Paralinguistic Attitude Recognition for Spoken Dialogue Systems

Kouki Miyazawa and Zhi Zhu and Yoshinao Sato
Fairy Devices Inc.
{miyazawa,zhu,sato}@fairydevices.jp

Abstract

Although paralinguistic information is critical for human communication, most spoken dialogue systems ignore such information, hindering natural communication between humans and machines. This study addresses the recognition of paralinguistic attitudes in user speech. Specifically, we focus on four essential attitudes for generating an appropriate system response, namely agreement, disagreement, questions, and stalling. The proposed model can help a dialogue system better understand what the user is trying to convey. In our experiments, we trained and evaluated a model that classified paralinguistic attitudes on a reading-speech dataset without using linguistic information. The proposed model outperformed human perception. Furthermore, experimental results indicate that speech enhancement alleviates the degradation of model performance caused by background noise, whereas reverberation remains a challenge.

1 Introduction

In human dialogue, people communicate various messages through paralinguistic features of speech, such as prosody and voice quality. Speech can convey emotions and attitudes through paralinguistic features regardless of linguistic information. Humans can recognize four intentions, namely affirm, deny, ask for repetition, and filler, with high accuracy using only paralinguistic features (Ishi et al., 2008). Moreover, humans can convey six intentions, namely criticism, doubt, naming, suggestion, warning, and wish, through prosodic patterns irrespective of lexical meaning (Hellbernd and Sammler, 2016). The paralinguistic information transmitted in this manner can affect listener behavior.

By contrast, paralinguistic information is ignored by most spoken dialogue systems, which adopt a cascaded pipeline of automatic speech recognition (ASR) and a linguistic dialogue model. This restriction requires users to convey their messages using only linguistic information; otherwise, miscommunication can occur. The limited paralinguistic ability in spoken dialogue systems impedes natural communication with humans.

In this study, we address the challenge of enabling a spoken dialogue system to recognize attitudes expressed through paralinguistic features in user speech. Specifically, we focus on four attitude classes, namely agreement, disagreement, questions, and stalling. Table 1 lists these definitions. In the case of no confusion, the agreement, disagreement, question, and stalling classes are abbreviated as A, D, Q, and S, respectively. Among other paralinguistic information, the four attitudes are critical in determining the reaction of a system. These attitudes are typically accompanied by the four main types of boundary pitch movement at the end of prosodic phrases (Igarashi and Koiso, 2012). Using prosody is an effective way to control voice interactive devices (Zhang et al., 2022). We believe that spoken dialogue systems should also be able to recognize paralinguistic attitudes to communicate naturally with humans. Note that this study does not aim to comprehensively theorize the paralinguistic aspects of dialogue acts. The proposed model focuses on resolving the ambiguity that arises when spoken dialogue systems try to understand user speech by relying solely on lexical information and ignoring paralanguage.

Only one of the four attitudes is deemed to accompany a single utterance. This is understood by the fact that boundary pitch movement at the end of an utterance substantially affects the attitude.

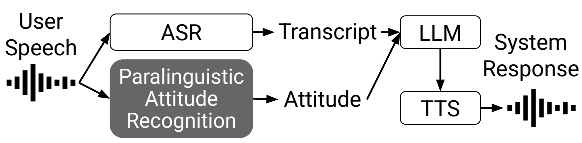

Figure 1: Example usage of the proposed model

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 137–142
May 27–30, 2025.

Table 1: Paralinguistic attitude classes

	Attitude	Expected reaction
Agreement	in favor, accept to continue	performing the approved action, moving on to the next
Disagreement	against, dissatisfied, request to stop	canceling the rejected action, asking for instructions
Question	not understand, confirm facts, listen back	answering the question, rephrasing the previous utterance
Stalling	thinking, worried, request to wait	waiting for instructions, providing additional information

Therefore, articulating multiple attitudes in a single utterance is challenging for most users of spoken dialogue systems. In other words, the paralinguistic attitudes investigated in this study are mutually exclusive and evoked in units of utterances.

We introduce one of the expected uses of our model, as illustrated in Fig. 1. An input user utterance is processed in parallel using an ASR model and a paralinguistic attitude recognition model. The transcription and the inferred attitude are subsequently passed on to a large language model (LLM). Finally, a text-to-speech (TTS) model synthesizes a system utterance according to the output generated by the LLM. Previous studies have explored methods to process paralinguistic cues in conjunction with transcripts by employing LLMs (Lin et al., 2024; Xue et al., 2024; Kang et al., 2024). A simple approach is to concatenate the transcript and class label in a prompt, for example: "*transcript* *<attitude>*."

2 Model

The network structure of the proposed model is listed in Table 2. The input feature of the proposed model is a waveform. The main part is a self-supervised learning (SSL) model called HuBERT-large (Hsu et al., 2021). The layer depth at which an embedding vector is obtained from the SSL model is optimized on the validation data, following (Zhu and Sato, 2023). The embedding vector yielded from the SSL model is averaged over time and passed to head layers that comprise two fully connected layers and a softmax layer. The output is the posterior probability of the attitude classes.

It is known that speech SSL models embed prosodic information in their hidden representations (Lin et al., 2023; de la Fuente and Jurafsky, 2024). Moreover, the explicit incorporation of pitch into the input features in our preliminary experiments did not enhance the model performance. Hence, we chose to use only the hidden representation of the HuBERT model.

Table 2: Model structure

Layer	Output size
HuBERT	$1024 \times T$
Mean pooling	1024
Fully connected	1024
Fully connected	1024
Softmax	4

T denotes the number of time frames.

We note that linguistic information was not used as an input feature. One reason for this choice is that a cascaded pipeline of ASR and the attitude recognition model cause considerable latency in generating a system response. To use linguistic information, the paralinguistic attitude recognition model should wait until the ASR model yields a transcription, inevitably causing additional latency. Thus, we made the model recognize attitude using only acoustic features to avoid hindering smooth communication. Another reason is that a spoken phrase can be accompanied by distinct intentions depending on its paralinguistic features regardless of linguistic information (Ishi et al., 2008; Tang et al., 2016; Hellbernd and Sammler, 2016). Therefore, linguistic features were not significant in recognizing the four paralinguistic attitudes. Another motivation was to avoid the domain dependence of linguistic features. Linguistic choices are affected by situations where dialogue occurs and the relationship between participants. Previous studies on paralinguistic information employing linguistic features focused on a specific domain, such as meetings (Ortega and Vu, 2018; Maltby et al., 2023) and news delivery (Takatsu et al., 2019). We used only acoustic features so that the model is useful in various domains.

3 Data

In this study, speeches read in Japanese by crowd workers and actors were used. Table 3 and Fig. 2 show the number of utterances and distribution of duration, respectively.

3.1 Crowd workers' speech

We used a Japanese reading speech dataset collected by (Sato and Miyazawa, 2023). In this section, we briefly review the dataset. It contains five sets of 63 scripts, including words, phrases, sentences, fillers, and back channels. A spoken sentence can be accompanied by a paralinguistic attitude regardless of its semantic content. Therefore, the same speaker read each script aloud with four attitudes in this dataset. In the recording process, 138 crowd workers read one script set of size 63 aloud with four paralinguistic attitudes. Another 20 crowd workers evaluated the utterances in which each speech was heard by two or three listeners. By using a statistical quality estimation method, 19,821 high-quality utterances were selected. This method estimates the quality of utterances from the speaker's intention and listeners' evaluations, while considering their reliability.

3.2 Actors' speech

In this study, we collected additional recordings using the same procedure. Six actors read a script set of size 63 aloud with four paralinguistic attitudes. Because we added a small number of recordings, the number of utterances per attitude was greater than 378. After recording, 31 crowd workers evaluated 384 randomly sampled utterances, each of which was heard by five listeners. We assumed that the attitudes intended by the actors were correct and used all the utterances without filtering.

The results are summarized in Table 4. The macro-F_1 score of the human perception of the actors' speech was 0.829.

4 Experiments

We trained and evaluated a paralinguistic attitude recognition model using a speech dataset of crowd workers and actors. All speech data were processed at a sampling rate of 16 kHz on a single channel.

Table 3: Number of utterances in the dataset

	Crowd workers	Actors
Agreement	8,581	470
Disagreement	976	378
Question	6,048	379
Stalling	4,216	379
Total	19,821	1,606

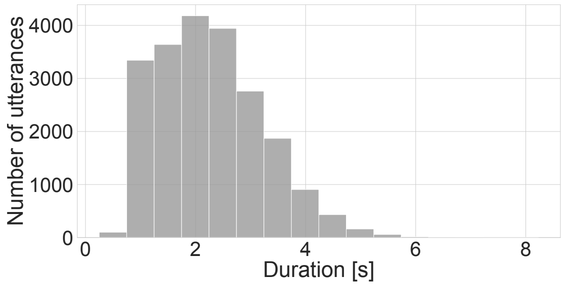

Figure 2: Duration distribution

The HuBERT model was frozen, and the head layers were fine-tuned during training. We measured the performance in terms of the macro-F_1 score using six-fold cross-validation. For each fold, the entire dataset was split into six sets, namely four for training, one for validation, and one for testing.

We augmented the training data four-fold by adding background noise and reverberation to improve model robustness. Noise signals were randomly selected from the DEMAND (Thiemann et al., 2013), MUSAN (Snyder et al., 2015), and FSD50K (Fonseca et al., 2022) datasets. The signal-to-noise ratio was randomly chosen from the uniform distribution from -10 to 10 dB. Room impulse responses were randomly sampled from the BIRD database (Grondin et al., 2020). Whether noise or reverberation was added during the test depended on the evaluation settings, as explained below.

We set the layer depth at which the HuBERT-

Table 4: Human perception of the actors' speech

		Perceived A	Perceived D	Perceived Q	Perceived S
Intended	A	451	10	5	14
Intended	D	2	339	133	6
Intended	Q	12	52	385	31
Intended	S	23	24	18	415

$F_1 = 0.829$

Table 5: Evaluation of the proposed model on the actors' speech

		Predicted			
		A	D	Q	S
Actual	A	453	4	1	12
	D	5	330	43	0
	Q	9	43	326	1
	S	20	0	4	355

$F_1 = 0.909$

Table 6: Evaluation of the proposed model on all the speech

		Predicted			
		A	D	Q	S
Actual	A	8744	13	106	188
	D	19	1049	286	0
	Q	203	192	6009	23
	S	234	0	13	4348

$F_1 = 0.912$

Table 7: Evaluation of the proposed model on all the speech in the noisy and reverberant conditions

Condition	Enhanced speech	F_1
Clean		0.912
Noisy		0.625
Noisy	✓	0.844
Noisy and reverberant		0.449
Noisy and reverberant	✓	0.492

large model yielded an embedding vector to 12 based on the validation data. This result is consistent with the findings of previous studies in which paralinguistic information was incorporated into the middle layers of the SSL models (Pepino et al., 2021; Li et al., 2022; Zhu and Sato, 2023).

4.1 Comparison with humans

We compared the performance of human listeners with our model on the actors' speech. In this experiment, we did not add noise or reverberation to the test data. The speech utterances of the crowd workers were not necessarily suitable for evaluating human perception because those on which the listeners disagreed were excluded during the filtering process. Therefore, we used the actors' speech to compare human perception with the proposed model. Table 5 presents the results of the model evaluation. The macro-F_1 score of the model measured using the actors' speech was 0.909.

We found that the proposed model outperformed human perception, as depicted in Tables 4 and 5. Moreover, the human confusion between the disagreement and question attitudes was reduced in the model prediction.

4.2 Evaluation of the model

Moreover, we evaluated the performance of our model on all the data (i.e., all the speech by the crowd workers and the actors). No noise or reverberation was introduced to the test data. For the actors' speech, we assumed the intended attitudes to be the ground truth. For the crowd workers' speech, we regarded the attitudes determined by the quality estimation method as the ground truth. Table 6 presents the results. The macro-F_1 score of the model evaluated using all data was 0.912.

No significant difference was observed between the model performance on the speech of the actors (F_1=0.909) and that of all the speakers (F_1=0.912). Therefore, the quality estimation method effectively selected quality speech.

4.3 Robustness to noise and reverberation

Real-world applications of spoken dialog systems are inevitably affected by noise and reverberation. Therefore, we evaluated model performance in noisy and reverberant environments using all the data. Specifically, we examined three conditions: (1) clean, (2) noisy, and (3) noisy and reverberant. The clean condition was identical to the one described in Section 4.2. Noise and reverberation were added in the same manner as the training data. Furthermore, we investigated the effects of speech enhancement. A state-of-the-art speech enhancement model, MP-SENet (Lu et al., 2023), is applied to the disturbed test data. The MP-SENet model simultaneously performs speech denoising and dereverberation. Table 7 presents the results.

In our experiment, noise and reverberation degraded the model performance even though data augmentation was used during training. The use of speech enhancement considerably improved model performance in the noisy condition. By contrast, the degradation due to reverberation was marginally mitigated. The results indicate that the influence of reverberation remains a challenge in paralinguistic attitude recognition. This can be explained by the fact that prosody, which is difficult to estimate in reverberant environments, is a key factor in communication through paralinguistic information.

5 Conclusion

This study addressed paralinguistic attitude recognition in user speech for spoken dialogue systems. Specifically, we focused on four essential attitudes for determining a system reaction, namely agreement, disagreement, questions, and stalling. We trained and evaluated the model using a reading-speech dataset of actors and crowd workers. The proposed model outperformed human perception when evaluating the actors' speech under a clean condition. Furthermore, the proposed model achieved almost the same performance on the crowd workers' speech after filtering by quality. Noise and reverberation degraded the model performance. Speech enhancement can alleviate the degradation caused by noise. However, the influence of reverberation remains a challenge. The use of paralinguistic attitude recognition enables spoken dialogue systems to understand what users convey through speech.

Finally, we discuss future research directions. We used a reading-speech dataset in this study. However, the manner in which attitudes are expressed through paralinguistic features varies depending on the situation in which speech utterances occur. Human speech directed to spoken dialogue systems is more diverse than reading speech but more controlled than casual everyday conversation. Therefore, we should investigate paralinguistic attitude recognition for speech directed to dialogue systems. Another direction is to clarify how to determine a system reaction, given an inferred user's paralinguistic attitude. Moreover, joint models of speech enhancement and paralinguistic attitude recognition should be examined to alleviate the degradation caused by reverberation.

References

Anton de la Fuente and Dan Jurafsky. 2024. A layer-wise analysis of mandarin and english suprasegmentals in ssl speech models. In *Interspeech*, pages 1290–1294.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 30:829–852.

François Grondin, Jean-Samuel Lauzon, Simon Michaud, Mirco Ravanelli, and François Michaud. 2020. BIRD: Big Impulse Response Dataset. ArXiv:2010.09930.

Nele Hellbernd and Daniela Sammler. 2016. Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, 88:70–86.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 29:3451–3460.

Yosuke Igarashi and Hanae Koiso. 2012. Pitch range control of japanese boundary pitch movements. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality. *Speech Communication*, 50(6):531–543.

Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, et al. 2024. Frozen large language models can perceive paralinguistic aspects of speech. ArXiv:2410.01162.

Yuanchao Li, Yumnah Mohamied, Peter Bell, and Catherine Lai. 2022. Exploration of a self-supervised speech model: A study on emotional corpora. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 868–875.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G. Ward. 2023. On the utility of self-supervised models for prosody-related tasks. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-Yi Lee, and Ivan Bulyko. 2024. Paralinguistics-enhanced large language modeling of spoken dialogue. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10316–10320.

Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. 2023. MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3834–3838.

Harry Maltby, Julie Wall, T Goodluck Constance, Mansour Moniri, Cornelius Glackin, Marvin Rajwadi, and Nigel Cannings. 2023. Short utterance dialogue act classification using a transformer ensemble. *UA Digital Theme Research Twinning (UA-DIGITAL)*.

Daniel Ortega and Ngoc Thang Vu. 2018. Lexico-acoustic neural-based models for dialog act classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6194–6198.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3400–3404.

Yoshinao Sato and Kouki Miyazawa. 2023. Statistical quality estimation for partially subjective classification tasks through crowdsourcing. In *Language Resources and Evaluation*, volume 57, pages 31–56. Springer.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A music, speech, and noise corpus. ArXiv:1510.08484.

Hiroaki Takatsu, Katsuya Yokoyama, Yoichi Matsuyama, Hiroshi Honda, Shinya Fujie, and Tetsunori Kobayashi. 2019. Recognition of intentions of users' short responses for conversational news delivery system. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1193–1197.

Yaodong Tang, Yuchen Huang, Zhiyong Wu, Helen Meng, Mingxing Xu, and Lianhong Cai. 2016. Question detection from acoustic features using recurrent neural network with gated recurrent unit. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6125–6129.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. 2013. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics*, volume 19. AIP Publishing.

Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Mengzhe Chen, Qian Chen, and Lei Xie. 2024. E-chat: Emotion-sensitive spoken dialogue system with large language models. In *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 586–590.

Xinlei Zhang, Zixiong Su, and Jun Rekimoto. 2022. Aware: Intuitive device activation using prosody for natural voice interactions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Zhi Zhu and Yoshinao Sato. 2023. Deep investigation of intermediate representations in self-supervised learning models for speech emotion recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.