Interactive Virtual Reality Speech Simulation System Using Autonomous Audience with Natural non-Verbal Behavior

Justin Andrew Liao, Nobuyuki Jincho, and Hideaki Kikuchi

Abstract—**Public speaking anxiety (PSA) is a fear of speaking in front of others. Most people experience a certain amount of anxiety in public speaking situation. This study aims to help people overcome PSA using an interactive VR simulation system with real-life scenarios. We present a multimodal VR speech simulation system using autonomous audience with natural non-verbal behavior to enhance users' sense of presence. Additionally, real-time multimodal feedback is produced by virtual audience based on users' public speaking behavior which automatically analyzed by multimodal sensors (e.g. microphone, motion capture, heart rate monitor). We perform an evaluation based on self-assessment questionnaires and biometry to investigate three study conditions: (I) control condition (baseline), (II) interactive virtual audience, and (III) virtual audience with natural non-verbal behavior. We divided participants into two groups with different conditions: interactive virtual audience condition (*n* = 7) and virtual audience with nature non-verbal behavior condition (*n* = 9). The results indicate that the usage of a virtual audience with natural non-verbal behavior increased a higher sense of presence and more anxiety-provoking.**

Index Terms—**Public speaking anxiety, virtual reality, autonomous virtual audience, non-verbal behavior.**

I. Introduction

Public speaking anxiety (PSA) is the most common social phobia among most people. In DSM-V[1], anxiety disorders are categorized into (1) panic disorder, (2) obsessive-compulsive disorder, (3) agoraphobia, and (4) social anxiety disorder (SAD). Since previous studies [1], [2] suggested that public speaking anxiety is a distinct subtype of SAD, we need to define SAD first. SAD is defined as the fear of interaction with other people that brings on self-consciousness arising from a fear of being closely watched, negatively judged and evaluated. As a result, SAD leads to avoidance of social interaction.

PSA is a fear of speaking in front of others. The fear causes clinically significant distress or impairment in social, occupational, or other important areas of functioning [3]. [4] indicated that around 85% of the general population experiences a certain amount of anxiety in public speaking situation, which might lead to underperformance in school or at work, disturbed interpersonal relationships or even truancy [5]. High anxious people in public speaking tend to have a higher heart rate [6]. Also, they are more self-focus, which means they lose track of their surroundings while presenting [7], [8]; moreover, they are the lack of self-confidence in being successful in public speaking [9]. [10] showed that virtual reality exposure therapy (VRET) improved on participants' reduction in public speaking anxiety levels.

This work is aim to help people overcome PSA through the use of implement of the autonomous audience with natural non-verbal behavior. We are constructing a multimodal VR speech simulation system to achieve two-phase goals: (I) Using autonomous audience with natural non-verbal behavior to construct a speech simulation system with a high sense of presence (II) The well-structured system used as an automatic speech training system which can provoke as much anxiety as in reality.

II. Related Work

A. Virtual Audience

Virtual audiences are being used as part of exposure therapy for individuals with SAD (Social anxiety disorder) [1] and speech anxiety [11] by exposing them to situations they fear in VR environment. [12] found that the speakers' anxiety levels differed when they faced a neutral static audience, a positive audience, and a negative audience, respectively. Studies on stress responses explored variations of stress tests using supportive and non-supportive audiences [13], [14]. [15] designed a flexible behavioral style which allows a human operator to set and control the virtual audience's behavioral styles. A listening agent with simulated backchannel (head nodding and smiling) can improve the rapport in the human-agent interaction [16].

B. Public Speaking Performance

Ref. [17] explored three different feedback strategies for public speaking, namely the use of (1) a non-interactive virtual, (2) a direct virtual feedback, and (3) a nonverbal feedback from an interactive virtual audience. Their experiments show that the interactive virtual audience improved public speaking skills as judged by experts.

Lastly, we defined the problems of the existing simulation system on PSA: (1) no real-time feedback based on multimodal information, (2) lack natural non-verbal behavior variability, and (3) focused on the small crowd of a conference room.

Manuscript received May 2, 2018; revised July 16, 2018. This work was supported by the Japan Society for the Promotion of Science KAKENHI under Grant Number 17KT0141.

Justin Andrew Liao, Nobuyuki Jincho, and Hideaki Kikuchi are with the Graduate School of Human Sciences, WASEDA University. 2-579-15 Mikajima, Tokorozawa, Saitama, Japan (e-mail: justinliao@toki.waseda.jp).

doi: 10.18178/ijmlc.2018.8.4.720

C. Research Questions

We identify three research questions before constructing our system:

RQ1: How the use of virtual audience with natural non-verbal behavior enhances users' presence?

RQ2: How the feedback strategies differ the varieties of users' anxiety levels?

RQ3: How to render hundreds of virtual audiences with the limited workload?

III. Implementation

A. System Overview

Our speech simulation system is composed of 2 Engines, 2 Generators, and 2 modules. Fig. 1 illustrates the framework of the integrated system and the composition of the autonomous audience. Fig. 1 illustrates the framework of the integrated system and the composition of the autonomous audience. This simulation system has been developed using Unity and models of auditorium and avatars' clothes are designed by using 3ds Max, SketchUp, and Adobe Fuse CC.

Additionally, we design our system to simulate a procedure setting of agents entering an empty auditorium gradually in order to measure the dynamic variation of anxiety.

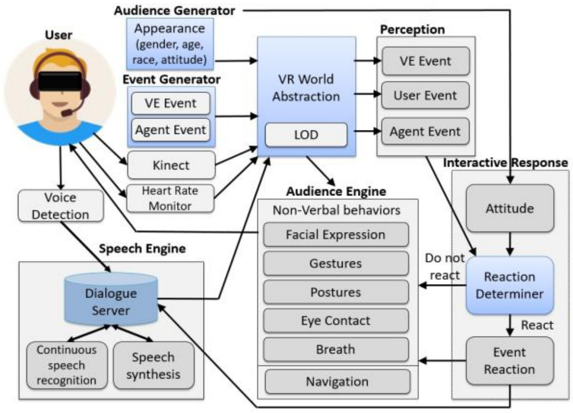

Fig. 1. The framework of the speech simulation system. The arrows (➔) in the diagram illustrate the information flow.

B. Research Question

RQ3. To render hundreds of virtual audiences, the native approach is to design as many avatars as possible. Whereas such an approach is difficult to achieve since it would require massive memory usage and computing load. Hence, we adopted Unity Multipurpose Avatar (UMA) and Level of detail (LOD) approaches [18] for our system. Our customized avatars were generated based on UMA which could share content across avatars using the same mesh and calculate all overlays to one single atlas. Additional, LOD made the rendering resolution adjustable according to the distance between the Agents and the User, and also the looking direction of the User.

IV. Methods and Materials

A. Experimental Design

We aimed to investigate how the use of virtual audience with natural non-verbal behavior enhances users' presence and users' anxiety levels by comparing with previous works [17]. To this effect, we had participants delivered 2 speeches with our speech simulation VE (cf. Fig. 2), i.e. we compared users' presence and users' anxiety levels, between each phase and both groups

In each speech session, there were 4 phases.

Phase 1: participations instructed to keep silent and standby for 3 minutes in real world (baseline)

Phase 2: participations instructed to keep silent and standby for 3 minutes in VR

Phase 3: follow-up phase 2, from virtual audience start entering the virtual auditorium to the end of the speech

Phase 4: silence time after VR in the real world (3 minutes)

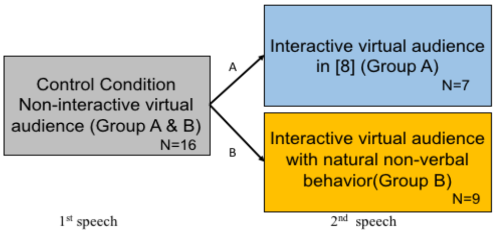
Control Condition Non-interactive virtual audience (Group A & B) N=16
A
B
Interactive virtual audience in [8] (Group A) N=7
Interactive virtual audience with natural non-verbal behavior(Group B) N=9
1st speech
2nd speech

Fig. 2. Participants were asked to present the same topic at the 1st speech, and a random topic at 2nd speech.

B. Procedures

This study protocol was approved by the ethics committee of the Waseda University. Participants were instructed to present two topics from [19] during 5-minute presentations (cf. Fig. 2). They were asked to prepare the slides for those presentations and send it before the day of the study. Before the first presentation, participants completed the questionnaires on Personal Report of Confidence as a Speaker Questionnaire (PRCS) [20], and Health Questionnaire to make sure subjects are in a normal health condition. In addition, participants were asked to wear a wireless heart rate monitor (myBeat-WHS-1) to measure the presenter's the anxiety levels in run-time. We divided participants into to 2 types of feedback strategies, one is interactive virtual audience [17] (Group A) and the other is an interactive virtual audience with nature non-verbal behavior (Group B). Each participant delivered 2 speeches (cf. Fig. 2). They were asked to present the same topic in front of a passive non-interactive virtual audience at first speech.

We compared the varieties of users' presence and users' anxiety between conditions 1) to 2), 2) to 3), and 1) to 3).

1) Control condition (Non-interactive virtual audience): No feedback during the speech.
2) Interactive virtual audience condition in [17]. Non-verbal feedback during speech
3) Interactive virtual audience with Natural non-verbal behavior condition. Nature non-verbal feedback during speech: higher variability of behaviors (e.g. yawning, falling asleep, using smartphone, applause)

In this study, the virtual audience was displayed using HMD (Oculus Rift). The slides made by participants were loaded into a virtual screen board, and participants can control it by Oculus Touch controller. Besides, Microsoft Kinect placed in front of the laptop capturing the body movement of the presenter. After each speech, the participants were asked to complete a Presence Questionnaire [21].

C. Participants and Dataset

In this study 19 people participated (paid 2500 JPY), 14 of which were recruited at a university and 5 were recruited from two companies. The dataset consisted of 11 males and 6 females with an average age of 23 years and standard deviation of 3.212 (3 data records had some technical problems leaving). 16 healthy participants were divided into two groups, with 7 to interactive virtual audience condition (Group A) and 9 to the interactive virtual audience with nature non-verbal behavior condition (Group B).

D. Data Acquisition

1) Self-assessment questionnaires

All participants completed the questionnaires on Personal Report of Confidence as a Speaker Questionnaire (PRCS) [20] before the first speech. PRCS is a commonly used to assess people's fear of public speaking. After each speech, the participants completed a 32-item Presence Questionnaire (PQ) [21] asking them to rate on a 5-point scale for measuring the sense of presence experienced in a virtual environment (VE).

2) Biometry

Participants were asked to wear a wireless heart rate monitor (myBeat-WHS-1) with the textile strap before the first speech. The R-R interval (RRI) time series were measured by heart rate monitor. Data acquisition was initiated 3 min after the start of each speech session to have stabilized hemodynamics.

We removed the outliers in the RRI time series by 2.5 SD above or below the mean in each phase. Besides, to eliminate individual differences, the mean RRI of *phase 1* has been viewed as 100% as a baseline to measure the relative variance of mean RRI of *phase 2* to *phase 4* (cf. Fig. 3). The differences between the mean of RRI's variation in each phase and the study groups were analyzed by a one-way analysis of variance.

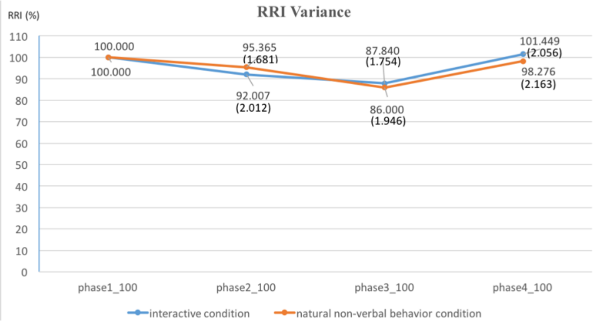

Fig. 3. Phase 1 based mean of RRI. The values inside parentheses represent standard error.

V. Results

A. Self-Assessment Questionnaires

A one-way analysis of variance between conditions was applied. We observed a statistically significant difference among conditions (cf. Table I). In PQ_{Q3} (*F* (1, 14) = 7.213, *p* = 0.018), " How natural did your interactions with the environment seem?", the presenters in the nature non-verbal behavior condition (*M* = 3.667, *SD* = 0.866) felt significantly more natural interactions with the environment than the interactive condition (*M* = 2.429, *SD* = 0.976). In PQ_{Q7} (*F* (1, 14) = 5.866, *p* = 0.030), "How natural was the mechanism which controlled movement through the environment?", participants in nature non-verbal behavior audience condition (*M* = 3.556, *SD* = 0.882) felt nature about the controlled movement significantly more than compare with interactive audience condition (*M* = 2.429, *SD* = 0.976). Furthermore, a significant difference among condition (II) and (III) was found, whether participants felt higher a sense of moving around (stage usage) in front of an audience with natural non-verbal behavior (PQ_{Q18}; *F* (1, 14) = 12.337, *p* = 0.004). The natural non-verbal behavior condition (*M* = 4.000, *SD*= 0.756) was rated as significant higher sense of body moving inside the VE than the interactive condition (*M* = 2.429, *SD* = 0.976).

TABLE I: Presence Questionnaires Results. * $P < 0.1$, ** $P < .05$

Condition (I) and (III) (within B Group)							
		Mean	Std. Deviation	Minimum	Maximum	F	Sig.
Q1	speech1	3.778	0.441	3.000	4.000	4.000	*
	speech2	3.333	0.500	3.000	4.000		
Q2	speech1	2.444	0.882	1.000	4.000	6.630	**
	speech2	3.667	1.118	1.000	5.000		
Q9	speech1	2.222	0.441	2.000	3.000	5.538	**
	speech2	1.556	0.726	1.000	3.000		
Q14	speech1	4.000	0.000	4.000	4.000	3.640	*
	speech2	4.375	0.518	4.000	5.000		
Q27	speech1	3.222	0.441	3.000	4.000	4.000	*
	speech2	3.667	0.500	3.000	4.000		
Condition (II) and (III) (between A, B Group)							
		Mean	Std. Deviation	Minimum	Maximum	F	Sig.
Q3	speech2_a	2.429	0.976	1.000	4.000	7.213	**
	speech2_b	3.667	0.866	2.000	5.000		
Q7	speech2_a	2.429	0.976	1.000	4.000	5.866	**
	speech2_b	3.556	0.882	2.000	5.000		
Q10	speech2_a	2.429	1.272	1.000	4.000	4.286	*
	speech2_b	3.667	1.118	2.000	5.000		
Q14	speech2_a	4.000	0.000	4.000	4.000	3.640	*
	speech2_b	4.375	0.518	4.000	5.000		
Q18	speech2_a	2.429	0.976	1.000	4.000	12.337	**
	speech2_b	4.000	0.756	3.000	5.000		
Q21	speech2_a	2.571	1.134	1.000	4.000	4.648	**
	speech2_b	3.625	0.744	3.000	5.000		

Besides, we found a significant difference between control condition and natural non-verbal behavior condition (within group B) in PQ_{Q2} (*F* (1, 16) = 6.630, *p* = 0.022). The natural non-verbal behavior condition (*M* = 3.667, *SD* = 1.118) was rated as significantly more responsively feedback to the actions that presenters performed than the control condition (*M* = 2.444, *SD* = 0.882).

B. Biometry

Here, we report the differences between conditions using the baseline (mean RRI of *phase 1*) to measure the relative variance of mean RRI of *phase 2* to *phase 4*. We observed a statistically significant difference (*F* (2, 29) = 6.677, *p* = 0.004) among of RRI variance from *phase 2* to *phase 3* between condition (II) and (III). The natural non-verbal behavior condition (*M* = -9.364, *SD* = 3.644) significantly more anxiety provoking than the interactive condition (*M* = -4.166, *SD* = 1.539).

Additionally, a significant difference is observed (*F* (1, 30) = 4.574, *p* = 0.041) from *phase 2* to *phase 3* that second speech (*M* = -7.090, *SD* = 3.889) significantly less anxiety provoking than first speech (M = -10.191, SD = 4.302).

VI. Discussion and conclusion

We have built a speech simulation virtual environment using the autonomous audience with natural non-verbal behavior. We discuss our results with respect to the research questions. **RQ1** Users' presence. First, We found that participants in the natural non-verbal behavior audience condition (Group B) felt more natural about the movement mechanism of the virtual audience and the interactions with the VE than the interactive audience condition (PQ_{Q1}, PQ_{Q7}). Second, the natural non-verbal behavior condition showed greater responsiveness than the control condition (PQ_{Q2}). **R Q2** users' anxiety levels. We observed that audience with natural non-verbal significantly more anxiety provoking than the interactive audience. Besides, the second speech had significant lower RRI variance than the first speech because of the practice effect. We will adjust our experiment procedures that the participants gave speech twice in a row. Overall, the usage of a virtual audience with natural non-verbal behavior increased the sense of presence and the anxiety-provoking. In the future study, we also plan to expand this study to automatically proximate the overall assessment of the speaking performance.

References

[1] American Psychiatric Association, "Diagnostic and statistical manual of mental disorders: DSM-5," *Am. Psychiatr. Assoc.*, p. 991, 2013.

[2] W. Eng, R. G. Heimberg, M. E. Coles, F. R. Schneier, and M. R. Liebowitz, "An empirical approach to subtype identification in individuals with social phobia.," *Psychol. Med.*, vol. 30, December 2000, pp. 1345–1357, 2000.

[3] C. B. Pull, "Current status of knowledge on public-speaking anxiety," *Curr. Opin. Psychiatry*, vol. 25, no. 1, pp. 32–38, 2012.

[4] M. C. E. Burnley, P. A. Cross, and N. P. Spanos, "The effects of stress inoculation training and skills training on the treatment of speech anxiety," *Imagin. Cogn. Pers.*, vol. 12, no. 4, pp. 355–366, 1993.

[5] S. R. Harris, R. L. Kemmerling, and M. M. North, "Brief virtual reality therapy for public speaking anxiety," *CyberPsychology Behav.*, vol. 5, no. 6, pp. 543–550, 2002.

[6] R. R. Behnke and L. W. Carlile, "Heart rate as an index of speech anxiety," *Speech Monogr.*, vol. 38, no. 1, pp. 65–69, 1971.

[7] J. A. Daly, A. L. Vangelisti, and S. G. Lawrence, "Self-focused attention and public speaking anxiety," *Pers. Individ. Dif.*, vol. 10, no. 8, pp. 903–913, 1989.

[8] W. Torsten and S. Scherer, "Automatic assessment and analysis of public speaking anxiety : A virtual audience case study," pp. 187–193, 2015.

[9] G. D. Bodie, "A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety," *Communication Education*, vol. 59, no. 1. pp. 70–105, 2010.

[10] W. Teramoto *et al.*, "'Spatio-temporal characteristics responsible for high 'Vraisemblance," *J. Virtual Real. Soc. Japan 15*, vol. 3, pp. 483–486, 2010.

[11] M. M. North, S. M. North, and J. R. Coble, "Virtual reality therapy: an effective treatment for the fear of public speaking," *Int. J. Virtual Real.*, vol. 3, no. 3, pp. 1–6, Dec. 2015.

[12] D.-P. Pertaub, M. Slater, and C. Barker, "An experiment on public speaking anxiety in response to three different types of virtual audience," *Presence Teleoperators Virtual Environ.*, vol. 11, no. 1, pp. 68–78, 2002.

[13] S. E. Taylor, T. E. Seeman, N. I. Eisenberger, T. A. Kozanian, A. N. Moore, and W. G. Moons, "Effects of a supportive or an unsupportive audience on biological and psychological responses to stress," *J. Pers. Soc. Psychol.*, vol. 98, no. 1, pp. 47–56, 2010.

[14] O. Kelly, K. Matheson, A. Martinez, Z. Merali, and H. Anisman, "Psychosocial stress evoked by a virtual audience: Relation to neuroendocrine activity," *CyberPsychology Behav.*, vol. 10, no. 5, pp. 655–662, 2007.

[15] N. Kang, W. P. Brinkman, M. B. Van Riemsdijk, and M. A. Neerincx, "An expressive virtual audience with flexible behavioral styles," *IEEE Trans. Affect. Comput.*, vol. 4, no. 4, pp. 326–340, 2013.

[16] L. Huang, L. P. Morency, and J. Gratch, "Virtual rapport 2.0," *Lecture Notes in Computer Science*, 2011, vol. 6895 LNAI, pp. 68–79.

[17] M. Chollet, T. Wörtwein, L.-P. Morency, A. Shapiro, and S. Scherer, "Exploring feedback strategies to improve public speaking: An interactive virtual audience framework," in *Proc. 2015 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. - UbiComp '15*, 2015. , pp. 1143–1154

[18] G. Ryder and A. M. Day, "Survey of real-time rendering techniques for crowds," *Comput. Graph. Forum*, vol. 24, no. 2, pp. 203–215, 2005.

[19] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation,"in *Proc. ISCA IEEE Work Spontaneous Speech Process. Recognit., Tokyo*, pp. 7–12, 2003.

[20] G. L. Paul, *Insight Vs. Desensitization in Psychotherapy: An Experiment in Anxiety Reduction*. 1966.

[21] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence Teleoperators Virtual Environ.*, vol. 7, no. 3, pp. 225–240, 1998.

Justin Andrew Liao was born in Ohio, USA, in 1991. He is a master's degree student studying at Waseda University, Tokyo, Japan. His major field includes virtual reality, augmented reality, and human agent interaction. He is a member of The Virtual Reality Society of Japan.

Nobuyuki Jincho was born in Tokyo, Japan, on June 23, 1976. He earned his Ph.D. in educational psychology from Waseda University, Tokyo, Japan in 2009.

He worked at Laboratory for language development in RIKEN Brain Science Institute, Japan. Now, he is assistant professor in School of Human Sciences, Waseda University. His research interest is language comprehension (reading and listening) and its development. Dr. Jincho is belonging to the Japanese Psychological Association, the Japanese Association of Educational Psychology, Japanese Cognitive Science Society, and German-Japanese Society for Social Sciences.

Hideaki Kikuchi earned his Ph.D. in information sciences from Graduate Schools of Science and Engineering, Waseda University, Tokyo, Japan in February 2002. His major field of study includes speech science, spoken dialogue system, and human agent interaction.

He is professor at the Faculty of Human Sciences, Waseda University, Japan. His previous job list: associate professor, Faculty of Human Sciences, Waseda University, Japan (September 2002-March 2012); visiting associate professor, Collaborative Research Unit, National Institute of Informatics, Japan (April 2008-); visiting researcher, Laboratory for Language Development, Brain Science Institute, RIKEN, Japan (September 2008-).

Prof. Kikuchi is belonging to the Japanese Society for Artificial Intelligence, the Acoustical Society of Japan, the Institute of Electronics, Information and Communication Engineers of Japan, and the Information Processing Society of Japan.