ORIGINAL RESEARCH
published: 16 March 2022
doi: 10.3389/fcomp.2022.815128

Modeling Japanese Praising Behavior by Analyzing Audio and Visual Behaviors

Toshiki Onishi [1†], *Arisa Yamauchi* [1†], *Asahi Ogushi* [2], *Ryo Ishii* [3], *Atsushi Fukayama* [3], *Takao Nakamura* [3] *and Akihiro Miyata* [2*]

[1] Graduate School of Integrated Basic Sciences, Nihon University, Tokyo, Japan, [2] College of Humanities and Sciences, Nihon University, Tokyo, Japan, [3] NTT Human Informatics Laboratories, NTT Corporation, Kanagawa, Japan

Praising behavior is considered to be verbal and nonverbal behaviors that expresses praise the behavior and character of the target. However, how one should use verbal and nonverbal behaviors to successfully praise a target has not been clarified. Therefore, we focus on attempts to analyze praising behavior in Japanese dialogue using verbal and nonverbal behaviors. In this study, we attempted to analyze the relationship between praising skills and human behaviors in Japanese dialogue by focusing on voice, head, and face behaviors. First, we created a new dialogue corpus in Japanese containing voice, head, and face behaviors from individuals giving praise (praiser) and receiving praise (receiver), as well as the degree of success of praising (praising score). Second, we developed machine learning models that uses features related to voice, head, and face behaviors to estimate praising skills to clarify which features of the praiser and receiver are important for estimating praising skills. Evaluation resulte showed that some audio features of the praiser are particularly important for estimation of praising skills. Our analysis results demonstrated the importance of features related to the zero-crossing rate, MFCCs of the praiser. Analyzing the features of high importance revealed that the praiser should praise with specific words that mean amazing or great in Japanese and the voice quality of the praiser is considered to be important for praising successfully.

Keywords: multimodal interaction, communication, praise, visual, audio

Edited by:
Atiqur Rahman Ahad,
University of Dhaka, Bangladesh

Reviewed by:
Ping Yang,
Western Sydney University, Australia
Swapnil Sayan Saha,
University of California, Los Angeles,
United States

Correspondence:
Akihiro Miyata
miyata.akihiro@acm.org

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

Received: *15 November 2021*
Accepted: *18 February 2022*
Published: *16 March 2022*

Citation:
Onishi T, Yamauchi A, Ogushi A, Ishii R, Fukayama A, Nakamura T and Miyata A (2022) Modeling Japanese Praising Behavior by Analyzing Audio and Visual Behaviors. Front. Comput. Sci. 4:815128. doi: 10.3389/fcomp.2022.815128

1. INTRODUCTION

Praising behavior is considered to be verbal and nonverbal behaviors that expresses praise directed at the behavior and character of the target (Brophy, 1981; Kalis et al., 2007; Jenkins et al., 2015). Praising behavior in Japanese culture is considered to be basically the same as in other cultures. Some studies in psychology and education have indicated that Japanese people have a tendency to praise others frequently and try to avoid being praised, due to the importance of humility. Also, praising behavior is not a simple one-way transmission from the praiser to receiver, but is a complex social communication in which the role of the receiver is just as critical as the role of the praiser (Henderlong and Lepper, 2002). To the best of our knowledge, how one should use verbal and nonverbal behavior to successfully praise a target has not been clarified. From the perspective of daily life, a person who has difficulty praising a target should learn how to improve their praising skills. From an engineering perspective, a system for judging an individual's skill in praising is difficult to create and an interactive agent for praising is difficult to develop. Based on the research background of this study, we attempt to clarify how verbal and nonverbal behavior can be used to praise successfully.

In this article, we attempt to demonstrate what behaviors are important to praising others successfully during dialogue. Our previous research reported that the nonverbal behaviors related head and face are useful for estimating praising scores (Onishi et al., 2020). Based on this research, we attempt to estimate praising scores by analyzing voice behaviors as well as head and face behaviors. We analyze the relationship between praising skills and human behaviors in dialogue by focusing on voice, head, and face behaviors to reveal what behaviors are important to praising others successfully during dialogue. We created a new dialogue corpus that includes voice, face, and head behavior information for praisers and receivers, as well as praising skills (Onishi et al., 2020). We develop a machine learning model that uses features related to voice, head, and face behaviors to estimate praising skills, allowing us to clarify which features of a praiser and receiver are important for estimating praising skills (Onishi et al., 2020). As in existing cases (Kurihara et al., 2007; Jokinen et al., 2013; Batrinca et al., 2016), our work has been carried out step by step; in this article, we introduce audio features to analyze praising behavior.

This study is a continuation of the past study (Onishi et al., 2020), however, there are major changes as follows. First, the definition of utterance scene was changed. This caused a difference in the number of scenes and the length of the scenes between this study and the previous study. Second, the definition of praising utterance scene was changed. While one annotator determined whether each utterance scene was a praising utterance scene in the previous study, five annotators determined whether each utterance scene was a praising utterance scene in this article. Third, ground truth labels were changed. While one participant (i.e., receiver) subjectively evaluated the praising skills of the dialogue partner (i.e., praiser) in the previous study, five annotators evaluated the praising skills of the praiser in this article. Lastly, features used for estimation were changed. In this manuscript, the audio modality was newly introduced.

The main contribution of this article is the clarification of which features of a praiser and receiver are important for estimating praising skills based on voice, head, and face behaviors. Specifically, we show the important features related to the zero-crossing rate, mel-frequency cepstral coefficients of the praiser.

2. RELATED WORK

Our study is to estimate praising skills using nonverbal behaviors. There have been many studies that have used verbal and nonverbal behaviors to estimate abilities and performance. Our study belongs to this category.

2.1. Estimation of Personality Traits

Many studies have been conducted to estimate the personality traits. Aran and Gatica-Perez (2013) investigated the prediction of personality traits of individuals participating in small group discussions. They used audio-visual nonverbal features in the experiment and showed that the extraversion trait can be predicted with high accuracy in a binary classification task. Batrinca et al. (2016) investigated automatic recognition of the "big five" personality traits (extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience) based on audio and video data collected in two scenarios: human-machine interaction and human–human interaction. Their findings showed that the relevance of each of the two scenarios when it comes to the degree of emergence of certain traits and the feasibility to automatically recognize personality under different conditions. Biel et al. (2012) investigated facial expression to predict personality traits in vlogs. They showed that their results were promising, specially for the case of the Extraversion impression. Lin and Lee (2018) proposed a framework that models the vocal behaviors of both a target speaker and their contextual interlocutors to improve the prediction performance of scores for 10 different personality traits in the ELEA corpus. Additionally, they observed several distinct intra- and inter-personal vocal behavior patterns that vary as a function of personality traits by analyzing the interpersonal vocal behaviors in the region of high attention weights. Pianesi et al. (2008) investigated the automatic detection of personality traits in a meeting environment using audio and visual features. Their result largely supported the idea that social interaction is an ideal context to conduct automatic personality assessment in Valente et al. (2012) investigated annotation and experiments toward automatically inferring speakers' personality traits in spontaneous conversations. Their result showed that speech activity statistics provide the best performance for the extraversion trait, prosodic features for the conscientiousness trait and interestingly, overlapping speech statistics provide best performances in case of neuroticism. Therefore, significant effort has been devoted to estimate the personality traits of individuals based on their verbal and nonverbal behaviors.

2.2. Estimation of Performance

Many studies have been conducted to estimate skills and performance. Chen et al. (2014) reported an automated multimodal scoring model for public speaking assessments. Ishii et al. (2018) investigated gaze behavior and dialog act categories during turn-keeping/changing activities based on empathy skill levels, which were measured using Davis's interpersonal reactivity index. Jayagopi et al. (2012) proposed a framework to define and extract group behavioral cues characterizing speaking and looking patterns in face-to-face interactions. Nguyen et al. (2014) proposed a computational framework to predict hirability in real job interviews automatically based on applicant and interviewer nonverbal cues extracted from audio and visual modalities. Okada et al. (2016) presented a computational analysis of individual communication skills that were assessed by 21 external raters with experience in human resource management. Park et al. (2014) proposed a computational approach for using verbal and nonverbal behavior from multiple modalities of communication to predict speaker persuasiveness in online social multimedia content and demonstrated that having prior knowledge regarding a speaker's sentiments partially contributes to predicting their level of persuasiveness. Ramanarayanan et al. (2015) presented a comparative analysis of three different feature sets for predicting different human-rated presentation proficiency scores. Sanchez-Cortes et al. (2011) proposed a computational framework to infer emergent leadership in newly formed groups based on nonverbal behaviors by combining speaking turns, prosodic features, visual activity, and motion. Soleymani et al. (2019) investigated verbal and nonverbal behaviors during intimate self-disclosure. Wörtwein et al. (2015) proposed using an interactive virtual audience for public speaking training. They focused on the automatic assessment of nonverbal behavior and multimodal modeling of public speaking behavior. Therefore, considerable efforts have also been devoted to estimating abilities and performance such as communication skills and empathy skills from verbal and nonverbal behaviors.

As a study related to praising behaviors, Onishi et al. (2020) reported that the nonverbal behaviors related head and face are useful for estimating praising skills. Based on this research, we attempt to estimate praising skills by using not only head and face behaviors but also voice.

3. RESEARCH GOALS

Many studies have focused on techniques for estimating abilities and performance such as personality traits, communication skills, and empathy skills, from verbal and nonverbal behaviors during dialogue and specific tasks. Additionally, some researchers have attempted to estimate abilities and performance based on entire dialogues. In contrast, we attempt to estimate the degree of success of each individual praising behavior during a dialogue. Because praising behavior is considered to vary siginificantly within an individual, we believe it is more appropriate to estimate the degree of success of each praising behavior, rather than the comprehensive human skill of praising behavior in a complete dialogue. In this study, as an attempt to analyze the relationship between praising skills and human behavior during dialogue, we focused on voice, head, and face behaviors. Our main research goal was to clarify which features of a praiser and receiver are important for estimating praising skills based on the basis of voice, head, and face behaviors. The approach in this study is shown in **Figure 1**. We designed this based on related studies (Chen et al., 2014; Ramanarayanan et al., 2015; Okada et al., 2016).

4. DIALOGUE CORPUS

4.1. Recording of Two-Party Dialogue

We created a new corpus of data that includes voice, head, and face behaviors of participants in two-party dialogues in Japanese, as well as evaluations of how successful praising was (Onishi et al., 2020). The participants in two-party dialogues were 34 university students in their twenties (28 males and 6 females). Since they were students of College of Humanities and Sciences in our institute, their academic backgrounds were diverse. They were divided into 17 pairs. Among the 17 pairs, 14 pairs included participants meeting for the first time, two pairs included acquaintances, and one pair included friends. The pairs of participants were assigned by the experimenter after confirming the participants' affiliations and schedules, so that the participants met for the first time as much as possible. There were 13 pairs of the same gender and 4 pairs of the opposite gender. To begin recording dialogues, we asked the participants to prepare two or more examples of things they had been working hard to accomplish with the intention of preparing material for the dialogues. The participants were seated facing each other and separated by 180 cm apart (**Figure 2**).

The dialogues were recorded by using a video camera to record each participant's head and face behaviors and a microphone to record each participant's voice. Each pair of participants (participants A and B) performed dialogues (1) to (3) in accordance with the experimenter's instructions.

(1) A self-introduction (5 min).
(2) Dialogue with participant A as a praiser and participant B as a receiver (5 min).
(3) Dialogue with participant B as a praiser and participant A as a receiver (5 min).

We recorded 17 pairs of dialogues (1) to (3) for a total of 255 min of two-party dialogues. Dialogue (1) (self-introduction) was not used in our analysis because many of the pairs were meeting for the first time and its purpose was simply to relieve the tension between participants. In dialogues (2) and (3), the receiver was instructed to discuss about the things that they had been working hard to accomplish. To ensure that the participants conversed naturally regarding a variety of topics, we also allowed them to discuss topics that they had not prepared beforehand. The praiser was instructed to praise the receiver. However, we allowed the participants to raise questions and react freely to avoid any unnatural dialogues that would have involved unilateral praising. This procedure was approved by our ethics committee.

4.2. Annotation of Dialogue Data and Evaluation of Praising Skills

We used ELAN (Brugman and Russel, 2004), which is a tool for annotating audio and video data, to annotate the utterance scenes to the recorded audio and video data manually. Utterance scene is a continuous utterance interval with a silence duration of less than 400 ms.

4.3. Evaluation of Praising Skills

After the dialogues were recorded, five annotators who did not participate in the recording of the two-party dialogue evaluated how successfully the praiser praised in each utterance scene. The annotators did not have any training or qualifications to avoid the influence of prior knowledge or preconceptions. They referred to the video data, and judged whether or not the praiser was praising the receiver. If the annotators judged that the praiser was praising the receiver, they used a seven point Likert Scale to indicate whether the praiser was successful in praising the receiver each utterance scene: 1 (I do not think the praiser is successfully praising) to 7 (I think the praiser is successfully praising). Next, we treat the utterance scenes which 3 or more annotators judged the praiser was praising the receiver as praising utterance scenes. A total

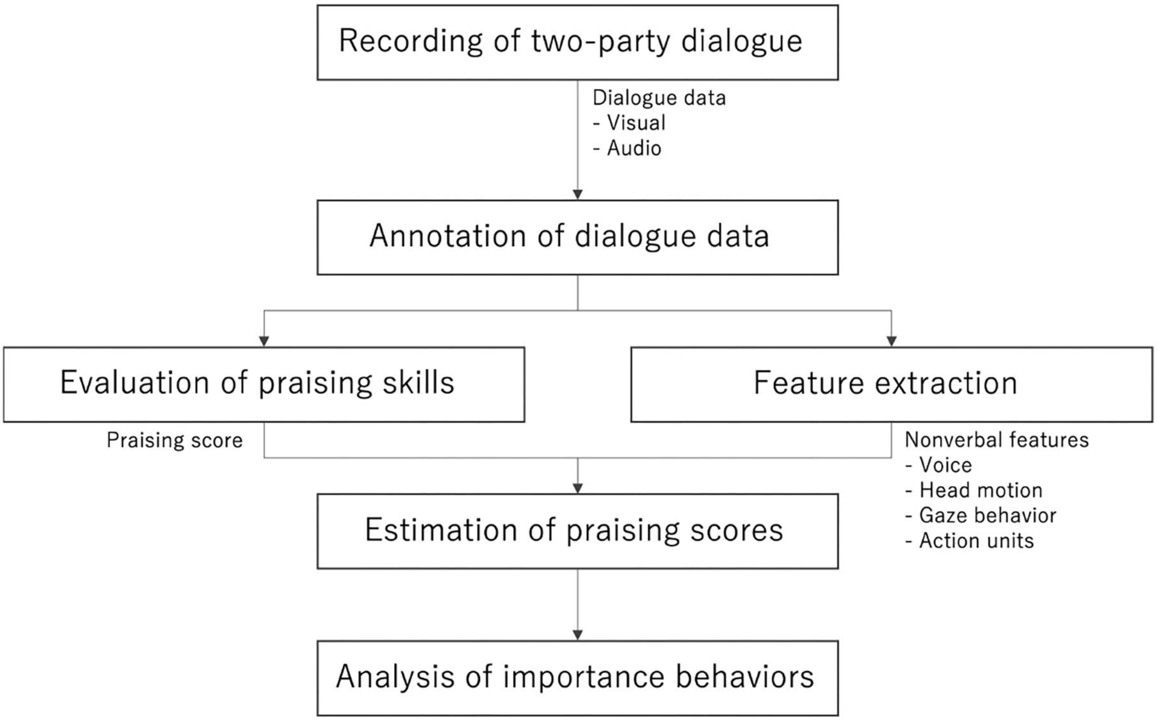
Recording of two-party dialogue
Dialogue data
- Visual
- Audio
Annotation of dialogue data
Evaluation of praising skills
Feature extraction
Praising score
Nonverbal features
- Voice
- Head motion
- Gaze behavior
- Action units
Estimation of praising scores
Analysis of importance behaviors

FIGURE 1 | Overall pipeline of our approach.

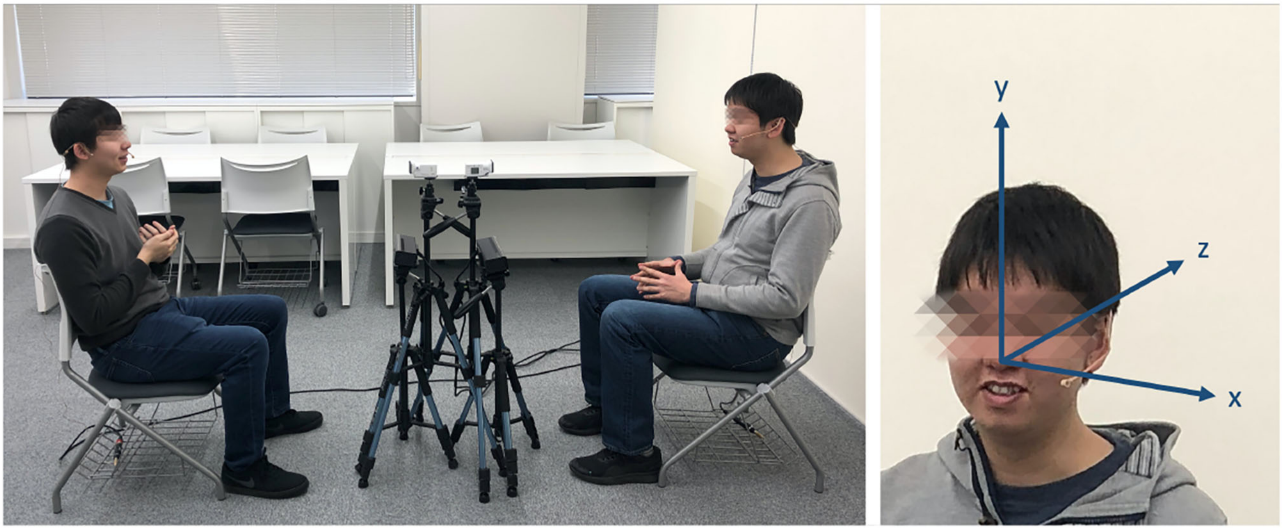

FIGURE 2 | Photograph of two-party dialogue (left) and coordinate system of the head (right).

of 228 praising utterance scenes were obtained. Additionaly, we treat the average value of the evaluation values of the annotators judged the praiser was praising the receiver as praising scores in each praising utterance scenes. We used the intra-class correlation coefficient (ICC) to evaluate the concordance rate of praising scores between annotators. After calculating the intra-class correlation coefficient for each combination of 3–5 annotators, and then calculating the weighted average considering the number of samples, the batch rate of the praiding score was ICC(2, k)=0.571. This result suggests that the praising scores are reliable data with a moderate concordance rate among annotators.

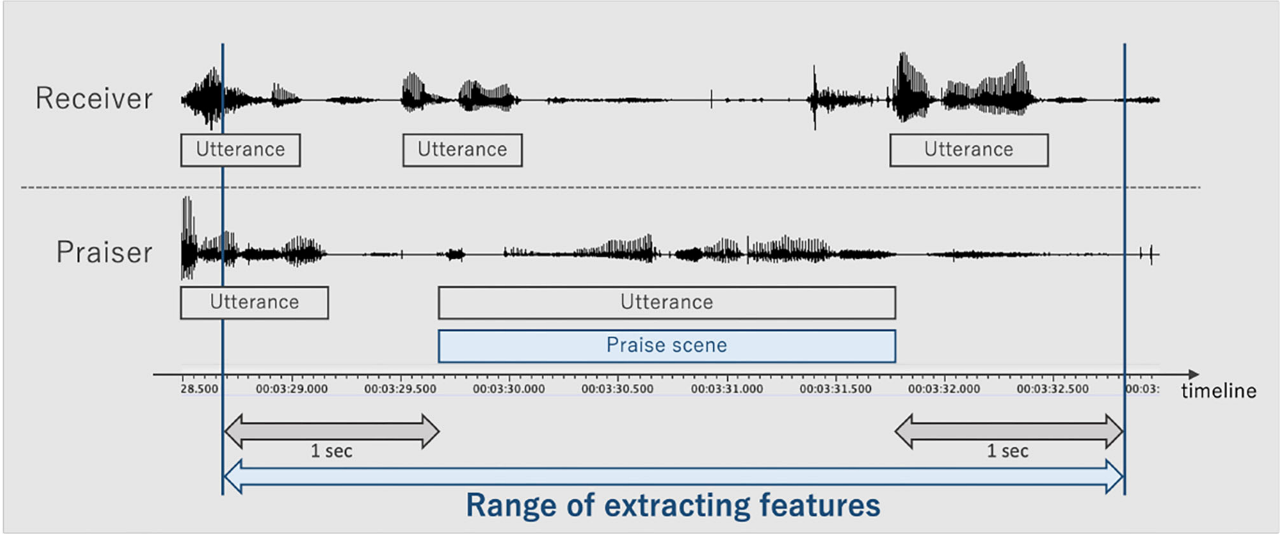

FIGURE 3 | Range for extracting features related to head and face behaviors.

5. ANALYSIS OF FEATURES THAT CONTRIBUTE TO PRAISING SCORES

We performed the analysis to clarify what kind of voice, head, and face behaviors of the praiser and receiver are important for praising successfully. Specifically, we extracted features related to the voice, head, and face behaviors, and developed a machine learning model that estimates the skills of praising. Based on the developed model, we analyzed what kind of behaviors of the praiser and receiver are important for praising the target successfully.

5.1. Feature Extraction

We extracted features related to the voice, head, and face behaviors of the praisers and receivers. Specifically, the features for voice were extracted from the audio data recorded by the microphones worn by participants using openSMILE (Eyben et al., 2010), which is a voice information processing tool. The features of head, gaze, and action units (Ekman and Friesen, 1978) were extracted from the video data captured by the video camera installed in front of the participants using OpenFace (Baltrušaitis et al., 2016), which is a face image processing tool.

Voice: We used the features provided as a standard set Schuller et al. (2009). We considered the maximum (_max), minimum (_min), range (_range), absolute position of the maximum/minimum (_maxPos/_minPos), arithmetic mean (_amean), slope/offset of a linear approximation (_linregc1/_linregc2), quadratic error computed as the difference between the linear approximation and actual contour (_linregerrQ), standard deviation (_stddev), skewness (_skewness), kurtosis (_kurtosis) of the root-mean-squared signal frame energy (pcm_RMSenergy), mel-frequency cepstral coefficients 1 to 12 (pcm_fftMag_mfcc), zero-crossing rate of the time signal (pcm_zcr), voicing probability (voiceProb), the fundamental frequency (F0) over the range of the utterance scene, and the first derivatives of all these features over the range of utterance scenes were also considered.

Head motion: We considered the variance (_var), median (_med), and 10th (_p10) and 90th percentile values (_p90) of the rotation angles about the *x*-axis (pose_Rx), *y*-axis (pose_Ry), and *z*-axis (pose_Rz) of the head over the time range of 1 s before and after the utterance scene (**Figure 3**). When the face was viewed from the video camera side, the *x*-axis was defined from the left to the right, the *y*-axis was defined from the bottom to the top, and the *z*-axis was defined from the front to the back. In Japanese culture, the behavior of moving the face down and returning to the front (nodding) expresses a positive meaning, so this behavior plays an important role in praising.

Gaze behavior: We considered the variance, median, and 10th and 90th percentile values of the angles about the *x*-axis (gaze_Ax) and *y*-axis (gaze_Ay) of gaze over the time range of 1 s before and after the utterance scene when the face was viewed from the video camera side, the *x*-axis was defined from the left to the right, and the y-axis was defined from the bottom to the top.

Action units: Action units are the fundamental actions of individual muscles or groups of muscles in the humen face (Ekman and Friesen, 1978). We considered the variance, median, and 10th and 90th percentile values of the intensities of action units (**Table 1**) used in OpenFace in over the time range of 1 s before and after the utterance scene.

These features were extracted from the praiser (praiser_) and receiver (receiver_). Specifically, these features were extracted from all utterance scenes and normalized each feature to have a mean value of zero and variance of one to align the values of the features. In the following, we define the features related

TABLE 1 | The list of action units.

Item	Content	Item	Content
AU01	Inner brow raiser	AU14	Dimpler
AU02	Outer brow raiser	AU15	Lip corner depressor
AU04	Brow lowerer	AU17	Chin raiser
AU05	Upper lid raiser	AU20	Lip stretcher
AU06	Cheek raiser	AU23	Lip tightener
AU07	Lid tightener	AU25	Lips part
AU09	Nose wrinkler	AU26	Jaw drop
AU10	Upper lip raiser	AU45	Blink
AU12	Lip corner puller		

voice behaviors as *the audio features*, and the features related head motion, gaze behavior, and action units as *the visual features*.

5.2. Estimation of Praising Scores

We developed a machine learning model that estimates the skills of praising by using one or combination of the audio features and visual features. Based on the above, we divided the praising scenes (228 scenes in total) into three classes: low class, medium class, and high class, and developed a classifier that estimates which class the praising score belongs to based on audio and visual features. In order to keep the number of praising scenes in each class as equal as possible, the praising score low to high classes were defined as follows. The threshold values were defined taking the number of scenes in the three classes to be as equal as possible and the number of scenes with the same praising score value not to exist in multiple classes.

- Low group: praising utterance scenes with a praising score of 3.8 points or less (82 scenes in total).
- Middle group: praising utterance scenes with a praising score greater than 3.8 points and less than 4.4 points (65 scenes in total).
- High group: praising utterance scenes with a praising score of 4.4 points or higher (81 scenes in total).

We used Random forests (Breiman, 2001), which can evaluate the importance of features, to develop estimation model. We tuned hyperparameters such as the learning rate and tree depth using Hyperopt (Bergstra et al., 2013). Feature selection was repeated until the model stopped improving by removing the least important features sequentially. The dataset was randomly divided into 90% training data and 10% test data. The task of estimating the class to which the test data belongs using a model trained on the training data was repeated 100 times.

5.3. Results of the Proposed Models

The mean values of each indicator are listed in **Table 2**. As the baseline, we used a model M0 that outputs low, medium, and high groups of praising scores with a probability of 36, 28, and 36% according to the proportion of each group in the dataset (chance level).

We performed a paired *t*-test on the *F*-values of model M0 and each model of M1 to M6. At this time, we performed the Shapiro-Wilk test on the *F*-values of M0 to M6 to check whether it follows a normal distribution. As a result, we confirmed that the values followed a normal distribution ($p > 0.05$), so we performed a parametric test. There are significant differences between model M0 and the proposed models M1 [$t(99) = -22.678, p < 0.01$], M2 [$t(99) = -18.031, p < 0.01$], M3 [$t(99) = -26.188, p < .01$], M4 [$t(99) = -10.021, p < 0.01$], M5 [$t(99) = -16.521, p < 0.01$], M6 [$t(99) = -10.682, p < 0.01$]. This leads us to consider that the models M1 to M6 that we proposed are able to obtain higher performance compared to the model M0. We performed a paired *t*-test on the *F*-values of the three models considering the behaviors of the praiser (model M1, M2, and M3). There are significant differences between the proposed models M1 and M2 [$t(99) = 4.218, p < .01$], M2 and M3 [$t(99) = -4.248, p < .01$]. This result suggests that using the audio features or both of the audio and visual features of the praiser yields higher performance than using the visual features. Additonally, this leads us to consider that the audio features of the praiser can be useful for estimation. Next, we performed a paired *t*-test on the *F*-values of the three models considering the behaviors of the receiver (model M4, M5, and M6). There are significant differences between the proposed models M4 and M5 [$t(99) = -6.026, p < .01$], M4 and M6 [$t(99) = -2.074, p < .05$], and M5 and M6 [$t(99) = 3.543$ $p < .01$]. This result suggests that using the visual features of the receiver yields higher performance than using the audio features or both of the audio and visual features. In addition, this leads us to consider that the visual features of the receiver can be useful for estimation. Finally, based on the above results, we constructed a model M7 using the audio features of the praiser and the visual features of the receiver and a model M8 using the audio and visual features of the praiser and the visual features of the receiver. We performed a paired *t*-test on the *F*-values of the each of two models M7, M8, and M1, M3, and M5. At this time, we performed the Shapiro-Wilk test on the *F*-values of M7 and M8 to check whether it follows a normal distribution. As a result, we confirmed that the values followed a normal distribution ($p > .05$), so we performed a parametric test. There is a significant difference between the proposed models M1 and M8 [$t(99) = -2.493, p < .05$], M3 and M8 [$t(99) = -2.347, p < .05$], M5 and M7 [$t(99) = -9.565, p < .01$], M5 and M8 [$t(99) = -11.599, p < .01$]. This result suggests that we could improve the performance of the model by adding the visual features of the receiver to the audio and visual features of the praiser, and we could estimate the praising skill to some reasonably from the model considering the audio and visual features of the praiser and the visual features of the receiver. Thus, we consider that audio features of praiser are important features for estimating the praising skills in addition to the visual features of praiser and receiver clarified in the previous study (Onishi et al., 2020).

5.4. Discussion of Important Features

The top 10 features in terms of importance for model M8 are presented in **Figure 4** to highlight important features for praiser and receiver. As described in **Figure 4**, important features for the praiser are related to pcm_zcr, pcm_fftMag_mfcc.

The praiser_pcm_zcr_sma indicates the number of times that the amplitude of the sound passes through the zero level in a

TABLE 2 | Evaluation results for the proposed models.

	Praiser		Receiver		Recall	Precision	*F*-value
	Audio	Visual	Audio	Visual			
M0					0.350	0.131	0.188
M1	✓				0.543	0.526	$0.508^{M0^*,M2^*}$
M2		✓			0.469	0.469	0.444^{M0^*}
M3	✓	✓			0.540	0.532	$0.510^{M0^*,M2^*,M6^*}$
M4			✓		0.328	0.335	0.305^{M0^*}
M5				✓	0.405	0.437	$0.387^{M0^*,M4^*,M6^*}$
M6			✓	✓	0.367	0.362	$0.332^{M0^*,M4\dagger}$
M7	✓			✓	0.569	0.551	0.539^{M5^*}
M8	✓	✓		✓	0.575	0.582	$0.548^{M1\dagger,M3\dagger,M5^*}$

* *Indicates significance level of 1% or less (p < .01),* † *Indicates significance level of 5% or less (p < .05).*

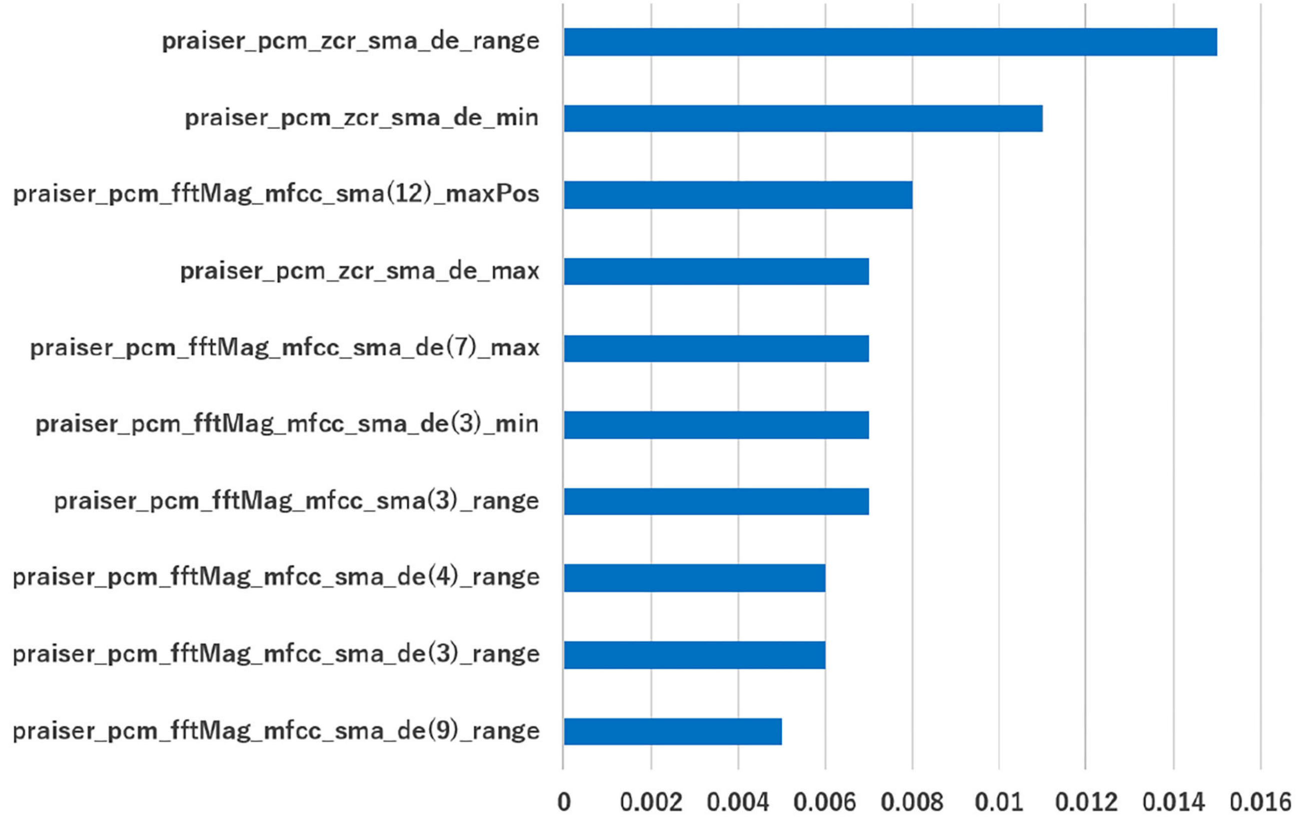

FIGURE 4 | The top 10 features in terms of importance in model M8. The *x*-axes represent the importance of the features.

frame of the praiser's voice. The value of this feature increases when it is voiceless sound. This indicates that voiceless sounds are included in praiser utterances. Actually, many praiser utterances were confirmed to be in Japanese, including voiceless sounds, such as *"sugoidesune."* Some of the participants commented that they tried to use these words to convey their praises. Therefore, we consider that praising a partner with words such as *"sugoidesune"* is important in order to fully praise their successes. However, the number of scenes containing words such as *"sugoi"* was 64 out of 82 cases (78.0%) in the high group, 44 out of 65 cases (67.7%) in the medium group, and 45 out of 81 cases (55.6%) in the low group. Thus, we consider that the detection of specific keywords alone is not enough to estimate the praising skills. In the future, we plan to analyze this feature in more detail by relating it to verbal features.

The praiser_pcm_fftMag_mfcc_sma indicates the acoustic characteristics of the vocal tract in consideration of human hearing. This feature is considered to represent the voice quality of an individual. Therefore, the voice quality of the praiser, such as strength and pitch, is considered to be important for praising successfully. Additionally, we plan to investigate what kind of vocal tract characteristics are effective for successful praising.

Based on these insights, we confirmed the importance of voice, head, and face behaviors to be successful praising. In the future, we will clarify more precisely how to use each modality for successful praising.

From **Figure 3**, the overlapped utterances of the praiser and the receiver in some scenes, the information about overlapped utterances and pauses (differences between utterances) may be useful for estimating the praising skills. First, a total of 89 scenes of overlapped utterances in praising scenes. In detail, the number of overlapped scenes was 33 out of 82 (40.2%) in the high group, 21 out of 65 (32.3%) in the medium group, and 35 out of 81 (43.2%) in the low group. The mean duration of overlapped utterances was 0.870 seconds in the high group, 1.020 seconds in the medium group, and 1.740 seconds in the low group. Secondly, the mean length of the pause between utterances in the praising scenes was 0.781 seconds for the high group, 0.895 seconds for the medium group, and 1.133 seconds for the low group. Based on the above two points, we consider that whether or not utterances are overlapped, the time information of overlapped utterances, and the length of pauses between utterances are useful for estimating the praising skills. In the future, we plan to use these information as features to estimate the skill of praise and examine the improvement of the estimation accuracy.

6. CONCLUSION

In this study, we attempted to analyze the relationship between praising skills and human behaviors in dialogue by focusing on voice, head, and face behaviors. We developed a machine learning model that uses features related to voice, head, and face behaviors to estimate praising skills and clarified which features of a praiser and receiver are important for estimating praising skills. We could estimate praising skills from the audio and visual features of the praiser and the visual features of the receivers. The experimental results demonstrated the importance of features related to the zero-crossing rate, mel-frequency cepstral coefficients. Analyzing the features of high importance revealed that a praiser should use the specific words that mean amazing or great in Japanese to achieve more successful praising. In addition, we also revealed that the voice quality of the praiser.

Although the above findings were obtained in this study, there are some limitations. First, the dataset size is not large. However, 17 pairs of dialogue data yielding 228 praising utterance scenes allowed us to construct valid machine learning models and conduct statistical analysis for an initial step of this research activity. Specifically, our machine learning models improved the accuracy compared to the baseline model (M0). In the future, we need to expand the dialogue data in order to develop this study. Second, to reduce the effects of shame, each dialogue was conducted with no one in the room other than the participants. However, it cannot be confirmed that this effect was completely eliminated. Third, we unified the dialogue conditions, meaning the social role of each participant was the same. Forth, the features of the praiser and receiver were extracted at the same time. Therefore, the behavior of the receiver is the behavior while being praised by the praiser. In fact, the receiver's behavior may appear after being praised by the praiser. In the future, we would like to increase the pattern of the extraction range of the features related to behavior and verify what kind of behaviors are important. Fifth, we used features extracted by openSMILE to investigate the relationship between voice behavior and praising skills. In the future, we need to analyze the voice behavior in detail. We are planning to use another method for extracting voice features to conduct a more detailed analysis. Based on these factors, there is room for additional improvement in the accuracy of the developed model. Therefore, we plan to improve accuracy by incorporating additional modalities, such as verbal behavior and gestures, and clarifying their importance for successful praising. Additionally, we plan to analyze praising behaviors among people with various roles. Finally, we plan to analyze and clarify the communicative effects of praising behaviors.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the Ethics Committee prohibits the disclosure of the dataset because it contains personal information. Requests to access the datasets should be directed to AM, miyata.akihiro@acm.org.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by College of Humanities and Sciences, Nihon University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TO, AY, and AO mainly contributed to definition of the utterance scene, annotates all the dialogue data, and analysis of visual and audio features. Particularly, TO and AY contributed the most to this research by working on feature extraction and machine learning model construction. RI, AF, TN, and AM contributed to dialogue recording, experimental design, and data analysis. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

This work is based on an earlier work: Analyzing Nonverbal Behaviors along with Praising, in Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), © ACM, 2020. http://dx.doi.org/10.1145/3382507.3418868.

REFERENCES

Aran, O., and Gatica-Perez, D. (2013). "One of a kind: inferring personality impressions in meetings," in *Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI'13)* (New York, NY), 11–18.

Baltrušaitis, T., Robinson, P., and Morency, L. (2016). "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision (WACV'16)* (Lake Placid, NY), 1–10.

Batrinca, L., Mana, N., Lepri, B., Sebeand, N., and Pianesi, F. (2016). Multimodal personality recognition in collaborative goal-oriented tasks. *IEEE Trans. Multimedia* 18, 659–673. doi: 10.1109/TMM.2016.2522763

Bergstra, J., Yamins, D., and Cox, D. (2013). "Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms," in *Proceedings of the 12th Python in Science Conference (SciPy'13)*, (Austin, TX), 13–20.

Biel, J., Teijeiro-Mosquera, L., and Gatica-Perez, D. (2012). "Facetube: predicting personality from facial expressions of emotion in online conversational video," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI'12)* (New York, NY), 53–56.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45. doi: 10.1023/A:1010933404324

Brophy, J. (1981). Teacher praise: a functional analysis. *Rev. Educ. Res.* 51, 5–32.

Brugman, H., and Russel, A. (2004). "Computational analysis of persuasiveness in social multimedia," in *Proceedings of the 4th International Conference on Language Resources and Language Evaluation (LREC'04)* (Istanbul), 2065–2068.

Chen, L., Feng, G., Joe, J., Leong, C., Kitchen, C., and Lee, C. (2014). "Towards automated assessment of public speaking skills using multimodal cues," in *Proceedings of the 16th ACM International Conference on Multimodal Interaction (ICMI'14)* (Istanbul), 200–203.

Ekman, P., and Friesen, W. (1978). *Manual for the Facial Action Coding System.* Palo Alto, CA: Consulting Psychologists Press.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). "opensmile - the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th International Conference on Multimedia* (Firenze), 1459–1462.

Henderlong, J., and Lepper, M. (2002). The effects of praise on children's intrinsic motivation: A review and synthesis. *Psychol. Bull.* 128, 774–795. doi: 10.1037/0033-2909.128.5.774

Ishii, R., Otsuka, K., Kumao, S., Higashinaka, R., and Tomita, J. (2018). "Analyzing gaze behavior and dialogue act during turn-taking for estimating empathy skill level," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI'18)* (Boulder, CO), 31–39.

Jayagopi, D., Sanchez-Cortes, D., Otsuka, K., Yamato, J., and Gatica-Perez, D. (2012). "Linking speaking and looking behavior patterns with group composition, perception, and performance," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI'12)* (Santa Monica, CA), 433–440.

Jenkins, L., Floress, M., and Reinke, W. (2015). Rates and types of teacher praise: a review and future directions. *Psychol. Schools* 52, 463–476. doi: 10.1002/pits.21835

Jokinen, K., Furukawa, H., Nishida, M., and Yamamoto, S. (2013). Gaze and turn-taking behavior in casual conversational interactions. *ACM Trans. Interact. Intell. Syst.* 3, 1–30. doi: 10.1145/2499474.2499481

Kalis, T., Vannest, K., and Parker, R. (2007). Praise counts: using self-monitoring to increase effective teaching practices. *Prevent. School Failure Alternative Educ. Children Youth* 51, 20–27. doi: 10.3200/PSFL.51.3.20-27

Kurihara, K., Goto, M., Ogata, J., Matsusaka, Y., and Igarashi, T. (2007). "Presentation sensei: a presentation training system using speech and image processing," in *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI'07)* (New York, NY), 358–365.

Lin, Y., and Lee, C. (2018). "Using interlocutormodulated attention blstm to predict personality traits in small group interaction," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI'18)* (New York, NY), 163–169.

Nguyen, L., Frauendorfer, D., Mast, M., and Gatica-Perez, D. (2014). Hire me: computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Trans. Multimedia* 16, 1018–1031. doi: 10.1109/TMM.2014.2307169

Okada, S., Ohtake, Y., Nakano, Y., Hayashi, Y., Huang, H., Takase, Y., and Nitta, K. (2016). "Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI'16)* (New York, NY), 169–176.

Onishi, T., Yamauchi, A., Ishii, R., Aono, Y., and Miyata, A. (2020). "Analyzing nonverbal behaviors along with praising," in *Proceedings of 22nd ACM International Conference on Multimodal Interaction (ICMI'20)* (New York, NY), 609–613.

Park, S., Shim, H., Chatterjee, M., Sagae, K., and Morency, L. (2014). "Computational analysis of persuasiveness in social multimedia," in *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI'14)* (New York, NY), 50–57.

Pianesi, F., Mana, N., Cappeletti, A., Lepri, B., and Zancanaro, M. (2008). "Multimodal recognition of personality traits in social interactions," in *Proceedings of the 10th ACM International Conference on Multimodal Interaction (ICMI'08)* (New York, NY), 53–60.

Ramanarayanan, V., Leong, C., Chen, L., Feng, G., and Suendermann-Oeft, D. (2015). "Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring," in *Proceedings of the 17th ACM International Conference on Multimodal Interaction (ICMI'15)* (New York, NY), 23–30.

Sanchez-Cortes, D., Aran, O., Mast, M., and Gatica-Perez, D. (2011). A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Trans. Multimedia* 14, 816–832. doi: 10.1109/TMM.2011.2181941

Schuller, B., Steidl, S., and Batliner, A. (2009). "The interspeech 2009 emotion challenge," in *10th Annual Conference of the International Speech Communication Association* (Brighton), 312–315.

Soleymani, M., Stefanov, K., Kang, S., Ondras, J., and Gratch, J. (2019). "Multimodal analysis and estimation of intimate self-disclosure," in *Proceedings of the 21st ACM International Conference on Multimodal Interaction (ICMI'19)* (New York, NY), 59–68.

Valente, F., Kim, S., and Motlicek, P. (2012). "Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus," in *Thirteenth Annual Conference of the International Speech Communication Association* (Portland, OR).

Wörtwein, T., Chollet, M., Schauerte, B., Morency, L., Stiefelhagen, R., and Scherer, S. (2015). "Multimodal public speaking performance assessment," in *Proceedings of the 17th ACM International Conference on Multimodal Interaction (ICMI'15)* (New York, NY), 43–50.

Conflict of Interest: RI, AF, and TN are employed by NTT Corporation.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.