Modality and Contextual Differences in Computer Based Non-verbal Communication Training

Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura
Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, Nara, Japan.
{hiroki-tan, ssakti, neubig, tomoki, s-nakamura}@is.naist.jp

***Abstract*—The number of people who have trouble with social skills and communication is now greater than ever for a variety of reasons. An extreme case of these traits is characteristics of autism spectrum disorders. Our objective is to enable people with social and communication difficulties to improve their social and communication skills. Previous reports have shown modality and contextual information is useful for emotion recognition. This paper examines modality and contextual differences in the context of the existing social skills training aid NOCOA+. In NOCOA+, there are four types of modalities: audiovisual, visual, audio, and verbal. It also has three types of contextual information: no context, prior 5 seconds context, and prior 10 seconds context. We performed two experiments that examined the effect of modality and context on social skills training tools. The results elucidated characteristics of non-verbal behaviour incorporating each sensory modality, and also found that contextual information is helpful to infer non-verbal behaviour.**

Keywords*—*non-verbal behaviour, autism spectrum disorders, contextual information, modality

I. Introduction

Socialization and communication are important factors influencing human social life. People who have trouble with social skills and communication have recently been increasing due to environmental and/or inherent reasons [1]. Some papers mention the extreme case of these traits is autism spectrum disorders (ASD) [2], a set of developmental disorders characterized by social interaction and communication difficulties, as well as unusually narrow, repetitive interests [3]. Given the impact of these problems on every-day life, there has been considerable interest in tools to both identify the degree of these difficulties and allow for training to improve social and communication skills.

One of the central psychological themes in autism is empathizing. Empathizing is a set of cognitive and affective skills we use to make sense of and navigate the social world. It is well established that empathy, particularly as manifested in emotion recognition and mental state recognition, is a core difficulty in people with ASD. One of the goals of social skills training (SST) is to train empathizing ability [4]. There have also been a number of studies on tools to automate the training and testing of empathizing ability. In particular, we have proposed a tool NOCOA+, which is an application to train and test empathizing by using non-verbal information [5]. In this paper, we focus on two factors influencing this sort of SST, specifically modality and context.

With regards to modality, previous research has found that communication difficulties span different sensory modalities, both visual and auditory. Some reports mention that both visual and audio information is important to recognize basic and complex emotion [6], [7]. However these papers did not compare modalities from same recorded video, and the effect of modality on recognition of other non-verbal behaviour such as intention and/or partner information has not been reported.

With regards to context, while our previous work utilized short term non-verbal signals (e.g. less than 5 seconds), other reports have mentioned that context influences emotion recognition [8]. While most people tacitly understand what context is, they find it hard to clearly define. Brown *et al.* [9] define context as location, identities of the people around the user, the time of day, season, temperature, etc. Ryan *et al.* [10] define context as the user's location, environment, identity and time. Dey [11] enumerates context as the user's emotional state, focus of attention, location and orientation, date and time, objects, and people in the user's environment. In these definitions, the common contextual factor is time. There is also a report mentioning that time information is important to identify human temporal emotion [12].

This paper is relevant to the field of cognitive infocommunications (CogInfoCom [13]) because it merges human cognitive capabilities with infocommunications. Moreover, this paper is related to the field of augmented cognition which aims to extend the user's abilities via computational technologies, while taking into account cognitive aspects such as limitations in attention, memory and learning capabilities [14]. A framework called NOCOA+ targets the enhancement of social skills in e.g., patients suffering from ASD.

In this paper, we first describe the NOCOA+ framework for social skills training. Then, we describe the collection and incorporation of data from several sensory modalities as well as data considering context. We also report two experiments examining the effect of modality and contextual differences on computer-based non-verbal communication training.

II. Computer-Based Non-verbal Communication Training

We have proposed an iPad application NOCOA+ (NOnverbal COmmunication for Autism plus), which is a communication aid to help measure and train social skills [5]. In this section, we describe the application and mention several principles that contributed to its design.

A. Social Skills Training

SST is a form of behaviour therapy widely used by teachers, therapists, and trainers to help persons who have difficulties relating to other people, such as ASD. For example, Bauminger's intervention [4] focused on teaching interpersonal problem solving, affective knowledge, and social interaction. Results show participants were more likely to initiate positive social interaction with peers after treatment; in particular, they improved eye contact and their ability to share experiences with and show interest in peers. In addition, after treatment, participants can provide more examples of complex emotions.

B. Social Skills and Communication Evaluation

While SST generally covers a wide variety of social situations, we argue that it is also useful to identify important social skills and develop focused training regimens for these skills. In order to identify particularly important social skills which should be measured and trained for people with ASD, we first focused on one widely used method for pre-screening for autism and other social difficulties: the autism spectrum quotient (AQ) test [15]. Autism is a spectrum condition that has a broad range of clinical characteristics ranging from mild to severe. There are several methods such as the AQ for measuring a person's position on the autistic spectrum in both people with and without autism. The AQ test is made up of 10 questions assessing 5 different areas: social skill, attention switching, attention to detail, communication, and imagination.

Independently of the AQ, the diagnosis criteria of autism includes "marked deficits in nonverbal and verbal communication used for social interaction [16]." In our previous work [17], we performed experiments to confirm the important non-verbal factors contributing to communication skills, and their relationship with AQ. To do so, we used factor analysis, which is commonly used to elucidate the factors contributing to scores on a psychometric test. To collect data, we first asked 21 Japanese students to take the English version of the AQ to measure two of the original five areas: social and communication skills (with a total of 20 statements). Next, we performed a factor analysis using individual responses to each question on the AQ questionnaire to determine several important factors for social and communication skills. We found 5 factors:

1) intention & interest.
2) politeness/impoliteness & new friends.
3) social places & situations.
4) chit-chat & feelings.
5) other factors.

Finally, we selected the first two factors (intention & interest, and politeness/impoliteness & new friends) as the non-verbal behaviour to be trained and tested by NOCOA+. In the description below, we abbreviate these as *intention* and *partner information*.

Vinciarelli *et al.* [21] define the terms non-verbal signal and non-verbal behaviour. Non-verbal behaviours are unobservable and long-lasting (e.g. emotion, rapport, and personality), while non-verbal signals are observable and short-lasting (e.g. facial expressions, prosody, and posture). In this study, intention and partner information are defined as non-verbal behaviour, and non-verbal signal is used to infer the non-verbal behaviour.

C. NOCOA+ design

Based on the important non-verbal behaviour identified in the previous section, the next step is to incorporate this into SST. To do so, we adopted a quiz format, where the user of NOCOA+ must choose from several categories of intention and partner information. Specifically, the user listens to an utterance, and has to guess the category for both intention and partner information. Data used in NOCOA+ was recorded for 4 participants and utterances were detected by speech features. Detected utterances automatically provide audio, visual, and audiovisual data. We also created segments including context information, the 5 seconds and 10 seconds prior to the actual utterance. NOCOA+ has two modes, training mode and test mode.

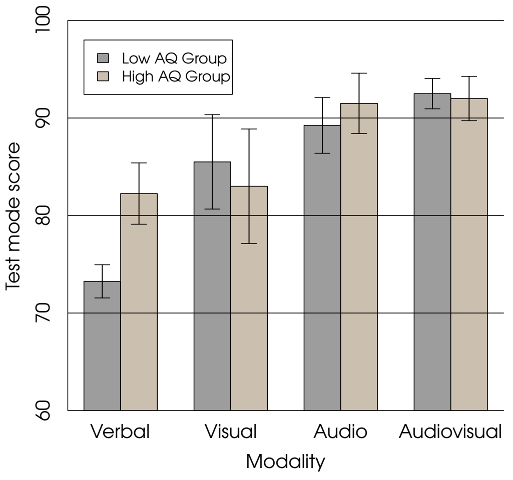

Fig. 1. Modality differences for the low and high AQ groups with standard error bars.

Training mode is designed to enhance users' socialization and communication skills. The training mode provides two types of training, "listen to a large number of examples" which means the user listens to utterances as many as possible and "check the rules" which means the user sees a description about how to solve the question. The former is developed to enable user to learn using a statistical-based training regimen, and the latter is developed to allow user to learn systematic rules. The user can select the preferred mode from the training menu.

In the test mode, 10 questions are provided, and the user's non-verbal communication skills are measured according to the number of questions answered correctly. The maximum score of test mode is 100, and random selection would result in a score of 51. It has 2 types of generalization levels: closed, where testing is performed using data that were included in the training mode, and open, where data is not included in the training mode. The test mode has three types of difficulty level according to the accuracy rate achieved by 10 subjects. Accuracy rate of each difficulty level is as follows; easy: 81-100%, normal: 51-80%, hard: 0-50%. We also prepare four types of modalities, audiovisual, visual, audio, and verbal (where the first author of this paper transcribed the speech in the audiovisual data and read it in a flat tone without emotion). The audiovisual data and contextual data were labeled by 3 annotators. A total of 109 utterances were chosen for which all 3 annotators agreed for use in NOCOA+.

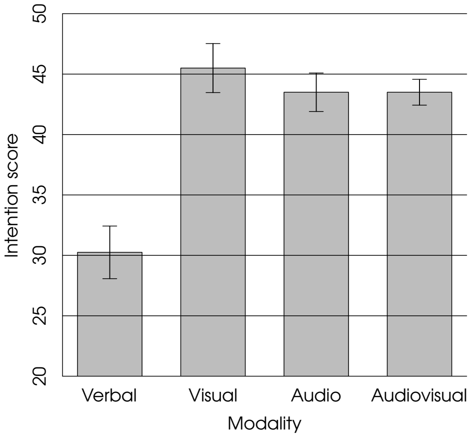

Fig. 2. Modality differences in terms of intention score with standard error bars.

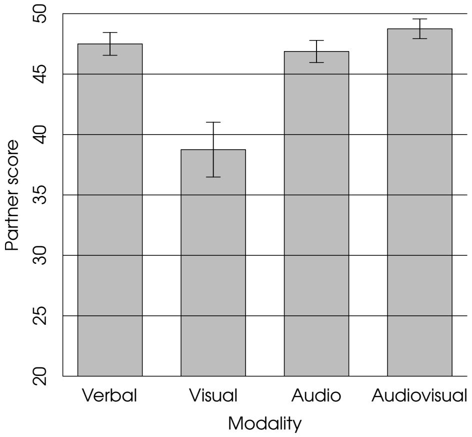

Fig. 3. Modality differences in terms of partner information score with standard error bars.

III. Experiment1: Modality Differences

In previous reports, there has been no examination on the effect of modality differences on the identification of non-verbal behaviour such as intention and partner information. In this section, we examine the effect of modality differences on the ability of people with relatively low and high AQ to identify non-verbal information. We use NOCOA+, which includes several modalities, in this experiment.

A. Effect of AQ

1) Method: We set two hypotheses. First, reaction time, latency between the beginning of the stimulus and the user's decision, is related to AQ values because this type of task is conducted by intuition [18]. Second, the effect of modality on the ability to identify non-verbal information is dependent on the AQ value. To verify the above hypotheses we perform experiments using the testing mode of NOCOA+ and use the t-test to measure statistical significance. A total of eight participants are recruited for the experiment. Four participants are selected as the "low AQ group" and the other four participants are selected as the "high AQ group." The low AQ group had a total score ranging from 1 to 9 with average score of 3.75 on the AQ subareas social skills and communication (out of 20). The high AQ group ranged from 12 to 18 with an average score of 14.75.

We had both groups take the test mode. They answer ten questions randomly selected from the easy difficulty level, which include four modalities, audiovisual, audio, visual, and verbal. The closed data was used, and scores were averaged.

2) Result: With respect to our first hypothesis, the averaged reaction time of the low AQ group was 12.15 (sd. 4.13) and the averaged reaction time of the high AQ group was 11.17 (sd. 4.76). Thus, we can see that the reaction time between the two groups was not significant. Next, Figure 1 shows the test mode score of each modality. The result shows that there is no tendency difference between low and high AQ group. Scores of audio, visual, and verbal decrease compared to audiovisual in both groups. In terms of visual, there is a lot of variation in both groups, demonstrating that scores from visual have individual differences or related to other factors. In the verbal setting, it is significantly different in comparison with audiovisual. However, in terms of the verbal, the low AQ group scored relatively low compared to the high AQ group. It possibly indicates that the low AQ group have difficulty in inferring non-verbal behaviour by using only verbal information, and more reliant on audio or visual cues than the high AQ group.

B. Characteristics of intention and partner information

1) Method: Based on the fact that we found no tendency differences related to AQ value in the previous subsection, we set a hypothesis that characteristics of intention and partner information are different. To verify the above hypothesis we analyzed the score for intention and partner information separately for the eight members of the both groups and used one-way ANOVA (analysis of variance) to measure statistical significance.

2) Result: Figure 2 and 3 indicate that there are significantly differences in each modality's score in terms of both intention and partner information. In the case of visual, a relatively large number of errors are found in the partner information category, and in the case of verbal, a large number of errors are found in the intention category. The ANOVA shows F(3,28)=10,975, p<0.001 for intention score and F(3,28)=15.745, p<0.001 for partner information score respectively which shows the above differences are significant. In case of intention, visual score is even slightly higher than audio and audiovisual. This is in line with previous work that has shown that visual information is more effective for identifying emotions [19].

IV. Experiment2: Contextual Differences

While it has been reported that contextual information is important to identify human temporal emotion [12], this result has not been adapted to other non-verbal behaviours. In this section, we clarify the benefit of contextual information in

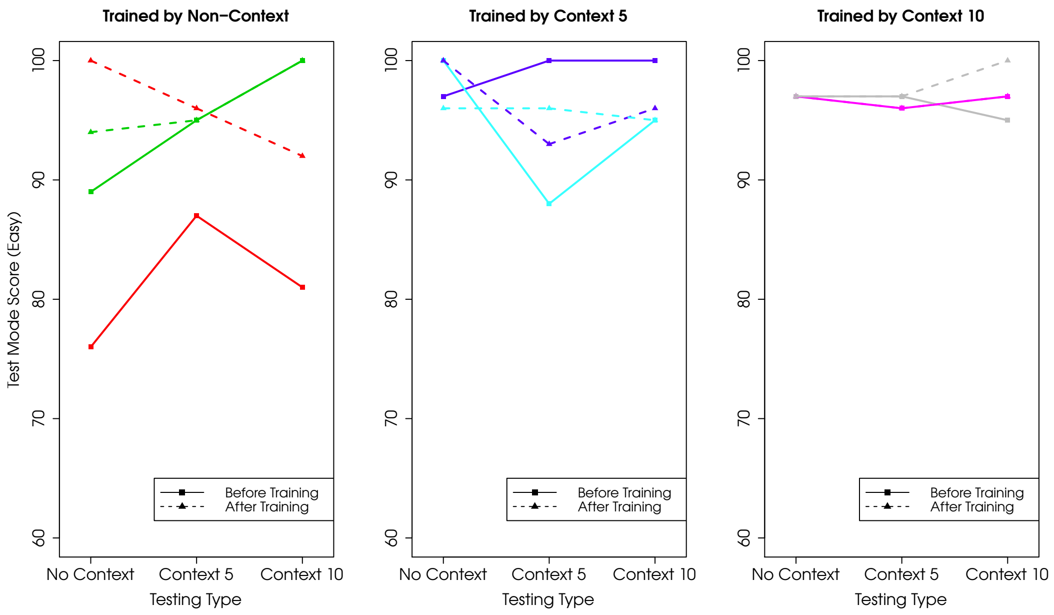

Fig. 4. Pre-training and post-training scores. The solid line indicates pre-training scores and the dotted line indicates post-training scores. Each color indicates a different participant.

TABLE I. UTTERANCES WITH PERCENTAGE OF ERROR RATE LESS THAN 20%.

	Percentage of error rate less than 20%
No Context	43%
Context 5	63%
Context 10	70%

the form of the content directly proceeding the utterance. We use a version of the NOCOA+ test mode that includes several contextual levels in this experiment.

A. Method

We set three types of contextual level, no context, 5 seconds context, and 10 seconds context. We hypothesized the following: First, contextual information can help the subjects answer questions. Second, people trained with no context improve their score in no context, and those trained with other contexts improve in the respective context level. Third, people trained by any contextual information improve their score in no context questions.

We collected data corresponding to each level of contextual information similar to [5]. Three types of difficulty level were set; easy, normal, and hard according to the criterion mentioned in section II-C. To categorize difficulty levels, 10 members of the general population answered all questions two times using full contextual information. Next, we performed experiments using six members of the general population. The procedure is as follows: 1. Divide participants into 3 groups of contextual levels according to their AQ value (e.g. one relatively high AQ person for each contextual level). 2. Have participants practice how to use NOCOA+. 3. Test open generalization level and easy difficulty level for each contextual level (before training). 4. Have each group perform training for the specified contextual level for both types of training "listen to a large number of examples" and "check the rules." 5. Have the participant take a break for 30 minutes. 6. Repeat procedure number 3 (after training).

B. Result

In Table I, we show the percentage of utterances with an error rate less than 20%. From the results, we can see that this value is correlated to the contextual level, which means contextual information helps people to infer the correct answer.

The initial score is not significantly different. Each averaged score is as follows: no context; 92.7, context 5; 93.8, context 10; 94.7 respectively. However, one participant trained with no context scored relatively low, because he/she has a high AQ subarea value (score: 18) compared to the other participants, and AQ value is related to the initial test mode score [5]. Figure 4 shows the scores between pre-training and post-training. We can see that five of the six participants improve their score while one participant decreasing slightly. The increases are significant according to the t-test, t=-1.72, p<0.1. However, the third hypothesis that training with contextual information is effective for other no context questions is not found to be the case in this experiment.

V. CONCLUSION

In this paper, we first described the NOCOA+ framework for social skills training. Then, we described the collection and incorporation of data from several sensory modalities as well as data considering context. We reported two experiments identifying the effect of modality (experiment 1) and contextual differences (experiment 2) on social skills training.

In experiment 1, we found reaction time and modality tendency is not related to AQ value, and confirm that there are differences in each modality's score in the cases of both intention and partner information. The AQ value is relatively related to the score of the verbal setting. The results show that people (especially the low AQ persons) have difficulty correctly inferring the other's intention by the content of speech, and people have difficulty correctly inferring the other's partner information by the only visual signals. People have insufficient opportunities for social communication by without audio signals, and it implies difficulties inferring the other's partner information by the only visual cues. We may consider visual modality plus some forms of augmentation such as a device for recognizing Non-Audible Murmur [20] to understand the partner information as well as the content of speech.

In experiment 2, we found contextual information is a helpful for answering questions, and training with no context and contextual information is effective particularly for matched testing conditions. However, training with contextual information is not effective for non-context level, which we hypothesize is because non-verbal signals typically last for a short time [21]. Further investigation is needed to clarify this question. For example, while we focus on 5 seconds context and 10 seconds context, the effect of other types of contextual information need to be elucidated.

Acknowledgment

The author thanks Dr. Hidemi Iwasaka of Nara University of Education, for his kind advice and support.

References

[1] GOLEMAN D., 2007. Social intelligence. Arrow Books.

[2] BARON-COHEN S., RICHLER J., BISARYA D., GURUNATHAN N., & WHEELWRIGHT S., 2003. 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 358: 361-374.

[3] KANNER L., 1943. Autistic disturbances of affective contact. Nervous Child 2: 217-250.

[4] BAUMINGER N., 2002. The facilitation of social-emotional understanding and social interaction in high-functioning children with autism: Intervention outcomes. Journal of autism and developmental disorders 32: 283-298.

[5] TANAKA H., SAKTI S., NEUBIG G., TODA T., & NAKAMURA S., 2013. Non-verbal Communication Training with an Interactive Multimedia Application, In Proc. The 5th Asian Conference on Education, Oct. 2013. (to appear)

[6] GOLAN O., BARON-COHEN S., HILL J. J., & RUTHERFORD M. D., 2007. 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 37: 1096-1106.

[7] GOLAN O., BARON-COHEN S., & GOLAN Y., 2008. The 'Reading the Mind in Films' task [child version]: Complex emotion and mental state recognition in children with and without autism spectrum conditions. Journal of Autism and Developmental Disorders 38: 1534-1541.

[8] BARETTT L.F., MESQUITA B., & GENDRON M., 2011. Context in emotion perception, Current Directions in Psychological Science, pp. 286-290.

[9] BROWN P.J., BOVEY J.D., & CHEN, X., 1997. Context-Aware Applications: From the Laboratory to the Marketplace. IEEE Personal Communications, 4(5), pp. 58-64.

[10] RYAN N., PASCOE J., & MORSE, D., 1997. Enhanced Reality Fieldwork: the Context-Aware Archaeological Assistant, Computer Applications in Archaeology.

[11] DEY A.K., 1998 Context-Aware Computing: The CyberDesk Project. AAAI Spring Symposium on Intelligent Environments, Technical Report SS-98-02, 51-54

[12] El KAILOUBY R., ROBINSON P., & KEATES, S., 2003. Temporal context and the recognition of emotion from facial expression, In Proc. HCI International Conference.

[13] BARANYI P., & CSAPO A., 2012. Definition and synergies of cognitive infocommunications. Acta Polytechnica Hungarica 9: 67-83.

[14] SCHMORROW D., 2005. Foundations of Augmented Cognition. Lawrence Erlbaum Associates.

[15] BARON-COHEN S., WHEELWRIGHT S., SKINNER R., MARTIN J., & CLUBLEY E., 2001. The Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. Journal of Autism and Developmental Disorders 31: 5-17.

[16] AMERICAN PSYCHIATRIC ASSOCIATION, 2013. The diagnostic and statistical manual of mental disorders, 5. Washington, D.C.: American Psychiatric Association.

[17] TANAKA H., SAKTI S., NEUBIG G., TODA T., CAMPBELL N., & NAKAMURA S., 2012. Non-verbal cognitive skills and autistic conditions: An analysis and training tool, In Proc IEEE CogInfoCom.

[18] GLADWELL M., 2005. Blink: The power of thinking without thinking, Little, Brown.

[19] AMANO R., OHTA A., IIDA A., 2007. Building a web based communication aid for autism person [in Japanese], IEICE, pp89-94.

[20] MORIGUCHI T., TODA T., SANO M., SATO H., NEUBIG G., SAKTI S., & NAKAMURA S.,2013. A Digital Signal Processor Implementation of Silent/Electrolaryngeal Speech Enhancement based on Real-Time Statistical Voice Conversion. Proc. of INTERSPEECH, pp.3072-3076.

[21] VINCIARELLI A., PANTIC M., & BOURLAND, H., 2009. Social signal processing: Survey of an emerging domain, Image and Vision Computing, 27(12), pp. 1743-1759.