NAIST-IS-DD1261007

Doctoral Dissertation

Automated Social Skills Training through Affective Computing

Hiroki Tanaka

February 4, 2015

Department of Media Informatics
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Hiroki Tanaka

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Yuji Matsumoto	(Co-supervisor)
Associate Professor Tomoki Toda	(Co-supervisor)
Assistant Professor Sakriani Sakti	(Co-supervisor)
Assistant Professor Graham Neubig	(Co-supervisor)
Professor Hidemi Iwasaka	(Nara University of Education)

Automated Social Skills Training through Affective Computing*

Hiroki Tanaka

Abstract

Social communication skills are important factors influencing human life, and the number of people who have trouble with these skills have recently been increasing for a variety of reasons. In this thesis, a computer-based training system to enhance human social communication skills is proposed. Computers have several advantages as tools for social skills training in that computerized environments are predictable, consistent, and free from social demands. One of the central psychological themes in communication difficulties is empathizing, which is a set of cognitive and affective components. In this thesis, several computer-based training methods to train both cognitive and affective skills are proposed.

For the cognitive component, I developed mobile applications ("NOCOA" and "NOCOA+") that use multiple modalities to help users recognize non-verbal behaviors. I confirmed the effectiveness of a method for predicting autistic traits by using these systems, examined the effect of modality differences, and evaluated the effectiveness of computer-based intervention.

For the affective component, first I automatically classified the types of laughter which are difficult to identify for people with communication difficulties. Then, I compared narrative stories of children with autism spectrum disorders to those with typical development. I found group differences in speech and language features. For classification, using linguistic cues and prosody, I analyzed the important feature sets and their effects on accuracy of identifying children with autism. The results suggest that the proposed method can effectively distinguish children with autism spectrum disorders from those with typical development.

* Doctoral Dissertation, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1261007, February 4, 2015.

Finally, I attempt to automate the process of social skills training by a dialogue system named "Automated social skills trainer," which provides the social skills training through human-computer interaction. The system includes a virtual avatar that recognizes user speech and language information and gives feedback to users to improve their social skills. Its design is based on conventional group or individual social skills training performed by human participants including defining target skills, modeling, role-play, feedback, reinforcement, and homework.

Keywords:

Autism spectrum disorders, computer-based training, affective computing, nonverbal behaviors, narrative

感情コンピューティングを用いた ソーシャルスキルトレーニングの自動化*

田中 宏季

内容梗概

社会的コミュニケーションのスキルは，人間の生活を営んでいくうえで大変重要である．様々な要因により，近年社会的コミュニケーションに困難がある人々が増えていると報告されている．本論文では，コンピュータを用いたコミュニケーション能力の向上手法を提案する．コンピュータを用いたトレーニングの有利な点として，社会とは無関係，一貫性がある，予測可能，ということが挙げられる．また心理学の知見から，社会的コミュニケーションのスキルは，共感する能力に起因すると言われている．共感とは，認知と表出の 2 つの要素に分解することができる．本論文では，認知と表出のスキルをトレーニングするための幾つかのアプリケーションを提案する．

まず認知に関しては，モバイルアプリケーションとして，NOCOA と NOCOA+ を開発した．これは複数のモダリティにより，ユーザが非言語性言動を認識するのを助けるアプリケーションである．これらのアプリケーションを使用することにより，社会的コミュニケーションスキルの自動予測，モダリティの違いの調査，コンピュータを使用したトレーニングの評価を実施した．

次に表出に関しては，コミュニケーション障害のある人が笑い声の種類分けを手動で行うことが難しいことから，人間の笑い声の自動識別に取り組んだ．次に，自閉スペクトラム症児と定型発達児のナラティブ発話の比較を実施した．これにより言語と音声において，自閉スペクトラム症児と定型発達児での差を確認した．また，これらの特徴量から両者の識別問題に取り組んだ．

最後に，人間が行うソーシャルスキルトレーニングを模倣したアプリケーションである自動ソーシャルスキルトレーナを開発した．本システムは，アバターと

* 奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 博士論文, NAIST-IS-DD1261007, 2015 年 2 月 4 日.

のインタラクションによって実施され，アバターはユーザの音声および言語情報を認識し，ユーザがソーシャルスキルを改善するためにフィードバックを行う．自動ソーシャルスキルトレーナの設計は従来の個別ソーシャルスキルトレーニングに基づいており，課題設定，モデリング，ロールプレイ，フィードバック，正の強化，宿題を含んでいる．

キーワード

自閉スペクトラム症, コンピュータベーストレーニング, 感情コンピューティング, 非言語性言動，ナラティブ

Acknowledgements

First, I would like to acknowledge Professor Satoshi Nakamura, who has always been full of inspiration and support. The problem that I attempted to solve in this thesis was considered undeveloped, and I faced a number of problems. Professor Nakamura still felt that it was an important problem, and encouraged me to help people with communication difficulties. Besides being a great mentor, Professor Nakamura cared for my well-being.

I would like to express my appreciation to my external committee Professor Hidemi Iwasaka. He believed in me and wanted me to succeed as a scientist and teacher.

I would like to acknowledge my thesis adviser Professor Yuji Matsumoto for his invaluable comments to my work.

I would also like to express my appreciation to Professor Tomoki Toda. He gave me helpful comments from the viewpoint of supporting handicapped persons. He also taught me about logical thinking and presentation skills.

I would like to acknowledge Professor Sakriani Sakti for her constant guidance and consideration through my doctoral course. Specifically, she made the cognitive communication group meeting of Augmented Human Communication Laboratory effective.

I also express my deepest appreciation to Professor Graham Neubig for his support and encouragement. He always encouraged me to elaborate my doctoral work. He also showed me a good model as scientist and engineer.

Professor Shrikanth Narayanan and Theodora Chaspari of SAIL, USC guided my work during my internship program. Their comments have always been a big source of inspiration.

Professor Nick Campbell of NAIST and TCD gave me advices and interesting talks regarding social interaction, and he also patiently taught me how to study during my master's course.

I was very glad to have the opportunity to communicate with Kazuyo Iida of Nara Autism Society. I thank her generous support of my work and for giving me opportunity to involve children and parents in this research. It was also a great opportunity to join in a summer program of children with ASD.

I also express my deepest appreciation to all children, parents and participants

in this study. They made me realize the most important thing in my work.

I thank to all members of Augmented Communication Laboratory for fruitful discussion and playing futsal together. The time spent with all of you will always remain very memorable to me.

I would like to thank my mother, father and brother for their supports. Communication with them always remind me an important thing in life.

I also would like to thank my friends of Lighthouse Christian Church for their unconditional love. I feel blessed to have your love. Let me end with humble acknowledgment of God, the Almighty, for His encouragement throughout my research work.

Contents

List of Figures

List of Tables

1. Introduction

People with social and communication difficulties have recently been increasing due to various reasons [1]. Many people have difficulties or are anxious in social interactions such as relating to friends, presentations and job interviews. Persistent social skill deficits impede those afflicted with them from forming relationships or succeeding in social situations. The extreme example of people with these difficulties are those with autism spectrum disorders (ASD). ASD is a set of neuro developmental conditions characterized by social interaction and communication difficulties, as well as unusually narrow and repetitive interests [2]. People with ASD also have recently been increasing due to genetic and/or environmental factors [3]. Technology for both identifying the degree of these difficulties and developing a learning tool for social and communication skills could help people overcome these barriers.

One of the psychological themes in social skills is empathizing. Empathizing is the capacity to attribute mental states, such as feelings, thoughts, and intentions to other people, and to respond to their mental states with an appropriate emotion. Empathizing is a set of cognitive (recognizing) and affective (emotional expressiveness) components we use to make sense of and navigate the social world [4]. The cognitive component of empathy is also referred to as theory of mind [5]. It is well established that emotion recognition and mental state recognition are core difficulties in people with ASD [1, 6]. Neuroimaging studies of emotion recognition from faces also reveal that people with ASD show less activation compared to those with typical development (TD), who does not have ASD, in brain regions central to face processing, such as the fusiform gyrus [7]. There is also evidence of reduced activation in brain areas that play a major role in emotion recognition, such as the amygdala, when individuals with ASD process socioemotional information [7, 8]. The affective component of empathy, emotional expressiveness, is important in situations such as interviews and presentations [9, 10]. However, there are fewer established methods to evaluate this skills. Teaching empathizing is important treatment of social and emotional training [11]. However, these training programs typically do not focus specifically on systematically teaching.

In contrast to empathy difficulties, individuals with communication difficulties show good and sometimes superior skills in "systemizing" [12]. Systemizing is the

drive to analyze or build systems, to understand and predict the behavior of events in terms of underlying rules and regularities. It may be easier for those with social impairments to use computers than interact directly. Donna Williams, who has autism, wrote a book entitled "Nobody nowhere" [13], in which she stated

> *"The comprehension of words works as a progression, depending on the amount of stress caused from fear and the stress of relating directly. At best, words are understood with meaning, as with the indirect teaching of facts by a teacher or, better still, a record, television, or book. In my first three years in the special class at primary school, the teacher often left the room and the pupils responded to the lessons broadcast through an overhead speaker. I remember responding to it without the distraction of coping with the teacher. In this sense, computers would probably be beneficial for autistic children once they had the skills to use one."*

Affective computing is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena [14] (e.g. affect recognition and expression). Several affective computing studies have been proposed. For example, Kaliouby and Robinson applied affective computing techniques to support people with ASD [15]. They developed "The emotional hearing aid" which includes automatic human mental state recognition by video input. Vision-based processing of video input of the face is combined with topdown models of actions (e.g. head up, and lip depress), displays (e.g. head nod, and lip corner pull) and mental states (e.g. agreement, and concentrating), to infer the likelihood of each of the mental states generating the observed facial behavior.

Taking into consideration the issues mentioned above, in this thesis I attempt to examine following questions.

1) Are applications that focus on the measurement and training of cognitive skills effective to quantify and improve social skills?

2) Is affective computing useful to detect social signals and atypical utterance?

3) Are applications that focus on the measurement and training of affective skills effective to quantify and improve social skills?

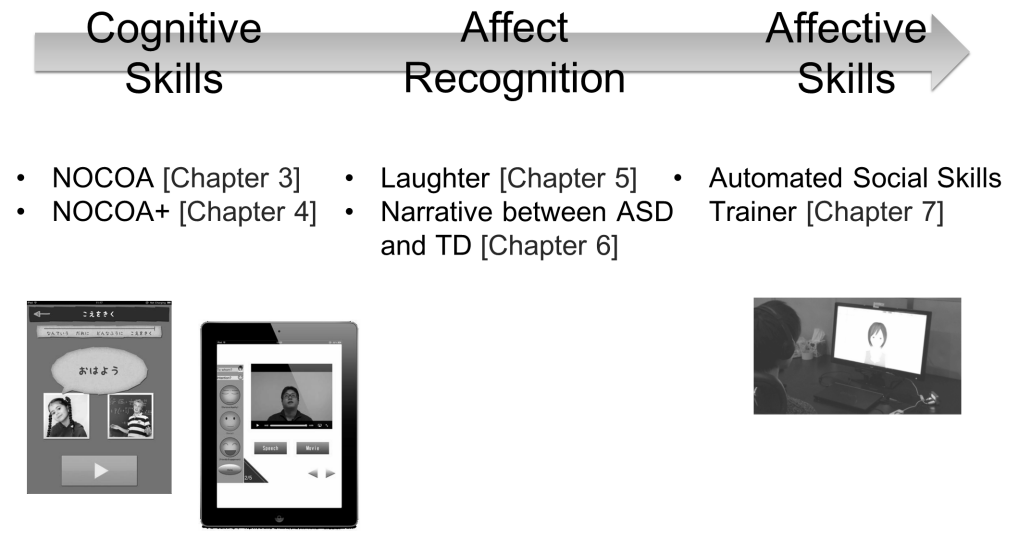

Figure 1. Relationship between the chapters.

I summarize the relationship of each chapter in Figure 1. In chapter 2, I introduce conventional methods to measure and improve social skills, and previously proposed computer-based applications. Chapter 3 proposes an application NOCOA that is aimed to train non-verbal recognition skills. Chapter 4 proposes NOCOA+ which is updated version of NOCOA, and examine modality and contextual differences. In chapter 5, I focus on automatic classification of types of laughter using affective computing. In chapter 6, I distinguish children with ASD and children with TD from their narrative speech. Finally, chapter 7 proposes automated social skills training which is aimed to enhance human social communication skills in terms of the affective component.

2. Previous Work

2.1 Introduction

In this chapter, first I introduce instruments to measure social communication skills, and conventional social skills training (SST) framework as well as its target skills. I also describe the use of computers to train social skills, which has flourished in the last decade. I summarize several applications in terms of cognitive and affective skills.

2.2 Degree of Social Communication Skills: Autistic Traits

One of the factors influencing the ability to empathize is the severity of ASD. Autism is a spectrum condition [16] that has a broad range of clinical characteristics ranging from mild to severe. There are several methods for measuring a person's position on the autistic spectrum.[1] However, most of these take a large amount of time to complete and there are few Japanese versions.

Baron-Cohen *et al.,* proposed the autism-spectrum quotient (AQ) [17], which is a self-administered screening instrument taking 5-10 minutes to complete, and that can be used for children from four years of age through to adulthood. Wakabayashi *et al.,* translated the AQ into Japanese [18]. AQ is made up of 10 questions assessing 5 different areas: social skill; attention switching; attention to detail; communication; imagination, in which one statement scores one point if the respondent records the abnormal or autistic-like behavior either mildly or strongly. Thus individuals score in range of 0-50. However, a high AQ score alone is not a reason to be referred for a diagnosis. The AQ measures how many autistic traits an individual shows, and can be used across the general population, not only with people who are suspected of having ASD.

It should be noted that among members of the general population, autistic conditions are widely distributed along a "spectrum." It is reported that autism occurs more often in families of physicists, engineers, and mathematicians, and there is a link between engineering and autism. Likewise, the AQ was tested among 840 students in Cambridge University, and the result showed that scientists

[1] http://www.autismresearchcentre.com/arc_tests

scored higher than both humanities and social scientists [19, 20]. This confirmed the association between science/maths skills, and autistic traits.

2.3 Social Skills Training

Conventional SST is an established method developed by Liberman for schizophrenics to reduce their anxiety and discomfort in social interaction, and obtain appropriate skills [21]. SST is often performed with multiple sessions, and each session focuses on the training of one target skill for one or two hours. It is well known that SST can be used to effectively improve social skills for people with social disorders and ASD [11].

SST can be classified into individual (one to one training) and group (one to many or many to many training) settings. One advantage of group SST is that it enables participants to observe other participants' behaviour and also receive feedback from others. On the other hand, the advantage of individual SST is that the training can be relaxed and comfortable for participants, and that lessons can be tailored to the individual's needs.

As shown in Figure 2, SST is generally based on the following steps: defining target skills, modeling, role-play, feedback, reinforcement, and homework. We briefly describe these steps as follows:

- **Defining target skills:** The major social problem is identified, and the skills to be trained are decided based on this problem. In order to figure out the major problems, the participants and trainers work together through discussion or trainers decide the target skills. Once the target skills are decided the trainers decide the goal after intervention. In this step, trainers sometimes use related books to help participants understand target skills and the goal of SST. Examples of target skills include presentation skills, job interview skills, self-introduction skills, or skills regarding how to decline another's offer or request.

- **Modeling:** Before participants are asked to perform an interaction, trainers act as a model, demonstrating the skill that the participants are focusing on so that participants can see what they need to do before attempting to do

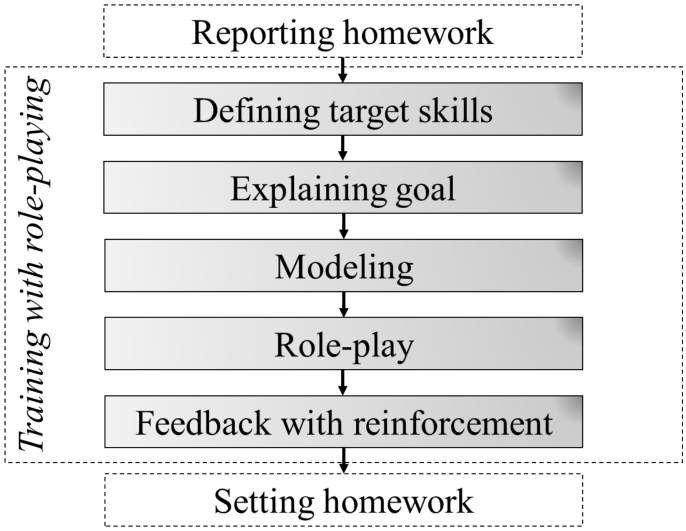

Figure 2. Human-to-human SST framework.

it themselves. For example, trainers may show a good story telling example using appropriate verbal and non-verbal cues.

- **Role-play:** Participants are asked to role-play. For example, participants tell their experience to the trainer. This allows the participants to practice their own skills in the target situation. Trainers observe participants' social skills subjectively, but mainly focus on voice quality, amplitude, facial expression, eye-gaze and other factors. This practice is a very important aspect of SST.

- **Feedback:** Trainers provide feedback at the end of role-playing (in the case of group SST, participants also receive feedback from other participants). This feedback helps participants to identify their strengths and weaknesses. For example, trainers may tell the participant that the role-play was very good because he/she used appropriate voice amplitude.

- **Reinforcement:** Trainers give positive reinforcement, praising the participant about their achievement of targeted behaviors, in addition to feedback. This is a very important aspect of the SST, because the participants often

do not have confidence in social interaction, and tend to have low self-esteem. Therefore, positive encouragement helps build confidence in social environments.

- **Homework:** Trainers set little homework challenges that participants are required to do in their own time throughout the week. For example, trainers may ask the participant to tell their story to friends or family, and let the trainer know about the result.

By performing this training, participants can learn better social skills in a number of different ways, a core aspect contributing to the effectiveness of training. However, it should be noted that the human trainer plays a very involved role in the majority of these steps. As a consequence, SST requires professional or at least well-trained trainers satisfying the above abilities (e.g. being able to perform modeling and give appropriate feedback comments). The number of skilled trainers is small, and thus the number of participants joining SST program is restricted and applications are competitive.

2.4 Examples of Target Skills

Trainers have to decide the target skills after discussion. For example, the Center for Special Needs Education at the Nara University of Education have implemented SST program[2] which is one year program including following set of skills: 1. self-introduction, 2. recognizing facial expressions and situations, 3. predicting result of action, 4. inviting friends, 5. asking reasons, 6. asking for help, 7. rewarding other people, 8. anger management, 9. declining and expressing an opinion, and 10. summary of learned skills. Other possible target skills are listening, asking for permission, joining an activity, waiting your turn, apologizing, accepting consequences.[3] On the other hand, a set of five core competencies is widely accepted within the community as a good description of the general goals: there are self-awareness, self-management, social awareness, relationship skills, and responsible decision making [22]. Figure 3 shows the relationship between

[2] http://nara-edu-csne.org/web/outline/

[3] http://www.projectachieve.info/

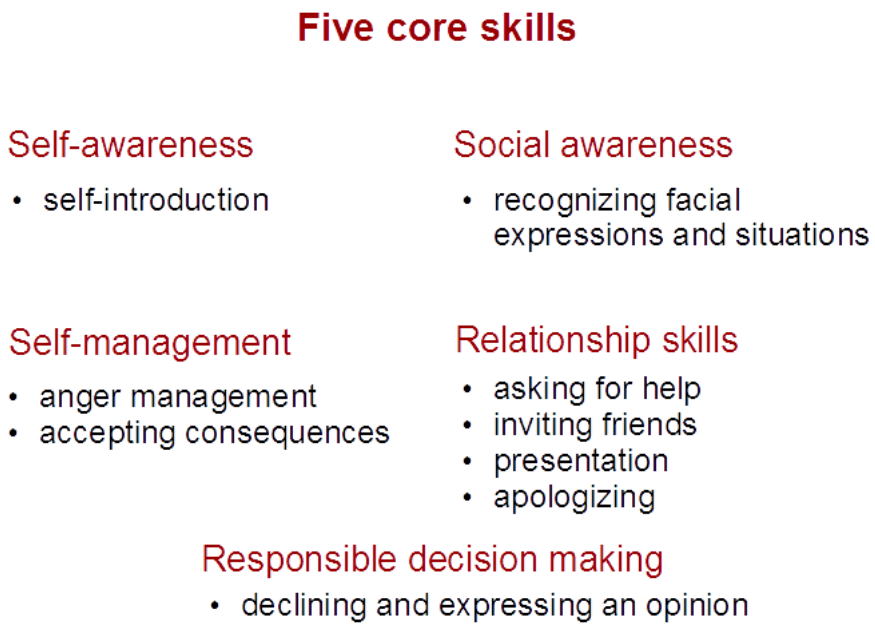

Figure 3. Five core skills [22] and its examples.

five core skills and its examples. Each target skill may relate to cognitive and/or affective components.

2.5 Existing Computer-based Training

As I mentioned in section 2.3, the number of skilled trainers is small, and the number of participants joining SST program is restricted. In contrast, the use of computers to aid people with communication difficulties has flourished in the last decade for several reasons. First, the computerized environment is predictable, consistent, and free from social demands, which people with ASD may find stressful. Users can also 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 [23, 24]. Most of these works have been proposed to support people with ASD (see review [25]). However, most of these applications tend

to be rather specific as skills in Figure 3 (e.g. focusing only on recognition of facial expressions from still photos) and have not been scientifically evaluated [26]. In addition, all of these embrace still only single aspects of SST (e.g. modeling or role-play in Figure 2).

First, I summarize applications that train cognitive skills. An application "FEFFA" was proposed to help users recognize emotion from still pictures of facial expressions and strips of the eye region [27]. "Emotion Trainer" teaches emotion recognition of four emotions from facial expressions [28]. "Lets Face It" teaches emotion and identity recognition from facial expressions [29]. Golan and Baron-Cohen [6] proposed a training tool "Mind Reading DVD" 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 (Figure 4).[4] They found that a computer-based 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. Generalization to questions not included in training is still difficult. However, when 8-11 year old children with ASD used it, improved generalization was found [26]. Virtual Environments (VE) form another domain with immense possibilities for people with ASD and related social difficulties. VE are artificial computer generated three dimensional simulations. The user can operate in realistic scenarios to practice social skills, conversations, and social problem solving. Parsons and their colleagues [30, 31] investigated a qualitative case-study approach to report observations of two adolescent boys with ASD, gathered during a series of sessions using a virtual cafe' and bus environment (Figure 5). For example, the user has to find a place to sit on the bus, and scenarios include asking someone to move their bags so that they can sit down, and standing when there are no seats available.

Next, I summarize applications that train affective skills. There were fewer reports of computers programs teaching emotional or non-verbal expressiveness. One example is ASC-Inclusion project [32], which aims to help children with ASD by allowing them to learn how emotions can be expressed and recognized through playing games in a virtual world (Figure 6). This includes automatic face analysis, voice analysis, and gesture analysis to evaluate emotional expressiveness. They

[4] http://www.cl.cam.ac.uk/research/rainbow/emotions/mrm.html

Figure 4. Image from Mind Reading DVD.

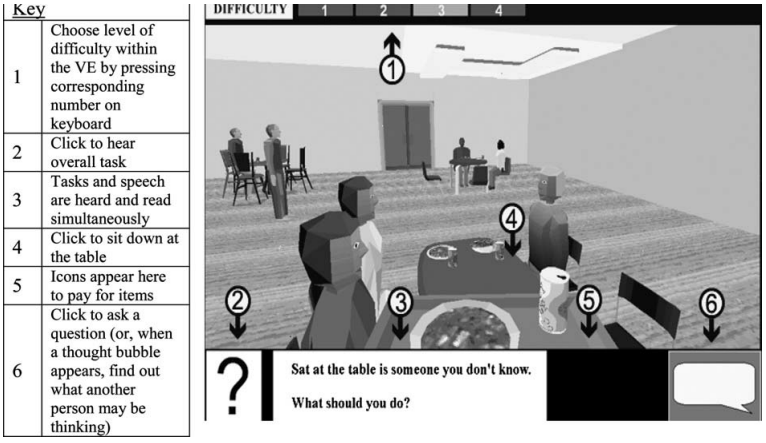

Figure 5. Screen shot of cafe' virtual environment.

focus on temporal basic emotions [33] such as angry, happy, sad, and afraid rather than context and complex emotions. There has been one previous work on an automated conversational coach named "Mach" [10], which is a dialogue systems aimed to train people for improving interview skills through real-time feature

Figure 6. Expression game with face, voice, and body gesture analysis.

detection and feedback. They achieved 1) a realistic task involving training real users, 2) formative affective feedback that provides the user with useful feedback on the behaviors that need improvement, and 3) the interpretation or recognition of user utterances to drive the selection of backchannels or formative feedback in a virtual world (Figure 7). They confirmed that participants significantly improved in subjective score using the system compared to a control group. They also found that there are significant gender differences for training effect. While, this work is an excellent first step, it did not faithfully follow the conventional SST framework.

2.6 Summary

In this chapter, I introduced instruments to measure degree of social skills, conventional SST, and some applications to enhance social skills. There have been a large amount of applications in terms of cognitive component, and most of these applications tend to be rather specific as skills. On the other hand, there were fewer applications teaching emotional or non-verbal expressiveness. In addition, previous computer-based trainings did not follow conventional SST framework.

Figure 7. The Mach interviews a participant and provides feedback.

3. NOCOA: A Computer-Based Training Tool for Social and Communication Skills That Exploits Non-verbal Behaviors

3.1 Introduction

The characteristics of autism includes deficits in nonverbal behaviors [34]. The term "non-verbal" is referred to not only emotion, but also partner information, intention, situation, age, sex and other factors.

In this chapter, I attempt to; first make clear relationship between non-verbal behavior skills and autistic conditions, second; develop a training method of social and communication skills. To evaluate the first goal, I evaluate the adult's autistic traits to confirm the non-verbal factors contributing to social and communication skills, which include, but are not restricted to emotion. To achieve second goal, and develop a mobile application reflecting the result of this analysis of a AQ tendencies. The mobile application allows users to measure autistic traits automatically, and enables people with social and communication difficulties to improve non-verbal cognitive skills for use in the real world.

3.2 Assessment of Communication Skills

Non-verbal information includes various factors (e.g. eye contact, intention, gesture, and sex). The objective of this section is to confirm the important non-verbal factors contributing to communication skills as measured by using AQ. To do so, I use Factor Analysis, which is commonly used to elucidate the factors contributing to scores on a psychometric test. To collect data, I 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).

The Cronbach's coefficient alpha is commonly used as a measure of the internal consistency or reliability of a psychometric test score for a sample of examinees. It is calculated with following formula:

$$\alpha = \frac{m}{m-1}\left(\frac{\sum_{j=1}^{m}\sigma_j^2}{\sigma_x^2}\right) \tag{1}$$

where m is the number of components (m-items), σ_x^2 is the variance of the observed total test scores, and σ_j^2 is the variance of component j for the current sample of persons. This resulted in Cronbach's coefficient alpha value of 0.73 (> 0.7), indicating that the test is reliable.

Next, I perform a factor analysis to determine several important factors for social and communication skills based on the AQ. Based on using principal component analysis (PCA) and the chi-square value I finally set 5 factors. Table 1 shows the loadings and the proportion of variance from the first factor to the fifth factor (the cumulative proportion of variance is up to 65%). Each individual factor's contribution ratio is not high, even for the first factor. Next, I perform an analysis with the promax method, which is an alternative non-orthogonal (oblique) rotation method that is effective when there are highly correlated factors. This reveals the following:

1. the first factor is largely related to intention and interest.
2. the second is related to politeness or impoliteness as well as new friends.
3. the third is related to social places and situations.
4. the fourth is related to chit-chat and feelings.
5. the fifth is other factors.

To confirm the degree to which each factor is effective for evaluating communication and social skills, I calculate Pearson's r value between each factor's total score. Table 2 reveals that the first five factors are sufficient to measure social and communication skills. As a result I selected the first two factors (intention & interest, and politeness/impoliteness & new friends) as non-verbal information. Finally these represent intention and partner information.

3.3 Classification of Natural Speech

In this section, I performed classification of natural speech data according to previous section. The categorized utterances can be used to measure and learn non-verbal cognitive skills.

Table 1. Factor analysis using the promax rotation method. Columns 1 to 5 show the loadings and the proportion of variance from the first to the fifth factor. Rows show the AQ question number and the statements. Underlines show larger values than 0.6.

No.	Statement	Factor loadings				
		1	2	3	4	5
	[intention, interest]					
45	I find it difficult to work out people's intentions.	1.308		-0.294		-0.191
35	I am often the last to understand the point of a joke.	0.687	-0.12		0.143	-0.109
15	I find myself drawn more strongly to people than to things.	0.613	0.263		-0.117	0.112
1	I prefer to do things with others rather than on my own.	0.571	0.436	0.138		
	[polite, new friend]					
22	I find it hard to make new friends.		0.869	0.114		0.187
7	Other people frequently tell me that what I've said is impolite, even though ...		-0.722			0.282
27	I find it easy to "read between the lines" when someone is talking to me.	0.159	-0.701		0.124	-0.129
47	I enjoy meeting new people.	0.161	0.524	-0.147	0.124	0.153
26	I frequently find that I don't know how to keep a conversation going.		0.515		0.189	-0.243
	[social place and situation]					
13	I would rather go to a library than a party.	-0.19	0.159	1.079		
48	I am a good diplomat.	-0.117	-0.225	0.734	0.201	0.768
18	When I talk, it isn't always easy for others to get a word in edgeways.	0.364	0.314	0.396	-0.179	
11	I find social situations easy.	0.281	-0.29	0.372		
	[chit-chat, feeling]					
31	I know how to tell if someone listening to me is getting bored.			-0.325	0.833	
17	I enjoy social chit-chat.			0.366	0.735	0.108
38	I am good at social chit-chat.	-0.212	0.128	0.309	0.531	-0.248
44	I enjoy social occasions.	0.384		0.175	0.492	
36	I find it easy to work out what someone is thinking or feeling just by looking ...	0.282		-0.213	0.475	0.219
	[others]					
33	When I talk on the phone, I'm not sure when it's my turn to speak.	-0.378	0.365		0.135	0.851
39	People often tell me that I keep going on and on about the same thing.	0.358	-0.283	-0.144	-0.317	0.552
	SS loadings	3.125	3.085	2.591	2.283	2.097
	Cumulative var.	0.156	0.31	0.44	0.554	0.659

3.3.1 Natural Conversational Speech Corpus

The FAN subset of JST/CREST Expressive Speech Processing (ESP) corpus[5] was recorded over a period of five years, and consists of over 600 hours of every-day conversational speech collected from a female volunteer, who used a high-quality head-mounted microphone to record her speech to a small mini-disc recorder. This corpus features a large amount of speech from various situations, including simple, repetitive and unstructured talk that shows how people actually speak in everyday situations. I prepared a total of 5,367 short utterances from the FAN database.

[5] http://www.speech-data.jp/corpora.html

Table 2. Correlation coefficient between factors and social and communication skills (***indicates $p < .001$, **indicates $p < .01$, and *indicates $p < .05$ by t-test).

			1	2	3	4	5
	Social and communication skill	1.00	0.73**	0.60***	0.79***	0.67***	0.08
1	Intention, interest		1.00	0.31	0.55**	0.25	-0.004
2	Polite or impolite, new friend			1.00	0.27	0.32	0.07
3	Social place and situation				1.00	0.44*	-0.11
4	Chit-chat, feeling					1.00	-0.27
5	Others						1.00

3.3.2 Communication Skill Categorization

Based on the result of the factor analysis described in previous section, I decide to use the first two factors plus another factor for content of conversation, which is essential for speech communication. The resulting axes are content of the utterance, partner information, and intention. The utterances were classified into one of the 3 types of categories by 3 Japanese students (male, ages: 23, 24, and 24). The final total number of content of the utterance was 27. The final categories of partner information was "friend" and "teacher", and intention was "derisive", "social", and "friendly." The categorization procedure is as follows.

The numbers of categories of partner information and intention in each axis were determined subjectively, bottom up, and only utterances that the 3 students all agreed upon were left in the database. As a result, utterances (content: 2, partner: 3, intention: 6) are chosen. For partner and intention, the annotators separated 60 randomly chosen utterances into the categories family, teacher, and friend. For these three categories, the agreement value was only 50%, which indicates low agreement. To resolve this problem I merged the family and friend categories, as the error rate between these two categories was the highest. In terms of intention there were 6 categories from bottom-up, and Cohen's multi-Kappa statistics were 0.32, which indicates low agreement. Thus I calculated Euclidean

distance between the clusters (which similar to error rate), and employed re-clustering. As a result, Cohen's multi Kappa statistics rose to more than 0.6. The final three categories were: derisive, social, and friendly.

3.4 Mobile Application

Finally I developed a mobile application named NOCOA (NOn-verbal COmmunication for Autism), which reflects the above result of overall AQ tendencies and classification of short utterances.

3.4.1 Voice Conversion

The speech used in NOCOA is converted to sound more like a child's voice considering that my final target is autistic children, and to protect FAN's privacy. I used the software MacSynthTransformer, which allows for changing the pitch and envelope of speech. 4 Japanese students (3 Male, 1 Female) listened to the two varieties of speech (original speech and converted speech), and confirmed that the speech quality was not reduced.

3.4.2 Facial Images

Next I prepared facial images for each category. As described in previous section, I chose three axis: 27 types of contents of the utterance, 3 types of intention or interest of talk, and 2 types of partner information. Both actual pictures (chosen via yourstock.com) and illustrations were prepared for each category, and the use of pictures or illustrations can be chosen by the user.

3.5 Structure

This subsection explains two modes of NOCOA.

3.5.1 Listening Mode

In listening mode, users touch the screen to choose the content, choose from two types of partner information, and then choose from three types of intention. If there is no available sound candidate, the photos will be blank. Finally the user

can see the result they chose on the play screen, and can listen to the appropriate sound. The maximum number of sounds in each category is 4, and the sound is played randomly.

3.5.2 Test Mode

NOCOA also has a test mode, which is able to measure users' intention and partner information cognitive skills. The user listens to the voice, and then chooses the appropriate face and partner. The test mode score is calculated by using agreement of 10 Japanese students in each category. 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 a social situation. In both partner information and intention the maximum score of each question is 5. The test mode score is calculated after answering 10 questions, so 100 is the best score. The 10 question set is chosen at random each time.

Here, computer-based intervention used drawings of photographs for training, rather than more lifelike stimuli. This might have made generalization harder than if more ecologically valid stimuli were used. Thus test mode also has three generalization levels:

1. Closed data: testing was performed using voices that were included in the listening mode but faces were presented using a different person.

2. Open data: faces and voices were not included in the training, but the content was the same as in the training.

3. Long sentences: faces and voices were not used in the training, and the content was not included in the listening mode, because the main utterances used in training are short.

3.6 Experimental Evaluation

3.6.1 Method

I performed an experimental evaluation of the correlation between AQ score and test mode score in members of the general population. I perform this evaluation

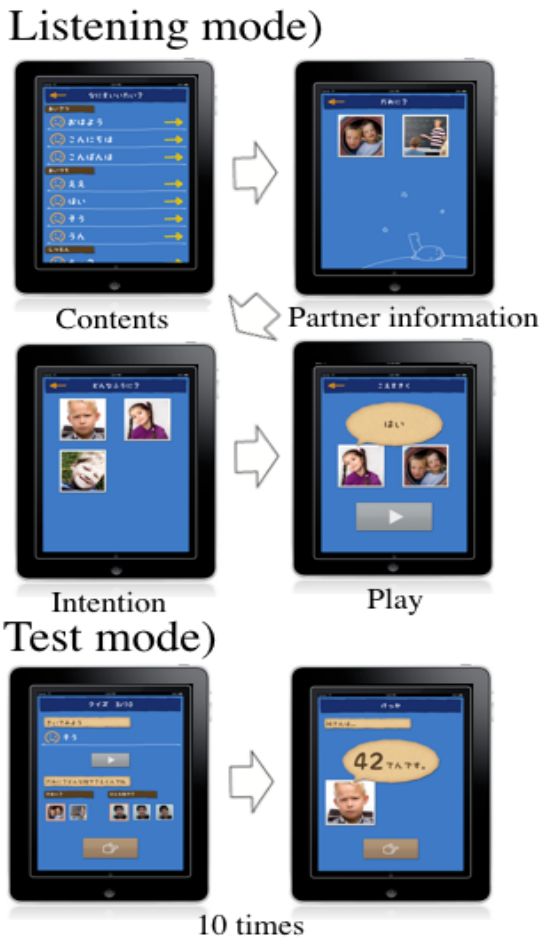

Figure 8. Two modes of NOCOA, listening mode and test mode. Both modes were developed systematically.

because my tool was developed for people who have difficulties with social and communication skills to measure their non-verbal cognitive skills and to systematically learn how to identify non-verbal information.

The procedure of the experiment is as follows: 19 Japanese participants were recruited (mean age is 25.0, 18 males and 1 female). They came to the laboratory one by one, and took the AQ test. After finishing, I checked the understanding of the concept of facial images. I confirmed that all participants did not have difficulty in understanding the concepts. Then, participants took generalization level 1 (closed data) of test mode on NOCOA two times, and the average score

of two trials was calculated.

I also tested efficacy of listening mode with several Japanese students (training group) who scored below average (mean age is 23.0). They used listening mode for 20 minutes, and the control group waited for the same 20 minutes. After 20 minutes both groups used test mode with the three generalization levels.

3.6.2 Results

First, I measured relation between test mode score and AQ, the correlation coefficient between AQ and averaged test mode score was 0.70 (see Figure 9). This reveals that large variations in the ability to recognize non-verbal and partner information exist in the general population, and these variations are significantly related to autistic traits. Note that despite the fact that the participants had not been diagnosed with Asperger syndrome or high-functioning autism and have average or above average IQ, their range of AQ scores was wide and well correlated with test mode score.

I also tested efficacy of listening mode with two Japanese students who scored below average (mean age is 23.0) and participated in training. Figure 10 shows that after using listening mode for 20 minutes, their score also improved above 10 points in the case of generalization level one. As a result of training that I found they maintained high scores in both open and long utterances (see Figure 11).

Here, we have to consider other factors to verify that the result is reliable. Because the FAN corpus is recorded by a person with a Kansai accent, I calculated the averaged test mode score between people with Kansai accent and people without Kansai accent in participants. The results showed with Kansai accent participants achieving 82.1 (11 people), and participants without a Kansai accent participants achieving 82.9 (8 people). The result shows no difference between accent types.

3.7 Summary

In this chapter, I confirmed the relationship between non-verbal cognitive skills and AQ by using speech output with visual hints, and examined prospective

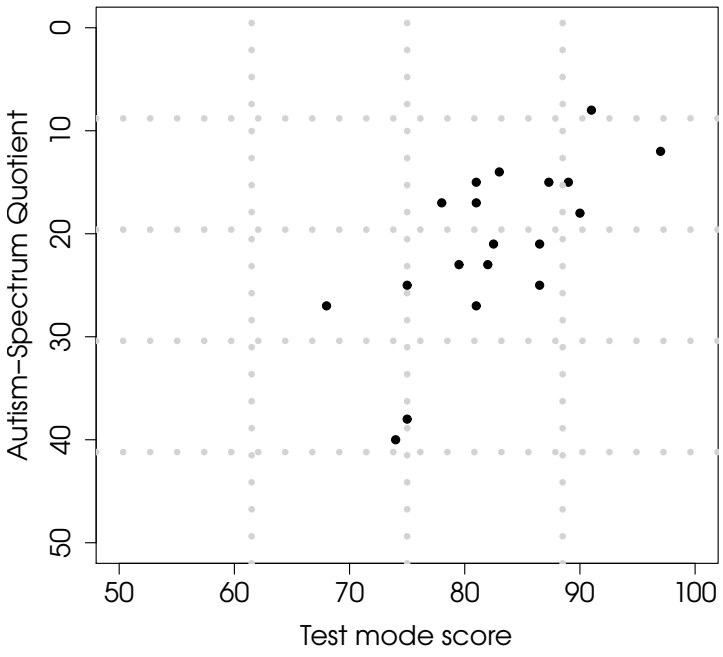

Figure 9. Correlate between test mode score and AQ score (r=0.70, $p < .01$) by 19 Japanese adults.

intervention through teaching non-verbal information, intention and partner information. Classification of utterances was based on the AQ, and my analysis revealed that it is an effective way to measure social and communication skills. According to factor analysis, I confirmed three important axes. The number of categories of partner information and intention on each axis was determined subjectively and bottom up. I conducted a subjective experiment with members of the general population, and confirmed that this tool is useful. As a result of experiment, correlation between AQ score and test mode score was 0.7 ($p < .01$) for 19 Japanese adults. This showed that ASD severity is significantly related to test mode score even in the Japanese adult group. It also reveals that in the general population, where the range of AQ scores was wider, the more autistic traits one possesses results in recognition of non-verbal information being more difficult. In

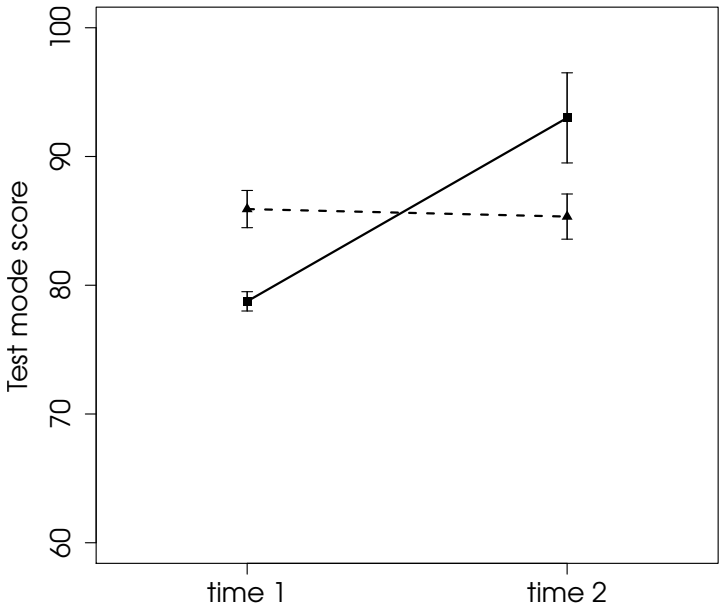

Figure 10. The test mode scores between before 20 minutes training (time 1) and after the training (time 2) with standard error bar. The dotted line shows control, and the solid line shows training group.

addition several Japanese students had difficulty distinguishing utterances compared to other members. However their test mode score was improved by using listening mode for 20 minutes. They also maintained high scores even in unseen open questions and long sentences.[6]

[6] I have been distributing NOCOA through the Apple AppStore for educational use since February 2012. (see http://itunes.apple.com/ph/app/nocoa/id501936653?mt=8)

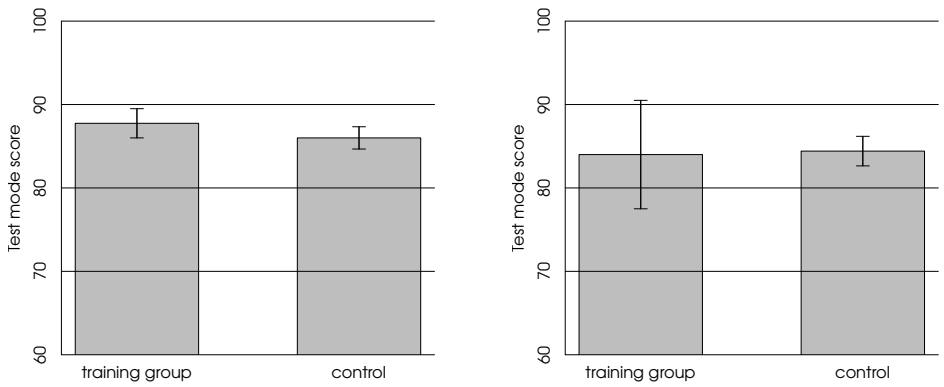

Figure 11. The score between training group and control. A left figure shows the result in generalization level two, and a right one shows the result in generalization level three.

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

4.1 Introduction

In the previous chapter, I proposed a tool NOCOA, which is a communication aid application to help test and train non-verbal behaviors. While the overall design of NOCOA proved advantageous, 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 [35, 36]. In addition, other reports have mentioned that conversational context influences emotion recognition [37], 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 [38, 39, 40, 41]. In most previous definitions, the common contextual factor is time, so I focus on temporal context.

In this chapter, I propose an updated application NOCOA+ that uses utterances in several modalities and context. I attempt to answer two major questions about computer-based training of non-verbal information left unanswered by the NOCOA framework: what is the effect of incorporating data from multiple modalities, and what is the effect of using data including conversational context? I do so by collecting and incorporating data from several sensory modalities, as well as data considering context. Based on this data, I perform a series of four experiments examining 1) the effect of temporal context on the ability to recognize social signals in testing contexts, 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 AQ and non-verbal behavior recognition skills measured by NOCOA+, and 4) the effectiveness of computer-based training in improving social skills.

4.2 Categorization of Non-verbal Behavior

Non-verbal behavior includes various factors (e.g. eye contact, intention, gesture, and gender). The experiments with NOCOA confirmed the important non-verbal

factors contributing to communication skills, and their relationship with the English version of the AQ [17]. Finally, I found two important factors: (a) intention & interest. (b) politeness/impoliteness & new friends, and selected these two factors as the non-verbal behaviors to be trained and tested by NOCOA+. In the description below, I abbreviate these as intention and partner information. 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, which are the same categories as NOCOA.

4.3 Recording and Annotation

I 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 I am able to collect video samples of "derisive," "social," and "friendly" utterances in the intention category, I 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, I 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 [17]). 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 Snack Tcl/Tk toolkit.[7] Detected utterances were automatically divided into speech and video. I also created utterances including temporal context information from the 5 s and 10 s prior to the actual utterance.

[7] http://www.speech.kth.se/snack/

Next, I annotated the recorded movies with correct category labels. In video recording, I prepared a total of 1200 audiovisual utterances. Because annotators are required to have good social skills to recognize non-verbal behaviors, I 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 for which all three annotators agreed on both partner and intention information were chosen for use in NOCOA+.

4.4 Design of NOCOA+

Using these movie samples, I 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 [16]. 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 have correlate with the AQ, and I have used this to guide my design. Second, systematic design: while individuals with ASD have difficulty in socialization and communication, they also show good and sometimes even superior skills in "systemizing" [12]. To create an application that satisfies these desiderata, I adopted a quiz format, where the user of the application must choose from several categories of intention and partner information.

4.4.1 Training Mode

Training mode was designed to enhance user's socialization and communication skills. Baron-Cohen *et al.* [12] speaks of the extreme male brain theory of individuals with ASD, which states that people with ASD prefer a rule-governed way. In contrast, previous work mentioned a large number of inputs was also needed to train the social skills [6]. Thus, I expanded training mode to provide two types of training, "listen to a large number of examples" and "check the rules." The former was developed to enable user to learn using a statistical-based training

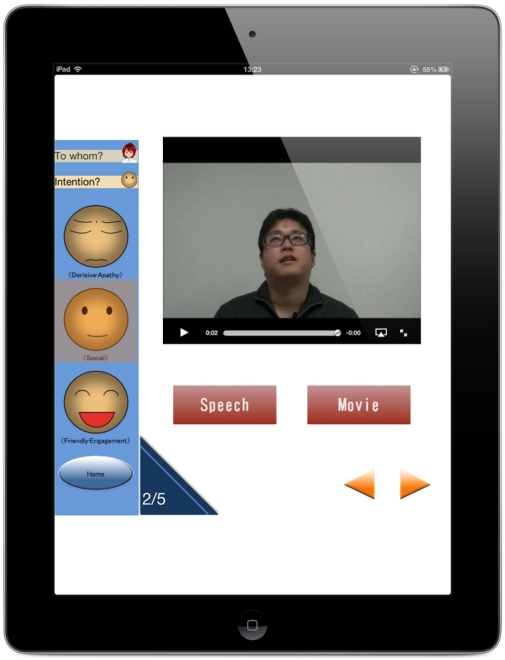

Figure 12. Screenshot of the training mode interface in English version.

regimen, which means that user can listen to and watch utterances for training (see Figure 12). 79 utterances were selected from the total of 109 utterances as a closed training set. In the latter rule-based training regimen, I created explanations of the eye-movement, prosody, and posture rules that provide hints about the correct answer, and the user can see the description. For example, "People in derisive situation tend to speak with short duration and lower pitch, and look down." The explanation was checked by two other people. The user can select the preferred training regimen from the training menu.

4.4.2 Test Mode

In the test mode quiz, 10 questions for measuring the user's non-verbal communication skills are provided. The user watches a video of an utterance, and then attempts to guess the intention and partner information corresponding to the utterance (Figure 13). The test mode score is calculated by using agreement in

Figure 13. Screenshot of the test mode interface in English version. The movie stimulus is displayed, and then the user selects the appropriate intention and partner information.

each category with the answer given by the annotators. 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.

The test mode score is calculated after answering 10 questions, with 100 the best score. The 10 question set is chosen at random each time. The questions have the two types of generalization levels shown below: (a) closed: testing is performed using data that was included in the training mode, (b) open: testing is performed using data that was not included in the training mode.

In the following sections, I 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 my experiments. Written informed consent was obtained from all subjects before the experiments.

4.5 Experiment 1: Difficulty Level and Contextual Differences

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

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

4.5.1 Method

I used the NOCOA+ test mode including three contextual levels: no context, 5 seconds context, and 10 seconds context. First, I collected data corresponding to each level of contextual information. Three types of difficulty levels were set; easy, normal, and hard according to the criterion mentioned before. To categorize difficulty levels, 10 participants (8 males and 2 females, mean age: 23.7) answered all questions twice using full contextual information.

4.5.2 Results

In Figure 14, I show the percentage of utterances with an error rate less than 20%. I can see that this value was related to the contextual level. This result indicates that contextual information helped people to infer the correct answer.

4.6 Experiment 2: Modality Differences

In the second experiment, I investigated the modality differences to recognize non-verbal information. I set a hypothesis that modality of stimulus has an effect on the ability to identify non-verbal information. To verify the hypothesis, I performed experiments using the testing mode of NOCOA+.

4.6.1 Method

I recruited a total of 14 participants (11 males and 3 females, mean age: 22.5) for the experiment. 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 I 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.

I set a hypothesis that characteristics of intention and partner information are different. To verify the hypothesis I analyzed the score for intention and partner information separately, and used one-way ANOVA to measure statistical significance. I also performed a pairwise comparison using Bonferroni's method [42].

4.6.2 Results

Figure 15 indicates that there were significant differences in each modality's score in terms of intention and partner information. Mean and SD values for intention score is as follows: verbal is 30.4 (SD: 5.14), visual is 44.1 (SD: 5.1), audio is 42.4 (SD: 4.4), and audiovisual is 43.2 (SD: 2.7). For partner score, mean and SD values is as follows: verbal is 47.5 (SD: 3.3), visual is 38.6 (SD: 5.0), audio is 46.8 (SD: 3.2), and audiovisual is 46.2 (SD: 4.1). The ANOVA showed [$F(3,28)=29.64$, $p < .01$] with $\eta^2 = 0.63$ for intention score and [$F(3,28)=15.77$, $p < .01$] with $\eta^2 = 0.48$ 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 ($p < .01$),

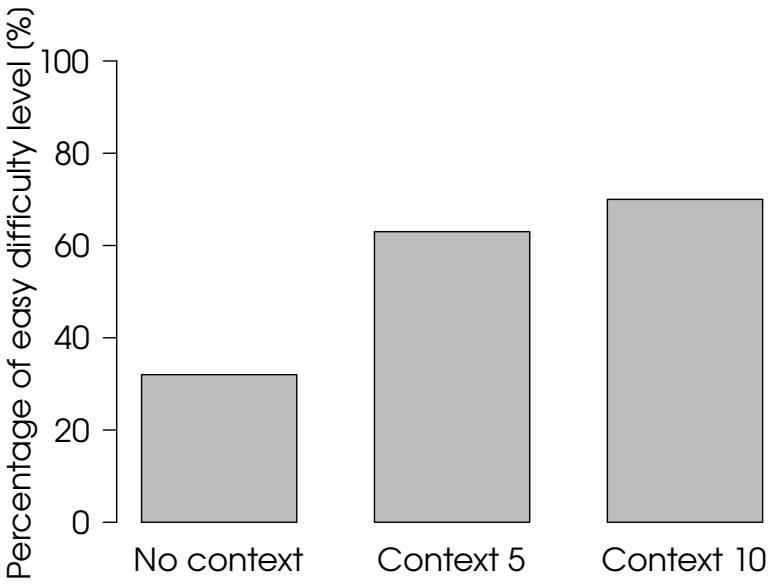

Figure 14. Utterances with percentage of error rate less than 20%.

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

4.7 Experiment 3: Relationship of Autistic Traits

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

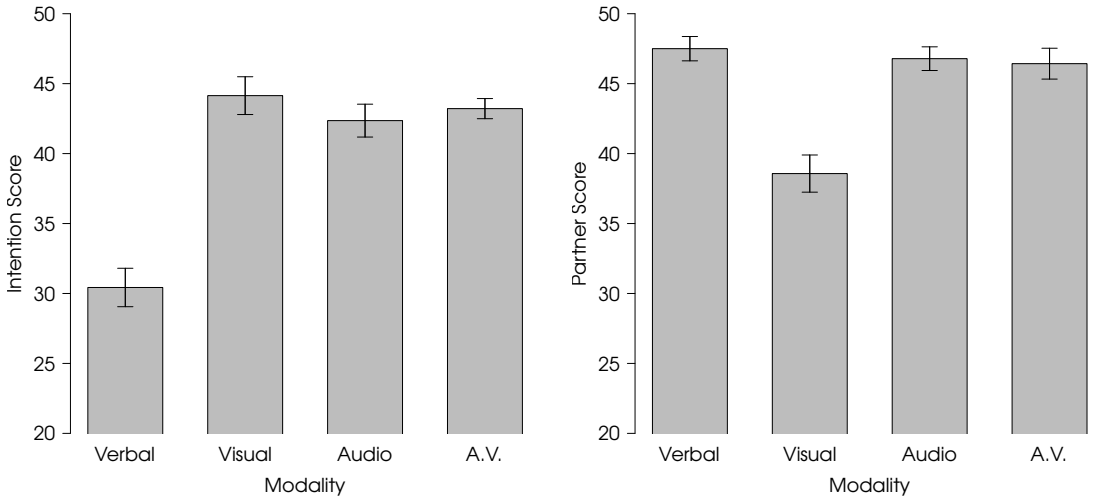

Figure 15. Modality differences in terms of intention and partner score with standard error bars. A.V. indicates Audiovisual.

4.7.1 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 [18], and the sum of the two AQ subareas (communication and social skill) was measured. I calculated the relationship and correlation coefficients between NOCOA+ score and AQ, and performed a linear regression analysis.

4.7.2 Results

Figure 16 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 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 Figure 16 shows, there is

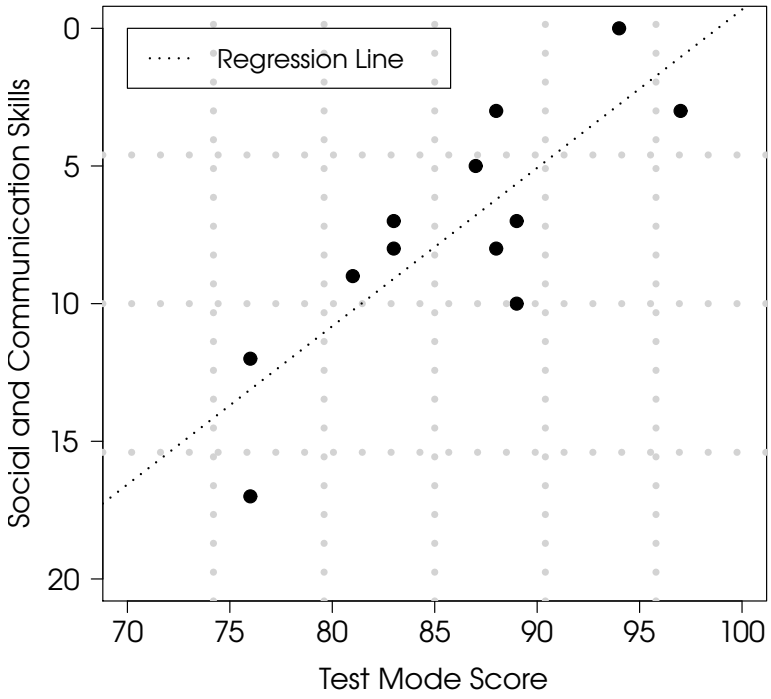

Figure 16. Relationship between the sum of social and communication AQ scores and test mode score of NOCOA+ with a regression line.

a correlation between the sum of the AQ subareas and averaged test mode score with a correlation coefficient of 0.82 ($p < .01$). I 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.

4.8 Experiment 4: Training Effect

In the forth experiment, I investigated whether computer-based training results in an increase in ability to recognize non-verbal information. I 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, I investigated whether users are able to maintain high scores even in unseen open questions.

4.8.1 Method

I recruited 12 participants (11 males and 1 female, mean age: 23.0). 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. The significant differences were tested by Student's t-test.

4.8.2 Results

Note that 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, $p > .1$]. Almost all participants were able to complete training on all utterances in 20 minutes. Figure 17 shows the improvement of test mode score before and after 20 minutes. In terms of difficulty level easy (left side of Figure 17), the improvement in score is 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 Figure 17), the improvement in score 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 significant difference [$t(10)$=1.90, $p < .05$]. For normal difficulty level,

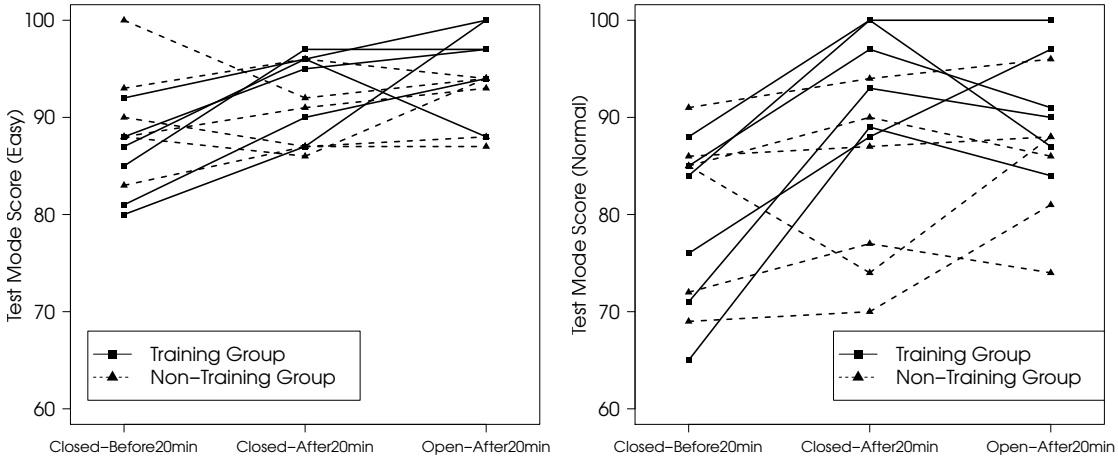

Figure 17. Test mode score before and after training. The left figure indicates difficulty level easy, and the right figure indicates difficulty level normal. Dotted lines indicate scores of the non-training group, and solid lines indicate scores of the training group. Pre and post 20 minutes (closed data) is shown as well as post 20 minutes (open data). Each line indicates a different participant.

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 it has a tendency [$t(10)=1.54$, $p < .1$].

I 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 I confirmed effectiveness by systematic training in both audio data and visual data.

4.9 Summary

Previous research has found that social and communication difficulties exist across different sensory modalities, both visual and auditory. I proposed a training tool NOCOA+ that uses utterances in several modalities and context. I used NOCOA+ to examine this result from the viewpoint of computer-based training, which was not only audio data, but also visual data.

In this chapter, I recruited a total of 48 participants and implemented a series of four experiments. I first analyzed contextual and modality differences. I found that contextual information was helpful for answering questions. I also found the effect of modality in each modality's score in the cases of both intention and partner information. In addition, I investigated the relationship of autistic traits measured by the AQ and non-verbal behavior recognition skills measured by NOCOA+, and the effectiveness of computer-based training. The results showed that a relationship between the AQ scores of communication and socialization subcategories and non-verbal communication skills, and participants significantly improved in score through computer-based training.[8]

[8] NOCOA+ has been distributed in Apple store as an educational application (https://itunes.apple.com/us/app/nocoa+/id622502354?ls=1&mt=8).

5. Automatic Classification of Affective States in Natural Conversational Speech

5.1 Introduction

My first focus to examine affective computing is laughter, which is one of the most common nonverbal vocalizations in social conversation [43]. Laughter is also reported as the most frequently annotated acoustic nonverbal behavior in meeting corpora [44] where 8.6% of the time a person vocalizes in a meeting is spent on laughing and 0.8% is spent on laughing while talking. Laughter is a universal and prominent feature of human communication [45], and expressed by both vocal and facial expressions. It is a powerful affective and social signal [46]. There is no culture where laughter is not found. Although children with ASD or other social difficulties have difficulties identifying types of laughter [47], current dialogue systems and computer-based training do not take into account laughter [6]. In a seminal study of the segmentation of laughs, Trouvain considers laughter as articulated speech, where at the low level there are sound segments that are either vowels or consonants [48]. At the next higher level, there are syllables consisting of sound segments. The next higher level deals with larger units such as phrases which are made up of several syllables. Owren [49] recommends the term 'bout' for the longer sequence, and 'call' for the individual syllables; I will adopt these terms in this study.

Some earlier work on the automatic segmentation of laughter has been reported in the literature. Khiet P. Truong *et al.* [50] reported automatic laughter segmentation in meetings. They performed laughter vs speech discrimination experiments comparing traditional spectral features and acoustic phonetic features, and concluded that the performance of laughter segmentation can be improved by incorporating phonetic knowledge into the models. Kennedy and Ellis [51] focused on joint laughter in meetings, which means participants (more than just one) laugh simultaneously [52, 53, 54], and they obtained detection results with a correct accept rate of 87% and a false alarm rate of 13% by using Support Vector Machines.

Types of laughter vary in natural conversational speech, and some classi-

fications have been reported in the literature regarding different categories of laughter. Most types of laughter were discussed in [55], and the major work is the discrimination of laughter into two types, voiced and unvoiced, based on acoustics [56, 57]. Laurence *et al.* [58] deal with a study of laughs in spontaneous speech and explore the positive and negative valence of laughter towards their global aim of detecting emotional behavior in speech. The conclusion of their acoustic analysis is that unvoiced laughs are more often perceived as negative and voiced segments as positive. Previous work in the literature has also discussed whether laughter patterns can be defined through stereotypes [59, 48, 60]. However, laughter is not simply positive or negative, or even defined by stereotypes; it is quite usual for people to infer different degrees of emotion and engagement based on its perceptions, and it is common for people to make use of social laughter in sophisticated social interaction. In this study I tested perceptual types of laughter to determine the main characteristics of laughter in social interaction by reference to the above previous studies.

Automatic classification of four phonetic types of laughter in a natural-speech conversation corpus was conducted by Campbell *et al.* [61], based on perceptual impressions of laughter, in which a laughter episode is considered as a sequence of speech-like phonetic segments (after Bachorowski *et al* [59]). The work described 4 different laughter types: voiced, chuckle, breathy and nasal, and modeled each laugh as composed of different combinations of these segments using Hidden Markov Models (HMMs) statistical classification. The study reported an automatic discrimination using 3 to 15 states with HMMs for 4 functions of laughter (hearty, amused, satirical, and polite). In categorizing emotional classification the work achieved 76% accuracy. However because of the hidden nature of the statistical modeling the report did not provide explicit details about which specific acoustic features contributed to the various categorizations of the laughter.

I report progress towards developing a sensor module that categorizes types of laughter for application in dialogue systems or SST situations. In the present study I only make use of the audio information but recognize that facial expression also carries an important channel of communicative information [62, 63]. This chapter reports a study of laughs in a corpus of human-human dialogues recorded

from Japanese telephone conversational speech [64]. I employed a corpus of natural spontaneous speech where laughter occurred naturally as a consequence of the dialogue interaction. I specifically avoid the use of contrived laughter or even specifically elicited laughs since they may not be representative of natural spontaneous interaction.

In the following sections I first provide details of the corpus, then introduce two Experiments. Experiment 1: a perceptual test by Japanese students to determine the number and types of easily discriminated laughter, and Experiment 2: describing the acoustic feature extraction, presenting the results of an analysis of the main acoustic features and finally reporting a classification of type of laughter using statistical methods.

5.2 Data: Natural Types of Laughter

I used two types of Japanese corpora. First, the Expressive Speech Processing (ESP) corpus[9] was used for this study. This data includes natural conversational speech with laughter. The speech data were recorded over a period of several months, with paid volunteers coming to an office building in a large city in western Japan once a week to talk with specific partners in a separate part of the same building over an office telephone. While talking, they each wore a head-mounted Sennheiser HMD-410 close-talking dynamic microphone and recorded their speech directly to DAT (digital audio tape) at a sampling rate of 48 kHz. They did not see their partners or socialize with them outside of the recording sessions. Partner combinations were controlled for sex, age, and familiarity, and all recordings were transcribed and time-aligned for subsequent analysis. Recordings continued for a maximum of eleven sessions between each pair which were numbered consecutively as session 01 to session 11. The additional eleventh session was only used in the case of absence of one of the volunteers from one of the regular sessions but provided useful additional material. Each conversation lasted for a period of thirty minutes. In all, ten people took part as speakers in the recordings, five male and five female. Six were Japanese, two Chinese, and two native speakers of American English. All were resident and working in Japan

[9] http://www.speech-data.jp/corpora.html

at the time. The speech data were transferred to a computer and segmented into separate files, each containing a single utterance for manual transcription by professional transcribers. Laughs were marked with a special diacritic, and laughing speech was also bracketed to show which sections of ordinary speech were spoken with a laughing voice. Laughs were transcribed using the Japanese Katakana phonetic orthography, wherever possible, alongside the use of the identifying symbol. The present analysis focuses on speakers JMA (age 20s) JMB (age 20s), EMA (age 20s), EFA (age 20s), CMA (age 30s), and CFA (age 20s) to confirm that the same types of laughter are common across different native language groups. The other speakers are all female and similar to the speaker FAN in terms of age, sex, and native language, and thus I selected one female speaker as representative for the present analysis. JMC is omitted because his speech data is insufficient. The initial letters J, C and E indicate native speaker of Japanese, Chinese, and English respectively, M or F indicates the gender of speaker, and A or B indicates the session group of speakers as used for a different experiment.

Second, data from speaker FAN (age 30s) was also used in this report. The FAN subset of the ESP corpus was recorded over a period of five years with everyday conversational speech collected from a single female volunteer wearing high-quality head-mounted microphones, recording her speech to a small MiniDisc recorder as she went about her daily life. This part of the corpus features a lot of speech in various situations and much simple, repetitive and unstructured talk that illustrates how we spontaneously speak in everyday situations. Speaker FAN was a young female Japanese who personally provided more than 600 hours of usable speech material. Because I was not able to enter into contractual agreements with her various interlocutors, only the voice of FAN herself has been transcribed or analyzed. While this material is less useful for the analysis of conversational interaction, it provides valuable insights into the range of voice qualities and speaking styles used by one person throughout her daily life.

The study reported here includes two perceptual experiments. The first tested for perceptual types of laughter using Japanese students as subjects listening to the natural conversational speech recordings. I used these results to confirm the classification into the most easily perceived classes of laughter in the corpus. The

second tested the degree to which opinions were shared between respondents in the initial classification. For both experiments I predicted the following:

1. In social communication, people do not use hearty laughter with high frequency, rather they typically express polite social laughter (Experiment 1);

2. There are some important acoustic features that can be used to distinctively classify the types of laughter; these includes laughter specific parameters such as the number of the calls; and

3. Automatic classification of laughter is possible at rates greater than chance in both closed and open tests (Experiment 2).

5.3 Experiment 1: Main Types of Laughter

This experiment concerned the annotation of types of laughter found in the ESP corpus and I chose conversations between JMA and JMB, and JMA and EFA as illustrative.

5.3.1 Method

I recruited 20 Japanese students (age 23 to 26), and they downloaded wav files from three of the thirty-minute sessions (JMA-EFA; session 03, JMA-JMB; session 03, and JMA-JMB; session 11). Male speaker JMA is the common factor here, and I noticed that his utterance and laughter would change depending on the partner information and the number of sessions (i.e., 'familiarity') [65, 66]. Annotators were free to select one from the list of three conversations for annotation, and were required to categorize both JMA's and partner's laughter. 8 students choose JMA-EFA; session 03, 6 students choose JMA-JMB; session 03, and 6 students choose JMA-JMB; session 11.

I determined types of laugher by reference to previous work [61, 67], as 'mirthful', 'polite', 'derisive', and 'others' because this research utilises spontaneous speech data, and thus derisive laughter is sometimes included in the corpus. Because hearty laugh and amused laugh in [61] were sometimes difficult to distinguish, these were both included under the category of mirthful laughter. The ESP

corpus has been richly transcribed and subjects worked from phonetic laughter transcriptions such as 'hahaha', 'hihihi', or 'huhuhu'.

The instruction page for the annotation exercise was created in html and students carried out annotations following these instruction in their own space, either at home or in the laboratory. The resulting annotation was sent to the first and second author by E-mail.

5.3.2 Results

The 20 annotator agreement was measured by Multi Cohen's kappa-coefficient which calculates agreement beyond chance by distinguishing the observed agreement (A_{obs}) from the agreement by chance (A_{ch}), according to the following:

$$\kappa = (A_{obs} - A_{ch}) / (1 - A_{ch}) \tag{2}$$

I implemented pair-wise kappa for all annotator pairs, and obtained a kappa value 0.46, which corresponds to moderate agreement according to the scale proposed by [68]. It must be noted that low kappa scores do not necessarily mean low agreement [69]: if the annotators share certain assumptions of the data, their chance agreement is higher, and the above formula gives smaller kappa values.

As a result, I found that mirthful and polite laughs account for 90 percent of all laughs in these samples of human social interaction and only a very small number of derisive laughs were heard. Approximately 8% of the time when a person vocalizes in natural dialogue is spent on laughing (Table 3). The table shows counts of labels both for laughs and laughing speech, though I omit any results for laughing speech from this study because of its linguistic complexity.

Experiment 1 was carried out to determine which types of laughter were most readily perceived by typical Japanese students, and I confirmed hypothesis 1; In social communication, people do not use hearty laughter with high frequency, rather they typically express polite social laughter. Since people with autism perceive polite laughter as mirthful laughter [47], a sensor module which classifies polite laughs is considered beneficial for SST situations. My research is directed to this goal.

The main types of laughter in these recordings were determined to be polite and mirthful (accounting for 90% of the laughs), and the number of other types

Table 3. An example of counts of four types of laughs (mirthful, polite, derisive, and others) and non-laughs in a representative thirty minute conversation between two males (JMA and JMB). I found that mirthful and polite laughs account for 90 percent of all laughs in this social interaction and only a very small number of derisive laughs were heard.

type	count	prop.	cumulative prop.
non-laughs	6999	none	none
polite	579	66%	66%
mirthful	244	28%	94%
derisive	49	5%	99%
others	4	1%	100%

of laughter is too small to be integrated into a sensor module reflecting social functions. Thus I take the majority vote of the observers, and categorized two basic types: mirthful laughs henceforth labeled 'm' and polite laughs labeled 'p' for use in Experiment 2.

5.4 Experiment 2: Classification of Types of Laughter

This experiment concerned an analysis of the acoustic parameters of the two types of laughter I defined above, and was implemented in classification of natural laughs by using Support Vector Machines, a widely-used high-performance statistical classifier.

5.4.1 Segmentation and Annotation

In Experiment 1, I determined two types of laughter that are common in Japanese social conversation, polite and mirthful. Experiment 2 utilized this result and two small classes (derisive and others) are removed because these is not enough data to use. I explored the variation as the number of speakers was increased. Table 4 shows the number of laughs used for this Experiment. For the analysis of acoustic features I used speakers JMA, JMB, and FAN, and for the test of cross-prediction

Table 4. Showing the number of laughs in each category.

	JMA	JMB	FAN	EMA	EFA	CMA	CFA
mirthful	129	127	196	5	44	57	5
polite	138	135	136	65	22	13	60

by Support Vector Machine the speakers JMB, FAN, EMA, EFA, CMA, and CFA were selected to evaluate the generalization ability of the classifier.

The choice of partner is important in classifying these two types of laughter; in this report the frequent speakers, JMA and JFA, who talk with almost all others were chosen. Thus, I select the following sessions; CMA-JFA, EFA-JFA, EMA-JFA, JMA-JFA, CFA-JMA, CMA-JMA, EFA-JMA, EMA-JMA, and JMB-JMA. The ESP corpus has rich transcription of all utterances and laughter segmentation was performed using linguistic label time-stamp information. An annotator manually labelled each laugh thus excised into either polite or mirthful categories according to the results obtained from Experiment 1.

5.4.2 Acoustic Feature Extraction

The prosodic acoustic features for each laugh were calculated by a software programme I wrote using the Snack speech processing Toolkit, part of the Tcl/Tk programming language. Explicit prosodic features were included for analysis. Overall classification accuracy from the mfcc alone is less than that obtained when using higher-level prosodic features such as F0, amplitude, duration, and their derivatives. In addition to these fundamental prosodic parameters, spectral tilt or shape parameters and positional parameters (fvcd, ppct, and fpct) were estimated to facilitate voice quality descriptions and to encode the acoustic dynamics of the laughter.

The features I tested were measures of fundamental frequency, speech amplitude, and spectral tilt. For fundamental frequency and power, I calculated the mean, maximum, and minimum values measured across each laugh (fmean, fmax, fmin, pmean, pmax, and pmin), as well as the position of the maximum in relative percentage values within each speech waveform (fpct, and ppct). I estimated

Table 5. Extracted acoustic features. The prosodic acoustic features for each laugh were calculated using the Snack speech processing Toolkit.

Features	Explanation
fmean	mean value of fundamental frequency
fmax	maximum value of fundamental frequency
fmin	minimum value of fundamental frequency
fpct	the position of the f0 maximum in relative percentage values
pmean	mean value of power
pmax	maximum value of power
pmin	minimum value of power
ppct	the position of the power maximum in relative percentage values
h1h2	the difference between the first harmonic and the second harmonic
h1a3	the difference between the first harmonic and the third formant
h1	the amplitude of first harmonic
a3	the amplitude of third formant
fvcd	the amount of voicing it contained
duration	duration of the laugh

spectral tilt from the difference between the first harmonic and the amplitude of the third formant (h1a3), and by the difference between the first harmonic and the second harmonic (h1h2), as well as taking into account the amplitude of first harmonic (h1) and third formant (a3) respectively. I also measured duration of the laugh (dn) as well as the amount of voicing it contained (fvcd) [70].

I extracted 'No.Call' (The number of calls in a bout) as a further feature for my analysis. The call unit segmentation is implemented by use of an mfcc 3-state Hidden Markov Model with, which achieved over 87% accuracy for each of the four call types within a bout (voiced, ingressive, chuckle, and nasal) as reported in [71]. I calculated the correlation coefficient between duration and No.calls of JMA and obtained a correlation of 0.91 ($p < .01$ (signif)). Although highly correlated I consider the number of calls to be a relevant parameter in my modeling as it may distinguish between many short calls and few longer ones each having the same

overall bout duration. Actually, approximately 1 % of accuracy rate is changed according to the inclusion each of these features against each speaker in my pilot experiment.

Two further dynamic parameters 'F0moveAB' and 'F0moveAN' were also extracted. As Figure 18 shows, these parameters need $F0avg2a$ which represents average logarithm of pitch within a first (A) call, and $F0tgt2b$ (Second (B) call) and $F0tgt2n$ (Final (N) call) which represents the pitch target at the end of each call by a simple regression coefficient. Pitch change between the first and the second call (F0moveAB) is calculated $F0avg2a - F0tgt2b$, and that between the first and the final call (F0moveAN) is also calculated $F0avg2a - F0tgt2n$. When there is one call within a bout, I set these dynamic parameters to zero.

5.4.3 Statistical Analysis Tool

This section reports a statistical analysis of human laughter which was annotated as either polite or mirthful, using parameter reduction by means of Principal Component Analysis and Classification Trees. An automatic classification of the two types of laughter is reported in this section. The statistical analyses were performed using the free public-domain software package R.[10] Specifically, I used the additional option package 'tree' for Classification Tree and package 'e1071' for the Support Vector Machine analysis.

5.4.4 Principal Component Analysis

I split the data into training and test set (JMA; training: 206, test: 61, JMB; training: 191, test: 71, FAN; training: 270, test: 61), and the number of label 'p' and 'm' are balanced in each set. I ensure that the test material does not appear anywhere except in a validation experiment.

Figures 19 and 20 show plots of the two types of laughter in terms of each acoustic representation for all data of the speaker JMA. From these plots I infer that type of laughter can be readily characterized by use of these acoustic features and will show the extent to which this can be achieved. Figure 19 shows first 8 parameters of JMA, and for example that 'p' (polite) is characterized by

[10] http://www.r-project.org/

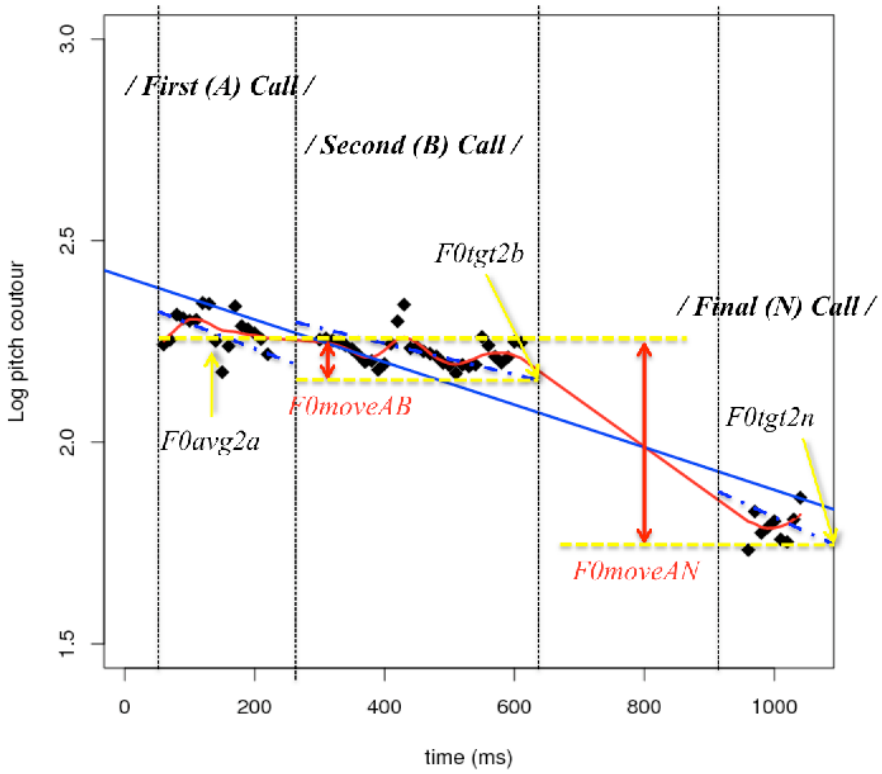

Figure 18. Log pitch contour and extracting method of two dynamic parameters F0moveAB and F0moveAN. These parameters need $F0avg2a$ which represents average logarithm of pitch within a first (A) call, and $F0tgt2b$ (Second (B) call) and $F0tgt2n$ (Final (N) call) which represents the pitch target mark at the end of each call by a simple regression coefficient. Pitch change between the first and the second call (F0moveAB) is calculated $F0avg2a - F0tgt2b$, and that between the first and the final call (F0moveAN) is calculated $F0avg2a - F0tgt2n$.

relatively low maximum power, and that 'm' (mirthful) is characterized by relatively high maximum power. Most laughs are in the region of high maximum power and there is considerable spread of laugh categories across the whole of fmean-pmax dimensional feature space. Figure 20 shows last 8 parameters and note that 'p' (polite) is characterized by relatively high h1a3 value, which is a spectral tilt parameter representing differences in voice quality, and 'm' (mirthful)

is characterized by relatively high duration and high No. calls. Since the data from speakerJMB and FAN show almost the same distribution as that of speaker JMA, their figures are omitted here (no individuals difference were found).

Principal Component Analysis (PCA) was used for analyzing and maximizing the combination of acoustic features across the speakers. The result from speakerJMB and FAN show almost the same as that of speaker JMA, and thus I report the PCA result for training data of the subject JMA. The proportion of variance from the first component to the fifth component (cumulative proportion of variance up to 70%) from a PCA rotation of these acoustic features shows that each component's contribution ratio is not individually high, even for the first component, for all speakers. Table 6 shows the result of JMA's factor loadings. It reveals that the first principal component is largely related to fundamental frequency and No.call, the second to power, the third to spectral slope, and the fourth to F0moveAB.

5.4.5 Decision Trees

Decision Trees are a very useful tool for confirming finer details of contributing factors within the three parameters of fundamental frequency and power, min, max, and mean, that emerged from the principal component analysis.

I employed both Classification Trees and Support Vector Machines in my modeling; the former being relatively weak at classification but very useful for examining the contribution of the individual factors, and the latter being perhaps the strongest statistical classifier available for general use.

Figures 21 shows the results of growing and pruning a decision tree having 10 leaves for speaker JMA. Detailed formation of each tree differs according to speaker, but the important acoustic parameters are similar. These can be used to classify laughs according to a cascade of IF-THEN rules, giving total accuracy of 77% (JMA), 74% (JMB), and 90% (FAN) respectively. Classification tree accuracies were measured for each test dataset. By observing the upper part of the tree, fmean, pmax, ppct, and dn (duration), the principal contributing features used to classify the two types of laughs can be determined.

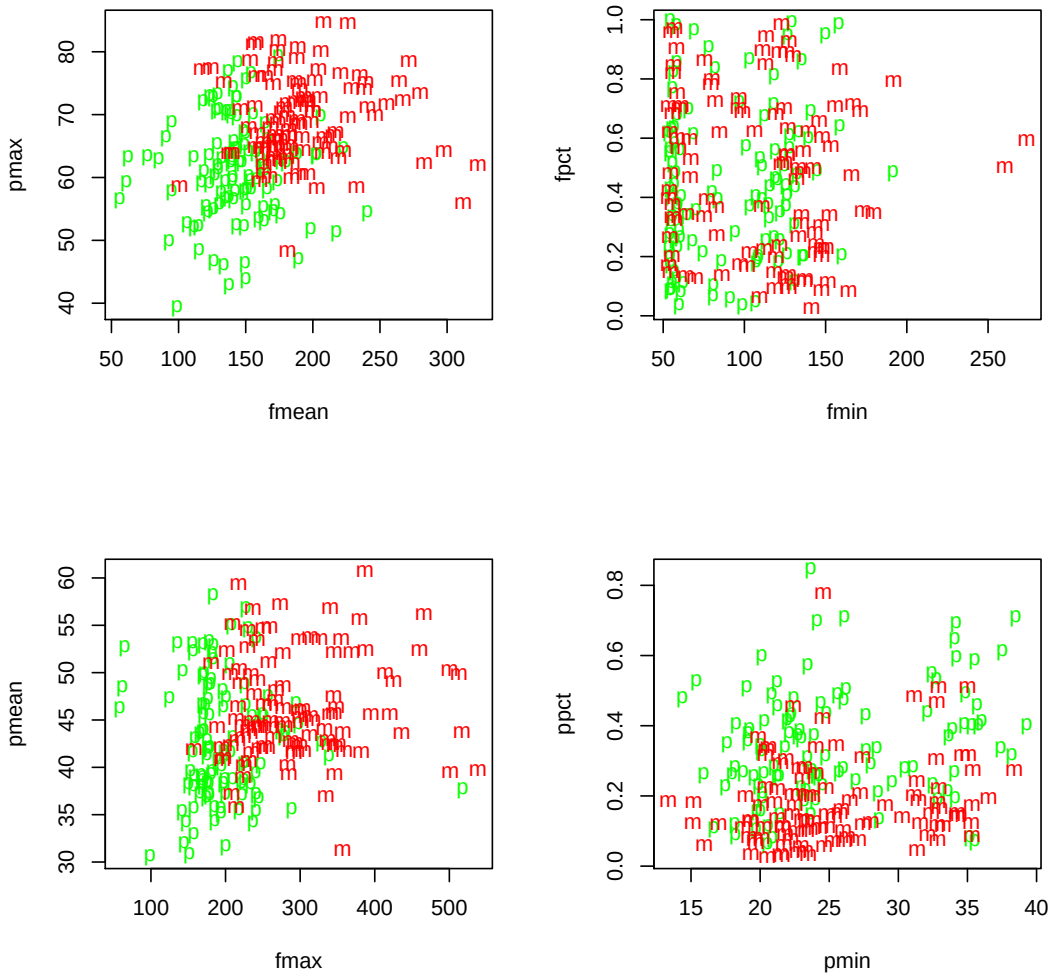

Figure 19. Showing first 8 parameters. JMA shows for example that 'p' (polite) is characterized by relatively low maximum power, and that 'm' (mirthful) is characterized by relatively high maximum power. Most laughs are in the region of high maximum power and there is considerable spread of laugh categories across the fmean-pmax dimensional feature space.

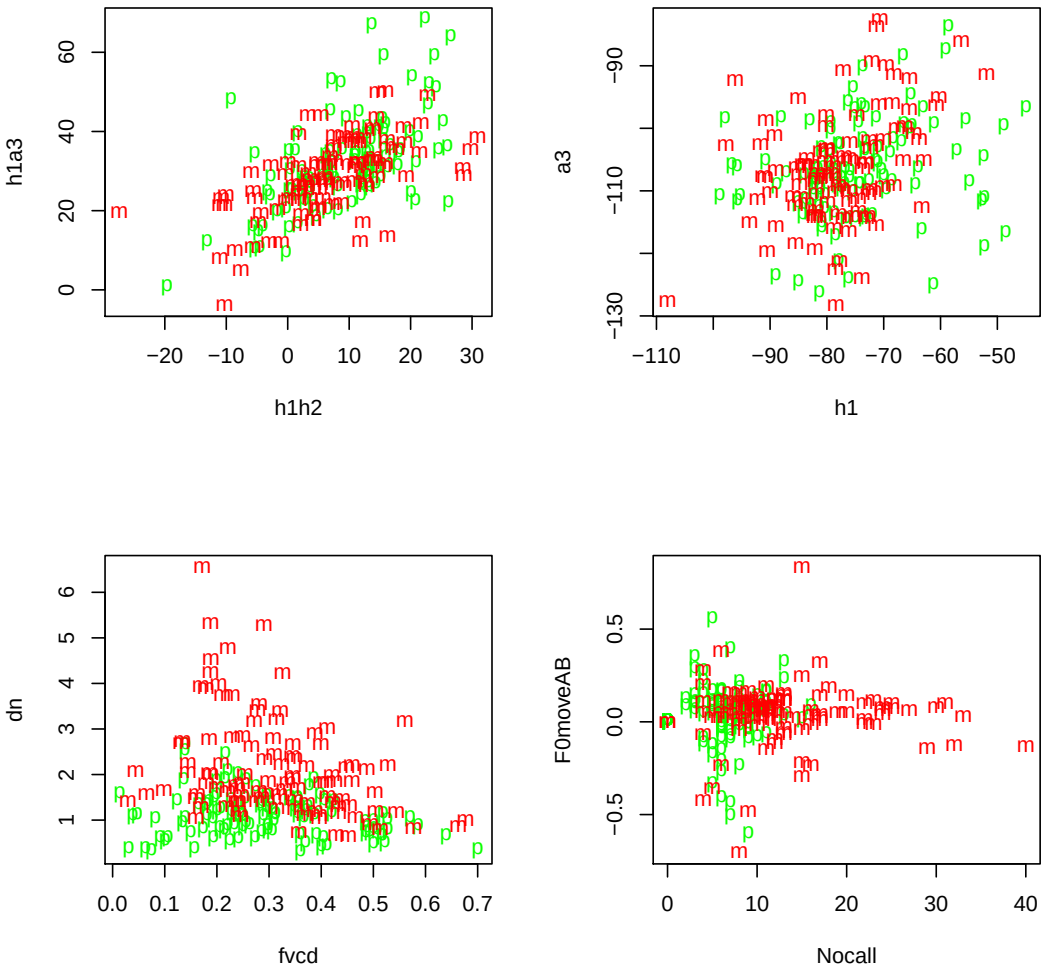

Figure 20. JMA shows a different distribution of categories across the different last 8 feature space which indicates that 'p' (polite) is characterized by relatively high h1a3 value, which is spectral tilt parameter correlated to voice quality, and 'm' (mirthful) is characterized by relatively high duration and high No. calls.

Table 6. Loadings of Principal Component Analysis. This reveals that the first principal component is largely related to fundamental frequency and No.call, the second to power, the third to spectral slope, and the fourth to my measure of prosodic activity F0moveAB.

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
fmean	0.406			-0.139	-0.360
fmax	0.426				-0.149
fmin	0.128	0.199	0.183	-0.249	-0.548
fpct	0.122			0.345	
pmean		-0.507			
pmax	0.223	-0.430			
pmin		-0.484			-0.180
ppct	-0.221		0.353	0.104	-0.184
h1h2	-0.258		-0.391		-0.158
h1a3	-0.312		-0.426		-0.154
h1	-0.287	-0.276	-0.312		-0.312
a3		-0.369	0.179	0.149	-0.184
fvcd		-0.202	0.167		0.494
dn	0.355		-0.411	0.113	
No.call	0.355		-0.382	0.106	0.148
F0moveAB				-0.620	
F0moveAN		-0.111		-0.580	0.177

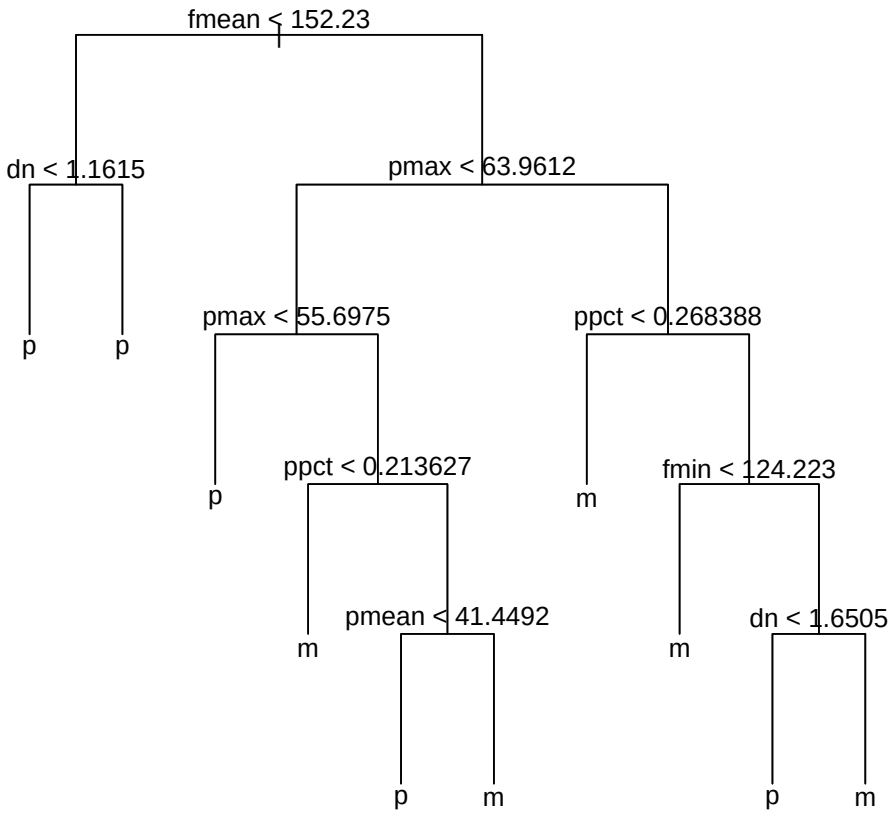

Figure 21. The Classification Tree for predicting laughs from JMA - with 10 leaves, using a different set of parameters and parameter ordering from that determined for JMA, starting from fmean, then taking into account dn (duration) and pmax (maximum power).

5.4.6 Both Speaker-dependent and Speaker-independent Classification by Support Vector Machine

Support Vector Machines are high-performance statistical classifiers. The SVM Type is C-classification, and kernel type is linear. Other system parameters are set to cost: 1, and gamma: 0.0625. The result of automatic discrimination using 15-fold closed (i.e., train and test on the same speaker) cross validation for JMA's mirthful (m) and polite (p) laughs I obtained 85% total accuracy. Training a Support Vector Machine on the same data gives a much more successful result. For JMB and FAN the same classification is implemented and total accuracy is 80% (JMB) and 92% (FAN). I split the data into training and test sets. The result of automatic discrimination on the test set for mirthful (m) and polite (p) laughs shows that I obtained JMA; 75% (F-measure = 0.76), JMB; 87% (F-measure = 0.88), FAN; 84% (F-measure = 0.84) respectively.

In the speaker-independent classification, I trained with JMA (using training set) and tested with JMB and FAN (using test set). Two speaker's classification rates are JMB: 90% and FAN: 67% respectively. Good categorization was possible for JMB, however for FAN, classification rates is relatively low. It is probably caused by overfitting due to high dimensional acoustic parameters and thus I try to implement parameter reduction.

5.4.7 Parameter Reduction

Having a large number of predictor features usually results in better classification accuracy, but often at the cost of generalizability. Accordingly, I performed a Principal Component Analysis and used Classification trees to reduce the number of features used in the final model. As I mostly inspect the Table 6 and Figure 21, the important features were selected. the optimal combination of features was chosen from the first or second principal components, and from those featuring most commonly in the upper part of the Classification trees. I was able to confirm the usefulness of seven important acoustic features; fmean (or fmax), pmax, ppct, h1a3, duration, No.call, and F0moveAB. The other parameters were omitted from the set of acoustic features used for the final training of the Support Vector Machines.

5.4.8 Classification by Support Vector Machine post Parameter Reduction

Following the above parameter reduction, I used a Support Vector Machine to predict the most likely category for each laugh token from its acoustics. The result of automatic discrimination using 15-fold cross validation for JMA's mirthful (m) and polite (p) laughs I obtained 86% total accuracy. For JMB and FAN the same classification is implemented and total accuracy is 86% (JMB) and 89% (FAN). It shows relatively high accuracies for each speaker compared to pre parameter reduction. The result of automatic discrimination using the test dataset for mirthful (m) and polite (p) laughs I obtained JMA; 79% (F-measure = 0.78), JMB; 89% (F-measure = 0.89), FAN; 79% (F-measure = 0.81) respectively.

5.4.9 Cross Prediction (speaker-independent) post Parameter Reduction

Table 7 shows the results of an open test across speakers, JMB, FAN, EMA, EFA, CMA, and CFA (mixing different native language and gender groups), training with JMA and testing with the others. All speaker's classification rates are over 70% (JMB: 85%, FAN: 74%, EMA: 93%, EFA: 79%, CMA: 86%, CFA: 86%). Good categorization was possible for each speaker by using the seven acoustic features described above. However, difference in speaker-independent results before and after parameter reduction is not statistically significant. Therefore, I conclude that feature reduction could not actually help to significantly improve results.

I performed an error analysis for these SVM results restricted to polite tokens of two speakers, EMA (English male speaker) and CFA (Chinese female speaker) who represent difference of both gender, native languages, and age. A Student's t-test was conducted for each of the seven acoustic parameters. As a result, I found that pmax parameter differs between true (same test and training sample) polite laughter and error (false prediction or difference between test and training sample). According to this test, EMA's mean pmax "error" polite laughter: 54.99, "true" polite laughter: 71,56 ($p < .01$) and CFA's mean pmax "error" polite laughter: 45.99, "true" polite laughter: 60.40 ($p < .01$). This may be due to microphone impact noise since the power parameter was not normalized in the

Table 7. An open test across speakers, JMB, FAN, EMA, EFA, CMA, and CFA, training with JMA and testing with the others. Each speaker's classification rates are over approximately 70%. The rows are true classes and the columns show predicted classes.

JMB	mirthful	polite		EMA	mirthful	polite		CMA	mirthful	polite
mirthful	36	5		mirthful	4	1		mirthful	52	5
polite	6	24		polite	4	61		polite	5	8
FAN	mirthful	polite		EFA	mirthful	polite		CFA	mirthful	polite
mirthful	29	1		mirthful	43	1		mirthful	3	2
polite	15	16		polite	13	9		polite	7	53

extraction process.

5.5 Summary

This chapter evaluated classification of natural laughter for engagement sensing in natural speech data. In Experiment 1 I observed several types of laughs (mirthful, polite, derisive, and others) in a natural speech corpus, and two predominant types of laughter (polite vs. mirthful) were defined and categorized from a manual examination of the data and by perceptual labeling carried out by 20 Japanese subjects. In social communication (for Japanese at least, but probably more generally), people do not use hearty laughter with the same frequency that they utter polite laughs. I found that human laughter includes various laughs in conflict with stereotyped laughter [72], and I found many instances of the various types of laughter in my spontaneous Japanese speech.

I reported an analysis of the acoustic features of these laughs. Global prosodic and laughter-specific acoustic features were extracted for the two types of laughter. These parameters were analyzed by Principal Component Analysis and Classification Trees to reduce the number of parameters. As a result of the analysis, I confirmed seven contributing acoustic features; mean value of fundamental frequency (fmean), maximum value of power (pmax), the position of the power maximum in relative percentage values (ppct), the difference between the first

harmonic and the third formant (h1a3), duration of the laugh (dn), The number of calls in a bout (No call), and Pitch change between the first and the second call (F0moveAB). For both parameter plots and statistical analysis I found a difference between the two main types of laughter.

A Support Vector Machine was trained and tested using these seven features, and total classification accuracy was confirmed to be at least 85% with cross validation for speaker JMA. As a result of statistical analysis I reduced the number of parameters to seven dimensions. By observing the output of a principal component analysis and by use of decision trees some strong predictor parameters were chosen. After parameter reduction, open speaker tests across different discourse modes achieved approximately 70%.

I found certain individual differences and some strong similarities between people and tested both open and closed prediction methods. By reducing the number of parameters and using only the strongest and most general predictors I was able to obtain good results on cross-prediction tests for variety of speakers (cross-culture and personality). However, in case of EFA, her accuracy was low compared to other speakers in ESP corpus. She seemed to be nervous during recording and thus she often laughs in a state of embarrassment that is difficult to classify into polite or mirthful laughs. Furthermore, speaker FAN data was recorded in various very different conditions as I mentioned in the introduction. That I can predict the type of laughter for her speech, when training on more constrained examples, indicates that this parameter reduction achieved high accuracy and allows high generalization.

I justifies my belief that prosodic parameters are sufficiently and statistically different in the two types of laughter and that Machine learning can classify them efficiently. This laughter detection is currently being integrated into a device to help people with autism spectrum disorders [47], who have difficulties understanding certain types of social functions.

6. Linguistic and Acoustic Features for Automatic Identification of Autism Spectrum Disorders in Children's Narrative

6.1 Introduction

The American Psychiatric Association defines the two characteristics of ASD as: 1) persistent deficits in social communication and social interaction across multiple contexts, and 2) restricted, repetitive patterns of behavior, interests, or activities [2]. In particular, the former deficits in social communication are viewed as the most central characteristic of ASD. Thus, quantifying the degree of social communication skills is a necessary component of understanding the nature of ASD, creating systems for automatic ASD screening, and early intervention methods such as SST and applied behavior analysis [73].

There are a number of studies finding differences between people with ASD and people with typical development (TD), who has not ASD. In terms of deficits in social communication, there have been reports describing atypical usage of gestures [74], frequency of eye-contact and laughter [75], prosody [76, 77], voice quality [78, 79], delay responses [73], and unexpected words [80]. In this chapter, I particularly focus on the cues of ASD that appear in children's language and speech.

In the case of language, Newton [81] analyze blogs of people with ASD and TD, and found that people with ASD have larger variation of usage of words describing social processes, although there are no significant differences in other word categories. In the case of speech, people with ASD tend to have prosody that differs from that of their peers [34], although Mccann [76] note that prosody in ASD is an under-researched area and that where research has been undertaken, findings often conflict. Since then, there have been various studies analyzing and modeling prosody in people with ASD [82, 79, 83, 84, 85]. For example, Kiss [83] find several significant differences in the pitch characteristics of ASD, and report that automatic classification utilizing these features achieves accuracy well above chance level. To my knowledge, there is no previous work integrating both language and speech features to identify differences between people with

ASD and TD. However, it has been noted that differences in personality traits including introversion/extroversion can be identified using these features [86].

In this chapter, I perform a comprehensive analysis of language and speech features mentioned in previous works, as well as novel features specific to this work. In addition, while previous works analyzed differences between people with ASD and TD, I additionally investigate whether it is possible to automatically distinguish between children with ASD or TD using both language and speech features and a number of classification methods. I focus on narratives, where the children serving as my subjects tell a memorable story to their parent [9]. Here, the use of narrative allows us to consider not only single-sentence features, but also features considering interaction aspects between the child and parent such as pauses before new turns and overall narrative-specific features such as words per minute and usage of unexpected words. Given this setting, I perform a pilot study examining differences between children with ASD and TD, the possibilities of automatic classification between ASD and TD, and the differences between American and Japanese children.

6.2 Data Description

As a target for my analysis, I first collected a data set of interactions between Japanese children and their parents. In collecting the data, I followed the procedure used in the creation of the USC Rachel corpus [87]. The data consists of four sessions: doh (free play), jenga (a game), narrative, and natural conversation. The first child-parent interaction is free play with the parent. The child and parent are given play doh, Mr. Potato Head, and blocks. The second child-parent interaction is a jenga game. Jenga is a game in which the participants must remove blocks, one at a time, from a tower. The game ends when the tower falls. The third child-parent interaction is a narrative task. The child and parent are asked to explain stories in which they experienced a memorable emotion. The final child-parent interaction is a natural conversation without a task. These child-parent interactions are recorded and will enable comparison of the child's interaction style and communication with their parent. Each session continues for 10 minutes. During interaction, a pin microphone and video camera record the speech and video of the child and the parent.

Table 8. Subjects' age and diagnosis

Subject	A1	A2	A3	A4	T1	T2
Age	10	10	10	13	10	12
Diagnosis	ASD	ASD	ASD	ASD	TD	TD

In this chapter, I use narrative data of four children with ASD (male: 3, female: 1) and two children with TD (male: 1, female: 1) as an exploratory study. The intelligence quotient (IQ) for all subjects is above 70, which is often used as a threshold for diagnosis of intellectual disability. Each subject's age and diagnosis as ASD/TD is provided in Table 8. In the narrative session, each child and parent speaks "a memorable story" for 5 minutes in turn, and the listener responds to the speaker's story by asking questions. After 5 minutes, the experimenter provides directions to change the turn.

In this chapter, I analyze the child-speaking turn of the narrative session in which the parent responds to the child's utterances. All utterances are transcribed based on USC Rachel corpus manual [87] to facilitate comparison with this existing corpus. In the transcription manual, if the speaker pauses for more than one second, the speech is transcribed as separate utterances. In this chapter, I examine two segment levels, the first treating each speech segment independently, and the second handling a whole narrative as the target. When handling each segment independently, I use a total of 116 utterances for both children with ASD and TD.

6.3 Single Utterance Level

In this section, I describe language and speech features and analysis of these characteristics towards automatic classification of utterances based on whether they were spoken by children with ASD or TD. I hypothesize that based on the features extracted from the speech signal I am capable to classify children with ASD and TD on a speech segment level, as well as on narrative level after temporally combining all the segment-based decisions.

Table 9. Description of language and speech features.

Language	**Features**
General descriptor	Words per sentence (WPS) Words with more than 6 letters Occurrences of laughter
Sentence structure	Percentage of pronouns, conjunctions, negations, quantifiers, numbers
Psychological proc.	Percentage of words describing social, affect, cognitive, perceptual, and biological
Personal concerns	Percentage of words describing work, achievement, leisure, and home
Paralinguistic	Percentage of assent, disfluencies, and fillers
Speech	**Features**
Pitch	Statistics of sd and cov
Intensity	Statistics of sd and cov
Speech rate	Words per voiced second
Voice quality	Amplitude of a3 Difference of the h1 and the h2 Difference of the h1 and the a3

6.3.1 Feature Extraction

I extract language and speech features based on those proposed by [86]. Extracted features are summarized in Table 9. I also add one feature not covered in previous work counting the number of occurrences of laughter.

6.3.2 Language Features

I use the linguistic inquiry and word count (LIWC) [88], which is a tool to categorize words, to extract language features. Because a Japanese version of

LIWC is not available and there is no existing similar resource for Japanese, I implement the following procedures to automatically establish correspondences between LIWC categories and transcribed Japanese utterances. First, I use Mecab[11] for part-of-speech tagging in Japanese utterances, translate each word into English using the WWWJDIC[12] dictionary, and finally determine the LIWC category corresponding to the English word. Among the language features described in Table 9, I calculate sentence structures, psychological processes, and personal concerns using LIWC, and other features using Mecab. Here, I do not consider language-dependent features and subcategories of LIWC.

6.3.3 Speech Features

For speech feature extraction, I use the Snack sound toolkit [70]. Here, I consider fundamental frequency, power, and voice quality, which are effective features according to previous works [76]. I do not extract mean values of fundamental frequency and power because those features are strongly related to individuality. Thus, I extract statistics of standard deviation (fsd, psd) and coefficient of variation (fcov, pcov) for fundamental frequency and power. I calculate speech rate, which is a feature dividing the number of words by the number of voiced seconds. Voice quality is also computed using: the amplitude of the third formant (a3), the difference between the first harmonic and the second harmonic (h1h2), and the difference between the first harmonic and the third formant (h1a3).

6.3.4 Projection Normalization

For normalization, I simply project all feature values to a range of [0, 1], where 0 corresponds to the smallest observed value and 1 to the largest observed value across all utterances. For utterance i, I define the value of the jth feature as v_{ij} and define $p_{ij} = \frac{v_{ij} - min_j}{max_j - min_j}$, where p_{ij} is the feature value after normalisation.

[11] https://code.google.com/p/mecab/

[12] http://www.edrdg.org/cgi-bin/wwwjdic/wwwjdic?1C

Table 10. Difference of mean values between ASD and TD based on language and speech features from children's utterances. Each table cell notes which of the two classes has the greater mean on the corresponding feature (*: $p < .05$, **: $p < .01$).

WPS	**6 let.**	**laughter**	**adverb**	**pronoun**	**conjunctions**	**negations**	**quantifiers**	**numbers**	**social**
-	ASD*	-	-	-	-	-	-	-	TD**
affect	**cognitive**	**perceptual**	**biological**	**relativity**	**work**	**achievement**	**leisure**	**home**	**assent**
TD**	TD*	-	-	-	-	-	-	-	ASD**
nonfluent	**fillers**	**fsd**	**fcov**	**psd**	**pcov**	**speech rate**	**a3**	**h1h2**	**h1a3**
-	ASD*	TD**	TD*	-	-	-	-	-	ASD**

6.3.5 Characteristics of Language and Speech Features

In this section, I report the result of a *t*-test, principal component analysis, factor analysis, and decision tree using the normalised features. I use R[13] for statistical analysis.

Table 10 shows whether utterances of children with ASD or TD have a greater mean on the corresponding feature. The results indicate that the children with ASD more frequently use words with more than 6 letters (e.g. complicated words), assent (e.g. "uh-huh," or "un" in Japanese), and fillers (e.g. "umm," or "eh" in Japanese) significantly more than the children with TD. In contrast, the children with TD more frequently use the words words categorized as social (e.g. friend), affect (e.g. enjoy), and cognitive (e.g. understand) significantly more than the children with ASD. In addition, there are differences in terms of fundamental frequency variations and voice quality (e.g. h1a3). In particular, I observe that the children with ASD tend to use monotonous intonation as reported in [34]. I do not confirm a significant differences in other features.

Next, I use principal component analysis and factor analysis to find features that have a large contribution based on large variance values. As a result of principal component analysis, features about fundamental frequency, power, and h1a3 have large variance in the first component, and the feature counting perceptual words also has large value in the second component. To analyze a different

[13] http://www.r-project.org

aspect of principal component analysis with rotated axes, I use factor analysis with the varimax rotation method. Figure 22 shows the result of factor analysis indicating that features regarding fundamental frequency and power have large variance. In addition, other features such as speech rate, a3, and h1a3 also have large variance. Here, I can see that for features such as statistics of fundamental frequency (fsd and fcov) and power (psd and pcov), the correlation coefficient between these features are over 80% ($p < .01$). For correlated features, I use only standard deviation in the following sections.

I also analyze important features to distinguish between children with ASD and TD by using a decision tree. Figure 23 shows the result of a decision tree with 10 leaves indicating that speech features fill almost all of the leaves (e.g. fsd is a most useful feature to distinguish between ASD and TD). In terms of the language features, I confirm that WPS and perceptual words are important for classification.

6.3.6 Classification

In this section, I examine the possibility of automatic identification of whether an utterance belongs to a speaker with ASD or TD. Based on the previous analysis, I prepare the following feature sets: 1) language features (Language), 2) speech features (Speech), 3) all features (All), 4) important features according to the t-test, principal component analysis, factor analysis, and decision tree (Selected), 5) important features according to the t-test that are not highly correlated (T-Uncor). The feature set of T-Uncor is as follows: 6 let., social, affect, cognitive, fillers, assent, fed, and h1a3. I also show the chance rate, which is a baseline of 50% because the number of utterances in each group is the same, and measure accuracy with 10-fold cross-validation and leave-one-speaker-out cross-validation using naive Bayes (NB) and support vector machines with a linear kernel (SVM). In the case of leave-one-speaker-out cross-validation, I use T-Uncor because the number of utterances without one speaker is too small to train using high dimensional feature sets.

Table 11 shows the result indicating that accuracies with almost all feature sets and classifiers are over 65%. The SVM with Selected achieves the best performance for the task of 10-fold cross-validation, and the SVM with T-Uncor

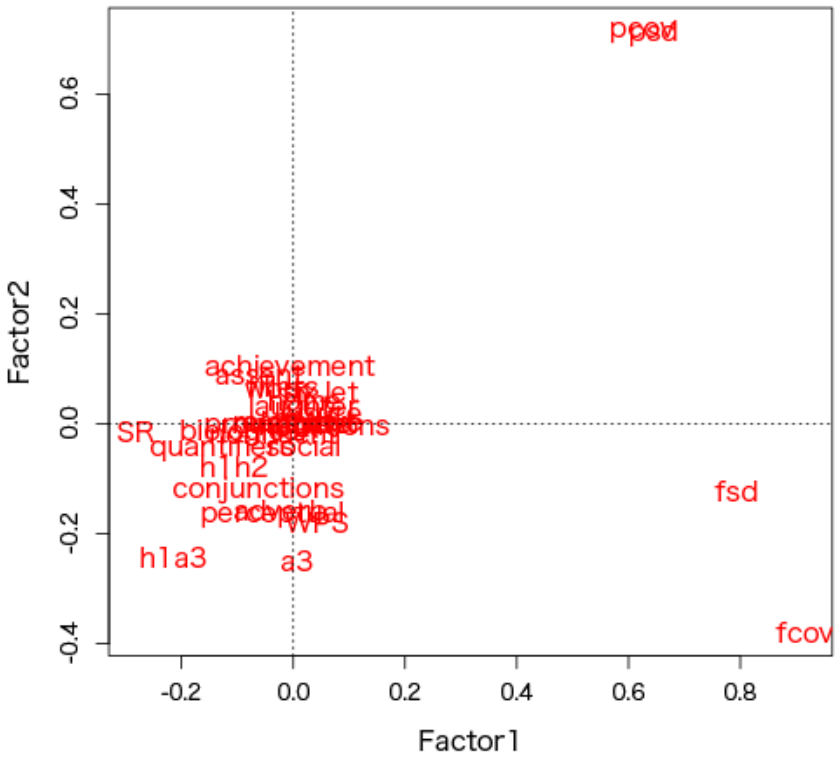

Figure 22. Factor analysis with varimax rotation method. First and second factors are indicated.

achieves 66.7% for the task of leave-one-speaker-out. The accuracy for the task of leave-one-speaker-out on each speaker A1 to T2 is as follows: 78%, 60%, 53%, 51%, 82%, and 78%.

6.4 Narrative Level

In this section, I focus on the features of entire narratives, which allows us to examine other features of child-parent interaction for a better understanding of ASD and classification in children with ASD and TD. Each following subsection describes the procedure of feature extraction and analysis of characteristics at the narrative level. I consider pauses before new turns and unexpected words, which are mentioned in previous works, as well as words per minute.

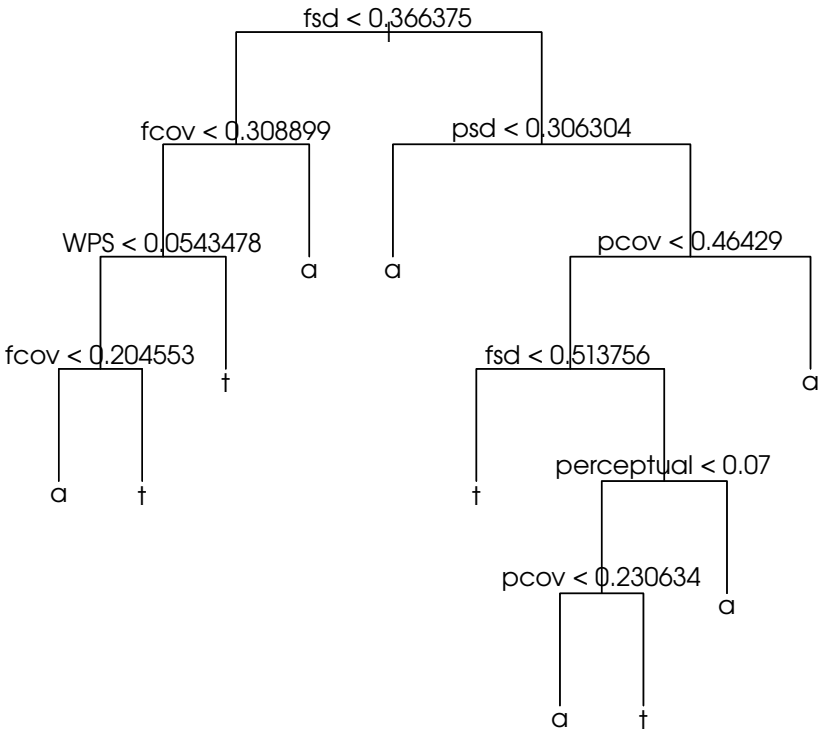

Figure 23. Decision tree with 10 leaves (a: ASD, t: TD).

Table 11. Accuracy using Naive Bayes and SVM classifiers. The p-value of the t-test is measured compared to baseline (chance rate) (†: $p < .1$, *: $p < .01$)

Feature set	Accuracy [%]		
	Baseline	**NB**	**SVM**
Language	50.0	62.2†	70.3*
Speech		57.6	67.6*
All		65.0†	68.8*
Selected		67.4*	71.9*
T-Uncor		67.8†	68.1†
Per-Speaker	50.0	65.5†	66.7†

6.4.1 Pauses Before New Turns

Heeman et al., [73] reported that children with ASD tend to delay responses to their parent more than children with TD in natural conversation. In this chapter, I examine whether a similar result is found in interactive narrative. I denote values of pauses before new turns as time between the end of the parent's utterance and the start of the child's utterance. I do not consider overlap of utterances. I test goodness of fit of pauses to a gamma and an exponential distribution based on [89], because the later is a special case of gamma with a unity shape parameter, using the Kolmogorov-Smirnov test.

Figure 24 shows a fitting of pauses to gamma or exponential distributions, and I select a better fitted distribution. All subjects significantly fit ($p > .6$). As shown in Figure 24, I confirm that children with ASD tend to delay responses to their parent compared with children with TD. To reflect this information in my following experiments in automatic identification of ASD in narrative, I extract the expectation value of the exponential distribution

Heeman *et al.,* [73] also reported the relationship of the parent's previous utterance's type (question or non-question) and the child's pauses. I examine the relationship between the parent's previous question's type and pauses before new turns. For each of the children's utterances, I label the parent's utterance

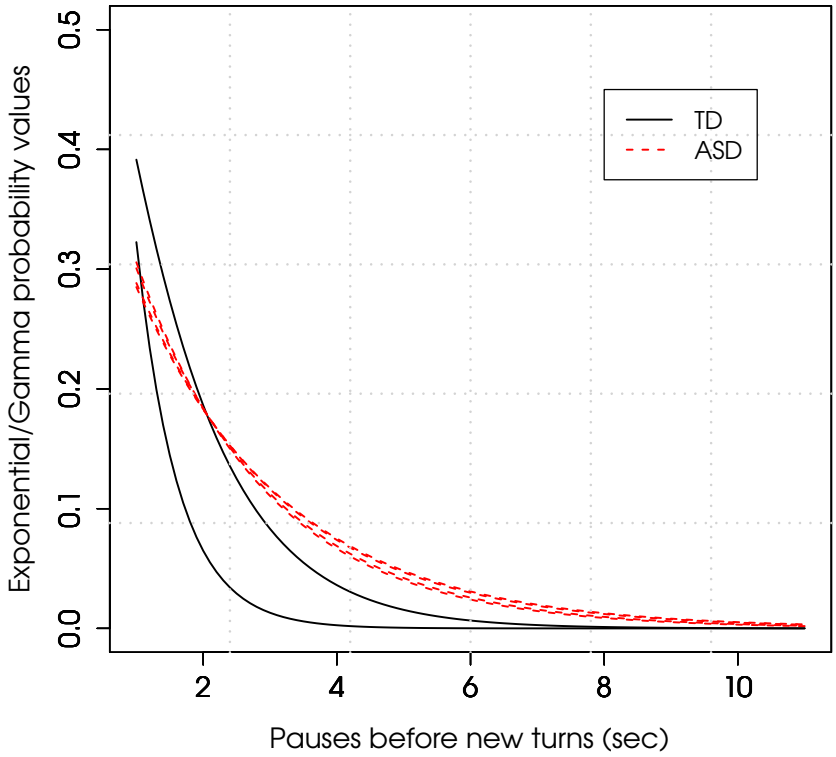

Figure 24. Gamma/Exponential pause distributions with parameters computed using Maximum Likelihood Estimation (MLE) for children with ASD and TD.

that directly precedes as either "open question," "closed question," or "non-question", and I calculate pause latency. Closed-questions are those which can be answered by a simple "yes" or "no," while open-questions are those which require more thought and more than a simple one-word answer. As shown in Table 12, children with ASD tend to delay responses to their parent to a greater extent than children with TD. I found no difference between open and closed questions, although a difference between questions and non-questions is observed. These results are consistent with those of previous work [73] in terms of differences between questions and non-questions.

Table 12. Relationship of pauses before new turns and parents' question types. The mean value and standard deviation are shown.

Question type	TD	ASD
Closed-question	0.47 (0.46)	1.61 (1.87)
Open-question	0.43 (0.34)	1.76 (1.51)
Non-question	0.95 (1.18)	2.60 (3.64)

Table 13. Mean value of words per minute.

Subj.	Averaged WPM
A1	18.25
A2	86.75
A3	23.75
A4	115.5
T1	99.25
T2	103.5

6.4.2 Words Per Minute

I analyze words per minute (WPM) in children with ASD and TD to clarify the relationship between ASD and frequency of speech. I use a total of 5 minutes of data in each narrative, and thus the total number of words are divided by 5 to calculate WPM. Table 13 shows the result. The data in this table indicates that some children with ASD have a significantly lower speaking rate than others with TD, but it is not necessarily the case that ASD will result in a low speaking rate such as the case of Asperger's syndrome [90].

6.4.3 Unexpected Words

Characteristics of ASD include deficits in social communication, and these deficits affect inappropriate usage of words [80]. I evaluate these unexpected words using two measures, term frequency-inverse document frequency (TF-IDF) and log odds ratio. I use the following formulation to calculate TF-IDF for each child's narrative i and each word in that narrative j, where c_{ij} is the count of word j in narrative i. f_j is the number of narratives from the full data of child narratives containing that word j, and D is the total number of narratives [80].

$$tf - idf_{ij} = (1 + \log c_{ij}) \log \frac{D}{f_j}$$

The log odds ratio, another measure used in information retrieval and extraction tasks, is the ratio between the odds of a particular word, j, appearing in a child's narrative, i. Letting the probability of a word appearing in a narrative be p_1 and the probability of that word appearing in all other narratives be p_2, I can express the odds ratio as follows:

$$\text{odds ratio} = \frac{\text{odds}(p_1)}{\text{odds}(p_2)} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

A large TF-IDF and log odds score indicates that the word j is very specific to the narrative i, which in turn suggests that the word might be unexpected or inappropriate. In addition, because the overall amount of data included in the narratives is too small to robustly analyze these statistics for all words, I also check for the presence of each word in Japanese WordNet[14] and determine that if it exists in WordNet it is likely a common (expected) word. Table 14 shows the result of TF-IDF, log odds ratio, and their summation, and I confirm that there is no difference between children with ASD and TD. This result is different from that of previous work [80]. The children in the previous study were all telling the same story, and one possible explanation for this is due to the fact that in this work I do not use language-constricted data such as narrative retelling, and thus differences due to individuality are more prevalent.

[14] http://www.omomimi.com/wnjpn/

Table 14. TF-IDF, log odds ratio, and their summation.

Subj.	TF-IDF	Log-odds	T+L
A1	0.50	1.01	1.52
A2	0.58	0.49	1.08
A3	0.66	1.23	1.89
A4	0.66	0.31	0.96
T1	0.74	0.49	1.23
T2	0.62	0.44	1.06

6.4.4 Classification

In this section, I examine the possibility of automatic classification of whether an interactive narrative belongs to children with ASD or TD. Because the total number of subjects is small (n=4 for ASD, n=2 for TD), I perform classification with a K-NN classifier with K=1 nearest neighbour. As features, I compute the features mentioned in Section 6.3.1, and use the average over all utterances as the features for the entire narrative. Finally, I use pauses before new turns (expectation value of the exponential distribution), WPM, TF-IDF, log odds ratio, 6 let., social, affect, cognitive, assent, fillers, fsd, h1a3, and calculate accuracy with leave-one-speaker-out cross-validation.

As a result, I achieved an accuracy of 100% in classification between ASD and TD on the full-narrative level, which shows that these features are effective to some extent to distinguish children with ASD and TD. However, with only a total of 6 children, my sample size is somewhat small, and thus experiments with a larger data set will be necessary to draw more firm conclusions.

6.5 Comparison of American Data and Japanese Data

As all my preceding experiments have been performed on data for Japanese child-parent pairs, it is also of interest to compare these results with data of children and parents from other cultures. In particular, I refer to the USC Rachel corpus

Table 15. In the case of USC Rachel corpus, bootstrap on difference of means between short (S) and long (L) pauses based on linguistic features from child's and parent's utterances (†: $p < .1$, *: $p < .01$). Each table cell notes which of the two types of pauses has greater mean on the corresponding feature.

Subj.	Child				Parent		
	WPS	conj.	affect	nonflu.	adverb	cogn.	percept.
S1	L*	L*	S*	-	L*	L*	L*
S2	L*	L*	S†	L*	L*	L*	L*
S3	L*	L†	-	S†	L*	L*	L*
S4	-	-	-	L*	L*	L*	L*
S5	L†	-	-	-	L*	L*	L*
S6	L*	-	S*	-	L*	L*	-
S7	L†	-	S†	-	L†	-	-
S8	L*	-	-	-	L*	L*	L*
S9	-	-	-	S†	L*	L*	L*

[87] (the subjects are nine children with ASD) for comparison. Using the USC Rachel corpus, there is a report mentioning the relationship of parent's and child's linguistic information and pauses before new turns [89]. In this chapter, I follow this work using Japanese data. The USC Rachel corpus includes a session of child-parent interaction, and the same transcription standard is used. I extract pauses before new turns, and short and long pauses are differentiated based on the 70th percentile of latency values for each child individually. I investigate the relationship between the parent and child's language information based on features used in Section 6.3.1, and short and long pauses.

Tables 15 and 16 show greater mean values, statistically significant according to bootstrap significance testing, on the means of the two pause types. By observing the values in the table, I can see that the trends are similar for both American and Japanese children. However, in terms of WPS, there is a difference. The American ASD children have greater means for WPS in the case of long pauses, while Japanese children have greater means for WPS in the case of short pauses.

In the Japanese corpus, I observe that WPS is larger in the case of short

Table 16. Bootstrap for pause differences in the Japanese corpus.

Subj.	Child				Parent		
	WPS	conj.	affect	nonflu.	adverb	cogn.	percept.
A1	S*	-	-	-	S*	L*	-
A2	S†	-	S*	-	L*	L*	L*
A3	S†	-	-	-	L*	L*	L*
A4	S*	-	-	-	-	-	-

pauses. As I noticed that the child often utters only a single word for responses that follow a long pause, I analyzed the content of these single word utterances. As shown in Figure 25, for example, A1 tends to use a word related to assent when latency is long, and A4 tends to use a word related to filler, assent or others when latency is long. Though there are individual differences, I confirm that the Japanese children with ASD examined in this study tend to delay their responses before uttering one word. These characteristics may be related to the parent's question types and the child's cognitive process, and thus I need to examine these possibilities in detail.

6.6 Summary

ASD are developmental disorders characterized as deficits in social and communication skills, and they affect both verbal and non-verbal communication.

Previous works measured differences in children with and without ASD in terms of linguistic and acoustic features, although they do not mention automatic identification using integration of these features. In this chapter, we perform an exploratory study of several language and speech features of both single utterances and full narratives. Using narrative data, I examined features mentioned in a number of previous works, as well as a few novel features. I confirmed about 70% accuracy in an evaluation over single utterances, and some narrative features also proved to have a correlation with ASD.

We also examined the differences between American and Japanese children and found significant differences with regards to pauses before new turns and

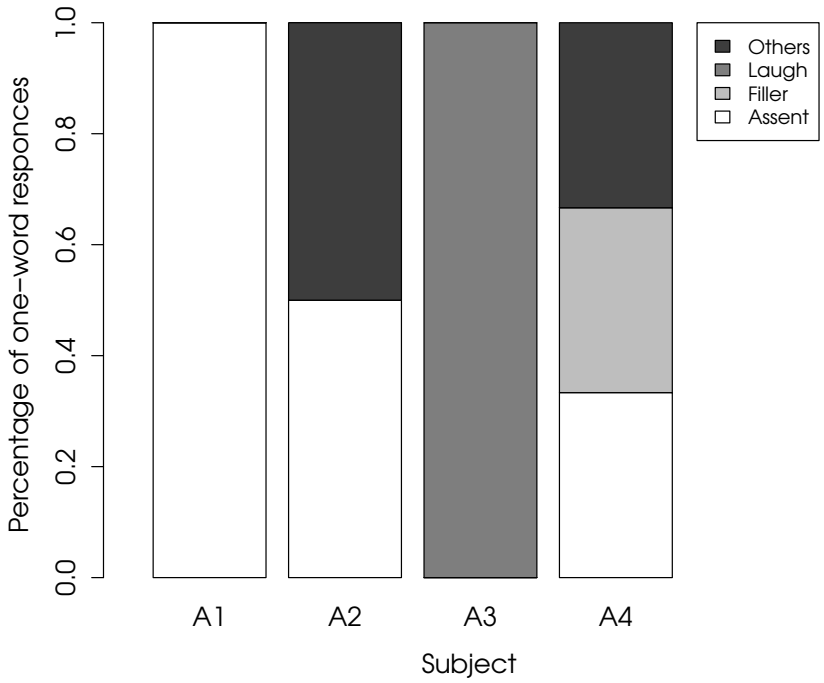

Figure 25. The language category of one-word responses in the case of a long pause.

linguistic cues.

7. Automated Social Skills Trainer

7.1 Introduction

SST is a general cognitive behavior therapy to train social skills for people who have difficulties in social interaction, and is widely used by teachers, therapists, and trainers [11]. However, SST requires well-trained teachers, so the number of participants joining SST programs is restricted and applications are competitive. If part or all of the SST process could be automated, it would become easier for those requiring SST to receive it anywhere and anytime.

There has been one previous work on automated conversational coaches [10], which are dialogue systems aimed to train people for improving interview skills through real-time feature detection and feedback. They achieved 1) a realistic task involving training real users, 2) formative affective feedback that provides the user with useful feedback on the behaviors that need improvement, and 3) the interpretation or recognition of user utterances to drive the selection of backchannels or formative feedback. While this work is an excellent first step, it did not faithfully follow the traditional SST framework, omitting steps such as modeling of human behavior [91].

As conventional SST is a well-established method to improve human behavior and has clear goals and definitions for each module in the framework, my motivation is to follow the traditional SST framework as closely as possible.

I propose a novel tool that tries to replicate conventional SST using a systematic and computer-based design. I develop a dialogue system named "automated social skills trainer," which is an application including video modeling of human behavior and real-time behavior detection as well as data visualization to help people improve their social skills (Figure 26). I investigate whether it is possible to help people who have difficulties in social interaction improve their social skills using an automated system which can be used anywhere, anytime.

This chapter presents the related works, system design, modeling, and evaluation of the automated social skills trainer. Two experimental evaluations show that social skill is related to automatically extracted features and has a relationship to autistic traits, and that some participants improve in social skill using the automated social skills trainer.

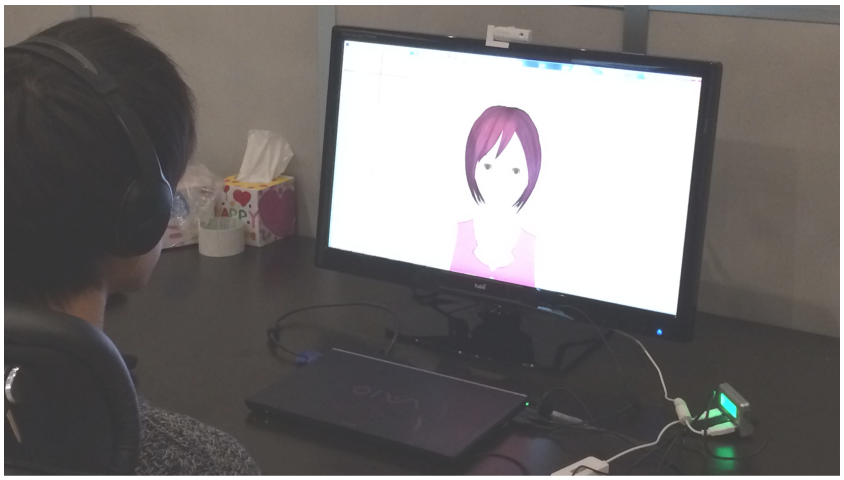

Figure 26. SST with the automated social skills trainer.

7.2 Automated Social Skills Training

In this section, I describe my proposed automated social skills trainer following the conventional individual SST framework. I replicate human-to-human individual SST by using a spoken dialogue system. While one disadvantage of individual SST is that there is no chance to see other participants' behavior, I can provide a surrogate by playing video of others on the computer screen.

Table 17 shows the correspondence between conventional SST and my proposed method. In the following few pages, I describe each proposed module corresponding to a step in conventional SST.

- **Defining target skills:** Ideally, I would define specific target skills for each user through an initial interaction with the system. I plan to tackle this in future work, but for the time being I focused on a single target skill that has been shown to be widely applicable: story-telling or narrative ability. Story telling/narrative is a task of telling memorable stories, and is related to social interaction skills such as presentations and job interviews [9]. It has also been shown useful to distinguish children with ASD and children with typical development. In the step of defining this goal, The system tells users that "This application will help you learn to tell stories well, and after

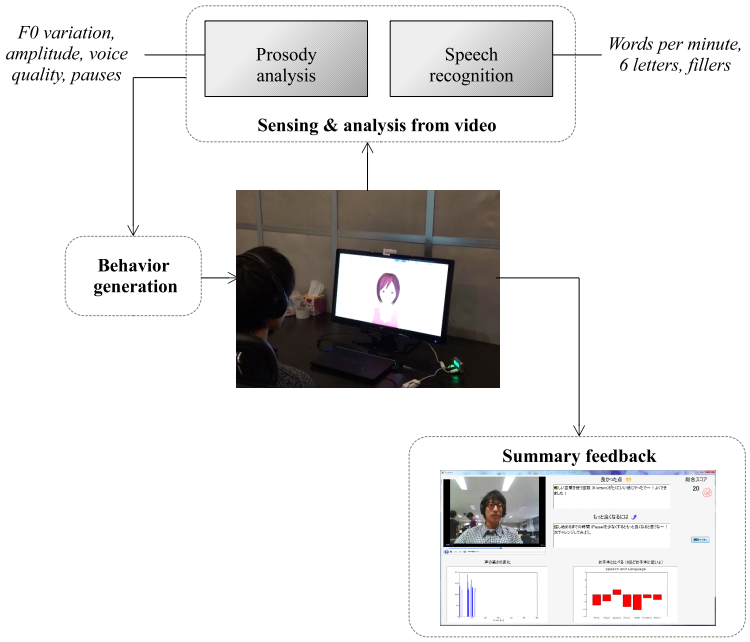

Figure 27. The automated social skills trainer framework.

Table 17. Correspondence of conventional SST and my proposed method.

Conventional SST	Proposed Method
Defining target skills	Story telling/narrative
Modeling	Recorded model video
Role-play	Conversation with an avatar
Feedback	Generating of visual feedback
Reinforcement	Generating of positive comments
Homework	Spoken instructions

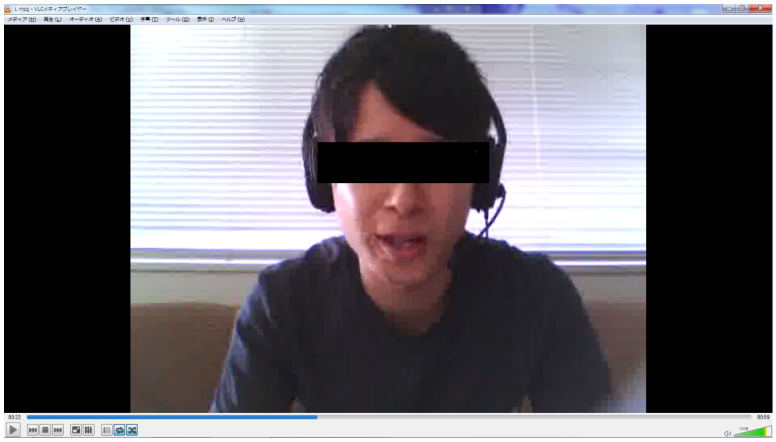

Figure 28. An example of video modeling.

training you will have more fun telling stories."

- **Modeling:** Users can watch a recorded model video (Figure 28) before the role-playing. The recorded models are people who have relatively good narrative skills according to subjective evaluation. Users can watch and imitate the good examples.

- **Role-play:** The main part of the proposed system is the role-playing, which is performed through interaction with an avatar. When the user says "start role-playing," the system says "Please tell me your recent memorable story" The role-playing starts after the avatar's question, and continues for one minute. During this time, the avatar nods its head, and the system detects and analyzes language and speech features automatically. In this work, I focus on features that could differentiate between people with and without ASD as described in previous chapter: F0 variation, amplitude, voice quality, pauses, words per minute, words more than 6 letters, and fillers. I show a list of these features as follows:

  F0 variation: F0 indicates fundamental frequency, or pitch of voice. F0 variation therefore corresponds to the amount of variety in pitch, with less variety corresponding to a more monotone voice. We use these

features because it has been widely noted that people with ASD have prosody that differs from that of their peers [34, 82, 84, 85]. For instance, Kiss *et al.,* [83] found several differences in the fundamental frequency characteristics of people with ASD.

Amplitude: In human-to-human SST, trainers often focus on volume of voice because both overly small and loud voices are not appropriate for many social situations.

Voice quality: People with ASD often exhibit abnormal voice quality, often described as more clear than their peers. Bonneh *et al.,* [79] quantified speech abnormalities in terms of the properties of the voice quality and was able to identify children with ASD with more than 80% accuracy.

Pauses: There are reports finding that children with ASD tend to delay responses to their parent more than children with typical development in natural conversation [73].

WPM: There is a report that speaking rate was strongly correlated to interview skills [10].

Words more than 6 letters: Children with ASD use more complicated or unexpected words than typically developing children, and deficits of ASD affect inappropriate usage of words [80].

Fillers: The frequency of filler usage is important in story telling or presentation. Too frequent use of fillers disturbs listener focus on the contents of speech.

- **Feedback:** The system displays summary feedback according to detected features. The feedback includes comments, the user's video, the parameters compared to model speaker, and the overall narrative score. The user can objectively confirm their strengths and weaknesses.

- **Reinforcement:** In addition to simply listing scores, the system also chooses good parts of the interaction, and gives positive feedback encouraging the targeted behavior.

Figure 29. The avatar used in the automated social skills training system.

- **Homework:** The system tells users to "Please tell your story to others throughout the week, and let me know about it." However, it should be noted that I did not evaluate SST across sessions in this work, though I plan to do so in the future.

Through this framework, I can replicate to some extent conventional individual SST with the spoken dialogue system replicating each module in the framework.

7.3 Implementation Details

The automated social skills trainer system works on a regular laptop, which processes the audio input in role-playing. The processed data is used to generate the behaviors of the avatar that interacts with and provides feedback to users.

The role-playing, feedback, and reinforcement consist of three modules: behavior generation, sensing & analysis from video, and summary feedback as shown in Figure 27. The following subsections describe the modules in detail. It should be noted that the target language of my system is Japanese, and all data creation and experiments are performed with native Japanese speakers.

7.3.1 Data Creation and Subjective Evaluation

As a first step towards building my system, I collected model data of people with relatively high levels of social skills. This video data is used both in the modeling module and for predicting scores in the feedback module. I collected data from a total of 19 people. Using this data, I assigned each dialogue an overall narrative skill score based on subjective evaluation, and the top five people were selected as models. Subjective evaluation was performed by having two raters watch the recorded participants' narrative and answer a questionnaire. This process is described in more detail in the following Experiment 1.

7.3.2 Dialogue Agent

The automated social skills trainer was developed using MMDAgent,[15] which is a Japanese spoken dialogue system integrating speech recognition, dialogue management, text-to-speech, and behavior generation. MMDAgent works as a Windows application. I selected a character who is similar to an actual human, as I hope that this will make it easier for the user to generalize learned skills in a real situation. The avatar is displayed from the front, and there are no distractions in background (Figure 29). The user can operate and interact with the avatar by using speech throughout the training. All dialogue system utterances were created using templates written by the first author.

In addition, the system performs a number of behaviors to keep the user engaged. It blinks its eyes once every three seconds, and reacts to users during the role-playing. When the system recognizes an utterance, after a few seconds the system nods its head. The blinking and nodding behavior motions were created by MikuMikuDance.[16]

[15] http://www.mmdagent.jp/

[16] http://www.geocities.jp/higuchuu4/

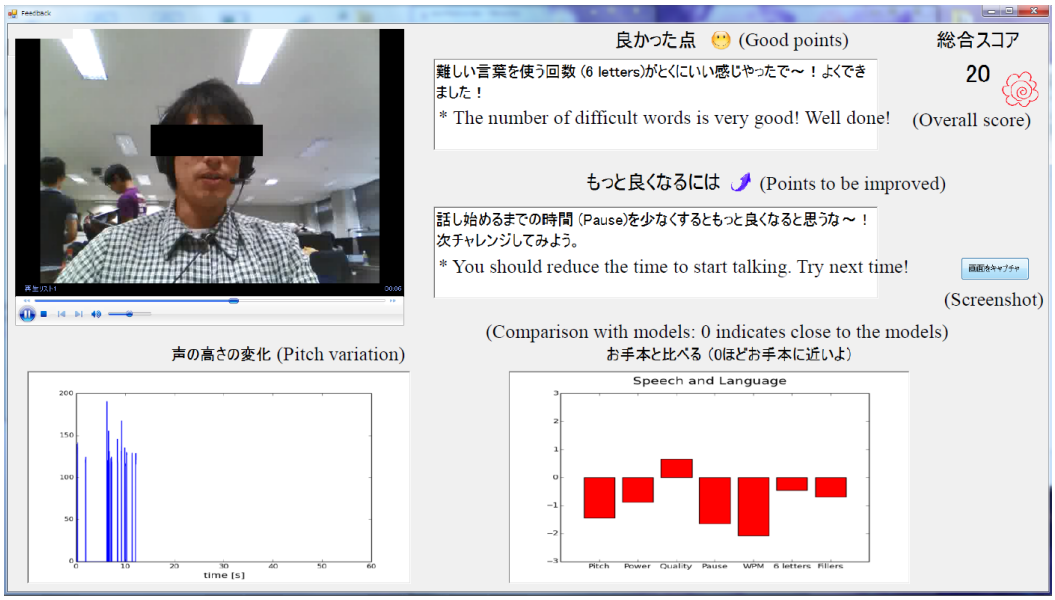

Figure 30. The summary feedback provided by the automated social skills trainer.

7.3.3 Sensing and Analysis from Video

To calculate the linguistic related features, I performed automatic speech recognition (ASR) using the Julius dictation kit.[17] I used Mecab[18] for part-of-speech tagging in Japanese utterances. For speech feature extraction, I used the Snack sound toolkit.[19]

I describe the implementation of features as follows: In this chapter, I used the coefficient of variation for fundamental frequency with a minimum pitch of 100 Hz. I did not use mean, maximum and minimum values because there are individual and gender differences in terms of these features. I used the mean value of amplitude. I extracted the spectral tilt by calculating the difference between the first harmonic and the third formant (h1a3) as a feature expressing voice quality. I calculated values of pauses before new turns as time between the end of the avatar's utterance and the start of the user's utterance. I analyzed words per

[17] http://julius.sourceforge.jp/index.php

[18] https://code.google.com/p/mecab/

[19] http://www.speech.kth.se/snack

minute (WPM) which is related to frequency of speech. In the automated social skills trainer, the narrative continues for one minute, and I counted the number of words in one narrative. Words more than six letters may be related to complicated words [88], and thus I extracted these as a feature. I calculated percentage of fillers such as "umm," or "eh" in Japanese. This feature automatically extracted by using the output of Mecab.

7.3.4 Summary Feedback

Based on the calculated features, I provide feedback to the users about their social skills (Figure 30). My goal was to design visualizations so that it would be easy for users to understand and interpret their narrative skill. The summary feedback provides following information.

- **User video:** Participants can watch the recorded video and audio in the narrative. In doing so, the user can confirm their speech contents, facial expression, posture and so on [10], which are not analyzed automatically in the current version of the automated social trainer.

- **Overall score:** The system displays the predicted overall score, which motivates the user to practice more and improve their score. I predict the overall score using the multiple regression method on a scale of 0 to 100.

- **Pitch variation:** Participants can see their pitch movement corresponding to the time. This also shows a visualization of how frequently they spoke.

- **Comparison with models:** The system visualizes the comparison of extracted features between the user's current narrative and model persons' narratives in terms of z-score, which is a statistical measurement of a score's relationship to the mean in a group of scores. The users are informed that they should attempt to emulate the model in all aspects.

- **Good points:** The system generates positive comments that reinforce the user's motivation with encouraging words [11]. The comments are generated based on the features that have values close to those of the models.

- **Points to be improved:** The system generates comments about points to be improved for next trial. The comments are generated based on the features that have values that differ from the models.
- **Screenshot:** Participants can save the feedback by clicking a button, and this is used for checking improvements over the course of training.

7.4 Experiment 1: Defining Model Persons

To evaluate the effectiveness of the proposed automated social skills trainer, I performed two experiments. In the first experiment, I sought to answer the following questions:

1) Does narrative skill relate to linguistic, acoustic, and other information?
2) Is there a difference between talking to humans (human-human interaction: HHI) and talking to avatars (human-computer interaction: HCI) in terms of narrative?
3) Does narrative skill relate to autistic traits?
4) Are the extracted features effective for identifying narrative skill?

The result of first experiment is used in the data collection and summary feedback of the automated social skills trainer.

7.4.1 Procedure

I recruited 19 graduate students (16 males and 3 females), all of whom were native Japanese speakers.[20] All subjects used the proposed system and were told that their speech and video would be recorded. A webcam (ELECOM UCAM-DLY300TA) placed on top of the laptop and headset (ELECM HS-HP168K) recorded the video and audio of participants. I recorded not only HCI but also

[20] Note that the Research Ethic Committee of my institution has reviewed and approved both this and the following experiments. Written informed consent was obtained from all subjects before the experiments.

an HHI setting in which the first author listened to speaker's story and nodded his head according to the speaker's utterances.

To get a grasp of each subject's social skills independent of the proposed system, or the narrative setting in general, I also administered a social skills test for each subject. Specifically, I measured the sum of subareas score for communication and social skills of Japanese version of the AQ [17, 18] which is a standard tool to measure autistic traits with a total of 50 questions including 5 subareas.

Next, I had raters watch the interactions of each participant and rate their narrative skill. Although it would be ideal for raters to be professional social skills trainers, they are few and far between, so it is difficult to recruit them for the experiment. Thus, as a proxy, we selected raters from members of the general population. Because raters are required to have good social skills to recognize users' non-verbal expressiveness, we selected two people (male and female) with good social skills as annotators. Specifically, the annotators were selected to have low sums of the AQ subarea scores for communication and social skills (where lower indicates better social skills). The sums of both areas were 1 and 4, which is lower than the mean value of 7.6 for Japanese students [18]. The raters did not participate in the experiments as subjects. The raters did not know the recorded participants, and were trained by rating several examples prior to the evaluation. Two raters watched recorded participants' narrative for both HHI and HCI, and answered a questionnaire,[21] which is based on [10]. The questionnaire included the following items related to the participant's overall narrative performance and use of non-verbal cues such as intonation, amplitude, and lexicon usage, rated on a scale of 1 (not good, not appropriate, or small (few)) to 7 (good, appropriate, or large (frequent)).

Q1. Overall narrative skill

Q2. Concentration

Q3. Friendliness

[21] https://docs.google.com/forms/d/1AQRc1sAQQooEt7zY89H7aJQz KFf8zqGH4u-nCCVwnGs/viewform?c=0&w=1&usp=mail_form_link

Q4. Attractiveness

Q5. Speaking rate

Q6. Usage of fillers

Q7. Intonation

Q8. Voice quality

Q9. Amplitude

Q10. Usage of easy words

7.4.2 Agreement

The two raters' agreement was measured by Cohen's kappa-coefficient. The two raters answered the ten questions for each speaker's narrative. We analyzed agreement for the question of overall narrative skills. For each rater, each subject was assigned a class of being either above or below the average score for the rater, and agreement between the classes was used to calculate the coefficient. The Kappa coefficient of two classes for two raters was 0.580, which corresponds to moderate agreement according to the scale proposed by [68].

7.4.3 Correlation between Questions

Figure 31 shows that the correlation matrix of each question. For Q6, because usage of fillers can be assumed to be inversely proportional to social skill, I inverted the ratings before measuring correlation. The result showed that questions especially asking about the speech and language features were significantly related to overall narrative skill. On the other hand, questions asking about concentration were not related to overall skills and other features. We can also see that questions related to speech and language features, excluding Q5, were correlated each other.

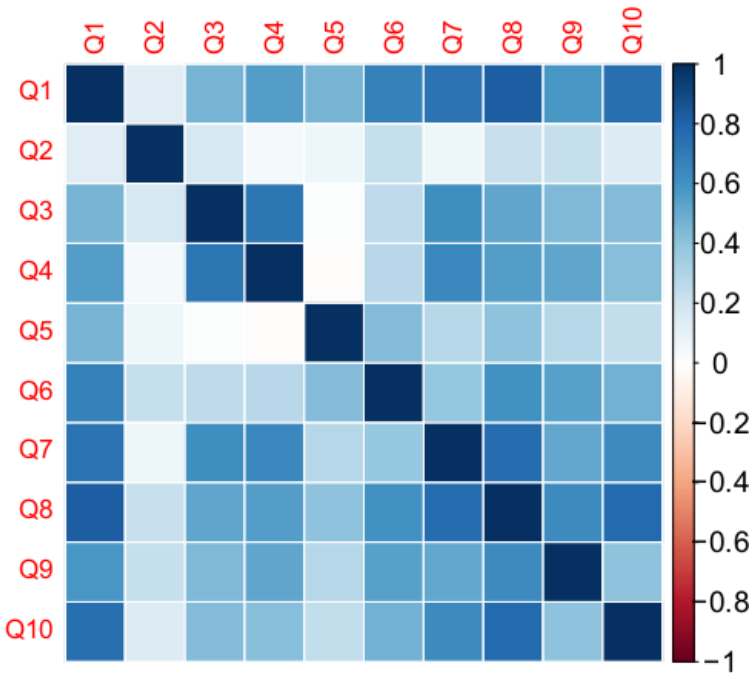

Figure 31. Pearson's r correlations between various questions. Color indicates the strength of statistically significant correlations, and white indicates zero. Rows and columns represent the questions in the same order, so the diagonal is self-correlation.

7.4.4 Differences between Human and Computer Interaction

In this section I examine the difference between HHI and HCI. Averaged rater scores for HHI and HCI are shown in Figure 32. We can see that there were differences between HHI and HCI, and raters' scores of narrative skill in HHI were slightly higher than HCI. However, I did not find a statistical difference ($p > .05$) by Student's t-test. It is likely that if differences exist between interaction with my proposed system and interaction with an actual human in terms of overall narrative skills, they are small.

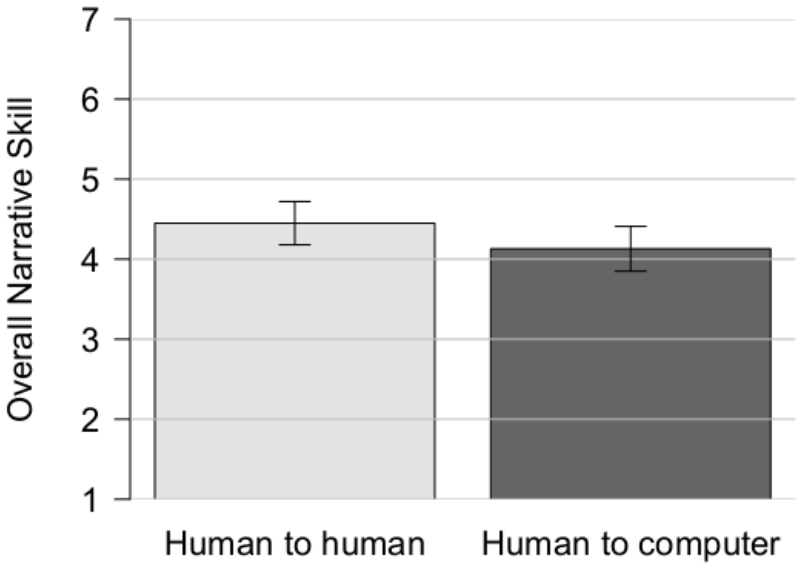

Figure 32. The difference of raters' scores between HHI and HCI. Error bars indicate standard error.

7.4.5 Model People and Autistic Traits

Based on the raters' scores, I determined the top 5 of 19 subjects to be my models for additional experiments including the modeling step of SST. As shown in Figure 33, the median value of the AQ was 1 in the case of model persons, and 13 in the case of the others. This indicates that there is also a strong relationship between the raters' assessment of narrative skill and the subjects' answers on the AQ test.

7.4.6 Regression

Finally, I calculated the statistical differences of the automatically extracted features between model persons and others using Student's t-test. I found that

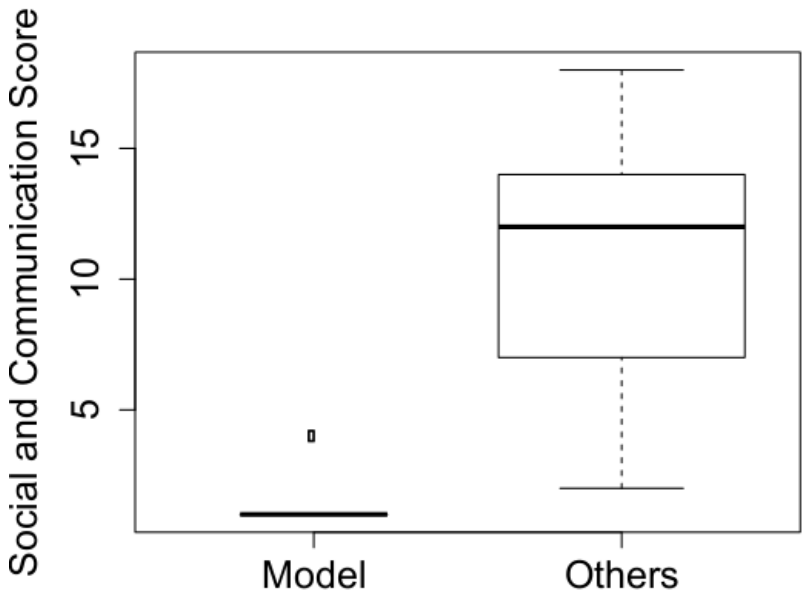

Figure 33. The ranges of the AQ for model persons and others. Zero indicates high social and communication skills, and 20 indicates low social and communication skills.

WPM, words of more than 6 letters, and amplitude were significantly different between the groups ($p < .05$), and other features were not significantly different ($p > .05$). Thus I used these three features to predict overall narrative skill using the multiple regression method. Finally I found that the correlation between the predicted value and actual value was 0.51 ($p < .05$). This regression model was integrated into the system as the feedback module's overall score.

7.5 Experiment 2: Social Skills Training

In the second experiment, I examined whether the automated social skills trainer is effective to train social skills, specifically:

1) How effective is the automated social skills trainer in helping users improve their narrative skills?

2) Do users find the automated social skills trainer easy to use and helpful?

7.5.1 Procedure

I recruited a total of 30 graduate students (22 males and 8 females) all of whom were native Japanese speakers, different from those who participated in the first experiment. Participants first entered the experiment room, and were given instructions by the first author. All subjects were told that their speech and video would be recorded. The webcam (ELECOM UCAM-DLY300TA) placed on top of the laptop and headset (ELECM HS-HP168K) recorded the video and audio of participants.

I separated participants into 3 groups: the reading book group, the video modeling group, and the feedback group. The reading book group, which serves as a control, read two types of social skills books which were related to story telling/narrative skills. The book titles were "Social skills training: collection of cases" and "The easiest guide to presentations" (in Japanese). The video modeling group and the feedback group used the automated social skills trainer for their training. The video modeling group only watched the model videos, while the feedback group performed role-play and received automated feedback. The feedback group can also watch the model videos. As shown in Figure 34, all subjects spoke their narratives to the agent (pre), received training for 50 minutes, and spoke their narratives to the agent again (post).

The same two raters from the first experiment evaluated the subjects' narrative skill by answering overall narrative skill (Q1 of the first experiment) rated on a scale of 1 to 7. Raters did not know subjects, and the order of the pre- and post-training narratives was randomized to prevent bias. I averaged the two raters' scores and calculated improvement in score (post - pre) for each group.[22] Note that the initial scores of each group were not significantly different ($F[2,25]=0.90$, $p > .05$).

[22] Among the 30 participants, one subject each in the reading book and feedback group did not have sufficient time to train for 50 min, so I omitted these subjects.

The effect of intervention type was analyzed using one-way analysis of variance (ANOVA). Post-hoc comparisons between the feedback group and the reading book group, and the feedback group and the video modeling group involved Bonferoni's method.

After using the automated social skills trainer, the feedback group answered a questionnaire to evaluate usability and effectiveness of the system.[23] The questionnaire included the following items related to the system usability and training effect, rated on a scale of 1 (disagree) to 7 (agree). The users were also asked to provide comments about each question.

Q1. The system was easy to use.

Q2. I would like to use this system frequently.

Q3. The trainer looks like a human.

Q4. Watching my own video and feedback were useful.

Q5. Watching model video was useful.

7.5.2 Agreement

I calculated agreement according to the same procedure described in the previous section. The Kappa coefficient of two classes for two raters was 0.638, which indicates good agreement based on [68]. The agreement of the two raters was almost the same as the first experiment.

7.5.3 Training Effect

Figure 35 shows the improvement of overall narrative skills in each group. These results show that intervention type significantly affected the change in raters' scores ($F[2,25]=4.67$, $p < .05$) according to ANOVA. Comparisons showed that the change in raters' scores of participants in the feedback group who used the automated social skills trainer was significantly higher than the reading book

[23] https://docs.google.com/forms/d/1Qhe1UvXrZlHvOfY5YewQD1d2pwue5APhrNiLHl3Iasc/viewform?c=0&w=1

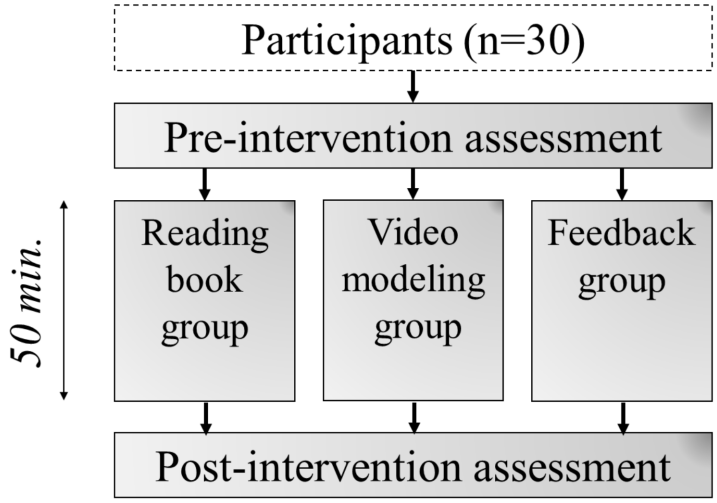

Figure 34. Study design and participant assignment to experimental groups in the second experiment.

group ($p < .05$). The difference between the video modeling group and the feedback group, and the video modeling group and the reading book group was not judged as statistically significant ($p > .05$).

Figure 36 shows the improvement of overall narrative skills and initial scores in each group. The correlation coefficient between overall narrative skills prior to training and improvement was -0.438 ($p < .05$) showing a weak negative correlation. This is a natural result, because people who have difficulties in social interaction have more space to improve.

7.5.4 Subjective Evaluations

The paragraphs below describe findings from the participants' subjective evaluations of the automated social skills trainer and their feedback on their experience. I analyzed qualitative and quantitative results to represent user experience and system usability.

- **The system was easy to use:** The usability of the automated social skills trainer was rated an average of 5.4 (SD = 0.9). Most participants

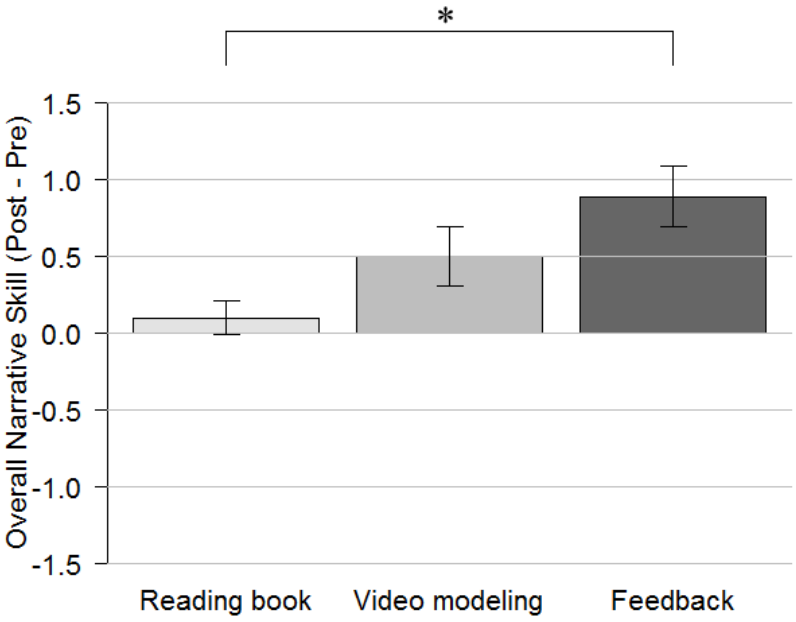

Figure 35. The overall narrative score of each group. Error bars indicate standard error (*: $p < .05$).

found the system is easy to use.

"It is easy to operate the system using only speech. My voice was recognized and I felt comfortable."

"The content of training was separated according to purpose (e.g. modeling, feedback, and homework), and it was easy for me."

- **I would like to use this system frequently:** The question regarding whether the user would like to use the system again was rated an average of 5.0 (SD = 0.7). Most participants would like to use the system frequently.

  "I would like to talk to system with more variation. I want to use this system every day, and also record a life log."

  "It is interesting to watch my score with helpful comments."

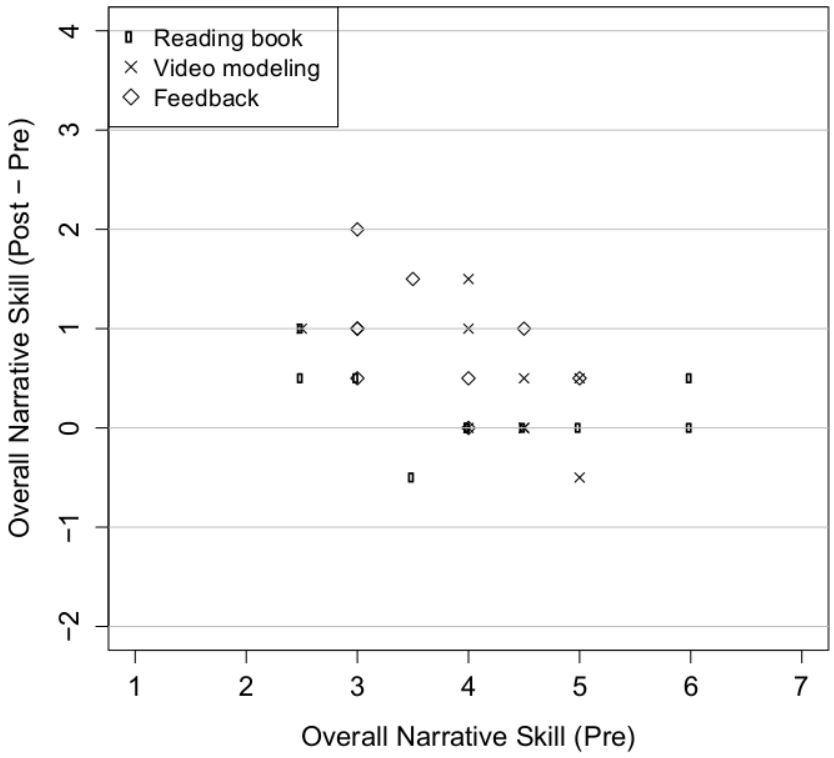

Figure 36. The relationship between initial and improvement in scores.

- **The avatar looks like a human:** The question about whether the avatar looks like a human was rated an average of 4.8 (SD = 0.4). Some participants thought the character looks like a human.

  "I thought avatar's behavior is natural, and I did not feel unnaturalness in interaction."

  "I felt like I spoke to a real human."

  However, some participants thought the avatar did not seem like a human specifically in terms of speech synthesis.

"I felt the synthesized speech is robot-like, the intonation was unnatural."

- **Watching my own video and feedback were useful:** According to the participant's responses to the questionnaire, the feedback and watching the user's own video was rated an average of 5.6 (SD = 1.1). Most participants thought the feedback and video were useful.

  "It was easy to train my skill because the system indicated the points to be improved."

  "I was happy to be encouraged."

  "Conversation is abstract, but the system displayed the concrete values. It is very interesting and helpful."

- **Watching model video was useful:** Overall, participants rated their preference toward watching the models' video an average of 5.2 (SD=1.5), suggesting the usefulness of model video.

  "After watching the role model, I easily started to talk because of the good reference."

  "I was interested in the variation of the good examples."

  However, the result also showed the SD value was large. Some participants did not say that model video was helpful.

  "I think I already had good skill so the modeling is not useful for me."

  "I would like to see the good points of the model persons."

7.6 A Case Study

One child (male) with ASD participated in this study and used the automated social skills trainer. A child was recruited from Nara Autism Society. His age was 11 and had a diagnosis of ASD. His developmental quotient was 90 based on the Kyoto scale of psychological development 2001 [92] (normal development of intellectual capacities). He received training as the feedback group of previous experiment.

As a result, his initial score was 1.5 and training score (post) was 2.5. The initial score is low compared to previous results. It is still not clear whether this is because of age, autistic traits or other factors. However, he improved 1.0 in score indicating that the training is helpful for him to enhance social communication skills. I also subjectively observed that he was very interested in using and talking to the system.

7.7 Summary

In this chapter, I developed a dialogue system named "automated social skills trainer" which provides SST in the context of human-computer interaction. The automated social skills trainer is based on conventional SST including defining target skills, modeling, role-play, feedback, reinforcement, and homework. I focus on story telling/narrative skill as a target skill. The system includes several modules: behavior generation, sensing & analysis from recorded video, and summary feedback. In this study, my focus was to identify the how effective the automated social skills trainer and that follows human-to-human SST as closely as possible. To evaluate effectiveness of the automated social skills trainer, I performed two experiments.

In my first experiment, I confirmed that relatedness of overall narrative skills and speech and language information, confirmed that there was no significant difference between HHI and HCI, set model persons according to the evaluation of two raters. However, this result is limited because I tested only one person as a conversant. I also found a relationship between observed narrative ability and AQ. Baron-Cohen and their colleagues reported that the AQ value was widely distributed among the member of general population and it is related to autistic traits [17]. My result showing a relationship between AQ and overall narrative skills was consistent to above report.

In my second experiment, I confirmed a training effect particularly for participants in the case of the feedback group rather than the reading book group. It showed that the system could help people who have difficulties in social interaction improve their social skills. The video modeling group also improved in their scores, which is consistent to the previous work [91]. The video modeling of others was also helpful in SST. In this experiment, I did not set a group that

watched their own video and did not watch the feedback. There is previous work reporting that subjects dislike looking at their own video during interview skill training and the skills did not change [10], so I plan to investigate these elements separately in the future. I also confirmed a weak negative correlation between initial narrative skills and improvements in scores. This shows that training effects are found more strongly in people who have difficulties in social interaction than others. In subjective evaluation, I confirmed most participants of the feedback group were satisfied with the system in terms of usability and the feedback.

8. Conclusion

8.1 Contribution

Social and communication skills are important for human life, and the number of people who have difficulties with these skills have been increasing. On the other hand, previous works imply that people who have problems with communication often have good or superior skills in using computers. Thus, the use of computers to aid people with communication difficulties has flourished in the last decade. However, most of these applications tend to focus on rather specific skills (e.g. emotion recognition from facial images). Meanwhile, there were fewer applications teaching emotional or non-verbal expressiveness. In addition, previous works did not follow conventional human-based SST framework. This thesis proposed computer-based approach to fill the above gap. I proposed several computer applications to help users train the cognitive and the affective skills as well as affective computing techniques to measure these skills.

For the cognitive component, I developed NOCOA, which uses speech modality. I confirmed the relationship between non-verbal recognition skills and the AQ by using speech output with visual hints, and examined prospective intervention through teaching non-verbal information, intention and partner information. I also proposed NOCOA+, which uses multiple modalities to recognize non-verbal behaviors. I confirmed a method for quantifying social communication skills by using these applications and the effect of context and modality differences, and evaluated the effectiveness of training.

For affect recognition, first I automatically classified the types of laughter which are difficult to recognize for people with communication difficulties. Finally I confirmed that speaker tests across different discourse modes achieved approximately 70% classification accuracy of two types of laughter. I also compared narrative stories of children with autism spectrum disorders to those with typical development. I found group differences in speech and language features. For classification, using linguistic cues and prosody, I analyzed the important feature sets and their effects on accuracy. The results suggest that my method can be used to effectively distinguish between autism spectrum disorders and typical development.

I applied affect recognition techniques to human affective skills training. I attempted to automate the process of SST by developing a dialogue system named automated social skills trainer, which aimed to provide the SST through human-computer interaction. The system includes a virtual avatar that recognizes user speech and language information and gives feedback to users to improve their social skills. Its design is based on conventional group or individual SST performed by human participants including defining target skills, modeling, role-play, feedback, reinforcement, and homework. The evaluation showed the system is useful to improve social skills. The system usability was also high according to the subjective evaluation. The automated social skills trainer has a possibility to add different target skills through same SST framework.

As a conclusion, the proposed applications helped people with communication difficulties enhance their social communication skills.

8.2 Future Directions

In this thesis, I proposed a computer-based approach following the conventional SST framework. We evaluated the system in terms of quantifying and improving social and communication skills. Finally, I summarize future directions to examine that this approach is helpful for people with social and communication difficulties.

8.2.1 ASD Recruiting and Larger Experiment

Most of this work implemented experiments with small number of participants, which makes it more difficult to confirm statistically significant effects. The number of participants of people with ASD is also few, and it is not clear that these results will simply apply to people with ASD. Thus, the training effect over a longer period and recruiting special-need populations such as people with ASD should be examined. I plan to perform larger scale experiments to investigate the potential of these effects. In addition, one potential direction for the future is consideration of individual differences as well as relationship between diagnosis and training effect.

8.2.2 Extension of Social Skills

It is also still not clear whether adding other target social skills (e.g. listed in Figure 3) in the automated social skills trainer is beneficial. One possible next step is to integrate skills that "how to interact when the listener looks bored," which is a combination of both cognitive and affective skills, and may be related to presentation or speaking skills.

8.2.3 Multimodality

Multi-modal feature analysis is also important. For example, I plan to incorporate eye-gaze and skin conductance (EDA signal) analysis into the system.

8.2.4 Generalization

I will more thoroughly examine SST from the viewpoint of HHI and HCI. In this thesis, I only examined interaction with one human, and effects of interaction partner is still not clear (e.g. familiar/unfamiliar persons). In addition, how the user can generalize learned skills to the real world is the remaining topic.

8.2.5 Modeling of Human Trainers

The recording of human-based SST is one important next step. I can measure and compare the effect of human-based SST and computer-based SST. The behaviors of professional trainer should be integrated into the automated social skills trainer.

References

[1] Baron-Cohen, S. Autism and Asperger syndrome. Oxford University Press, USA, 2008.

[2] American Psychiatric Association. Diagnostic and statistical manual of mental disorders (5th ed.). Washington, DC, 2013.

[3] Weintraub, K. Autism counts. Nature, vol.479, pp.22-24, 2011.

[4] Davis, H. Measuring individual differences in empathy: Evidence for a multidimensional approach. J. Personality and Social Psychology, vol.44, pp.113-126, 1983.

[5] Ozonoff, S., Miller, N. Teaching theory of mind: A new approach to social skills training for individuals with autism. J. Autism and developmental Disorders, vol.25, pp.415-433, 1995.

[6] Golan, O., Baron-Cohen S. Systemizing empathy: Teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia, Development and Psychopathology, vol.18, pp.591-617, 2006.

[7] Critchley, D., Daly, M., Bullmore, T., Williams, R., Amelsvoort, T., Robertson, M., Rowe, A., Phillips, M., Mcalonan, G. Howlin, P. The functional neuroanatomy of social behaviour. Brain vol.123, pp.2203-2212, 2000.

[8] Ashwin, C., Baron-Cohen, S., Wheelwright, S., Riordan, M., Bullmore, T. Differential activation of the amygdala and the social brain during fearful face-processing in Asperger syndrome. Neuropsychologia, vol.45, pp.2-14, 2007.

[9] Davis, M., Dautenhahn, K., Nehaniv, C., Powell, S. Towards an Interactive System Facilitating Therapeutic Narrative Elicitation. Proc. 3rd Conf. on NILE, 2004.

[10] Hoque, E., Courgeon, M., Mutlu, B., Martin, C., Picard, W. MACH: my automated conversation coach. Proc. 15th Conf. on UbiComp, pp.697-706, 2013.

[11] Bauminger, N. The facilitation of social-emotional understanding and social interaction in high-functioning children with autism: Intervention outcomes. J. Autism and Developmental Disorders, vol.32, pp.283-298, 2002.

[12] Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., Wheelwright, S. The systemizing quotient: an investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. Philosophical Trans. the Royal Society of London Series B: Biological Sciences, vol.358, pp.361-374, 2003

[13] Williams, D. Nobody nowhere. Jessica Kingsley Publishers Ltd., 1992.

[14] Picard, W. Affective Computing. The MIT press, 1997.

[15] Kaliouby, R., Robinson, P. The emotional hearing aid: an assistive tool for children with Asperger syndrome. Universal Access in the Information Society, vol.4, pp.121-134, 2005.

[16] Wing, L. Autistic spectrum disorders. British Medical Journal, vol.312, pp.327-328, 1996.

[17] Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Clubley, E. The Autism-Spectrum Quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. J. Autism and Developmental Disorders, vol.31, pp.5-17, 2001.

[18] Wakabayashi, A. Baron-Cohen, S., Wheelwright, S., Tojo, Y. The Autism-Spectrum Quotient (AQ) in Japan: a cross-cultural comparison. J. Autism and Developmental Disorders, vol.36, pp.263-270, 2006.

[19] Baron-Cohen, S., Wheelwright, S., Stott, C., Bolton, P., Goodyer, I. Is there a link between engineering and autism? Autism London, vol.1, pp.101-109, 1997.

[20] Baron-Cohen, S., Bolton, P., Wheelwright, S., Short, L., Mead, G., Smith, A., Scahill, V. Autism occurs more often in families of physicists, engineers, and mathematicians. Autism vol.2, pp.296-301, 1998.

[21] Wallace, J., Nelson, J., Liberman, P., Aitchison, A., Lukoff, D., Elder, P., Ferris, C. A review and critique of social skills training with schizophrenic patients. Schizophr Bull, vol.6, pp.42-63, 1980.

[22] Durlak, J., Weissberg, R., Dymnicki, A., Taylor, R., Schellinger, K. The impact of enhancing students' social and emotional learning: a meta-analysis of school-based universal interventions. Child development, vol.82, pp.405-432, 2011.

[23] Bishop, J. The Internet for educating individuals with social impairments. J. Computer Assisted Learning, vol.19, pp.546-556, 2003.

[24] Bernard-Opitz, V., Sriram, N., Nakhoda-Sapuan, S. Enhancing social problem solving in children with autism and normal children through computer-assisted instruction. J. Autism and Developmental Disorders, vol.31, pp.377-384, 2001.

[25] Kientz, J., Goodwin, M., Hayes, G., Abowd, G. Interactive Technologies for Autism. Synthesis Lectures on Assistive, Rehabilitative, and Health-Preserving Technologies, 2013.

[26] Golan, O., LaCava, P., Baron-Cohen, S. 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.

[27] Bolte, S., Hubl, D., Feineis-Matthews, S., Prvulovic, D., Dierks, T., and Poustka, F. Facial affect recognition training in autism: can we animate the fusiform gyrus? Behavioral neuroscience, vol.120, pp.211-216, 2006.

[28] Silver, M., Oakes, P. Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others. Autism, vol.5, pp.299-316, 2001.

[29] Tanaka, W., Wolf, M., Klaiman, C., Koenig, K., Cockburn, J., Herlihy, L., Brown, C., Stahl, S., Kaiser, D., Schultz, T. Using computerized games to teach face recognition skills to children with autism spectrum disorder: the lets face it! program. J. Child Psychology and Psychiatry, vol.51, pp.944-952, 2010.

[30] Parsons, S., Mitchell, P. The potential of virtual reality in social skills training for people with autistic spectrum disorders. J. Intellectual Disability Research, vol.46, pp.430-443, 2002.

[31] Parsons, S., Leonard, A., Mitchell, P. Virtual environments for social skills training: comments from two adolescents with autistic spectrum disorder. Computers & Education, vol.47, pp.186-206, 2006.

[32] Schuller, B., Marchi, E., Baron-Cohen, S., O'Reilly, H., Pigat, D., Robinson, P., Davies, I., Golan, O., Fridenson, S., Tal S., Newman, S., Meir, N., Shillo, R., Camurri, A., Piana, S., Stagliano, A., Bolte, S., Lundqvist, D., Berggren, S., Baranger, A., Sullings, N. ASC-Inclusion: Interactive emotion games for social inclusion of children with autism spectrum conditions. In Proc. 2nd International Workshop on Digital Games for Empowerment and Inclusion, 2014.

[33] Ekman, P. Facial expression and emotion. American Psychologist, vol.48, pp.384-391, 1993.

[34] Kanner, L. Autistic disturbances of affective contact. Nervous Child vol.2, pp.217-250, 1943.

[35] Golan, O., Baron-Cohen, S., Golan, Y. 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 and Developmental Disorders, vol.38, pp.1534-1541, 2008.

[36] Golan, O., Baron-Cohen, S., Hill, J., Rutherford, D. The 'reading the mind in the voice' test-revised: A study of complex emotion recognition in adults with and without autism spectrum conditions. J. Autism and Developmental Disorders, vol.37, pp.1096-1106, 2007.

[37] Barett, F., Mesquita, B., Gendron, M. Context in emotion perception. Current Directions in Psychological Science, pp.286-290, 2011.

[38] Brown, J., Bovey, D., Chen, X. Context-Aware Applications: From the Laboratory to the Marketplace. IEEE Personal Communications, vol.4, pp.58-64, 1997.

[39] Ryan, N., Pascoe, J., Morse, D. Enhanced reality fieldwork: the context-aware archaeological assistant. British Archaeological Reports, Oxford, 1998.

[40] Dey, K., Context-Aware Computing: The Cyber Desk Project. AAAI Spring Symposium on Intelligent Environments, Technical Report, pp.51-54, 1998.

[41] Kaliouby, R., Robinson, P., Keates, S. Temporal context and the recognition of emotion from facial expression. In Proc. HCI International Conference, 2003.

[42] Olive, D. Multiple Comparisons Among Means. J. the American Statistical Association, 1961.

[43] Petridis, S., Audiovisual Laughter Analysis. Ph.D. dissertation, University of London, 2011.

[44] Laskowski, K., Burger, S. Analysis of the occurrence of laughter in meetings. Proc. Interspeech, pp.1258-1261, 2007.

[45] Jung, E. The inner eye theory of laughter: Mindreader signals cooperator value. Evolutionary Psychology, vol.1, pp.214-253, 2003.

[46] Vinciarellia, A., Pantic, M., Bourland, H. Social signal processing: Survey of an emerging domain, Image and Vision Computing, vol.27, pp.1743-1759, 2009.

[47] Tanaka, H., Kashioka, H., Campbell, N. Laughter as a gesture accompanying speech - towards the creation of a tool for the support of children on the autistic dimension. In Proc. GESPIN, 2011.

[48] Trouvain, J., Schroder, M. How not to add laughter to synthetic speech. Proc. The Workshop on Affective Dialogue Systems, pp.229–232, 2004.

[49] Owren, J. Understanding Acoustics and function in spontaneous human laughter. Interdisciplinary Workshop on The Phonetics of Laughter, pp.4-5, 2007.

[50] Truong, P., Leeuwen, A., Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features. Interdisciplinary Workshop on The Phonetics of Laughter, pp.49-53, 2007.

[51] Kennedy, S., Ellis, W. Laughter detection in meetings. ICASSP 2004 Meeting Recognition Workshop, pp.118-121, 2004.

[52] Glenn, J. Current speaker initiation of two- party shared laughter, Research on Language & Social Interaction, vol.25, pp.139-162, 1991.

[53] Jefferson, G. A technique for inviting laughter and its subsequent acceptance declination. Everyday language: Studies in ethnomethodology, vol.79, pp.79-96, 1979.

[54] Kangasharju, H., Nikkot, T. Emotions in Organizations Joint Laughter in Workplace Meetings. J. business communication, vol.46, pp.100-119, 2009.

[55] Shimizu, A., Yutaka, K., Makoto, N. Hito wa naze waraunoka [In Japanese], Kodan-sha, 1994.

[56] Bachorowski, A., Owren, J. Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. Psychological Science, vol.12, pp.252-257, 2001.

[57] Hudenko, J., Stone, W., Bachorowski, A. Laughter differs in children with autism: An acoustic analysis of laughs produced by children with and without the disorder. J. Autism and Developmental Disorders, vol.39, pp.1392-1400, 2009.

[58] Laurence, D., Laurence, V. Positive and negative emotional states behind the laughs in spontaneous spoken dialogs. Interdisciplinary Workshop on The Phonetics of Laughter, pp.37-40, 2007.

[59] Bachorowski, A., Smoski, J., Owren, J. The acoustic features of human laughter. Acoustical Society of America, pp.1581-1597, 2001.

[60] Sundaramb, S., Narayanan, S. Automatic acoustic synthesis of human-like laughter. Acoustical Society of America, pp.527-535, 2007.

[61] Campbell, N., Kashioka, H., Ohara, R. No laughing matter. In Proc. Interspeech, pp.465-478, 2005.

[62] Carroll, M., Russel, A. Do facial expressions signal specific emotions? Judging emotion from the face in context. J. personality and social psychology, vol.70, pp.205, 1996.

[63] Gelder, B., Vroomen, J. The perception of emotions by ear and by eye. Cognition & Emotion, vol.14, pp.289-311, 2000.

[64] The Expressive Speech Processing corpus: www.speech-data.jp

[65] Campbell, N. Whom we laugh with affects how we laugh. Interdisciplinary Workshop on The Phonetics of Laughter, pp.61-65, 2007.

[66] Campbell, N. Differences in the speaking styles of a Japanese male according to interlocutor; showing the effects of affect in conversational speech. Differences, vol.12, 2007.

[67] Nishio, S., Koyama, K., Nakamura, T. Temporal differences in eye and mouth movements classifying facial expressions of smiles. In Proc. the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp.206-211, 1998.

[68] Rietveld, T., Hout, R. Statistical techniques for the study of language behavior. Mouton de Gruyter, 1993.

[69] Jokinen, K., Nishida, M., Yamamoto, S. Eye-gaze experiments for conversation monitoring. In Proc. the 3rd International Universal Communication Symposium, pp.303-308, 2009.

[70] Tcl/Tk Snack Toolkit www.speech.kth.se/snack/

[71] Tanaka, H., Campbell, N. Acoustic features of four types of laughter in natural conversational speech. In Proc. ICPHS XVII, pp.1958-1961, 2011.

[72] Provine, R., Yong, L. Laughter: A stereotyped human vocalization. Ethology, vol.89, pp.115-124, 1991.

[73] Heeman, A., Lunsford, R., Selfridge, E., Black, L., Santen, J. Autism and interactional aspects of dialogue. Proc. 11th SIGDIAL, pp.249-252, 2010.

[74] Marchena, A., Eigsti, I. Conversational gestures in autism spectrum disorders: asynchrony but not decreased frequency. Autism Research, vol.3, pp.311-322, 2010.

[75] Dawson, G., Hill, D., Spencer, A., Galpert, L., Watson, L. Affective exchanges between young autistic children and their mothers. J. Abnormal Child Psychology, vol.18, pp.335-345, 1990.

[76] McCann, J., Peppe, S. Prosody in autism spectrum disorders: a critical review. International Journal of Language & Communication Disorders, 2003.

[77] Paul, R., Augustyn, A., Klin, A., Volkmar, R. Perception and production of prosody by speakers with autism spectrum disorders. J. Autism and Developmental Disorders, vol.35, pp.205-220, 2005.

[78] Asgari, M., Bayestehtashk, A., Shafran, I. Robust and Accurate Features for Detecting and Diagnosing Autism Spectrum Disorders. In Proc. Interspeech, 2013.

[79] Bonneh, Y., Levanon, Y., Dean Pardo, O., Lossos, L., Adini, Y. Abnormal speech spectrum and increased pitch variability in young autistic children. Frontiers in Human Neuroscience, vol.4, 2011.

[80] Rouhizadeh, M., Prud'hommeaux, E., Roark, B., Santen, H. Distributional semantic models for the evaluation of disordered language. Proc. NAACL-HLT, pp.709-714, 2013.

[81] Newton, T., Kramer, I., Mcwentosh, N. Autism online: a comparison of word usage in bloggers with and without autism spectrum disorders. In Proc. the SIGCHI Conference on Human Factors in Computing Systems, pp.463-466, 2009.

[82] Bone, D., Black, P., Lee, C., Williams, E., Levitt, P., Lee, S., Narayanan, S. Spontaneous-Speech Acoustic-Prosodic Features of Children with Autism and the Interacting Psychologist. Proc. Interspeech, 2002.

[83] Kiss, G., Santen, H., Prud'hommeaux, T., Black, M. Quantitative Analysis of Pitch in Speech of Children with Neurodevelopmental Disorders. Proc. Interspeech, 2012.

[84] Kiss, G., Santen, H. Estimating Speaker-Specific Intonation Patterns Using the Linear Alignment Model. Proc. Interspeech, pp.354-358, 2013.

[85] Santen, H., Richard, S., Alison, H. Quantifying repetitive speech in autism spectrum disorders and language impairment. Autism Research, vol.6, pp.372-383, 2013.

[86] Mairesse, F., Walker, A., Mehl, R., Moore, K. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. J. Artificial Intelligence Research, vol.30, pp.457-500, 2007.

[87] Mower, E., Black, P., Flores, E., Williams, M., Narayanan, S. Rachel: Design of an emotionally targeted interactive agent for children with autism. In Proc. IEEE International Conference on Multimedia and Expo, pp.1-6, 2011.

[88] Pennebaker, W., Martha, F., Roger, B. Linguistic inquiry and word count (LIWC). LIWC [Computer software], 2005.

[89] Chaspari, T., Gibson, B., Lee, C., and Narayanan, S. Using physiology and language cues for modeling verbal response latencies of children with ASD. In Proc. ICASSP, pp.3702-3706, 2013.

[90] Asperger, H. Die "Autistischen Psychopathen" im Kindesalter. European Archives of Psychiatry and Clinical Neuroscience, vol.117, pp.76-136, 1944.

[91] Essau, A., Olaya, B., Sasagawa, S., Pithia, J., Bray, D., Ollendick, H. Integrating video-feedback and cognitive preparation, social skills training and behavioral activation in a cognitive behavioral therapy in the treatment of childhood anxiety. J. Affect Disorders, vol.167, pp.261-267, 2014.

[92] Ikuzawa, M., Matsushita Y., Nakase, A. Kyoto Scale of Psychological Development. Kyoto: Kyoto International Social Welfare Exchange Centre, 2001.

Publication List

Journal

1. Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi Nakamura, NOCOA+: Multimodal Computer-Based Training for Social and Communication Skills, IEICE transactions on Information and Systems. (Conditional acceptance)
2. Hiroki Tanaka, Nick Campbell, Classification of Social Laughter in Natural Conversational Speech. Computer Speech & Language, vol.28, pp.314–325, 2014.
3. Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Satoshi 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, pp.19–26, 2014. (Short note)

International Conference

1. Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, Satoshi Nakamura, Automated Social Skills Trainer, International Conference on Intelligent User Interfaces, 2015. (Accepted)
2. Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura, Linguistic and Acoustic Features for Automatic Identification of Autism Spectrum Disorders in Children's Narrative, ACL2014 Workshop on Computational Linguistics and Clinical Psychology, pp.88–96, Baltimore, June 2014.
3. Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura, Modality and Contextual Differences in Computer Based Non-verbal Communication Training, In Proc. 4th IEEE CogInfoCom, Budapest, Hungary, pp.127–132, Dec. 2013.

4. Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura, Non-verbal Communication Training with an Interactive Multimedia Application, In Proc. The 5th Asian Conference on Education, pp.392–402, Oct. 2013.
5. Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Nick Campbell and Satoshi Nakamura, Non-verbal Cognitive Skills and Autistic Conditions: An Analysis and Training Tool, In Proc. 3rd IEEE CogInfoCom, pp.41–46, Dec. 2012.
6. Hiroki Tanaka, Hideki Kashioka, and Nick Campbell, Laughter as a gesture accompanying speech - towards the creation of a tool for the support of children on the autistic dimension, In Proc. GESPIN, Bielefeld, Sep. 2011.
7. Hiroki Tanaka, and Nick Campbell, Acoustic features of four types of laughter in natural conversational speech, In Proc. ICPHS XVII, pp.1958–1961, Aug. 2011.
8. Hiroki Tanaka, Hideki Kashioka, and Nick Campbell, Analysis of Laughter for Autistic Children, COST Final Meeting, Dresden, Feb. 2011.

Domestic Conference

1. 田中 宏季, 中村 哲, 感情コンピューティングによるソーシャルスキルトレーニングの自動化, 発達障害支援研究会, Mar. 2015.
2. 田中 宏季，サクリアニ サクティ，グラム ニュービッグ，戸田 智基，根來秀樹，岩坂英巳，中村 哲， ソーシャルスキルトレーニングの自動化，電子情報通信学会技術研究報告，ET2014-61，pp.1–6, Nov. 2014. [研究奨励賞]
3. 田中 宏季，サクリアニ サクティ，グラム ニュービッグ，戸田 智基，中村 哲，自閉症スペクトラム児と保護者間のインタラクション分析，人工知能学会全国大会，May. 2014.
4. 田中 宏季，サクリアニ サクティ，グラム ニュービッグ，戸田 智基，中村 哲，物語発話からの自閉症スペクトラム障害児と定型発達児の語彙と韻律の特性分析，日本音響学会春期大会， pp.1487–1490, Mar. 2014.

5. 田中 宏季，自閉症スペクトラム児と定型発達児のナレーティブ発話分析，SIG-SLP シンポジウム ショート発表，Jan. 2014.
6. 中村 哲, 田中 宏季, 非言語情報読み取りスキルを用いた自閉症スペクトラム指数の測定, 発達障害支援研究会, Oct. 2013.
7. 田中 宏季，サクリアニ サクティ，グラム ニュービッグ，戸田 智基，中村 哲，Computer-Based Training による非言語コミュニケーションスキルの改善に関する検討，人工知能学会全国大会，Jun. 2013.
8. 田中 宏季，サクリアニ サクティ，グラム ニュービッグ，戸田 智基，中村 哲， 非言語情報読み取りスキルを用いた自閉症スペクトラム指数の測定，電子情報通信学会技術研究報告，IMQ2012-34-IMQ2012-91，pp.223–226, Mar. 2013.
9. 田中 宏季，サクリアニ サクティ，グラム ニュービッグ，戸田 智基，中村 哲．非言語認知スキルからの自閉症スペクトラム指数の自動測定．教育システム情報学会研究報告，vol.27, pp.44–46, Nov. 2012.
10. 田中 宏季，豊川 弘樹，ニック キャンベル，自閉症児のためのノンバーバル情報支援アプリケーションの開発，日本音響学会春季大会，Mar. 2012.
11. 田中 宏季，豊川 弘樹，藤田 朋希，Sakriani Sakti， 戸田 智基 ，ニック キャンベル，中村 哲，自閉症児のためのノンバーバル情報学習，表出支援 iPad アプリケーションの開発，第 34 回関西合同音声ゼミ，Dec. 2011.
12. 田中 宏季，柏岡 秀樹，ニック キャンベル，自閉症児支援に向けた笑い声のアノテーション結果分析，電子情報通信学会技術研究報告，SP2011-60, WIT2011-42，Oct. 2011．
13. 田中 宏季，柏岡 秀樹，ニック キャンベル，自閉症児支援に向けた自然対話音声の笑いの種類分析，日本音響学会秋季大会，Sep. 2011．
14. 田中 宏季，柏岡 秀樹，ニック キャンベル，自閉症児支援のための笑い分析，日本音響学会春期大会， Mar. 2011.

Talk

1. Hiroki Tanaka, Linguistic and Acoustic Features for Automatic Identification of Autism Spectrum Disorders in Children's Narrative, EC-NLP, Jun. 2014.
2. 田中 宏季, コンピュータを用いた自閉症児支援, JSPACC 手をつなぐ親の会定例会, Nov. 2013.
3. 田中 宏季, ICT を用いた教材活用のネットワーク構築, 平成 25 年度第 1 回特別支援教育セミナー, Aug. 2013.
4. 田中 宏季，Computer-Based Training による非言語コミュニケーションスキル改善に関する検討, 発達障害研究会, Feb. 2013.

Master's Thesis

Hiroki Tanaka, A Method for Supporting Children with Autism Spectrum Disorders using Paralinguistic Information, Feb. 2012.

Award

1. 電子情報通信学会　教育工学研究会　研究奨励賞 (2014)
2. 奈良先端科学技術大学院大学　優秀学生奨学（トップ奨学生）賞（2012）

Research Grant

1. (公財) 奈良先端科学技術大学院大学支援財団支援事業「教育研究活動支援（教育研究活動助成）」 言語非言語情報を用いた自閉症スペクトラ ム障害の自動測定支援技術の研究 (2014)
2. 田中 宏季, 豊川 弘樹, 藤田 朋希, 中村 哲, Creative and International Competitiveness Project 自閉症児のための 3G を含んだ意図理解・表出支援 iPad アプリケーションの開発 (2012)