IEICE TRANS. INF. & SYST., VOL.E98–D, NO.8 AUGUST 2015

PAPER

NOCOA+: Multimodal Computer-Based Training for Social and Communication Skills

Hiroki TANAKA[†a)], **Sakriani SAKTI**[†b)], ***Members***, **Graham NEUBIG**[†c)], ***Nonmember***, **Tomoki TODA**[†d)], ***and*** **Satoshi NAKAMURA**[†e)], ***Members***

SUMMARY Non-verbal communication incorporating visual, audio, and contextual information is important to make sense of and navigate the social world. Individuals who have trouble with social situations often have difficulty recognizing these sorts of non-verbal social signals. In this article, we propose a training tool NOCOA+ (Non-verbal COmmuniation for Autism plus) that uses utterances in visual and audio modalities in non-verbal communication training. We describe the design of NOCOA+, and further perform an experimental evaluation in which we examine its potential as a tool for computer-based training of non-verbal communication skills for people with social and communication difficulties. In a series of four experiments, we investigated 1) the effect of temporal context on the ability to recognize social signals in testing context, 2) the effect of modality of presentation of social stimulus on ability to recognize non-verbal information, 3) the correlation between autistic traits as measured by the autism spectrum quotient (AQ) and non-verbal behavior recognition skills measured by NOCOA+, 4) the effectiveness of computer-based training in improving social skills. We found that context information was helpful for recognizing non-verbal behaviors, and the effect of modality was different. The results also showed a significant relationship between the AQ communication and socialization scores and non-verbal communication skills, and that social skills were significantly improved through computer-based training.

key words: *computer-based training, multimodality, non-verbal behaviors, context information*

1. Introduction

Socialization and communication are important factors influencing human social life, but the number of people who have trouble with social skills and communication have recently been increasing for a variety of reasons [20]. It has been noted that the extreme case of these traits is autism spectrum disorders (ASD) [3], genetic disorders characterized by social interaction and communication difficulties, as well as unusually narrow, repetitive interests [1], [21]. Given the impact of these problems on everyday life, there has been considerable interest in tools to both identify the degree of these difficulties and allow for training tools to improve social and communication skills. One of the central psychological themes in ASD is empathizing [4]. Empathizing is a set of cognitive and affective skills that we use to make sense of and navigate the social world [12]. The cognitive component of empathy is referred to as "theory of mind" or "mindreading" and entails recognizing the mental state of others. The affective component entails having an emotional response to this recognized mental state. It is well known that Social Skills Training (SST) can be used to effectively improve empathizing ability [2].

We have previously proposed a tool NOCOA [27], which is an application to help test and train non-verbal behaviors. NOCOA allows users to listen to an utterance, and guess intention (is the speaker friendly, sociable, or derisive?) and partner information (is the speaker conversing with a friend or someone senior such as a teacher?), allowing the user to improve their skills in recognizing this information. Previous work with NOCOA confirmed a correlation between non-verbal recognition skills and autistic traits, and examined prospectives for intervention through systematically teaching nonverbal behaviors. While the overall design of NOCOA has proven advantageous in the previous research, NOCOA used only short audio snippets for testing and training the ability to recognize non-verbal behaviors.

On the other hand, there are reports mentioning that not only audio, but also visual information is important to recognize basic and complex emotion [17], [18]. In addition, other reports have mentioned that conversational context influences emotion recognition [7], with potential contextual factors including location, identities of the people around the user, date, time of day, season, temperature, emotional state, and focus of attention [11], [14], [15], [24]. In most previous definitions, the common contextual factor is time, so we focus on temporal context.

In this work, we propose a method that improves the training of non-verbal information recognition skills by incorporating the audio, visual, and contextual information that has been shown to play an important role in recognizing basic and complex emotions. Specifically, we propose an updated application NOCOA+ that uses the multimodal and context information to help in training the ability to recognize non-verbal behaviors, as shown in Table 1. We do so by collecting and incorporating data from several sensory modalities, as well as data considering context. We perform a series of four experiments examining

1. the effect of temporal context on the ability to recog-

Manuscript received November 26, 2014.
Manuscript revised March 26, 2015.
Manuscript publicized April 28, 2015.

[†]The authors are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
a) E-mail: hiroki-tan@is.naist.jp
b) E-mail: ssakti@is.naist.jp
c) E-mail: neubig@is.naist.jp
d) E-mail: tomoki@is.naist.jp
e) E-mail: s-nakamura@is.naist.jp

DOI: 10.1587/transinf.2014EDP7400

Table 1 Comparison of previous works and this work.

	speech	*modality/context*
emotions	[Golan, 2007]	[Golan, 2008; Barett, 2011]
non-verbal	[Tanaka, 2013]	*this work*

nize non-verbal behaviors in testing context,
2. the effect of modality of presentation of social stimulus on ability to recognize non-verbal behaviors,
3. the correlation between autistic traits and non-verbal behavior recognition skills measured by NOCOA+,
4. the effectiveness of computer-based training in improving social skills.

This paper is an extension of work originally reported in [26]. We implemented four experiments with larger number of participants and discuss the results in more detail.

2. Related Work

The use of computers to aid people with communication difficulties has flourished in the last decade However, most applications tend to be rather specific (e.g. focusing only on emotion recognition of facial expressions from still photos) and are often not scientifically evaluated [19].

An application "FEFFA" was proposed to help users recognize emotion from still pictures of facial expressions and strips of the eye region [8]. "Emotion Trainer" teaches emotion recognition of four emotions from facial expressions [25]. "Lets Face It" teaches emotion and identity recognition from facial expressions [28]. Golan and Baron-Cohen [16] proposed a training tool "Mind Reading" which implements an interactive guide to emotions and teaches recognition of 412 emotions and mental states, systematically grouped into 24 emotion groups, and 6 developmental levels. Experiments found that this method can enable adults with ASD to learn mental state recognition, with an improvement of mental state recognition skills indicated during three months of intervention. However, learning skills that generalize beyond the stimuli used in training is still difficult. Although people with ASD improved their ability to recognize emotions from trained stimuli, they had difficulty in recognizing emotions from films in more realistic situations. This is in concert with reports that people with social communication difficulties have trouble in applying learned skills to unseen situations [2].

In previous work, the training typically tended to focus on skills of emotion recognition, and did not include non-verbal behaviors [27]. In this paper, we propose an application NOCOA+ that uses utterances in several modalities and context to help users recognize non-verbal behaviors.

3. Categorization of Non-verbal Behavior

Non-verbal behavior includes various factors (e.g., eye contact, emotion, intention, partner, gesture, and gender). We have previously performed a factor analysis [27] to confirm the important non-verbal factors [29] contributing to social and communication skills, and their relationship with autism-spectrum quotient (AQ), which is a standard method to measure social and communication skills [5]. We found five important factors: 1) intention & interest, 2) politeness/impoliteness & new friends, 3) social places and situations, 4) chit-chat and feelings, 5) other. We selected the first two factors (intention & interest, politeness/impoliteness & new friends) as non-verbal behaviors, and named these groupings as representing "intention" and "partner information" respectively. For example, an AQ question related to intention is "I find it difficult to work out people's intentions," and a question related to partner information is "other people frequently tell me that what I've said is impolite." These two factors were also used as the non-verbal behaviors to be trained and tested by NOCOA+. The categories of partner information were utterances spoken to a "friend" and utterances spoken to a "teacher," and categories for intention were utterances in a "derisive" situation, utterance in a "social" situation, and utterances in a "friendly" situation [27].

4. Recording and Annotation

We next recorded a number of videos representing each of the categories of non-verbal behavior defined in the previous subsection in as natural a manner as possible. In order to ensure that we are able to collect video samples of "derisive," "social," and "friendly" utterances in the intention category, we had each subject perform a conversation according to the following procedure: (a) read the sports section of the newspaper, (b) converse about the content of the article for 10 minutes, (c) read the society section of the newspaper, (d) converse for 10 minutes. The sports and society sections were expected to elicit friendly and derisive behaviors respectively. In addition, to make it easier to collect two types of partner information, we had each subject converse with both a close friend and a teacher.

In this study, four students (4 males, mean age: 23.5) acted as subjects, with each having a score of under 32 on the overall AQ test (the cut-off value of ASD [5]). A video camera (SONY HDR-CX560) was used, and placed in the middle of the two conversants to take frontal shots. A pin microphone (Olympus ME52W) was used for recording each person's speech data. Movie data and speech data are synchronized using the Windows movie maker, and each speech interval (utterance) was detected using the power value extracted by the Snack Tcl/Tk toolkit[†]. Detected utterances were automatically divided into speech and video. We also created utterances including temporal context information from the 5s and 10s prior to the actual utterance.

We annotated the recorded movies with correct category labels. In video recording, we prepared a total of 1200 audiovisual utterances without contextual information, and asked annotators to annotate them. Because annotators are required to have good social skills to recognize non-

[†]http://www.speech.kth.se/snack/

Table 2 Examples of selected utterances

Did you go skating?	Why did you start to play baseball?	We can know whether a company we are employed at is good after ...
Yes, I agree.	I played with one person and maybe he knew my name.	I think people watch figure skating only during big competitions ...
That is an overstatement.	I do not frequently watch figure skating in TV.	And, I think no-one does frequently watch that.

verbal behaviors, we selected three annotators for whom the sums of the AQ subarea scores for communication and social skills were low (the sum of both areas was one for all three annotators). The annotators labeled each utterance into friend, teacher, or others for partner information and into derisive, social, friendly, or others for intention respectively. A total of 109 utterances (9.1%[††]) for which all three annotators agreed on both partner and intention information were chosen for use in NOCOA+. The Cronbach alpha coefficient value was 0.89, indicating that the coding is reliable. We did not select utterances based on discussion between the annotators. Examples of selected utterances are listed in Table 2.

5. Design of NOCOA+

Using these movie samples, we next designed an application to test and train ability to recognize intention and partner information. NOCOA+ was designed according to several principles. First, correlation with AQ: one of the factors influencing the ability to empathize is the severity of ASD [31]. The AQ test is generally used for measuring a person's position on the autism spectrum in both people with and without ASD. Thus, non-verbal behaviors as tested by NOCOA+ should correlate with the AQ, and we have used this to guide our design. Second, systematic design: while individuals with ASD have difficulty in socialization and communication, they also show good and sometimes even superior skills in non-social areas such as "systemizing" [4]. Systemizing is the drive to analyze or build systems, to understand and predict the behavior of events in terms of underlying rules and regularities, and previous work has noted that learning materials can be presented in a manner that utilizes these systemizing skills for increased learning effect [16]. The use of computer-based training for individuals with ASD can take advantage of this systemizing tendency because computer-based environments are predictable, consistent, and free from social demands, which individuals with ASD may find stressful. Users can work at their own pace and level of understanding, and lessons can be repeated over and over again, until mastery is achieved. In addition, interest and motivation can be maintained through different and individually selected computerized rewards [9], [22]. To create an application that satisfies these desiderata, we adopted two types of training and a quiz format that includes computerized reward, where the user of the application chooses from several categories of intention and partner information, modality, contextual information, and difficulty levels.

[††]The chance rate was 2.7%

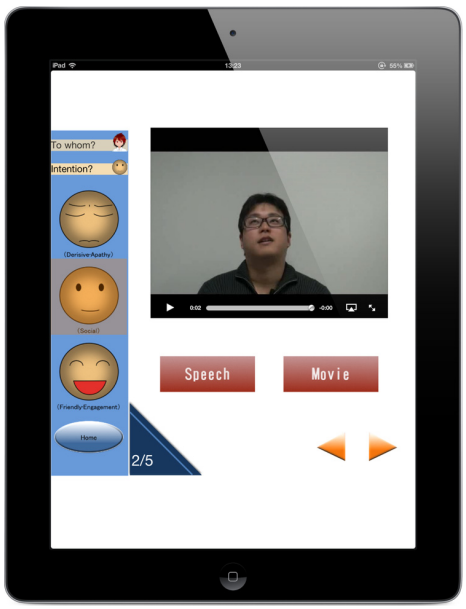

Fig. 1 Screenshot of the training mode in the English version. The user selects modalities (speech and/or movie) and non-verbal behaviors (intention and partner information).

5.1 Training Mode

Training mode was designed to enhance the user's socialization and communication skills. Baron-Cohen et al. [4] speaks of the extreme male brain theory of individuals with ASD, which states that people with ASD prefer things that function in a rule-governed way. In contrast, previous work mentioned that a large number of inputs were needed to train social skills [16], [27]. Thus, we designed training mode to provide two types of training, "listen to a large number of examples" and "check the rules." The former is a conventional method and developed to enable users to learn by listening to and watching utterances for training. 79 utterances were randomly selected from the total of 109 utterances as a closed training set used in training mode. Users can work at their own pace and level of understanding by selecting nonverbal behaviors, difficulty levels, types of modality, and contextual information (Fig. 1). The latter rule-based training regimen is a new approach. The first author created explanations of the eye-movement, prosody, and posture rules that provide hints about the correct answer by listening to the samples, and the user can see these descriptions. An example of this explanation is "People in derisive situation tend to speak with short duration and with lower variation of pitch, and look down." The explanation was reviewed and modified by two other people. The user can select the preferred training regimen from the training menu.

Fig. 2 Screenshot of the test mode interface in the English version. The movie stimulus is displayed, and then the user selects the appropriate intention and partner information.

5.2 Test Mode

In the test mode quiz, 10 questions for measuring the user's non-verbal recognition skills are provided. The 10 question set is chosen at random each time. The questions have the two types of generalization levels shown below: 1) closed: testing is performed using data that was included in the training mode, 2) open: testing is performed using data that was not included in the training mode. The user watches a video or listens to audio of an utterance, and then attempts to guess the intention and partner information corresponding to the utterance (Fig. 2).

For both partner information and intention the maximum score of each question is five. For partner information, the user gets a score of five when the correct partner is chosen and zero otherwise. For intention, the score for mistakes between derisive and social is two, between social and friendly is three, and between derisive and friendly is zero. The intention category's score penalty for mistakes between derisive and social is higher than for those between social and friendly because these are critical misses in social situation [27].

The test mode score is calculated after answering 10 questions, and 100 is the best obtainable score. During the test, the system does not show feedback. After the test, the system shows total score, intention score, partner score, and comments based on the score aimed to encourage the user. These scores are automatically sent to a web server and users can watch their ranking to maintain their motivation.

Table 3 Relationship between difficulty levels and contextual information.

	Easy [%]	Normal [%]	Hard [%]
No context	32	58	10
Context 5	63	37	0
Context 10	70	30	0

6. Experimental Evaluation

In this section, we describe a series of experiments that use NOCOA+ to evaluate contextual differences, modality differences, the relationship between NOCOA+ score and AQ, and the effect of training. The Research Ethic Committee of the Nara Institute of Science and Technology has reviewed and approved our experiments. Written informed consent was obtained from all subjects before the experiments.

6.1 Difficulty Level and Contextual Differences

We expanded the test mode by setting a difficulty level for each utterance. We did this by having participants other than the annotators use test mode. Three difficulty levels were set according to each question's accuracy rate: 1) easy, 2) normal, 3) hard. The accuracy rate of each difficulty level is easy: 81–100%, normal: 51–80% and hard: 0–50%.

In the first experiment, we clarify the benefit of temporal context information in the form of the content directly proceeding the utterance. We hypothesized that contextual information can help the subjects answer questions.

6.1.1 Method

We used the NOCOA+ test mode including three contextual levels: no context, 5 seconds context, and 10 seconds context, which indicate that the user watches not only the utterance itself, but also video from the 0s, 5s, and 10s prior to the actual utterance. First, we collected data corresponding to each level of context (in Sect. 4). Three types of difficulty level were set; easy, normal, and hard according to the criterion mentioned previously. To categorize difficulty levels, 10 participants (8 males and 2 females, mean age: 23.7) answered all questions with each level of contextual information. This experiment conducted using a within subjects design.

6.1.2 Results

In Table 3, we show the relationship between difficulty levels and contextual information. We can see that the percentage of each difficulty category is related to the contextual level. In the 5s and 10s contexts, more than 60% of questions were categorized as the easy difficulty level. This result indicates that contextual information helped people to infer the correct answer.

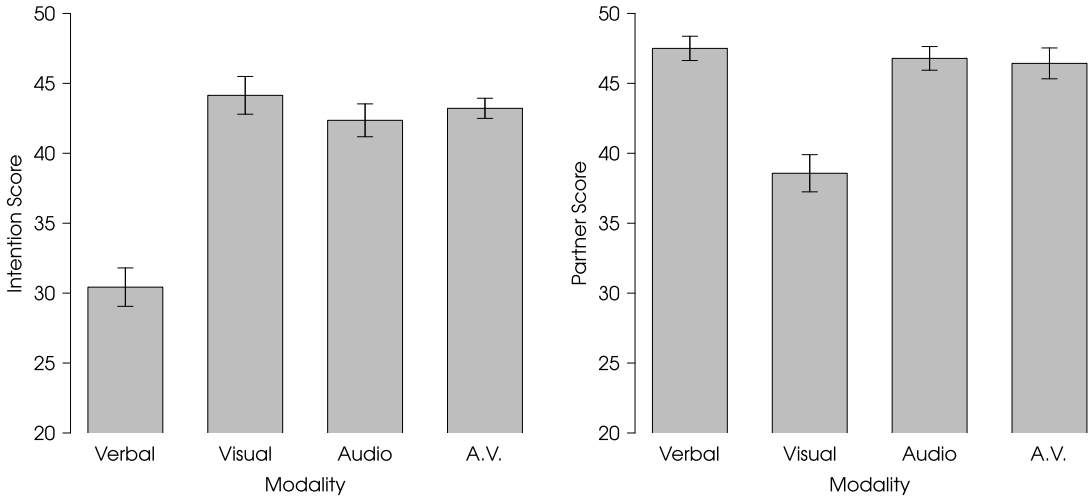

Fig. 3 Modality differences in terms of intention and partner score with standard error bars. A.V. indicates audiovisual.

6.2 Modality Differences

In the second experiment, we investigated the effect that modality differences have on recognition of non-verbal information. We set a hypothesis that modality of stimulus has an effect on the ability to identify non-verbal information, and performed experiments to test this hypothesis using the testing mode of NOCOA+.

6.2.1 Method

We recruited a total of 14 participants (11 males and 3 females, mean age: 22.5) for the experiment. This experiment is conducted using a within subjects design. Here, because we only sought to investigate the effect of modality differences, we controlled for difficulty level. Participants took the NOCOA+ test mode, and answered 10 questions randomly selected from the easy difficulty level, which include four modalities: audiovisual, audio, visual, and verbal (where the first author of this article transcribed the speech in the audiovisual data and read it in a flat tone without emotion). The closed data was used, and scores were averaged.

We set a hypothesis that characteristics of intention and partner information are different. To verify the hypothesis we analyzed the score for intention and partner information separately, and used one-way ANOVA to measure statistical significance. We also performed a pairwise comparison using Bonferroni's method [23].

6.2.2 Results

The results in Fig. 3 indicate that there were significant differences in each modality's score in terms of intention and partner information. The ANOVA showed [$F(3,52)=29.64$, $p < .01$, $\eta_p^2 = 0.63$] for intention score and [$F(3,52)=15.77$, $p < .01$, $\eta_p^2 = 0.48$] for partner information score respectively. In the case of the verbal modality, a large number of errors were found in the intention category, and in the case of the visual modality, a relatively large number of errors were found in the partner information category. Post-hoc comparison showed that in the case of intention the verbal score was significantly lower than audiovisual ($p < .01$), audio ($p < .01$) and visual ($p < .01$) scores, and in the case of partner information, the visual score was significantly lower than audiovisual ($p < .01$), audio ($p < .01$) and verbal ($p < .01$) scores.

The results showed that people have difficulty correctly inferring others' intention by only the linguistic information of speech, and people have difficulty correctly inferring others' partner information by only visual signals.

6.3 Relationship of Autistic Traits

In the third experiment, we investigated the relationship between the AQ score and non-verbal communication skills measured using NOCOA+.

6.3.1 Method

12 participants (11 males and 1 female, mean age: 23.1) performed the easy and normal difficulty levels with the closed data set using audiovisual data one time. The averaged score of the easy and normal difficulty levels was calculated. Finally, they took the Japanese version of AQ [30], and the sum of the two AQ subareas (communication and social skill) was measured. We calculated the relationship and correlation coefficients between NOCOA+ score and AQ, and performed a linear regression analysis.

6.3.2 Results

Figure 4 shows the results indicating the relationship of the sum of social and communication scores and test mode score of NOCOA+. The maximum score of test mode is

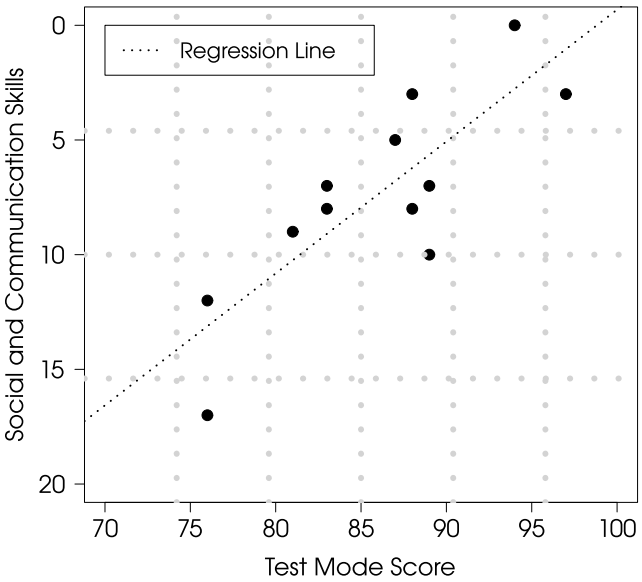

Fig. 4 Relationship between the sum of social and communication AQ scores and test mode score of NOCOA+ with a regression line.

100, and a high score indicates high non-verbal communication skills. On the AQ test, the maximum social and communication scores are each 10, and a high score indicates a high level of autistic traits. As Fig. 4 shows, there is a correlation between the sum of the AQ subareas and averaged test mode score with a correlation coefficient of 0.82 ($p < .01$). We also fitted a regression line using the least squares method with a coefficient of determination of 0.67.

These results confirmed that there is a strong relationship between the ability to recognize non-verbal information in video and the AQ subareas.

6.4 Training Effect

In the fourth experiment, we investigated whether computer-based training results in an increase in ability to recognize non-verbal information. We hypothesized that computer-based training is effective in allowing users to train their ability to recognize intention and partner information, and that the effectiveness is not related to difficulty and generalization level. To verify the hypothesis, we investigated whether users are able to maintain high scores even in unseen open questions.

6.4.1 Method

We recruited 12 participants (11 males and 1 female, mean age: 23.0). This experiment was conducted using a between-subjects design. The participants were randomly assigned to the training group (6 males) or the non-training group (5 males and 1 female). The mean value of initial scores of two groups were not significantly different for both easy difficulty level (training: 85.5 (SD: 4.5), non-training: 90.3 (SD: 5.6)) [$t(10)=-1.62$, $p > .1$] and normal difficulty level (training: 78.2 (SD: 9.0), non-training: 81.3 (SD: 8.7)) [$t(10)=-0.62$, $p > .1$], which is similar to the result of Sect. 6.3.2.

The procedure includes a training session in which the subject: (a) Enters a laboratory and receives a description by first author, (b) Practices how to use NOCOA+, (c) Performs the easy and normal difficulty levels using the closed data set one time, (d) Either uses training mode for 20 minutes (training group), or waits for the same 20 minutes (non-training group), (e) Repeats procedure (c) using test mode with open data as well. The training group is instructed to first use rule-based training and then use statistics-based training. Almost all participants were able to complete training on all utterances in 20 minutes. The absolute improvement in score ((e) score - (c) score) was calculated and averaged for each group. We also test whether the training group scored higher than the non-training group in the case of open data. The significant differences were tested by Student's t-test.

6.4.2 Results

Almost all participants were able to complete training on all utterances in 20 minutes. Figure 5 shows the improvement of test mode score before and after 20 minutes. In terms of difficulty level easy (left side of Fig. 5), the improvement in score was 8.0 (SD: 2.7) in the training group and −0.5 (SD: 4.7) in the non-training group respectively [$t(10)=3.86$, $p < .01$]. In terms of difficulty level normal (right side of Fig. 5), the improvement in score was 16.3 (SD: 5.4) in the training group and 0.8 (SD: 6.1) in the non-training group respectively [$t(10)=4.66$, $p < .01$].

In the case of open data, for easy difficulty level, the averaged score was 96.0 (SD: 4.5) in the training group and 91.7 (SD: 3.3) in the non-training group, indicating that the training group had a score significantly higher than that of the non-training group [$t(10)=1.90$, $p < .05$] (one-tailed test). For normal difficulty level, the averaged score was 91.5 (SD: 6.0) in the training group and 85.5 (SD: 7.4) in the non-training group, indicating that there is a tendency that the training group was higher than the non-training group [$t(10)=1.54$, $p < .1$] (one-tailed test).

Thus, we found that in both difficulty levels, 20 minutes of training was helpful for participants of the training group with both closed and open data, and we confirmed effectiveness by systematic training in both audio data and visual data.

7. Conclusion

In this paper, we proposed a training tool NOCOA+ that uses utterances in several modalities and context. We used NOCOA+ to examine computer-based social skills training that uses not only audio data, but also visual and contextual data. NOCOA+ was designed for systematic computer-based communication training, and thus users can work at their own pace and level of understanding by selecting modality, contextual information, and difficulty level. To

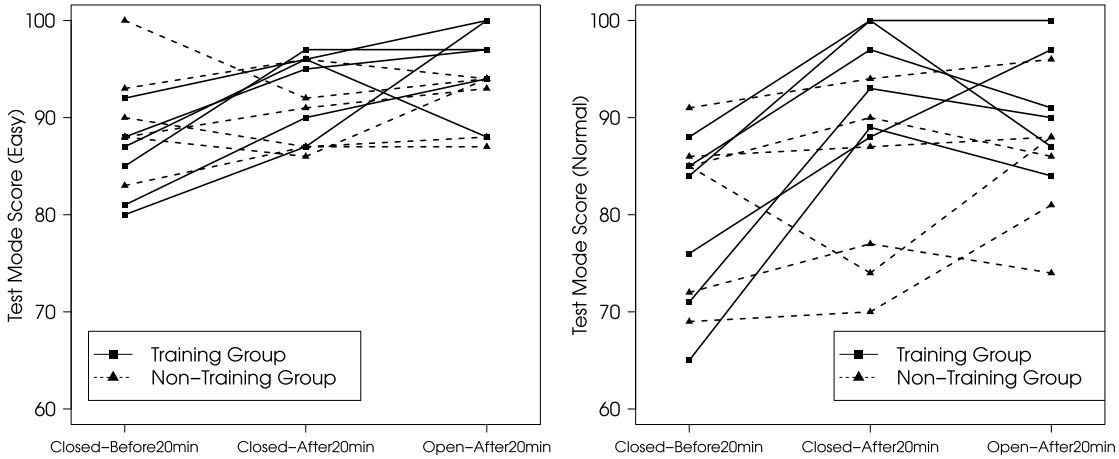

Fig. 5 Test mode score before and after training. The left figure indicates difficulty level easy, and the right figure indicates difficulty level normal. Dotted lines indicate scores of the non-training group, and solid lines indicate scores of the training group. Pre and post 20 minutes (closed data) is shown as well as post 20 minutes (open data). Each line indicates a different participant.

measure effect of training, we designed a test mode including 10 questions in closed and open sets. Users can be motivated by seeing their total score and generated comments.

For evaluation of NOCOA+, we recruited a total of 48 participants, and performed a series of four experiments: 1) We analyzed contextual differences, and found that contextual information was helpful for answering questions. This result showed similar tendencies to emotion recognition in previous work [7]. 2) We found that these were differences in each modality's score in the cases of both intention and partner information. We also confirmed that the audio modality, which was used in NOCOA, allowed users to accurately recognize non-verbal behaviors. 3) We investigated the relationship between autistic traits measured by the AQ and non-verbal behavior recognition skills measured by NOCOA+. The results showed a correlation between AQ scores of the communication and socialization subcategories and non-verbal communication skills. This result showed an improvement in the correlation coefficient of NOCOA+ ($r = 0.82$) compared with NOCOA ($r = 0.71$). 4) We found that participants significantly improved in score through computer-based training in terms of closed and open question sets.

Although each experiment was performed with a limited number of participants, we found that multimodality and context information is useful to accurately recognize non-verbal behaviors, and two types of training regimen have effective to improve social skills. In social communication, skills for recognition of non-verbal behaviors are one of the important components. These results also imply that it is better to take into consideration the effect of multimodality and context information than only using a single modality in communication training.

One potential direction for the future is consideration of individual differences (e.g., the relationship between tendency of mistakes and autistic traits) as well as the relationship between autistic traits and training effect. In addition, NOCOA+ proposed two types of training methods, "listen to a large number of examples" and "check the rules." Differences in the effect of these training methods in terms of social communication training should be examined in the future. NOCOA+ has been distributed in the Apple store as an educational application (https://itunes.apple.com/us/app/nocoa+/id622502354?ls=1&mt=8).

References

[1] American Psychiatric Association, "The diagnostic and statistical manual of mental disorders 5," Washington, D.C., 2013.
[2] N. Bauminger, "The facilitation of social-emotional understanding and social interaction in high-functioning children with autism: Intervention outcomes," J. Autism. Dev. Disord., vol.32, no.4, pp.283–298, 2002.
[3] S. Baron-Cohen, Autism and Asperger syndrome, Oxford University Press, USA, 2008.
[4] S. Baron-Cohen, J. Richler, D. Bisarya, N. Gurunathan, and S. Wheelwright, "The systemizing quotient: an investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences," Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, vol.358, no.1430, pp.361–374, 2003.
[5] S. Baron-Cohen, S. Wwheelwright, R. Skinner, J. Martin, and E. Clubley, "The Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians," J. Autism. Dev. Disord., vol.31, no.1, pp.5–17, 2001.
[6] V. Bernard-Opitz, N. Sriram, and S. Nakhoda-Sapuan, "Enhancing social problem solving in children with autism and normal children through computer-assisted instruction," J. Autism. Dev. Disord., vol.31, no.4, pp.377–384, 2001.
[7] L.F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," Current Directions in Psychological Science, vol.20, no.5, pp.286–290, 2011.
[8] S. Bölte, D. Hubl, S. Feineis-Matthews, D. Prvulovic, T. Dierks, and F. Poustka, "Facial affect recognition training in autism: can we animate the fusiform gyrus?" Behav. Neurosci., vol.120, no.1, pp.211–216, 2006.
[9] J. Bishop, "The Internet for educating individuals with social impair-

ments," J. Comput. Assist. Lear., vol.19, no.4, pp.546–556, 2003.
[10] S. Blte, S. Feineis-Matthews, S. Leber, T. Dierks, D. Hubl, and F. Poustka, "The development and evaluation of a computer-based program to test and to teach the recognition of facial affect," International Journal of Circumpolar Health, vol.61, pp.61–68, 2002.
[11] P. Brown, J. Bovey, and X. Chen, "Context-Aware Applications: From the Laboratory to the Marketplace," IEEE Pers. Commun., vol.4, no.5, pp.58–64, 1997.
[12] M.H. Davis, "Measuring individual differences in empathy: Evidence for a multidimensional approach," J. Pers. Soc. Psychol., vol.44, no.1, pp.113–126, 1983.
[13] V. Dekker, M.H. Nauta, E.J. Mulder, M.E. Timmerman, and A. de Bildt, "A randomized controlled study of a social skills training for preadolescent children with autism spectrum disorders: generalization of skills by training parents and teachers?," BMC psychiatry, vol.14, no.1, 2014.
[14] A. Dey, "Context-Aware Computing: The Cyber Desk Project," AAAI Spring Symposium on Intelligent Environments, Technical Report, pp.51–54, 1998.
[15] El Kailouby, P. Robinson, and S. Keates, "Temporal context and the recognition of emotion from facial expression," Proc. HCI International Conference, 2003.
[16] O. Golan, and S. Baron-Cohen, "Systemizing empathy: Teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia," Development and Psychopathology, vol.18, no.2, pp.591–617, 2006.
[17] O. Golan, S. Baron-Cohen, and Y. Golan, "The 'Reading the Mind in Films' task [child version]: Complex emotion and mental state recognition in children with and without autism spectrum conditions," J. Autism. Dev. Disord., vol.38, no.8, pp.1534–1541, 2008.
[18] O. Golan, S. Baron-Cohen, J.J. Hill, and M.D. Rutherford, "The 'reading the mind in the voice' test-revised: A study of complex emotion recognition in adults with and without autism spectrum conditions," Journal of Autism and Developmental Disorders, vol.37, no.6, pp.1096–1106, 2007.
[19] O. Golan, P. LaCava, and S. Baron-Cohen, "Assistive technology as an aid in reducing social impairments in autism," Growing Up with Autism: Working with School-Age Children and Adolescents, pp.124–142, 2007.
[20] D. Goleman, Social intelligence. Arrow Books, 2007.
[21] L. Kanner, "Autistic disturbances of affective contact," Nervous Child, vol.2, pp.217–250, 1943.
[22] D. Moore, P. McGrath, and J. Thorpe, "Computer-aided learning for people with autism – a framework for research and development," Innovations in Education & Training International, vol.37, no.3, pp.218–228, 2000.
[23] D. Olive, "Multiple Comparisons Among Means," Journal of the American Statistical Association, 1961.
[24] N. Ryan, J. Pascoe, and D. Morse, Enhanced reality fieldwork: the context-aware archaeological assistant. British Archaeological Reports, Oxford, 1998.
[25] M. Silver, and P. Oakes, "Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others," Autism, vol.5, no.3, pp.299–316, 2001.
[26] H. Tanaka, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Modality and contextual differences in computer based non-verbal communication training," Proc. 4th IEEE CogInfoCom, pp.127–132, 2013.
[27] H. Tanaka, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "NOCOA: A computer-based training tool for social and communication skills that exploits non-verbal behaviors," The Journal of Information and Systems in Education, vol.12, no.1, pp.19–26, 2013.
[28] J.W. Tanaka, J.M. Wolf, C. Klaiman, K. Koenig, J. Cockburn, L. Herlihy, C. Brown, S. Stahl, M.D. Kaiser, and R.T. Schultz, "Using computerized games to teach face recognition skills to children with autism spectrum disorder: the lets face it! program," Journal of Child Psychology and Psychiatry, vol.51, no.8, pp.944–952, 2010.
[29] A. Vinciarelli, M. Pantic, and H. Bourland, "Social signal processing: Survey of an emerging domain," Image and Vision Computing, vol.27, no.12, pp.1743–1759, 2009.
[30] A. Wakabayashi, S. Baron-Cohen, S. Wheelwright, and Y. Tojo, "The Autism-Spectrum Quotient (AQ) in Japan: a cross-cultural comparison," J. Autism. Dev. Disord., vol.36, no.2, pp.263–270, 2006.
[31] L. Wing, "Autistic spectrum disorders," British Medical Journal, vol.312, no.7027, pp.327–328, 1996.

Hiroki Tanaka received his B.E. from Asahikawa National College of Technology in 2010, and his M.E from Nara Institute of Science and Technology in 2012. He is currently Ph.D student at Nara Institute of Science and Technology, and researcher at the Research Center for Special Needs Education, Nara University of Education. His research interests in-clude education systems for non-verbal communication, and automatic measurement of communication skill.

Sakriani Sakti received her B.E degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received "DAAD-Siemens Program Asia 21st Century" Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR and U-STAR (2006-2011). She also served as a visiting professor of Computer Science Department, University of Indonesia (UI) in 2009-2011. Currently, she is an assistant professor of the Augmented Human Communication Lab, NAIST, Japan. She is a member of JNS, SFN, ASJ, ISCA, IEICE amd IEEE. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

Graham Neubig received his B.E. from University of Illinois, Urbana-Champaign, U.S.A, in 2005, and his M.E. and Ph.D. in informatics from Kyoto University, Kyoto, Japan in 2010 and 2012 respectively. He is currently an assistant professor at the Nara Institute of Science an Technology, Nara, Japan. His research interests include speech and natural language processing, with a focus on machine learning approaches for applications such as machine translation, speech recognition, and spoken dialog.

Tomoki Toda earned his B.E. degree from Nagoya University, Aichi, Japan, in 1999 and his M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively. He was a Research Fellow of JSPS in the Graduate School of Engineering, Nagoya Institute of Technology, Aichi, Japan, from 2003 to 2005. He was an Assistant Professor of the Graduate School of Information Science, NAIST from 2005 to 2011, where he is currently an Associate Professor. He has also been a Visting Researcher at the NICT, Kyoto, Japan, since May 2006. From March 2001 to March 2003, he was an Intern Researcher at the ATR Spoken Language Communication Research Laboratories, Kyoto, Japan, and then he was a Visiting Researcher at the ATR until March 2006. He was also a Visiting Researcher at the Language Technologies Institute, CMU, Pittsburgh, USA, from October 2003 to September 2004 and at the Department of Engineering, University of Cambridge, Cambridge, UK, from March to August 2008. His research interests include statistical approaches to speech processing such as voice transformation, speech synthesis, speech analysis, speech production, and speech recognition. He received the 18th TELECOM System Technology Award for Students and the 23rd TELECOM System Technology Award from the TAF, the 2007 ISS Best Paper Award and the 2010 ISS Young Researcher's Award in Speech Field from the IEICE, the 10th Ericsson Young Scientist Award from Nippon Ericsson K.K., the 4th Itakura Prize Innovative Young Researcher Award and the 26th Awaya Prize Young Researcher Award from the ASJ, the 2009 Young Author Best Paper Award from the IEEE SPS, the Best Paper Award (Short Paper in Regular Session Category) from APSIPA ASC 2012, the 2012 Kiyasu Special Industrial Achievement Award from the IPSJ, and the 2013 Best Paper Award (Speech Communication Journal) from EURASIP-ISCA. He was a member of the Speech and Language Technical Committee of the IEEE SPS from 2007 to 2009. He is a member of IEEE, ISCA, IEICE, IPSJ, and ASJ.

Satoshi Nakamura is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He also serves as a visiting professor of Collaborative Research Unit, National Institute of Informatics. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He was a project leader of the world first network-based commercial speech-to-speech translation service for 3-G mobile phones in 2007 and VoiceTra project for iPhone in 2010. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampoli Award 2012. He organized the International Workshop of Spoken Language Translation (IWSLT 2006) and Oriental Cocosda 2008 as a general chair. He also served as the program chair of INTERSPEECH 2010. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011 and IEEE Signal Processing Magazine Editorial Board Member since April 2012.