting 3: selecting from seven power-related features. The analysis window length was set to 8 sec.*

		Setting 1			Setting 2			Setting 3		
	Predictee	Selected features	MAE	r	Selected features	MAE	r	Selected features	MAE	r
un	F_0 med.	mean$_f$ SD$_f$ med$_p$	0.61	**0.56**	med$_f$ SD$_f$ max$_f$	0.7	0.39	med$_p$	0.66	0.48
	F_0 SD	mean$_f$ med$_p$	0.20	**0.24**	mean$_f$ max$_f$	0.2	0.17	med$_p$ SD$_p$	0.2	0.21
	Power med.	mean$_f$ SD$_f$ med$_p$ max$_p$	0.23	**0.55**	mean$_f$ SD$_f$	0.3	0.33	med$_p$ SD$_p$ max$_p$	0.28	0.47
	Power SD	mean$_f$ SD$_f$ mean$_p$	0.11	**0.44**	mean$_f$ max$_f$	0.11	0.28	med$_p$	0.11	0.34

Table 5: *Pearson's correlation coefficient between the predicted and observed prosodic features of backchannels. The feature selection was conducted with selection from all features (same as Setting 1 in Table 4) and selection from the same type of features (Setting 2 for prediction of F_0-related features, and Setting 3 for prediction of power-related features). The analysis window length was set to 8 sec.*

		Selected from all features (Setting 1)			Selected from the same type of features (Setting 2 or 3		
	Predictee	Selected features	MAE	r	Selected features	MAE	r
un	F_0 med.	mean$_f$ SD$_f$ med$_p$	0.61	0.56	med$_f$ SD$_f$ max$_f$	0.7	0.39
	F_0 SD	mean$_f$ med$_p$	0.2	0.24	mean$_f$ max$_f$	0.2	0.17
	Power med.	mean$_f$ SD$_f$ med$_p$ max$_p$	0.23	0.55	med$_p$ SD$_p$ max$_p$	0.28	0.47
	Power SD	mean$_f$ SD$_f$ mean$_p$	0.11	0.44	med$_p$	0.11	0.34
unun	F_0 med.	med$_f$ max$_f$ med$_p$	0.59	0.67	med$_f$ IQR$_f$ max$_f$	0.74	0.49
	F_0 SD	mean$_f$ med$_p$	0.15	0.48	min$_f$	0.16	0.31
	Power med.	mean$_f$ med$_f$ med$_p$	0.18	0.55	med$_p$	0.21	0.43
	Power SD	min$_f$ med$_p$	0.08	0.30	med$_p$	0.08	0.27
hai	F_0 med.	mean$_f$ mean$_p$ SD$_p$ min$_p$	0.80	0.31	max$_f$	0.91	0.16
	F_0 SD	IQR$_p$	0.24	0.08	min$_f$	0.24	0.03
	Power med.	mean$_f$ max$_p$ min$_p$	0.24	0.76	min$_p$	0.27	0.77
	Power SD	mean$_p$	0.11	0.61	mean$_p$	0.11	0.61

the proposed backchannel prosody control method from the perspective of how actively elicit the user's talk.

7. Acknowledgement

This study was supported by JSPS KAKENHI (19H05691) and JST Moonshot R&D Grant Number JPMJMS2011.

8. References

[1] C. Dideriksen, R. Fusaroli, K. Tylén, M. Dingemanse, and M. H. Christiansen, "Contextualizing conversational strategies: backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations," *Proceedings of CogSci'19* pp. 261–267, 2019.

[2] H. W. Park, M. Gelsomini, J. J. Lee, and C. Breazeal, " Telling stories to robots: The effect of backchanneling on a child's storytelling," *Proceedings of 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 100–108, 2017.

[3] J. P. Wolf, "The effects of backchannels on fluency in L2 oral task production," *System*, vol. 36, no. 2, pp. 279-294, 2008.

[4] J. A. Maddrell and G. S. Watson, "The influence of backchannel communication on cognitive load," *The Next Generation of Distance Education*, pp. 171–180, Springer, Boston, MA.

[5] M. Sannomiya, I. Yamakawa, A. Kawaguchi, and Y. Morita, "Effect of backchannel utterances on facilitating idea-generation in Japanese think-aloud tasks," *Psychological reports* vol 93, no. 1, pp. 41-46, 2003.

[6] K. Kobayashi, K. Funakoshi, T. Komatsu, S. Yamada, and M. Nakano, "Improving user experiences in talking to robots using

ASE-based backchannel feedbacks," Transactions of the Japanese Society for Artificial Intelligence. vol. 30, no.4, pp. 604–612.

[7] R. H. Gálvez, A. Gravano, S. Beňuš, R. Levitan, M. Trnka, and J. Hirschberg, " An empirical study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars," *Speech Communication*, vol. 124, pp. 46–67, 2020.

[8] H. Sugiyama, T. Meguro, Y. Yoshikawa, and J. Yamato, "Improving dialogue continuity using inter-robot interaction," *Proceedings of 27th IEEE International Symposium on Robot and Human Interactive Communication* pp. 105-112, 2018.

[9] S. Fujie, K. Fukushima, and T. Kobayashi, "A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information,". *Proceedings of ICARA International Conference on Autonomous Robots and Agents*, pp. 379–384, 2004.

[10] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction" *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* pp. 375–382, 2010.

[11] N. Hussain, E. Erzin, T. M. Sezgin, and Y. Yemez, " Speech driven backchannel generation using deep q-network for enhancing engagement in human-robot interaction," arXiv preprint arXiv:1908.01618, 2019.

[12] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, " Prediction of turn-taking using multitask learning with prediction of backchannels and fillers," *Proceedings of INTERSPEECH*, pp. 991—995, 2018.

[13] N. Ward, and W. Tsukahara"Prosodic features which cue backchannel responses in English and Japanese," *Journal of pragmatics*, vol. 32, no. 8, pp. 1177–1207.

[14] H. W. Park, M. Gelsomini, J. J. Lee, T. Zhu, , and C. Breazeal, "Backchannel opportunity prediction for social robot listeners," *Proc IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2308–2314, 2017.

[15] A. I. Adiba, T. Homma, T.and T. Miyoshi, " Towards Immediate Backchannel Generation Using Attention-Based Early Prediction Model," *Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7408–7412, 2021.

[16] T. Kawahara, T. Yamaguchi, K. Inoue, K. Takanashi, and N. Ward. "Prediction and generation of backchannel form for attentive listening systems," *Proceedings of INTERSPEECH*. pp. 2890–2894, 2016.

[17] R. Ishii, X. Ren, M. Muszynski, and L. P. Morency,"Multimodal and Multitask Approach to Listener's Backchannel Prediction: Can Prediction of Turn-changing and Turn-management Willingness Improve Backchannel Modeling?," *Proceedings of of the 21st ACM International Conference on Intelligent Virtual Agents* pp. 131–138.

[18] R. Ruede, M. Müller, S. Stüker, and A. Waibel "Yeah, right, uh-huh: a deep learning backchannel predictor," *Advanced Social Interaction with Agents* pp. 247–258, Springer, Cham. 2019.

[19] M. M. Willi, S. A. Borrie, T. S. Barrett, M. Tu, and V. Berisha, "A discriminative acoustic-prosodic approach for measuring local entrainment," arXiv preprint arXiv:1804.08663, 2018.

[20] J. Michalsky, H. Schoormann, and O. Niebuhr, "Conversational quality is affected by and reflected in prosodic entrainment," *Proceedings of Speech Prosody 9*, 2019.

[21] S. Beňuš, M. Trnka, E. Kuric, L. Marták, A. Gravano, J. Hirschberg, and R. Levitan, "Prosodic entrainment and trust in human-computer interaction," *Proceedings of 9th International Conference on Speech Prosody*, pp. 220–224, 2018.

[22] U. D. Reichel, K. Mády, and J. Cole,"Prosodic entrainment in dialog acts," arXiv preprint arXiv:1810.12646, 2018.

[23] M. Heldner, Je. Edlund, and J. B. Hirschberg, "Pitch similarity in the vicinity of backchannels," *Poce. Interspeech 2010*. 2010.

[24] R. Levitan, S. Beňuš, A. Gravano, and J. Hirschberg. "Entrainment and turntaking in human-human dialogue," *Proceedings of AAAI 2015 Spring Symposium on Turn-taking and Coordination in HumanMachine Interaction*, 2015.

[25] K. Inoue, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara, "An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions," *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue,* pp. 118–127, 2020.

[26] Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, and N. Yoshida,"Two-level annotation of utterance-units in japanese dialogs: An empirically emerged scheme," *Proceedings of LREC*, 2010.

[27] R. Levitan, and J. B. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," *Proceedings of Interspeech 2011*, pp. 3081–3084.

[28] C.T. Ishi, H. Ishiguro, N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality," *Speech Communication* vol. 50 no. 6, pp. 531–543, 2008.

[29] T. Kawahara, M. Uesato, K. Yoshino, and K. Takanashi. "Toward adaptive generation of backchannels for attentive listening agents," *Proceedings of International Workshop Spoken Dialogue Systems (IWSDS)*, 2015.

[30] A. Rosenberg and J. Hirschberg, "On the correlation between energy and pitch accent in read English speech," *Proceedings of Interspeech 2006*, pp. 1294–1298, 2006.