Estimation of Conversational Activation Level during Video Chat using Turn-taking Information.

Yurie Moriya
Tokyo University of Agriculture and Technology
2-24-16 Nakacho, Koganei, Tokyo
50011646135@st.tuat.ac.jp

Takahiro Tanaka
Tokyo University of Agriculture and Technology
2-24-16 Nakacho, Koganei, Tokyo
takat@cc.tuat.ac.jp

Toshimitu Miyajima
Tokyo University of Agriculture and Technology
2-24-16 Nakacho, Koganei, Tokyo
tomiyaji@cc.tuat.ac.jp

Kinya Fujita
Tokyo University of Agriculture and Technology
2-24-16 Nakacho, Koganei, Tokyo
kfujita@cc.tuat.ac.jp

ABSTRACT
In this paper, we discuss the feasibility of estimating the activation level of a conversation by using phonetic and turn-taking features. First, we recorded the voices of conversations of six three-person groups at three different activation levels. Then, we calculated the phonetic and turn-taking features, and analyzed the correlation between the features and the activity level. The analysis revealed that response latency, overlap rate, and speech rate correlate with the activation levels and they are less sensitive to individual deviation. Then, we formulated multiple regression equations, and examined the estimation accuracy using the analyzed data of the six three-person groups. The results demonstrated the feasibility to estimate activation level at approximately 18% root-mean-square error (RMSE).

Author Keywords
Conversational activation; estimation; phonetic feature; speech information processing.

ACM Classification Keywords
H.5.2. User Interface: Theory and methods

INTRODUCTION
The growth of high-speed broadband networks facilitates the development and spread of various remote communication systems, such as instant messaging tools, micro-blogs, and video chat systems. A video chat system provides an easy-to-use remote communication function that includes images of the users. However, privacy-sensitive users might be unwilling to display their own images. Furthermore, there is a risk of an unintentional leak of personal information from the background image. On the other hand, avatar-based communication systems, such as Second Life [18], are utilized for casual remote communication. The use of an avatar, as the surrogate of the user, allows the user to avoid unintentional transmission of personal information. However, the conventional avatars lack the ability to express nonverbal information, such as a gaze, facial expressions, and gestures. Vargas noted that 65% of the message is communicated via nonverbal channels [20]. Mehrabian reported the roles of facial expression, phonetic information, and verbal information are 55%, 38%, and 7%, respectively, if contradictory information is expressed via verbal and nonverbal information [9]. Therefore, the expression of adequate nonverbal information is needed for smooth and natural avatar-mediated communication. However, the nonverbal information is manually controlled in most of the avatar communication systems.

The manual control of all nonverbal information, such as gaze and gesture, is cumbersome and increases the user task load. Therefore, the automatic control methods of nonverbal information have been studied. Interactor, developed by Watanabe et al., revealed that the automatic control of nods and gestures using the vocal information of the speaker facilitates the conversation [22]. Miyajima and Fujita also proposed methods to control the gaze, facial expressions, and gestures of avatars based on phonetic features, and experimentally examined the effects of these features on communication [10]. In natural communication, the magnitude and velocity of gestures are thought to change with the level of enthusiasm in conversation. Therefore, it is expected that automatically controlling the magnitude and velocity of gestures to reflect the conversational activation level produces a more natural impression and facilitates communication. Saeki et al. studied a unified control algorithm used to control the multiple nonverbal information based on the activation level [17]. However, methods to estimate conversational activation levels have not been established.

Maeda et al. proposed a method to estimate the activation level based on gestures such as hand motions [7]. The proposed method was expected to work well with a camera-compatible environment; however, the use of a camera limits

APCHI'12, August 28–31, 2012, Matsue-city, Shimane, Japan.

Table 1. Relationship between emotions and phonetic features (summary of preceding studies).

	emotion				
	anger	joy	dislike	sad	laxation
pitch	high	high	low	low	low
intensity	high	high	low	low	-
utterance rate	very fast	fast	fast	slow	-
cadence	-	appear	-	-	disappear

the possibilities for the usage of the proposed method. On the other hand, voice is essentially required information for voice-based chatting and is less restrictive in terms of system setup. Therefore, we experimentally discuss the feasibility of automatically estimating the activation level during natural conversation based only on vocal information.

It is known that prosodic information of the voice reflects emotions [14]. Turn-taking features are also known to change to reflect emotions [13]. Therefore, we studied the relationship between the phonetic and turn-taking features and the activation level in conversation among groups of three subjects. The results indicated the correlation between the activation level and the response latency, and the overlap rate and the speech rate. Therefore, the off-line estimation of the activation level was also examined on the basis of these three indicators. It demonstrated the feasibility of activation level estimation that is robust against individual differences and environmental variations. The contribution of this study to CHI is to demonstrate the feasibility of calibration-free conversational activation level estimation. It is expected to allow voice communication systems, as well as avatar chat systems to control their functions and reflect the activation level of the users.

RELATED STUDIES

Activation Level and Phonetic and Turn-taking Features

Numbers of studies have pointed out the change of conversational voice features due to emotional variations. For example, intensity, pitch, and utterance rates are higher when speakers are angry or joyful [1, 12]. On the other hand, these three features are lower when the speakers are sad [12], and intensity and pitch are lower [12] and utterance rate is higher [3] when speakers are expressing dislike. In addition, the previous studies indicated a higher pitch and a strong cadence when speakers are joyful, and a lower pitch and a decrease of cadence when the speakers are relaxed [11]. The changes of prosody due to emotional variations are summarized in Table 1. As the emotions during the conversation stabilize or change gradually without a strong prompt, the emotion is expected to induce automatically detectable changes in a conversational voice.

A circumplex model of affect, proposed by Russel, represents emotions by using two dimensions of pleasure-displeasure and sleep-arousal[15]. In this circumplex model, anger and joy are considered as higher arousal level emotions. On the other hand, the arousal levels of dislike and sadness are considered as low. When we define the conversational activation level as the enthusiasm in the conversation, it is considered to be proportional to the arousal level in the circumplex model. Therefore, the activation level is expected to induce the observable effects in the conversational voice, and the estimation of activation level based on phonological information would be possible.

In addition to the prosodic changes, turn-taking features were also researched [16]. It has been pointed out that the overlap frequency of utterances also increases along with the activation level [13]. It has also been reported that turn-taking latency, which is the interval after one speaker finishes speaking and the other starts speaking, decreases in the case of an affirmative or an approval response [6]. Therefore, in the conversations in which the speakers expressed joy, the response latency could be a good indicator of the activation level. In addition, motion synchronization among the speakers was also indicated in continuous conversation, and prosodic synchrony occurred, especially in groups of women [13].

Avatar Control using an Estimated Activation Level

If the duration and speed of the avatar behavior, such as gesture and facial expression, are appropriately controlled in conjunction with the conversational activation level, avatar communication might be more natural. Therefore, we defined the degree of conversational activity as enthusiasm of the speakers during the conversation, and studied the automatic control algorithms of avatar behavior according to the activation level [17]. Sejima et al. reported that conversational impressions are improved by controlling the avatar behavior based on the estimated conversational activation level that is a weighted moving average of overlap frequency [19].

If the real-time estimation of the activation level from the conversational voice is possible, the automatic control of the avatar behavior would be feasible. The activation-based control will allow avatars to behave more naturally. However, a certain amount of conversational voice is required for accurate estimation. It causes an estimation delay from several seconds to several minutes. This delay issue is discussed in the discussion section.

The activation levels of the participants are individually different. The speaker's activation level is supposed to be reflected by the voice features. On the other hand, the listening participant, who say nothing, would also be influenced by the general mood of the group and might feel his ore her own activation level. Therefore, a parameter is required to describe the conversational mood of the entire group. Thus, in this study, with the objective of controlling an avatar to reflect the activation level of not only a speaker but also the group, we assume two activation levels: a group activation level and a personal activation level.

EXPERIMENTS

Methods

We conducted video chat experiments among groups of three subjects under three different conditions, and recorded the conversation voices. In principle, avatar-mediated communication has limited nonverbal expression functions compared to face-to-face or video communication, even with automatic

control mechanisms. Therefore, to limit the nonverbal communication channel, we limited the viewing range of the camera, which allowed the users to watch only the face and chest images of their partners. There were 18 subjects (nine male and nine female) in their twenties. The conversation groups were three three-male groups and three three-female groups. Each conversational group consisted of participants of the same gender who were acquainted with each other to minimize the influence of factors other than the test activity and to let the groups raise their activation level more easily.

We attempted to control the activation level of each conversation as low, medium, or high by imposing different constraints on the topics during the experiment. There were no other constraints, which allowed participants to converse naturally. The duration of the experiments was 15 minutes for each condition; 45 minutes in total. The sequence of the conditions was randomized to avoid order effect. After the conversation during each condition, the participants were required to evaluate the subjective activation level using a linear scale from 0 to 100. The participants scored their own personal activation level as well as the entire group ' s activation level, for the beginning, middle, and final parts of the conversation, as well as for the entire conversational period.

Experimental Conditions

To control the conversational activity at each of the three levels, we imposed several conditions on the talking topics for discussion. Before the experiments, topics related to news and social issues were collected as conversational topic candidates. Then, we interviewed the subjects to determine their interest and amount of knowledge of each topic. According to the answers from the interview, the topics for discussion were selected as follows:

- L-condition, envisaged low activity: We requested the subjects to seriously discuss topics related to news, in which all three subjects were least interested and had least knowledge. The conversation about other topics was strictly prohibited.

- M-condition, envisaged middle activity: We requested the subjects to discuss topics related to social issues of which they had some knowledge but little interest. The conversation about other topics was strictly prohibited.

- H-condition, envisaged high activity: We imposed no limitations on the topics for discussion. To help the participants choose discussion topics, the results of the survey questionnaire about the hobbies of each subject was disclosed to all the subjects. Furthermore, the subjects were instructed to enjoy the conversation.

During the L-condition, subjects had very little knowledge and interest for the discussion topic. The expected activation level was low because the subjects had difficulty in broadening the conversation. On the other hand, in H-conditions, there was no limitation on the topics, and the speakers were allowed to talk about their favorite topics. Therefore, the activation level was higher. In M-conditions, the subjects were able to broaden the conversation, because they had some knowledge about the topics. However they had less interest about the topics than during the H-condition. Therefore, the expected activation level was between the L and H conditions.

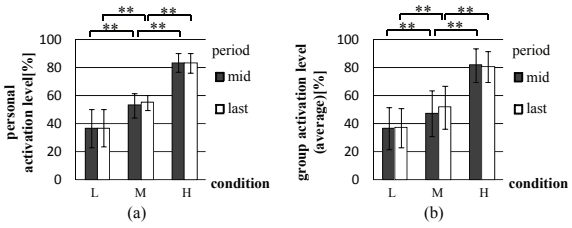

Figure 1. Evaluation of subjective activation level. (a) personal activation level, (b) group activation level (error bars represent standard deviations)

Summary of Experimental Results

To confirm that the activation levels were controlled as designed, we calculated the averages of the subjective activation levels. The results are shown in Figure 1. The data for the first five minutes of each discussion was excluded from the analysis because we observed that a similar conversation concerning what to speak about occurred in each condition independent of the imposed topics. The significant differences in each pair of conditions were observed in multiple comparisons, after the confirmation of the significant difference in variance analysis ($p<0.01$). The results indicated that the conversations could be controlled at three levels by imposing limitations on the discussion topics.

ANALYSIS OF THE RELATIONSHIP BETWEEN THE ACTIVATION LEVEL AND VOICE FEATURES

In this study, we analyzed 10 features in conformity with previous studies. The analyzed features that reflect the prosody are average sound pressure, average pitch, sound pressure variation, pitch variation, utterance rate, utterance length and synchrony of sound pressure. In addition, we analyzed the average of response latency, overlap rate, and speech rate features that reflect the structures of conversations. The features were experimentally examined as activation level indicators for automatic estimation.

Methods of Voice Features Calculation

We used the application software, Praat [2], for analysis, similar to the preceding studies [23]. The calculation interval of sound pressure and pitch was set to 30 ms. To calculate the turn-taking features, we clipped out the vocal durations by applying the experimentally determined threshold, which was five times the white noise level and one tenth of the average sound pressure. Because short silences among utterances are caused by sound fluctuation and breathing, we interpolated short silences similar to the manner in the preceding study [21]. In this study, the silent durations less than 0.3 s were interpolated. After the interpolation of the temporal silence, the utterances longer than 0.1 s were regarded as the actual utterances and others were regarded as noise. The calculation method for each feature is described as follows:

- Average sound pressure: The sound pressure levels, sampled at 30 ms intervals, were normalized from 0 to 1, by dividing by the maximum value for each speaker. Average sound pressure was calculated every 5 minutes.
- Average pitch: Average pitch was calculated every 5 minutes after the normalization.
- Sound pressure variation: Variation in sound pressure is expected to be larger in tonic utterances. Therefore, the standard deviation of the sound pressures for the utterances was calculated and normalized by the average value.
- Variation coefficient of pitch: Pitch also has a great variation during tonic utterances. The standard deviations of pitch for the utterance interval was calculated and normalized.
- Utterance rate: To calculate the rate of utterances, it was necessary to detect each syllable and count the number of the syllables. In this study, the syllables were approximated to the duration between the two troughs of the sound pressures [8]. The rate of utterance was approximated as the average of the duration of each syllable.
- Utterance length: The utterance duration for a sentence decreases as the speaking speed increases. Consequently, the utterance length is likely to change in reflection of the change of speaking speed. Therefore, the length of each utterance, which was clipped out as described previously, was calculated by subtracting the start time of the utterance from the end time.
- Sound synchrony: The cases, that showed a positive correlation among the three speakers in sound pressure, pitch, utterance rate, or utterance length were regarded as the synchrony of the subjects, as similarly regarded by previous studies [13]. The rate of the correlative utterances was calculated as the synchronic trend feature. Linear interpolation was applied to the non-utterance durations, similar to methods used in previous studies. The rate of positive correlation coefficients for each 50 utterances was calculated as the measure of synchrony.
- Overlap rate: In enthusiastic conversations, more speech overlaps occurred, such as laughing. Therefore, we determined the overlaps by detecting when subject began speaking while another person was speaking. Furthermore, the rate of the overlaps was calculated to eliminate the effect of the total number of the utterances [6].
- Average response latency: After the detection of the turn-taking, the intervals were calculated between the end of the utterance of the first speaker and the beginning of the utterance of the next speaker. The intervals are regarded as the response latency of the latter speaker. The response latency takes a negative value when overlap occurs. In enthusiastic conversations, the faster turn-taking rhythm was presumed to shorten the response latency.

  However, turn-taking did not occur in the utterances that completely overlapped another utterance, such as laughter or back-channel feedback. Therefore, we eliminated the cases in which the lately-begun utterance ended before the end of the turn-holding utterance. The latencies were calculated only for the cases without complete overlaps, which means turn-taking occurred.
- Speech rate: The personal speech rate was calculated as the rate of the utterance duration for which the system detected the utterance of the subject during the entire experiment.

Results of Analysis

The averages of the features of each speaker and each group for the last 10-minute periods of the conversations are shown in Figures 2 and 3. The speech rate of a group was defined as the rate of non-silent state time, during which at least one of the three participants is speaking. The group features, other than the speech rate, are the average values of the three subjects. Synchrony was not observed for both utterance rate and utterance length, but similar synchrony was observed for sound pressure and pitch. Therefore, we show only the results of the sound pressure. The groups were composed as follows:

- Group G1: Speakers a, b, and c
- Group G2: Speakers d, e, and f
- Group G3: Speakers g, h, and i
- Group G4: Speakers j, k, and l
- Group G5: Speakers m, n, and o
- Group G6: Speakers p, q, and r

Speakers a-i were male, and speakers j-r were female.

As a result of the analysis, similar to the previous studies, at a high activation level, the averages of the normalized sound pressures, the averages of the normalized pitches, the overlap rate, and the speech rate showed increasing tendencies, and decreasing tendencies were observed in the response latencies. On the other hand, the variations of sound pressure and pitch, the speaking rate, and the utterance length did not show significant correlation with the activation level. The synchrony was observed in only the female groups, similar to results reported in the previous study [13]. The Kruskal-Wallis test showed significant differences in five features: the average of normalized sound pressure and pitch, overlap rate, the average of response latency, and speech rate.

Depending on the activation level, an increase or a decrease was observed in the averages of normalized sound pressure and pitch, the overlap rate, the average response latency, and the speech rate for most of the subjects. However, the sound pressure and the pitch also showed larger individual differences even after the normalization. On the other hand, the overlap rate and the speech rate showed relatively smaller differences among the groups. The average response latency decreased to 0 s at a high activation level for most of the subjects, and both the group differences and the individual differences were smaller.

The changes of three turn-taking features related to the conversation structure of G1 are exceptionally smaller at a high

activation level. The difference of the activation level between L and H-conditions was approximately 50% for the five other groups; however, it was 20% for G1. In other words, it is considered that the subjects in G1 were less excited compared to those in other groups during the H-condition. This result might be due to the less effectiveness of topics control in G1.

Selection of Indicators for Estimation

Out of the analyzed features, the averages of the normalized sound pressure and the pitch, the overlap rate, the speech rate, and the average response latency showed tendencies to increase or decrease depending on the activation level. They showed good correlation with the activation level. The sound pressure and pitch showed larger individual differences even after the normalization. In addition, they also were affected by factors in the system environment, for example, microphone sensitivity. Therefore, the features must be calibrated before using them to control avatars, which could be a potential disadvantage for practical usage. On the other hand, the overlap rate, the average response latency, and the speech rate, which relate to the conversational structure, showed fewer individual and group differences during the H-condition. Because these three features are calculated on the basis of the relationship between time and utterance, they are not affected by the system environment or by the phonetic individual differences. Therefore, the estimation of the activation level without calibration is expected to be feasible.

We discuss the feasibility to estimate personal and group activation levels by using three features as an indicator: overlap rate, the average of response latency, and speech rate, as discussed in the following sections.

ESTIMATION

Examination of Estimation Equation

Relationship between the Activation Level and Conversational Indicators

To discuss the feasibility of estimating the activation levels using overlap rate, average response latency, and speech rate, we analyzed the correlation between each indicator and the subjective activation level. The scatter plots of the average values of the three indicators during the middle and the final conversational periods versus the subjective activation level of each speaker or each group are shown in Figure 4. Because of the individual variability of the subjective scale, the averages of the three speakers were utilized as the group activation levels, which was the subjectively scored activation level of the entire group. Furthermore, exceptionally long latencies were observed in some situations, for example, when the topics of discussion were exhausted. They reflect the state of the group, but are different from the conversational rhythm. Therefore, we eliminated the longer latencies using an experimentally decided threshold of 5 s. The averages were calculated after the exceptional values were eliminated. Table 2 shows the Pearson product-moment correlation coefficients between each indicator and activation level. The decorrelation test indicated significant correlation between the activation level and all indicators ($p<0.05$).

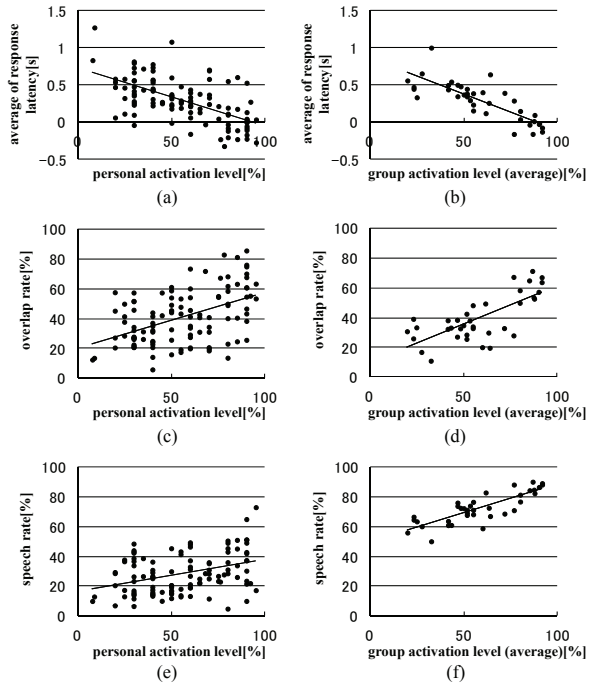

Figure 4. Scatter plots of activation levels and turn-taking features. (a,b) averages of response latency, (c,d) overlap rates, and (e,f) speech rates

The personal activation level, as seen in the scatter plot, showed a strong correlation with the average response latency. Moderate correlations were observed in the overlap rate and the speech rate, even their variability was relatively larger. The variabilities for the groups were smaller than those for the persons for all three indicators, and they showed a stronger correlation. In summary, there is some difference in accuracy, but the feasibility to estimate the activation levels of individuals and groups by using the three conversational indicators was demonstrated.

Determination of the Estimation Equation

The correlation has been confirmed between the activation level and each of the three indicators. The use of larger numbers of indicators is expected to improve the estimation robustness against the variation of the conversational patterns. Therefore, we performed multiple regression analyses using all three indicators and obtained multiple regression equations as estimation equations. The estimation equations for personal and group activation levels are shown as Equations 1 and 2, respectively.

Table 2. Correlation coefficients between activation levels and turn-taking features.

	overlap rate	average of response latency	speech rate
personal	0.48	−0.60	0.37
group	0.72	−0.79	0.82

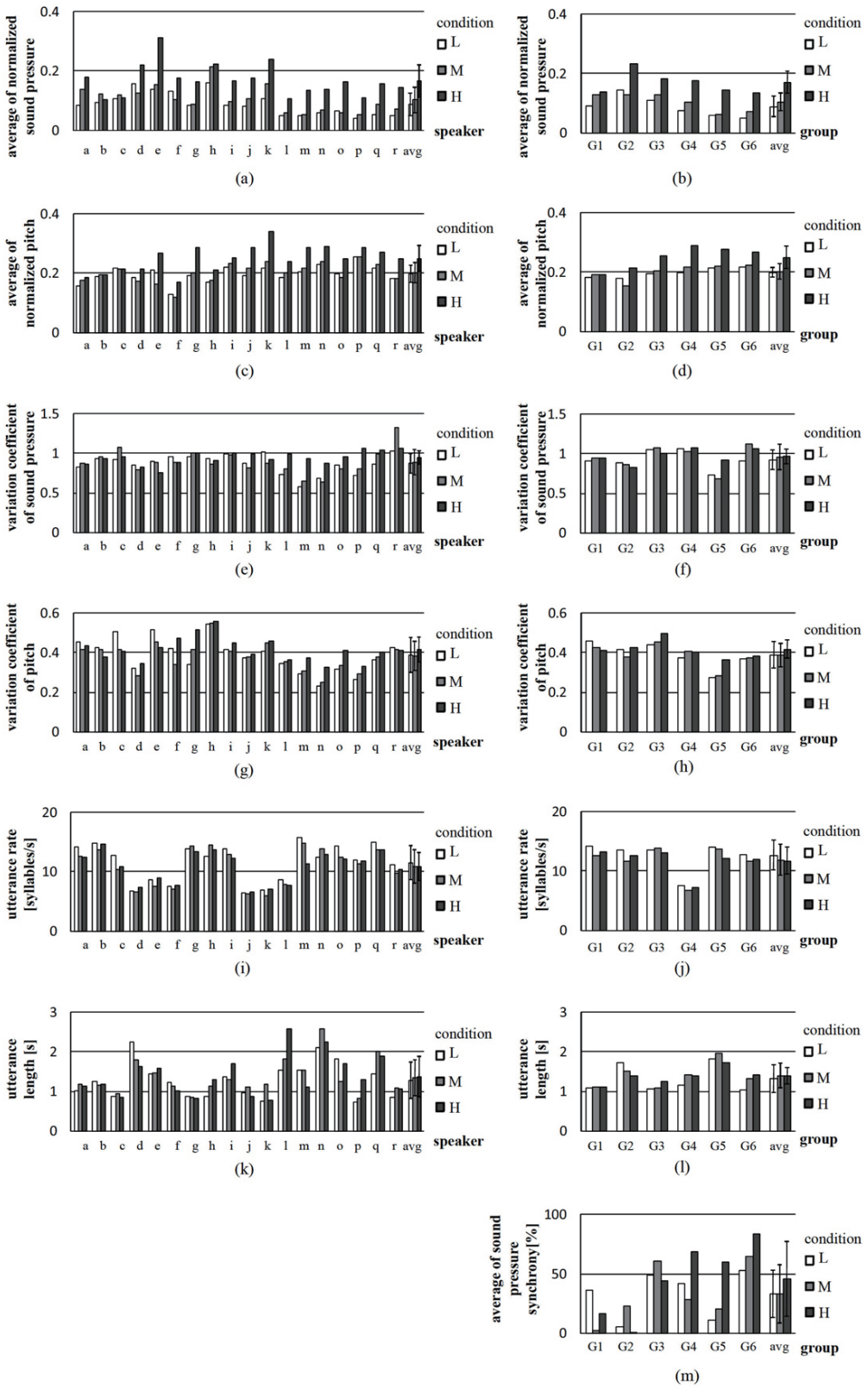

Figure 2. Results of phonetic feature analysis. (a,b) averages of normalized sound pressure, (c,d) averages of normalized pitch, (e,f) variation coefficients of sound pressure, (g,h) variation coefficients of pitch, (i,j) utterance rates, (k,l) utterance lengths, and (m) averages of normalized sound pressure synchrony

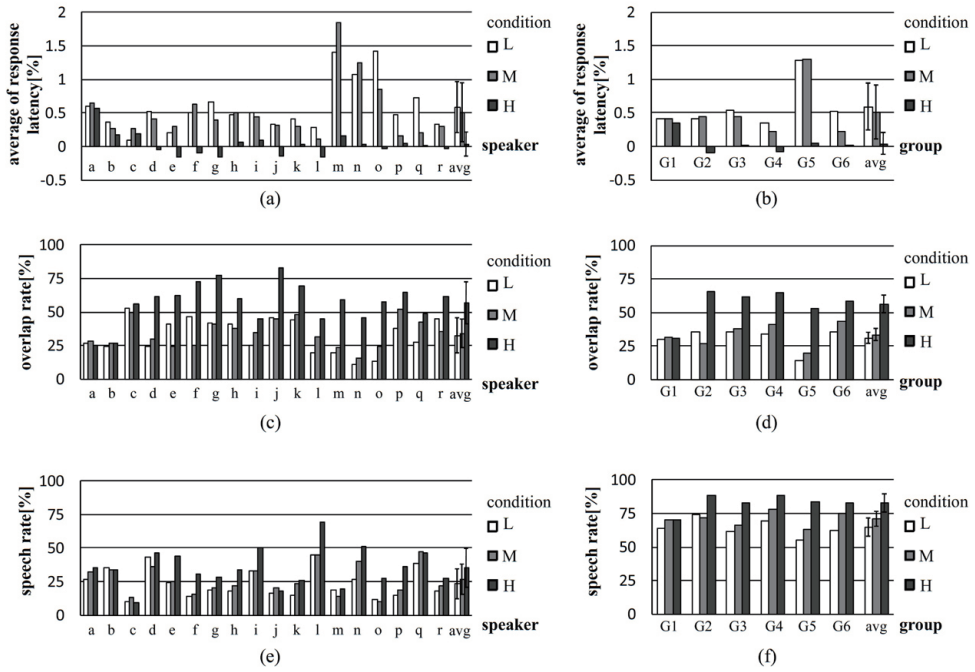

Figure 3. Results of turn-taking feature analysis. (a,b) averages of response latency, (c,d) overlap rates, and (e,f) speech rates

personal activation level

$$\begin{aligned} = \; & 0.276 * \text{overlap rate}[\%] \\ & -27.2 * \text{avg. of response latency[s]} \\ & +0.423 * \text{speech rate}[\%] + 41.1 \end{aligned} \quad (1)$$

group activation level

$$\begin{aligned} = \; & 0.363 * \text{overlap rate}[\%] \\ & -33.8 * \text{avg. of response latency[s]} \\ & +1.44 * \text{speech rate}[\%] - 22.0 \end{aligned} \quad (2)$$

Estimation Experiment

To verify the estimation capability, we performed off-line estimations of the personal and group activation levels of the analyzed conversational data of all six groups using Equations 1 and 2. The selected three indicators during the middle and final periods were utilized for estimation. The estimated values below zero and over 100 were limited to zero and 100. The root-mean-square errors (RMSE) of the estimated activation levels and the correlation coefficients between the subjective evaluation activation levels and estimated values are shown in Table 3.

The RMSE for personal and group activation levels were 18% and 12%. respectively. The estimation error for group activation levels was less than that for personal activation levels, but strong correlations were obtained in both cases. Therefore, in summary, the activation level estimation is highly feasible using conversational indicators without individual or group calibration.

DISCUSSIONS

Estimation Accuracy

In this study, the accuracy for the group activation level was high compared to that for personal activation level because of the following reasons: The subjective scales of persons essentially have individual variation. The effect of the individual variation was reduced in the group activation level evaluation because the scores of three subjects were averaged, while the personal activation level is the score of a person. Furthermore, the personal activation level indicator, such as overlap rate or average latency, tended to show a larger variance coefficient, especially for the relatively silent speakers, because the number of their utterances was small. In contrast, compensatory behaviors were observed in the groups. For example, participants sometimes intentionally spoke to avoid the awkward continuance of silence. This kind of behavior reduces the variance coefficient of the group speech rate. Therefore, a more robust estimation would be generally possible for group activation levels compared to the personal activation levels, similar to the proposed results.

To improve the accuracy, methods to reduce the individual differences are required, such as automatic calibration of the indicators or automatic setting of parameters as determined by machine learning. However, the deviation of the indicator might be also affected by dynamic factors, such as mood or

Table 3. RMSE and correlation coefficients for analyzed data set.

	RMSE	Correlation
personal	18%	0.60
group	12%	0.83

the subject ' s role in the conversation, as well as individual difference. The effectiveness of the parameter setting needs to be evaluated before applying it in various cases. In this study, we did not utilize sound pressure and pitch, because of larger individual differences. However, they could be candidates for activation level estimation parameters after additional processing to reduce the individual differences. The group activation level might be utilized for the improvement of the accuracy of the personal activation level. For example, the summation of the relative change of the personal activation level as an offset and the group activation level as a baseline might be effective for the persons whose activation level tends to be underestimated.

In this study, no mixed-gender group was studied to minimize the influence of the factors other than activity and to let the groups raise the activation level more easily. The turn-taking indicators, which were used for estimation, showed less gender difference compared to personal or group differences as seen in Figure 3. The mixture of gender might make conversation difficult in some situations. In these cases, the speech rates of the participants would become unbalanced, as seen in G4 in Figure 3. However, the estimated values of G4 were comparable to the other groups. Therefore, it is expected that the estimation of a mixed-gender group's activation level is feasible.

Control of Avatar Activation Level

In this study, each subject subjectively evaluated his or her own activation level. However, the self-judged activation level of a person does not necessarily correspond to the activation level evaluated by another person. For example, the activation level will be estimated to be lower than the self-judged level in the subjects whose voice and conversational behavior show less change even during enthusiastic conversations. The activation level, which is subjectively evaluated by another person, is likely to be lower as well as the automatically estimated value. Therefore, the use of the estimated activation level to control avatars might improve the degree of coincidence between the avatar behavior and the impression induced by the voice. The appropriateness of the automatically estimated activation level needs to be discussed by applying the estimated values to avatar behavior.

The activation levels were estimated on the basis of the subjects ' voices during a five-minute conversation. The use of voice for a longer duration imposes a time-delay to the behavioral change of the avatar. The proposed method will adequately reflect the gradual change of the mood in calm conversations. On the other hand, the avatars might not adequately behave during sudden changes in activity level caused by a joke or other factors. However, the shortening of the estimation duration will seriously affect the estimation accuracy. The temporal activation usually disappears during a shorter term. Therefore, the activation level appears to be modeled better by a temporal component and a continuous component. The proposed method appears suitable for the continuous component estimation. The supposed method for using a temporal component is the usage of the initial sound pressure of a responsive utterance, which was utilized for the control of an avatar ' s smile level [10].

LIMITATIONS

In this study, the experiments were carried out in groups of three people of the same gender. Therefore, the effects of the gender mixture and the number of the participants still need to be studied. A video chat system was used instead of an avatar voice chat system. However, the voice-driven automatic control of nonverbal expressions might affect the turn-taking behaviors of the users. It would be necessary to perform the experiments with an avatar voice chat system. Furthermore, the current study utilized voice data of five-minute intervals for estimation. It theoretically produces a delay in estimation. However, the shortening of the data for the estimation will increase estimation error, which was 18% in this study. The detection of a momentary change of the activation level is the next challenge.

CONCLUSIONS

In this study, we examined the relationship between the conversational activation level and the phonetic and turn-taking features, which were calculated from the voice recordings of the conversations in groups of three persons each. The response latency, the overlap rate and the speech rate showed strong correlation with both the personal and group activation levels. The off-line estimation was performed using a multiple regression equation on the basis of the three indicators. The feasibility of automatically estimating the activation level was demonstrated. The future work is to develop the real-time estimation algorithm and the application to avatar control.

ACKNOWLEDGEMENTS

This work was partly supported by the MEXT Fund for Promoting Research on Symbiotic Information Technology and JSPS KAKENHI.

REFERENCES

1. Bezooyen, R, V.: Characteristics and Recognizability of Vocal Expressions of Emotion, Foris Pubns USA (1984).
2. Boersma, P., and Weenink, D.: Praat: doing phonetics by computer (Version 5.2.21) [Computer program]. Retrieved from http://www.praat.org/.
3. Coleman, R, F., and Williams, R.: Identification of emotional states using perceptual and acoustic analyses, Transcripts of the 8th Symposium on Care of the Professional Voice, The Voice Foundation Vol. 1, pp. 77-83 (1979).
4. Hartmann, B., Mancini, M., and Pelachaud, C.: Implementing Expressive Gesture Synthesis for Embodied Conversational Agents; Gesture Workshop 2005: pp.188-199 (2005).
5. Kato, Y., Kato, S., and Akahori, K.: Effects of emotional cues transmitted in e-mail communication on the emotions experienced by senders and receivers, Computers in Human Behavior, Vol. 23, No. 4, pp.1894-1905 (2007).

6. Kawahara, T., Kawashima, H., Hirayama, T., and Matuyama, R.: ”Automated Information Concierge” based on Proactive Dialog and Information Retrieval, Magazine of the Information Processing Society of Japan, Vol. 49, No. 8, pp. 912-918 (2008).
7. Maeda, T., Takashima, K., Kajimura, Y., Yamaguchi, N., Kitamura, Y., Kishino, F., Masda, N., Daibo, I., and Hayashi, Y.: A study of nonverbal cues and atmosphere in three-person conversation (in Japanese), IEICE Technical Report, Vol. 109, No. 457, pp.73-78 (2010).
8. Maeran, O., Piuri, V., and Storti, G. G.: Speech recognition through phoneme segmentation and neural classification, Instrumentation and Measurement Technology Conference, 1997. IMTC/97. Proceedings. ” Sensing, Processing, Networking ” IEEE, Vol. 2, pp. 1215-1220 (1997).
9. Mehrabian, A.: Nonverbal communication. Aldine-Atherton, Chicago, Illinois (1972).
10. Miyajima, T., Fujita, K.: Control of avatar ’ s facial expression using fundamental frequency and sound pressure in multi-user voice chat system (in Japanese), Trans. Human Interface Society, Vol. 9, No. 4, pp. 503-512 (2007).
11. Moriyama, T., Saito, H., and Ozawa, S.: Evaluation of the Relation between Emotional Concepts and Emotional Parameters in Speech (in Japanese), Trans. IEICE, Vol. J82-DII, No. 4, pp. 703-711 (1999).
12. Murray, I. R., Arnott, J. L.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, J. Acoust. Soc. Am, Vol. 93, No. 2, pp. 1097-1108 (1993).
13. Nishimura, R., Kitaoka, N., and Nakagawa, S.: Analysis of Factors to Make Prosodic Change in Spoken Dialog (in Japanese), Journal of the Phonetic Society of Japan, Vol. 13, No. 3, pp. 66-84 (2009).
14. Nwe, T. L., and Foo, S. W.: Speech emotion recognition using hidden Markov models, Speech Communication, Vol. 41, pp.603-623 (2003).
15. Russell, J. A.: A circumplex model of affect, Journal of Personality and Social Psychology, Vol. 39, pp.1161-1178 (1980).
16. Sacks, H., Schegloff, E., and Jefferson, G.: A Simplest Systematics for the Organization of Turn-Taking for Conversation, Language, Vol. 50, No. 4, pp. 696-765 (1974).
17. Saeki, Y., Tanaka, T. and Fujita, K.: Unified Control of Avatar ’ s Motion Based on Conversation Activity (in Japanese), Correspondences on Human Interface, Vol. 11, No. 2, pp. 71-74 (2009).
18. Second Life Official Site. http://secondlife.com/.
19. Sejima, Y., Ishii, Y., and Watanabe, T.: A Virtual Audience System for Enhancing Embodied Interaction Based on Conversational Activity, Lecture Notes in Computer Science, Vol. 6772, pp. 180-189 (2011).
20. Vargas, F. M.: Louder than words − An Introduction to Nonverbal Communication, Iowa State University Press (1987).
21. Murakami, I., Katou, H., and Watanabe, T.: Patient-Nurse Communication Support System by Using Speech-Driven Embodied Characters Called InterActors (in Japanese), Proceedings of 67th National Convention of IPSJ, pp. 25-26 (2005).
22. Watanabe, T., Okubo, M., Nakashige, M., and Danbara, R.: InterActor: Speech-Driven Embodied Interactive Actor, International Journal of Human-Computer Interaction, Vol. 17, No. 1, pp. 43-60 (2004).
23. Werker, J. F., Pons, F., Dietrich, C., Kaiikawa, S., Fais, L., Amano, S.: Infant-directed speech supports phonetic category learning in English and Japanese, Cognition, Vol. 103, Issue 1, pp. 147-162 (2007).