Validity of a voice-based evaluation method for effectiveness of behavioural therapy

Shuji Shinohara[1], Shunji Mitsuyoshi[2], Mitsuteru Nakamura[2], Yasuhiro Omiya[1],
Gentaro Tsumatori[3], and Shinichi Tokuno[2,3]

[1]PST inc., Industry & Trade Center Building 905, 2 Yamashita-cho, Naka-ku, Yokohama,
Kanagawa, Japan
{shinohara,omiya}@medical-pst.com
[2]Graduate School of Medicine, The University of Tokyo, Industry & Trade Center Building
905, 2 Yamashita-cho, Naka-ku, Yokohama, Kanagawa, Japan
{mitsuyoshi,tokuno}@m.u-tokyo.ac.jp
{NAKAMURAM-EME}@h.u-tokyo.ac.jp
[3]National Defense Medical Collage, 3-2 Namiki, Tokorozawa, Saitama, Japan
{tsumagen, tokuno}@ndmc.ac.jp

Abstract. In this study, we used General Health Questionnaire 30 (GHQ30) and voice to evaluate the stress reduction effect of a stress resilience program, and examined the validity of stress evaluation by voice. We divided the subjects who participated in the program into two groups by the number of training sessions. The results showed a stress-reduction effect only in the group with more training sessions (more than 13 sessions) for both GHQ30 and voice-based indexes. Moreover, both indexes showed a highly negative correlation between the pre-training value and the difference between the post-training and pre-training values. This implies that the effect of the training is more evident for subjects with higher stress levels. The voice-based evaluation showed trends similar to those displayed by GHQ30.

Keywords: stress check, voice, vitality, GHQ30, Stress Resilience Program

1 Introduction

Mental health problems are serious issues in many developed countries [1], and economic costs such as medical expenses and poor performance at work are enormous [2]. Thus, there is a need for techniques that easily check depression state and stress, as well as ways to cure or reduce such conditions.

An example of screening methods for patients with mental health issues include self-administered psychological tests such as the General Health Questionnaire (GHQ) [3] and the Beck Depression Inventory (BDI) [4][5]. Methods to check stress levels by using saliva and blood have also been proposed [6]. Self-administered psychological tests are effective for early detection and diagnostic aids but suffer from reporting bias issues. Additionally, stress-check methods using saliva and blood are not as simple due to issues related to test cost and burdens on the examinees.

In contrast, analysis of patients' medical condition, stress and emotion using voice data has been attracting attention due to the widespread use of smartphones in recent years [7][8][9].

Voice-based evaluations with a smartphone are advantageous since they are non-invasive and can be conducted easily and remotely without any special equipment.

Studies on the relationship between mental disorders and voice characteristics include those which analysed depressed patients' speaking rates [10][11][12] and their switching pauses and percent pauses [12][13]. Additionally, a study used chaos analysis to measure the Lyapunov exponents and Kolmogorov entropy in the voices of patients with depression [14]. Other research used frequency analysis to show that the shimmer and jitter values of vowel sounds made by patients with depression are higher than those of healthy individuals, while the first and second formant frequencies are lower for patients with depression [15]. A study proposed new features derived from Teager energy operator for stress classification [16]. Moreover, another report [17] proposed a method to measure mental health status based on the envelope information within pitch and voice waveforms.

While the abovementioned studies can be applicable for depression diagnosis and assessing stress levels, resilience programs that incorporate yoga and breathing techniques have been developed to reduce stress and depression, and have been implemented on a trial basis [18]. Additionally, a behavioural therapy called Smart, Positive, Active, Realistic, X-factor thoughts (SPARX), which utilises fantasy role-playing games, has also been developed and shown to be effective in treating younger-generation patients with depression [19][20].

In this study, we used the GHQ30 and patients' voices before and after the stress resilience program to evaluate their stress levels, and examined the validity of the voice-based stress evaluation.

2 Materials and Methods

2.1 Samurai’s Group and Individual mental training (S-Gim)

S-Gim is a stress resilience program developed by the Japan Self-Defence Forces [18]. S-Gim aims to acquire six skills, consisting of yoga stretches, breathing, imagery, viewpoint control, self-disclosure methods, and ways of supporting others to control stress. Yoga stretch and breathing can lead to control the mind by controlling the body. These give how to relax under the stress. Imagery is a method of controlling the image biased own. This is how to regain confidence. Viewpoint control fix a habit that is easy to catch negative. This is how to be taken to the positive things. Self-disclosure method is a training that represents the inside of your own mind. This is how to ask for help well. Way of supporting others is a technique to save the crisis of colleagues. This is how to control stress as a team.

The program entails 15 minutes a day, five times a week, for a total of 50 sessions, to become capable of demonstrating these skills easily.

2.2 Measuring method for the effectiveness of S-Gim

In this study, we measured the effect of S-Gim using the GHQ30 and a vitality score obtained from the voice-based analysis. GHQ30 is a self-administered psychological test with 30 questions, which provides scores for general disorder trends, physical conditions, sleep disorders, social activity disorders, anxiety and dysthymia, suicidal ideation, and depression [1].

A vitality score is one of the indices of mental health status that can be obtained by analysing patients' voices. The word "vitality" can have different definitions and implications. Here, it can be summarized as a measure that is low for patients with depression and strokes, and high for healthy individuals. The vitality score is calculated from the sound pressure level at the nadir of the amplitude envelope of the patient's voice between syllables, the change in the number of zero crossings in the waveform, and the pitch detection rate. Roughly speaking, clear, discernible, and fast voices usually correspond to higher vitality scores [17].

2.3 Acquisition of voices

From 3rd October, 2012 to 18th February, 2013, S-Gim was carried out with approximately 100 members of the Japan Self-Defence Forces. We collected voice data and the self-administered GHQ30 psychological test data from the subjects before and after the program. Voices were recorded by an IC recorder ICR-PS502RM (Sanyo Electric, Osaka, Japan) placed about 15 cm from the subject’s mouth. The recording format was as follows: linear PCM, a sampling frequency of 44.1 kHz, 16-bit quantization, low recording level, and ZOOM for directivity switching. Moreover, the microphone auto level control, low cut filter, recording peak limiter, VAS setting, and automatic soundless partitioning were turned off. The subjects were asked to read 11 types of passages.

There were 59 members from whom we were able to obtain both the voice and GHQ30 data before and after S-Gim. This paper targets these 59 members for the analysis.

3 Results

3.1 Evaluation of the effect of S-Gim by GHQ30

The average GHQ30 score before S-Gim was 3.85 (SD=5.57, n=59). The average score after S-Gim was 2.85 (SD=4.25). Additionally, there were 17 subjects whose score before S-Gim was zero. The purpose of this study is to measure the effect of S-Gim. The 17 subjects who scored zero before S-Gim were excluded from further analysis because no measurable decrease in their GHQ30 scores was possible. On the remaining 42 subjects, the average GHQ30 score before S-Gim was 5.40 (SD=5.93, n=42). The average score after S-Gim was 3.81 (SD=4.66).

In order to examine the effect of S-Gim based on the number of sessions completed, we divided the 42 subjects into two groups: 22 subjects with fewer sessions completed (1-12 sessions), and 20 subjects more sessions completed (more than 13 sessions)[1]. The average number of sessions conducted for the two groups were 7.23 (SD=3.80) and 31.50 (SD=14.95), respectively.

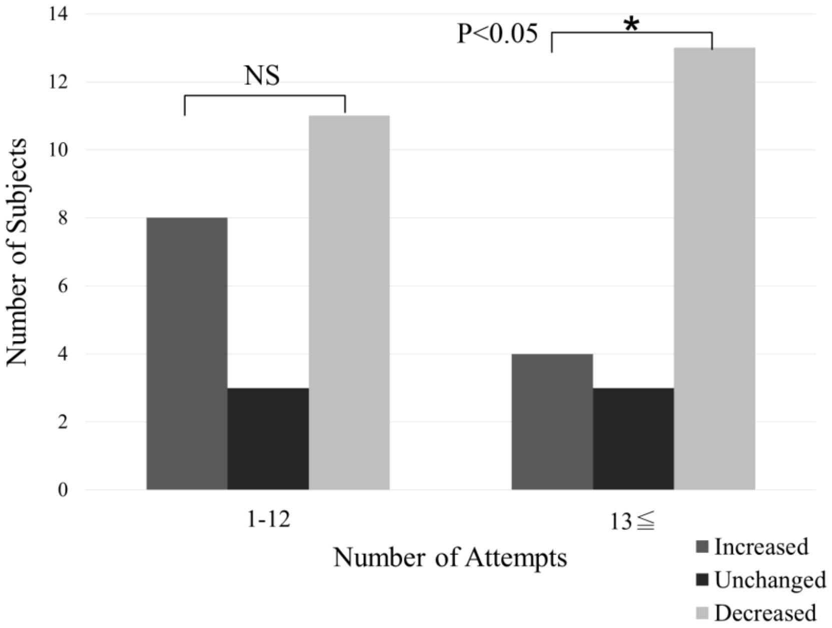

Fig. 1. Relationship between the number of S-Gim sessions attended and GHQ score change patterns. The bars on the left and the right represent the subjects who attended 1 to 12 sessions and more than 13 sessions, respectively. The vertical axis represents the number of subjects who experienced score change patterns (increased, unchanged, and decreased).

Fig. 1 shows the change in the GHQ30 scores for each group. The left group of bars shows the data for subjects who attended 1 to 12 sessions, while the right group of bars shows the data for subjects who participated in more than 13 sessions. The vertical axis shows the number of subjects who experienced each score change pattern (increased, unchanged, and decreased) before and after S-Gim. The proportions of subjects whose GHQ30 score decreased in each group were 50% and 65%, respectively. We performed a binomial test for the subjects whose scores increased

[1] We used 12 as cut-off criteria of two groups, because the value was the median of their training sessions.

and declined. The test results showed that there was no significant difference in the group that completed 1-12 sessions (p = 0.678)[2]. However, in the group that completed more than 13 sessions, there was a significant difference at the 5% level (p=0.0245).

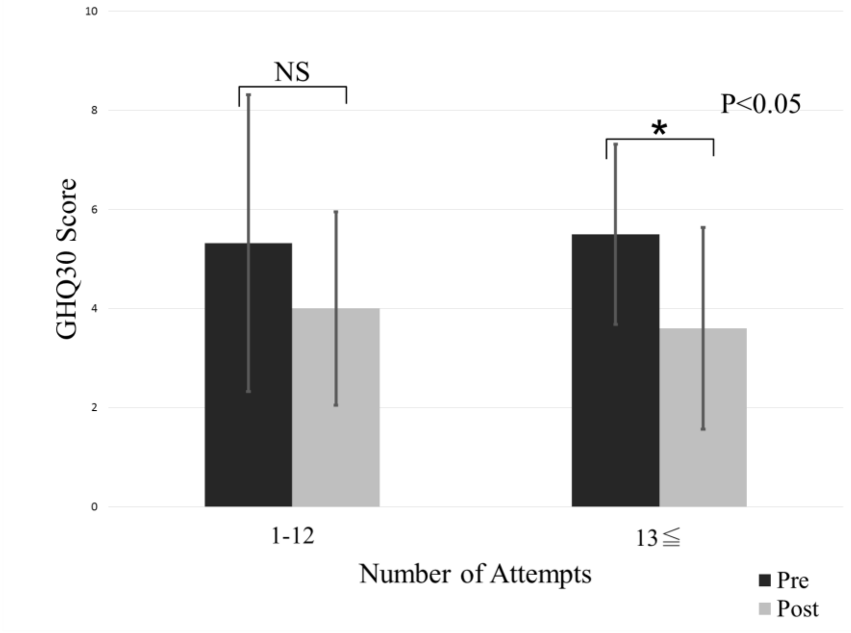

Fig. 2. Comparison between pre-and post-S-Gim GHQ30 average scores. The bars on the left and the right represent the subjects who attended 1 to 12 sessions, and more than 13 sessions, respectively. The vertical axis represents the GHQ30 score. The error bars represent the 95% confidence intervals. For the group with more than 13 sessions, there was a significant difference at the 5% level between the average scores before and after the training.

Fig. 2 shows the average GHQ30 scores for each group before and after conducting S-Gim. For the group that completed 1-12 sessions (n=22), the average GHQ30 scores before and after S-Gim were 5.32 (SD=7.17) and 4.00 (SD=4.67), respectively. A paired t-test showed no significant difference between the scores before and after the training (t(21)=0.904, p=0.376). In the group of subjects who completed more than 13 sessions (n=20), the average GHQ30 scores before and after S-Gim were 5.50 (SD=4.15) and 3.60 (SD=4.64). There was a significance difference at the 5% level before and after the training (t(19)=2.57, p=0.018)[3]. Additionally, there was a highly negative correlation between the pre-S-Gim GHQ30 score and the difference between

[2] We assumed that increases and decreases in the score would occur with the same probability if the S-Gim were not performed.

[3] We used the test function in Microsoft Excel 2010 for the tests.

the scores before and after the training (n=42, r=-0.662). That is, subjects with higher initial scores tended to show greater reductions in their scores.

3.2 Evaluation of the effect of S-Gim by vitality scores

The average vitality score before S-Gim was 7.15(SD=1.66, n=59). The average score after the training was 7.99 (SD= 1.38). The following comparison with the GHQ30 results only targeted the 42 subjects whose GHQ30 scores were 1 or higher at the time of conducting the training.

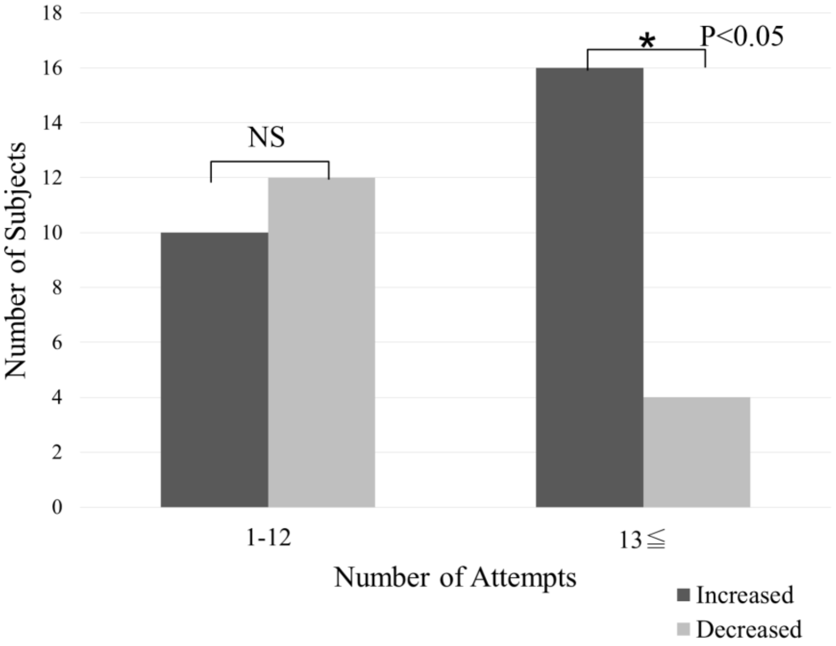

Fig. 3. Relationship between the number of S-Gim sessions and vitality score change patterns. The bars on the left and the right show the data for the subjects who attended 1 to 12 sessions and those who attended more than 13 sessions, respectively. The vertical axis represents the number of subjects who expereinced each pattern of vitality score changes (increased and decreased).

Fig. 3 shows the change in vitality scores in each group. The vertical axis represents the number of subjects who experienced each type of pattern (increased and decreased) of vitality score changes before and after S-Gim[4]. The proportions of

[4] Since vitality scores are continuous values, there was no subjects whose vitality score did not change.

subjects whose GHQ30 score increased in each group were 45% and 80%, respectively. A binomial test comparing the subjects with increased vitality scores and those with decreased scores showed no significant difference for the group whose members completed 1-12 sessions (p=0.832). In contrast, there was a significant difference at the 5% level for the group whose members completed more than 13 sessions (p=0.012).

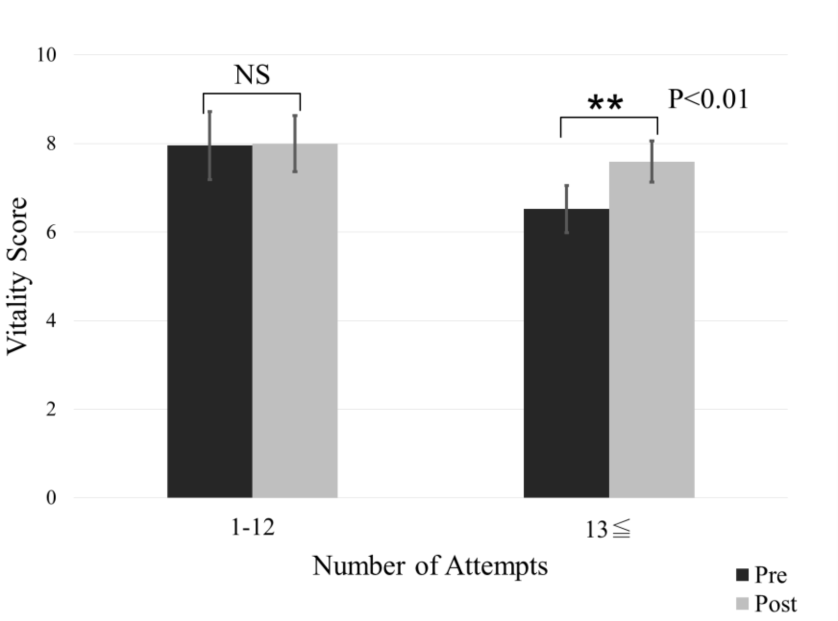

Fig. 4. Comparison of pre- and post-S-Gim vitality scores. The bars on the left and the right represent the subjects who attended 1 to 12 sessions, and those who attended more than 13 sessions, respectively. The vertical axis shows the vitality score. The error bars represent the 95% confidence intervals. There was a significant difference at the 1 % level between pre- and post-S-Gim vitality scores for the subjects who attended more than 13 training sessions.

Fig. 4 shows a comparison of the average vitality scores before and after S-Gim. The bars on the left and the right show the subjects who attended 1 to 12 sessions, and those who attended more than 13 sessions, respectively. For those who attended 1-12 sessions (n=22), the average vitality scores before and after S-Gim were 7.96 (SD=1.83) and 7.99 (SD=1.51), respectively. A paired t-test showed no significant difference between the scores before and after the training (t(21)=-0.085, p=0.933). For the subjects who completed more than 13 sessions (n=20), the average vitality scores before and after S-Gim were 6.52 (SD=1.22) and 7.59 (SD=1.07), respectively. There was a significant difference at the 1% level before and after the training (t(19)=-4.15, p=0.00054). Moreover, there was a highly negative correlation between the pre-S-Gim vitality scores and the difference between the pre- and post-S-Gim scores (n=42, r =-0.717). That is, subjects with lower vitality scores before training tended to increase their scores to a greater extent.

As these findings indicate, the subjects' vitality scores showed similar trends to the GHQ30 in terms of the effect of S-Gim. However, there was no direct correlation between GHQ30 scores and vitality scores (r = -0.022).

4 Discussion and Conclusion

In this study, we used a self-administered psychological test called the GHQ30, and vitality scores from a voice-based analysis, in order to evaluate S-Gim, a stress resilience program developed by the Japan Self-Defence Forces.

Fig. 1 shows that there were more subjects whose GHQ30 scores decreased after S-Gim in the group whose members attended more than 13 sessions (average number of sessions attended =31.50). Fig. 2 also shows that the scores themselves declined after the training. That is, the effect of S-Gim was confirmed in terms of the number of subjects and the average score. However, there was no effect in the group of subjects whose members attended less than 12 sessions (average number of sessions attended =7.23). This implies that a certain period of training is required to learn how to control stress through S-Gim. Additionally, there was a highly negative correlation between the pre-S-Gim GHQ30 score and the difference between the pre- and post-S-Gim scores. That is, subjects with higher stress levels experienced more apparent improvement in their stress levels through S-Gim.

Similarly to the GHQ30, we also evaluated the effect of S-Gim using an algorithm [17] that measures mental vitality levels from the subject's voice. As shown in Fig. 3 and Fig. 4, an effect of the training was observed in the group of subjects who completed more than 13 sessions. As for the GHQ30, there was a highly negative correlation between the pre-S-Gim vitality score and the difference between the pre- and post-S-Gim vitality scores.

The subjects' vitality scores showed similar trends to the GHQ30 in terms of the effect of S-Gim. However, there was no direct correlation between GHQ30 scores and vitality scores, which implies that GHQ30 and vitality scores do not necessarily evaluate the same characteristics. A study has reported success in overcoming reporting bias through voice-based analysis, albeit using different algorithms to those used here [21]. This indicates that the voice-based method might capture the difference between subjective and objective symptoms. A detailed analysis in this regard should be a future priority.

In this study, the vitality score was used to evaluate the effect of S-Gim. However, this measure can also be used to check mental health status, similarly to GHQ30. The vitality score can be measured from the voice, making it easier to administer than the GHQ30. Moreover, it is feasible to record daily changes in mental health easily by installing the system on smartphones. We are currently developing a smart phone application equipped with the vitality score algorithm.

Acknowledgements
We would like to express our sincere appreciation to Colonel Sota Shimozono of JGSDF Medical School and his staff members for their cooperation in sharing S-Gim and collecting data.

References

1. World Health Organization, The Global Burden of Disease: 2004 update, pp. 46--49. WHO Press, Geneva, Switzerland (2004)
2. Kessler, R. C., Akiskal, H. S., Ames, M., Birnbaum, H., Greenberg, P., Hirschfeld, R. M. A., Jin, R., Merikangas, K.R., Simon, G.E., Wang, P.S.: Prevalence and effects of mood disorders on work performance in a nationally representative sample of U.S. workers. Am. J. Psychiatry. 163(9), 1561--1568 (2006)
3. Goldberg, D.P., Blackwell, B.: Psychiatric illness in general practice: a detailed study using a new method of case identification. BMJ. 2(5707), 439--443 (1970)
4. Beck, A.T.: A systematic investigation of depression. Compr. Psychiat. 2(3), 163--170 (1961)
5. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. Arch. Gen. Psychiatry. 4(6), 561--571 (1961)
6. Suzuki, G., Tokuno, S., Nibuya, M., Ishida, T., Yamamoto, T., Mukai, Y., Mitani, K., Tsumatori, G., Scott, D., Shimizu, K.: Decreased plasma brain-derived neurotrophic factor and vascular endothelial growth factor concentrations during military training. PloS One 9(2), e89455 (2014)
7. Arora, S., Venkataraman, V., Zhan, A., Donohue, S., Biglan, K.M., Dorsey, E.R., Little, M.A.: Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study. Parkinsonism Relat. D. 21(6), 650--653 (2015)
8. Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P.J., Longworth, C., Aucinas. A.: EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In Proceedings of the 12th ACM international conference on Ubiquitous computing, pp. 281--290, (2010).
9. Lu, H., Rabbi, M., Chittaranjan, G. T., Frauendorfer, D., Mast, M. S., Campbell, A. T., Gatica-Perez, D., Choudhury, T.: Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 351-360. (2012).
10. Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., Snyder, P.J.: Voice acoustical measurement of the severity of major depression. Brain Cognition. 56, 30--35 (2004)
11. Moore, E. II., Clements, M., Peifert, J., Weisser, L.: Analysis of prosodic variation in speech for clinical depression. In: Proc. of the 25" Annual International Conf. of the IEEE EMBS, vol. 3, pp. 2925 -- 2928. IEEE Press, New York (2003)
12. Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., Geralts, D.S.: Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J. Neurolinguist. 20(1), 50--64 (2007)
13. Yang, Y., Fairbairn, C., Cohn , J.F: Detecting depression severity from vocal prosody. IEEE Trans. Affective Computing. 4(2), 142--150 (2013)
14. Shimizu, T., Furuse, N., Yamazaki, T., Ueta, Y., Sato, T., Nagata, S.: Chaos of vowel /a/ in Japanese patients with depression: A preliminary study. J. Occup. Health. 47(3), 267--269 (2005)

15. Vicsi, K., Sztaho, D.: Examination of the sensitivity of acoustic-phonetic parameters of speech to depression. In: IEEE 3rd International Conference on Cognitive Infocommunications, pp. 511--515. IEEE Press, New York (2012)
16. Zhou, G., Hansen, J.H.L., Kaiser, J.F.: Nonlinear feature based classification of speech under stress. IEEE Transactions on Speech and Audio Processing. 9(3), 201--216, (2001)
17. Shinohara, S., et al.: A mental health evaluation method using prosody information of voice. In preparation
18. Tokuno, S., et. al: Usage of emotion recognition in stress resilience program. In: Proc. of 40th ICMM World Congress on Military Medicine, (2013)
19. Merry, S.N., Stasiak, K., Shepherd, M.: The effectiveness of SPARX, a computerised self help intervention for adolescents seeking help for depression: Randomised controlled non-inferiority trial. BMJ. 344, 1--16 (2012)
20. Fleming, T., Dixson, R.: A pragmatic randomized controlled trial of computerized CBT (SPARX) for symptoms of depression among adolescents excluded from mainstream education. Behav. Cogn. Psychoth. 40, 529--541 (2012)
21. Tokuno, S., Mitsuyoshi, S., Suzuki, G., Tsumatori, G.: Stress evaluation using voice emotion recognition technology: A novel stress evaluation technology for disaster responders. Proc. XVI World Congress of Psychiatry. 2, 301 (2014).