Individual differences of emotional expression in speaker's behavioral and autonomic responses

Yoshiko Arimoto[1,2], *Kazuo Okanoya*[1,2,3]

[1]Okanoya Emotional Information Project, JST, ERATO, Saitama
[2]Brain Science Institute, Riken, Saitama
[3]Graduate School of Arts and Sciences, The University of Tokyo, Tokyo
ar@brain.riken.jp, okanoya@brain.riken.jp

Abstract

The goal of this study is to elucidate differences in speakers' emotional expressions in behavioral and autonomic responses. Verbal and non-verbal emotional behaviors of interlocutors were recorded during two types of dialogs (competitive and cooperative). Autonomic nervous system (ANS) activity (heart rate and skin conductance level) was also recorded as an internal measure of emotional reactions toward an interlocutor. To annotate the emotional states of speakers, the speakers who participated in the recording evaluated their own emotional states (arousal, valence and positivity) and their interlocutor's states along with the time course of the dialogs. The behavioral and autonomic emotional reactions were used as independent variables for speaker-independent and speaker-specific models to predict a speaker's emotion. The results demonstrate that speaker-independent models could explain each emotional state in a certain degree; in contrast, some speaker-specific models could explain each emotional state with moderate or high accuracy. Moreover, a comparison of the absolute standard partial regression coefficients of each variable of the models revealed that there are two types of emotional expression styles, one in which emotional behavioral expression is dominant and another in which emotional autonomic reaction is dominant.
Index Terms: Emotional speech, facial expressions, autonomic nervous system activity, spontaneous dialog

1. Introduction

Individual emotional reactions differ among speakers even if they feel the same emotion. Some react via their voice or facial expressions, whereas others react only through their autonomic nervous system (ANS) activity. When a speaker's emotion is expressed through their behavioral reactions, listeners can behave according with the speaker's emotional changes. However, listeners cannot take any action in response to a speaker's emotional changes when the speaker's experienced emotion is reflected only in autonomic responses; in this case, listeners cannot obtain any information regarding the speaker's emotion because the information is not be conveyed through any communication pathway. These individual differences in emotional expression among speakers have been previously noted and examined with traditional experimental techniques [1]. Emotional discharge theory claims that there is a negative correlation between external emotional responses in behavior and internal emotional responses in ANS activity. This negative correlation between behavior and ANS activity was explained by the sympathetic nervous system being activated by the suppression of behavioral emotional expression. It was suggested that there are two types of emotional expression styles. There are two distinctions based on emotional expression styles, an externalizer and an internalizer, according to [2]. An externalizer is a person whose emotional behavioral expression was encouraged but whose emotional autonomic nervous activity was suppressed. In contrast, an internalizer is a person whose emotional behavioral expression was suppressed but whose autonomic nervous activity was encouraged. This claim was based on the results of traditional psychological experiments that tested subjects' responses to various types and intensities of stimuli in standard laboratory settings. However, emotional expression styles in behavioral and autonomic responses in naturalistic settings have never been demonstrated. This work investigated the individual differences of emotional expression in speakers' behavioral and autonomic responses to determine whether there are several emotional expression styles, such as externalizer or internalizer, even in naturalistic communication settings.

2. Material

Fifty-two speakers ($mean\ age = 21, SD = 2.34$) participated in two types of dyadic dialogs with a friend of the same sex. Half of the 26 pairs of speakers were female pairs, and the remainder were male pairs. Each pair committed two types of tasks: one was a competitive task, and the other was a cooperative task. The competitive task was a Japanese word-chain game, cap verses (*Shiritori*). The cooperative task was a simple game in which each pair worked together to raise a score to 100. Because the pairs were not informed regarding how the scores could be increased, they had to discuss and cooperate with each other to determine how to raise their score.

The speakers' behavior (speech and facial expression) were recorded as a measure of external reaction during the dialogs. Simultaneously, ANS activity (using electrocardiography (ECG) and electrodermal activity (EDA)) was recorded as a measure of internal reaction. Each speaker in a pair sat in a soundproofed room connected with soundproofed glass. She faced her interlocutor over the glass, and they talked with each other using microphones and headphones. Their speech was recorded in individual channels of an audio stream with a sample rate for 48 kHz and 16-bit precision. To record a speaker's behavior, 4 CCD cameras took images of the speaker. One of the cameras focused on the speaker's face. Each speaker's behavior was recorded at 30 frames per second on images of 640×480 resolutions. For synchronous audio and visual recording, the audio-visual recording system Potato (Library Co. Ltd., Tokyo, Japan) was used. Electrocardiography (ECG) and electrodermal activity (EDA) were recorded using a Biopac MP150 system (Biopac Systems, CA, USA). Both signals were recorded at a sampling rate of 1 kHz. To record a speaker's ECG, three electrodes was placed 10 cm below the speaker's right collarbone and on the lower right and left sections of the speaker's ribcage. To record a speaker's EDA, two Ag-AgCl electrodes were attached to the volar surfaces of the index and

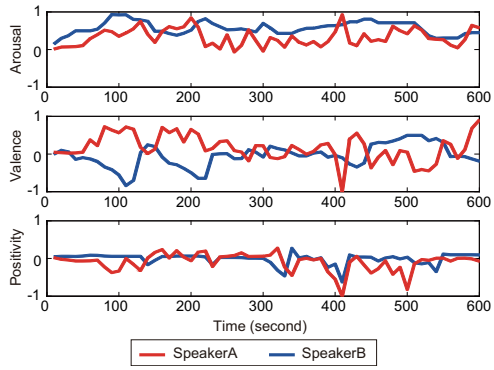

Figure 1: *Result of dynamic emotional state annotation.* Each panel shows the chronic flow of the emotional state of each speaker of one pair (for a competitive task).

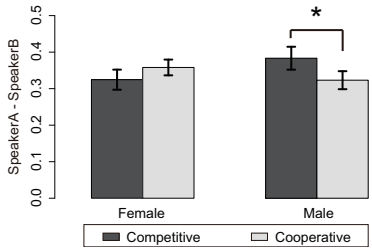

Figure 2: *Result of Bonferroni's multiple comparison test.*

middle toes of the left foot. For synchronous recording, the Biopac MP150 system started to record signals after it received a sync signal sent by Potato at the time of the first camera shutter activity.

3. Dynamic emotional state annotation

Speakers also conducted subjective evaluations for dynamic emotional state annotation using GTrace [3]. Speakers dynamically rated their own emotional state and their interlocutor's emotional state during the recorded audio-visual video sequences of 10-minute competitive- and cooperative-task dialogs. The target emotional states were arousal (aroused–sleepy), valence (pleasant–unpleasant) and positivity (positive–negative). The mean evaluated value of each emotional state was calculated at 10-second intervals as a measure of the dynamic emotional state of the speakers. Figure 1 shows a chronic flow of the emotional state of each speaker in one pair during a competitive-task dialog. It indicates that emotional states of each speaker were consistent in arousal but inconsistent in valence. It suggests that one of the pair felt pleasant, but the other felt unpleasant because each speaker competed with each other during a dialog.

A $2 \times 3 \times 2$ repeated measures analysis of variance (ANOVA) on each pair's mean absolute difference of each emotional state in each dialog was performed with the factors of task (competitive and cooperative), emotional states (arousal, valence and positivity) and gender (female and male). The result revealed a significant interaction between task and gender ($F(1, 24) = 6.04$, $p < 0.05$). Bonferroni's multiple comparison test revealed that a significant simple main effect of task on male speakers ($F(1, 24) = 4.99$, $p < 0.05$, see Fig. 2).

4. Feature calculations

Following [4–12], features of external and internal emotional responses were calculated. Facial and vocal features were calculated as features of external emotional responses, and features of heart rate and skin conductance were calculated as internal emotional responses. Mean values of heart rate and skin conductance were calculated at 10-second intervals. After calculating frame-based facial features and utterance-based vocal features, the mean values of each feature were calculated at 10-second intervals to adjust analysis units among modalities.

4.1. Vocal features (SPEECH)

Seven vocal features were extracted from segmented speech signals according to [4–6]. Each speech signal was segmented based on 200-ms inter-pausal units (IPUs). The features were the maximum and standard deviation of the fundamental frequency (F_0), the short-term power, the first cepstral coefficient, and the speaking rate. To avoid the influence of outliers, the maximum and standard deviation were calculated after removing the upper and lower 10% of the data. For pitch features, the fundamental frequency was extracted from speech signals using STRAIGHT [13]. For voice quality features, the first cepstral coefficients were extracted from the voiced speech signal. For speaking rate features, the number of morae was divided by the duration of each utterance after counting the number of morae in one utterance.

4.2. Facial features (FACE)

Facial feature detection software, FaceSDK 4.0 (Luxand, Inc., VA, USA), was used for extraction of facial points. Twenty-six facial points (4 points in the eyebrows, 10 in the eyes, 3 in the nose, 8 in the mouth and 1 on the chin) were adopted following [7–9]. Figure 3 shows the facial points used for a facial feature calculation. An affine transform based on two referential points of face positions extracted using FaceSDK was performed. First, a standard facial expression for each speaker that was selected by the experimenter was transformed to adjust for the different facial sizes among the speakers. Then, each facial point in each frame was transformed based on the adjusted standard facial expression to eliminate head motions in a sequence. To adjust in-plane rotation, each set of facial points was transformed based on an angle extracted using FaceSDK. Cubic interpolation was performed for frames from which facial points could not be extracted. For temporal smoothing, a 5-frame central moving average was calculated for each point after the cubic interpolation was performed.

Each angle θ between a line connecting two of the 26 facial points and a horizontal line was calculated according to

$$\theta_{i,j} = \arctan\left(\frac{p_{i,y} - p_{j,y}}{p_{i,x} - p_{j,x}}\right), \quad (1)$$

where p_i and p_j are two different facial points in the same frame ($i = 1, 2, 3, ...n$, $j = 1, 2, 3, ...n$, $i \neq j$, $n = 26$) and $p_{i,x}$ and $p_{i,y}$ are the coordinate points on the x axis and y axis of the i-th point, respectively. To describe how much the feature values changed relative to their values for the neutral face, delta features were calculated according to

$$\boldsymbol{\Delta\theta_t} = \boldsymbol{\theta_t} - \boldsymbol{\theta_b}, \quad (2)$$

where $\boldsymbol{\theta_t}$ is the vector of $\theta_{i,j}$ for the t-th frame ($t = 1, 2, 3, ...T$, $T = 18000$) and $\boldsymbol{\theta_b}$ is the vector of $\theta_{i,j}$ for the frame of the neutral face. The frames of the neutral face were selected by one of the authors from the frames that were recorded before or after the dialog. The number of facial features was 325.

To reduce the number of facial features, principal component analysis was performed on all facial features. Then, 37 principal components of facial features were obtained explaining 92% of the original facial features. The 37 components (PC1–37) were adopted as facial features for the following analysis.

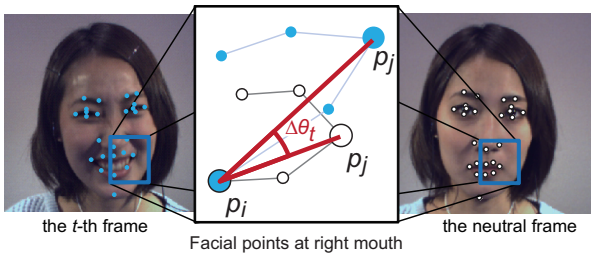

Figure 3: *Facial points and facial feature calculations.* An angle θ between two different facial points in the tth frame .

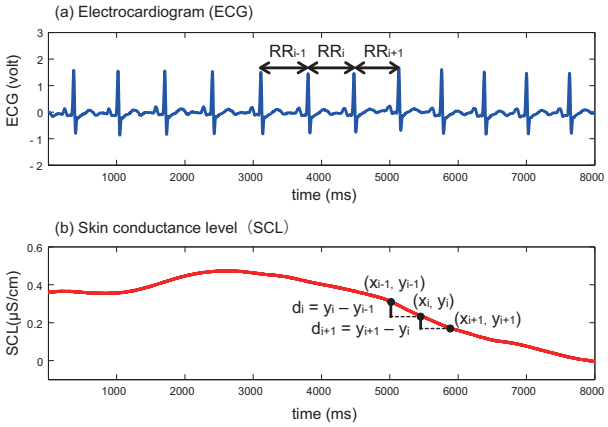

Figure 4: *ECG signal for a calculation of heart rate (a) and EDA signal for a calculation of skin conductance level (b).*

4.3. Heart rate (HR)

Five heart rate features were calculated from the ECG signal of each speaker following [10–12]. For the calculation of heart rate features, the heart rate (beat per minute) and RR intervals were extracted from the speaker's ECG signal using the Biopac MP150 system and were downsampled to 200 Hz. To normalize individual difference of absolute heart rate and RR intervals among speakers, the mean resting heart rate and RR interval were subtracted from the heart rate and RR intervals, respectively. As heart rate variability features, the followings were calculated: the average and standard deviation of the heart rate, the standard deviation of RR intervals (SDNN), the square root of the mean squared differences of successive RR intervals (RMSSD), and the proportion of the number of interval differences of successive RR intervals greater than 50 ms (pNN50); the last was calculated by dividing the number of such intervals by the total number of RR intervals during each 10-second period. Figure 4(a) shows an example of an ECG signal for heart rate feature calculations.

4.4. Skin conductance level (SCL)

Six skin conductance level features were calculated from the EDA signal downsampled to 200 Hz following [11, 12]. To normalize individual difference of absolute skin conductance levels among speakers, the mean resting skin conductance level were subtracted from the skin conductance level. These features were the average and standard deviation of the skin conductance level, the average and standard deviation of the derivative, the average of the derivative for negative value only, and the proportion of negative samples in the derivative compared with all samples. These six features were derived from both the raw SCL signals and the low-pass-filtered signals using a cut-off frequency of 0.2 Hz. Total number of features comes to 12. Figure 4(b) shows an example of an SCL signal.

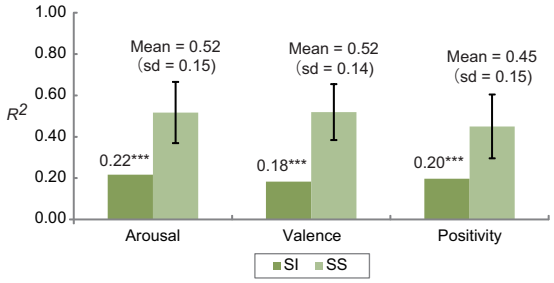

Figure 5: *Barplots of R^2 for the speaker-independent model (SI) and the mean R^2 and its SD for the speaker-specific models (SS).*

4.5. Correlations among features

No feature exhibited a strong correlation ($|r| \geq 0.75$) with features from a different modality. The SPEECH features and FACE features exhibited no strong correlations ($|r| \geq 0.75$) among their own modalities. Three SCL features, the average and standard deviation of non-filtered SCL and the average of its derivative, exhibited strong correlations ($|r| = 1$) with the corresponding feature calculated from the low-pass-filtered SCL. Those three features were not used in the following analysis.

5. Correlations between emotional states and each feature

To investigate the relationship between three emotional states and each feature, the correlation coefficients among them were calculated. Each emotional state exhibited a weak correlation with each feature. The features that exhibited the strongest correlation with two emotional states are all the SPEECH features, i.e., the maximum of the short-term power feature with valence ($r = 0.19$), the maximum of the first cepstral coefficient with arousal ($r = -0.16$). The FACE feature, PC1, showed the strongest correlation with positivity ($r = 0.15$). The ANS features did not have any correlation coefficients greater than 0.10 with any emotional state. Among the FACE features, 6 features (PC16, 17, 21, 26, 31, and 34), 4 features (PC1, 3, 17, and 35) and 5 features (PC1, 11, 14, 17, and 19) exhibited correlation coefficients greater than 0.10 with arousal, valence and positivity, respectively.

6. Multiple regression analysis on individual emotional expressions

6.1. Procedure

To elucidate speakers' differences of emotional expression in behavioral and autonomic responses, multiple regression analysis was conducted to model speaker-independent and speaker-specific emotional expression to predict speakers' emotions. The self-evaluated emotional states (arousal, valence and positivity) were used as the dependent variables, and the vocal, facial and ANS features were used as independent variables for the multiple regression analysis. A step-wise regression based on the Akaike information criterion (AIC) was performed. The bidirectional elimination approach was adopted for model selection.

To compare the results of multiple regression analysis between the speaker-specific models and the speaker-independent model, bar plots of R^2 of the speaker-independent model and the mean R^2 and its SD for the speaker-specific models are shown in Fig. 5.

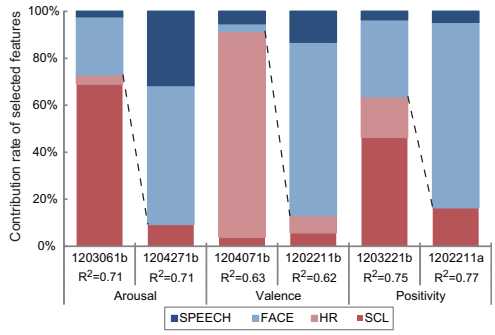

Figure 6: *Examples of the contribution rates of standard partial regression coefficients selected in some speaker-specific models.*

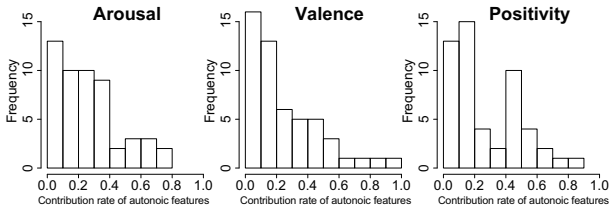

Figure 7: *Histograms of speakers based on the contribution rates of autonomic features selected in each speaker-specific model.*

6.2. Results

The R^2 values of the speaker-independent models for each emotional state were 0.22 (arousal), 0.18 (valence) and 0.20 (positivity), respectively (Fig 5). The models predicted each emotional state with moderate accuracy, with significance indicated by the F-test ($p < 0.001$).

The R^2 values of speaker-specific model for predicting each emotional state were greater than 0.45. The highest R^2 values were 0.84, 0.80 and 0.80 for arousal, valence and positivity, respectively. These results suggest that integrated use of behavioral and autonomic features strongly correlates each emotional state. In contrast, the lowest R^2 values were 0.21 (arousal), 0.19 (valence) and 0.19 (positivity). These models explained each emotional state well, with significance indicated by the F-test ($p < 0.001$). However, their prediction accuracies were moderate.

7. Discussion

There are two types of emotional expression styles, externalizer and internalizer, according to [1, 2]. An externalizer is a person whose emotional behavioral expression is encouraged but whose emotional autonomic nervous activity is suppressed. In contrast, an internalizer is a person whose emotional behavioral expression is suppressed but whose autonomic nervous activity is encouraged. To demonstrate whether there are several emotional expression styles, such as externalizer and internalizer, even in naturalistic communication settings, the rate of the absolute standard partial regression coefficient of each selected feature was calculated as the contribution rate for each speaker-specific model and compared with each selected feature within each of the 52 speakers. The rate of an absolute standard partial regression coefficient was calculated by dividing the absolute standard partial regression coefficient of each feature by the summation of them for all selected features. As a result, the models for which the contribution rate of the SPEECH and FACE features accounted greater than 50% became 44 (arousal), 45 (valence), and 44 (positivity). In contrast, the models for which the contribution rate of the SCL and HR features accounted for greater than 50% became 8 (arousal), 7 (valence), and 8 (positivity). This result indicated that whereas there were many speakers for which their emotions were reflected in their external reactions (vocal and facial expressions), there were speakers for which their emotions were also reflected in their internal reactions (autonomic nervous system activity).

Figure 6 shows the contribution rates of features in two of the speaker-specific models for each emotional state. The contribution rates were calculated for four groups (SPEECH, FACE, HR and SCL). Among the two models for each emotional state, one is such that the contribution rate of the SPEECH and FACE features accounts for more than that of the SCL and HR features (greater than 80%), and the other is such that the contribution rate of the SCL and HR features accounts more than that of the SPEECH and FACE features (greater than 60%). It was suggested that there would be two types of emotional expression styles, one in which emotions are strongly reflected in the speaker's external reactions (externalizer) and another in which emotions are reflected in the speaker's internal reactions (internalizer).

To demonstrate whether there are clear distinctions between externalizers and internalizers among the 52 speakers, histograms of speakers based on the contribution rate of autonomic features selected in each speaker-specific model are shown in Fig 7. The x-axis indicates the contribution rate of autonomic features. When the x-axis is 0.0, the speaker's emotion is only reflected in the speaker's external reactions. When the x-axis is 1.0, the speaker's emotion is only reflected in the speaker's internal reactions. According to the left panel of Fig. 7, the histogram of arousal exhibits two peaks in the distribution; one is near 0.0, and the other is near 0.6. This distribution suggests that there are two types of emotional expression styles (externalizer and internalizer) among speakers when they feel arousal. The right panel of Fig. 7 also indicates that the histogram of positivity exhibits two peaks in the distribution; one is near 0.0, and the other is near 0.5. The distribution suggests that there are also two types of emotional expression styles among speakers when they feel positive. One of emotional expression styles could be an externalizer; however, the other could not be an internalizer because its peak is near 0.5. This result implies that positivity was reflected in both external and internal reactions equally. This type of emotional expression style was previously described as a generalizer, whose emotions are reflected both in behavioral expressions and autonomic nervous activity, by [1]. The second peak at approximately 0.5 in positivity could correspond to groups of generalizers. In contrast, the histogram of valence exhibits a monomodal distribution for which the peak is approximately 0.0. This distribution suggests that valence was strongly reflected in external reactions and that there is only one emotional expression style (externalizer) with respect to valence.

8. Conclusions

This paper investigated individual differences in speakers' emotional expressions through behavioral and autonomic responses to demonstrate whether there are several emotional expression styles, such as externalizer and internalizer, even in naturalistic communication settings. As a result, it was found that there are two types of emotional expression styles for arousal and positivity (externalizer and either internalizer or generalizer)

9. Acknowledgements

We would like to thank Dr. Noriko Kondo from JST, ERATO, for her help organizing our large-scale and complex multimodal data.

10. References

[1] J. T. Cacioppo, B. N. Uchino, S. L. Crites, M. A. Snydersmith, G. Smith, G. G. Berntson, and P. J. Lang, "Relationship between facial expressiveness and sympathetic activation in emotion: A critical review, with emphasis on modeling underlying mechanisms and individual differences." *Journal of Personality and Social Psychology*, vol. 62, no. 1, pp. 110–128, 1992.

[2] H. E. Jones, "The galvanic skin reflex as related to overt emotional expression," *The American Journal of Psychology*, vol. 47, no. 2, pp. 241–251, 1935.

[3] "GTrace," https://sites.google.com/site/roddycowie/work-resources.

[4] C.-c. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, P. G. Georgiou, and S. S. Narayanan, "An Analysis of PCA-based Vocal Entrainment Measures in Married Couples' Affective Spoken Interactions," in *Proceedings of Interspeech2011*, 2011, pp. 3101–3104.

[5] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions ." in *Proceedings of Interspeech2011*, 2011, pp. 3081–3084.

[6] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, "Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment," *Acoustical Science and Technology*, vol. 33, no. 6, pp. 359–369, 2012.

[7] M. F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proceedings of the ninth international conference on Multimodal interfaces - ICMI '07*, 2007, pp. 38–45.

[8] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Proceedings of IEEE International Conference of Face and Gesture 2011*, 2011, pp. 921–926.

[9] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions." *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 42, no. 1, pp. 28–43, 2012.

[10] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, "Heart Rate Variability : Standards of Measurement, Physiological Interpretation, and Clinical Use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[11] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening." *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[12] S. Koelstra, C. Muhl, M. Soleymani, J.-s. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Y. Patras, "DEAP : A Database for Emotion Analysis Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[13] H. Kawahara, T. Takahashi, M. Morise, and H. Banno, "Development of exploratory research tools based on TANDEM-STRAIGHT," in *Proceedings of APSIPA ASC 2009 : Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*. Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, 2009, pp. 111–120.