Dimensional Mapping of Multimodal Integration on Audiovisual Emotion Perception

Yoshiko Arimoto1,2, Kazuo Okanoya1,2,3

1Emotional Arimoto1,2, Kazuo Okanoya1,2,3

1Emotional Arimoto1,2, Kazuo Okanoya1,2,3

1Emotional Arimoto1,2, Kazuo Okanoya1,2,3

1Emotional Arimoto1,2, Kazuo Okanoya1,2,3

Abstract

The aim of this research was to investigate what emotions are perceived from incongruent vocal and facial emotional expressions as an integrated emotional expression. Our approach is unimodal and bimodal perceptual emotional information mapping to dimensions of emotional space created with principal component analysis (PCA). Unimodal perception tests and a bimodal congruent/incongruent perception test were conducted with each stimulus in which professional actors expressed four emotions: anger, joy, fear, and sadness. Observers rated the intensity of six emotions (the four expressed emotions plus disgust and surprise) on a six-point scale. A PCA was performed with the scores of each stimulus to create a perceptual emotional space and to compare the difference between unimodal perception and bimodal perception. The results showed that some incongruent emotional expressions were significantly perceived as inconsistent emotions with expressed emotion.

Index Terms: emotional speech, facial expression, emotion perception, multimodal integration, principal component analysis

1. Introduction

Emotional information from several cues, such as facial and vocal expressions, is simultaneously perceived from several modalities, e.g., auditory and visual. In our daily life, each expression does not always convey the same emotion. Speakers frequently express their emotions with incongruent expressions through their voice and face, such as an angry voice with a smile, or a sad voice with a joyful face. Observers often perceive such incongruent expressions as a single emotion, or even as a third emotion that the speakers do not actually express. This raises questions about what combinations of incongruent emotional expressions are integrated to be perceived as a third emotion that is not expressed by a speaker, and what emotions are perceived from incongruent emotional expressions.

Previous research [1, 2, 3, 4, 5] has indicated that vision is dominant in multisensory perception and that audio information affects the perception of emotion from visual information (i.e., the rate of correct emotion recognition through visual stimuli dropped when incongruent emotional expressions were presented through audio stimuli). However, these findings do not deal with what combinations of emotional expressions convey a third emotion.

To address this issue, we conducted unimodal perception tests and a bimodal perception test with vocal and facial emotional expressions. We performed a principal component analysis (PCA) on the results obtained from the tests to visualize fluctuations between unimodal emotional perceptions and bimodal emotional perceptions and to investigate the possibility of perceiving a third emotion that is not expressed by either the voice or the face during bimodal stimulus. In a previous study, Mower et al. [7] studied the differences between unimodal and bimodal perception on each of the three continuous scales of Pleasure-Arousal-Dominance (PAD) dimension [8]. They clarified that audiovisual emotional perception was biased toward audio emotional expressions on each single dimension. We expanded their single dimensional approach to clarify a modality bias on emotion perception into a two-dimensional approach to explore the possibility of the perception of a third emotion.

One possible technological application of our research is emotion synthesis for real-life agent systems. Mower et al. [7] aimed to apply their research to a synthetic character. Hiruma [9] investigated the influence of incongruities between the emotional content of synthesized speech and the human face on emotion perception. Another technological application of our research is automatic emotion recognition. Previous studies have focused on more natural and spontaneous emotional expressions during a dialog [10, 11, 12], the results of which suggest that recognition accuracy decreases as the number of target emotions the system has to recognize increases [13]. One of the reasons for this is that a speaker might express incongruent emotions and an observer might perceive a single emotion from the speaker’s incongruent emotional expressions as a result of audio-visual integration in the perception process. If it is true, it will be necessary to design audiovisual emotion recognition systems which recognize a different emotion from ones that could be recognized with separate emotional information sources. The findings of our study will be helpful for designing audiovisual emotion recognition systems.

2. Audiovisual material

Audiovisual stimuli were collected for our perception tests. Four professional Japanese actors (two of each sex in their twenties or thirties) were asked to express four emotions: anger (ANG), fear (FEA), joy (JOY), and sadness (SAD). Neutral expressions (NEU). The four emotions were selected with reference to Ekman’s six basic emotions [14]. Each actor was instructed to perform each emotion at three levels of emotional intensity: weak, medium, and strong. Although Ekman proposed disgust and surprise as two of the basic emotions, he also pointed out that these two emotions can be experienced with other emotions, like disgust with anger, and surprise with fear, happiness, anger or others. Moreover, the facial expressions for surprise and fear particularly resemble each other. For this reason and for the sake of convenience, the actors in our study were not asked to exhibit disgust or surprise.

The vocal and facial expressions were recorded separately to enable the actors to concentrate on each unimodal emotional expression. For the audio stimuli, each actor was asked to ut-
A unimodal perception test of intensity ratings of six emotions was conducted for both audio and visual stimuli for two reasons: (1) to select one stimulus for each emotional expression for the following bimodal perception test, and (2) to assess whether observers could correctly perceive the emotion that the actor expressed.

3.1. Method

The unimodal perception test was organized into two blocks, audio only and visual only (in a counter-balanced order). Twenty-three Japanese participants (12 male and 11 female) took part in the two blocks of the perception test. The mean age of the participants was 36.7 (range = 24–63). They were instructed to rate all stimuli based on how strong each of the six emotions was perceived. The sampling frequency was 48 kHz with 16 bits per sample. For the visual stimuli, each actor was asked to create a natural expression starting from a neutral face and to maintain that expression for three seconds. They were asked to utter /e/ or /e:/ when they expressed the face so that we could synchronize their lip movement with their emotional speech when the bimodal stimuli were prepared. Each clip of the emotional facial expressions was 3 ms with a 500-ms margin before the start frame and after the end frame. The start frame was defined as the frame in which an actor started to move one of his/her facial features, e.g., eyes, eyebrow, nose, cheeks, mouth, or forehead. Facial expressions were recorded at 30 fps with a CCD camera at a 480 × 600 pixel resolution.

As the actors performed three levels of intensity of each emotion as many times as they satisfied, they were asked to choose the 5 best expressions among their own vocal or facial expressions of each intensity level of emotional expressions. They were also asked to choose the 13 best neutral expressions among their neutral expressions. The total number of stimuli of one modality for all actors was 292.

3. Unimodal perception test

A unimodal perception test of intensity ratings of six emotions was conducted for both audio and visual stimuli for two reasons: (1) to select one stimulus for each emotional expression for the following bimodal perception test, and (2) to assess whether observers could correctly perceive the emotion that the actor expressed.

3.2. Results and discussion

3.2.1. Stimuli selection

A Steel-type nonparametric multiple comparison test was performed for each stimulus, to select one best stimulus from the actor’s own selection. As a result, the stimuli, which yielded a significantly higher mean score of one perceived emotion corresponding to the expressed emotion than one of the other emotions corresponding to the three other non-intended emotions (p < 0.05), were selected as the stimuli that satisfactorily conveyed the expressed emotion. To make up for a shortage of 10 audio stimuli and 6 visual stimuli, the stimuli that showed the largest mean difference among the three other non-intended emotions were also selected. These stimuli were used in the following analysis and in the bimodal perception test.

3.2.2. Multi-emotional perception of unimodal Stimuli

Mean scores of the six perceived emotions were calculated for each audio/visual stimulus. Table 1 shows the mean intensities across 5 emotional expressions (in row) and 6 perceived emotions (in column) of either vocal or facial expression. The underlined numbers are the mean scores of the perceived emotions consistent with the expressed emotions of the stimuli. These were analyzed with a 6 × 5 ANOVA with the factors of perceived emotion (ANG, JOY, FEA, SAD, DIS, and NEU) and expressed emotion (ANG, JOY, FEA, SAD, and NEU). The results for audio stimuli showed significant main effects of per-
ceived emotion ($F(5, 282) = 9.24, p < 3.71e-08$) and expressed emotion ($F(4, 282) = 10.90, p < 3.14e-08$), as well as the interaction between the main effects ($F(20, 282) = 27.32, p < 2.2e-16$). The results for visual stimuli showed significant main effects of perceived emotion ($F(5, 282) = 7.67, p < 8.89E-07$) and expressed emotion ($F(4, 282) = 15.69, p < 1.31E-11$), as well as the interaction between the main effects ($F(20, 282) = 59.17, p < 2.2e-16$). This indicates that either vocal or facial emotional expression conveys different emotions based on a given expression. Dunnett-type parametric multiple comparison tests were conducted with the mean scores of each stimulus for further investigation. The intensity of emotions inconsistent with the expressed emotion were perceived as much weaker than those that were consistent. This held true for 3 of 4 emotional expressions. However, FEA emotional expressions of both modalities did not show significantly low mean scores at SUR perception in comparison with ones at FEA perception ($p < 0.05$). This indicates that surprise is perceived from either the vocal or facial expressions of fear. This result makes sense when we consider Ekman’s assumption (mentioned in Section 2).

### 4. Bimodal perception test

A bimodal perception test was performed with the bimodal stimuli of vocal and facial emotional expressions for two reasons: (1) to determine whether an incongruity between a vocal and facial emotional expression is perceived as an integrated emotional expression through audio and visual stimuli, and (2) to investigate what emotions are perceived from congruent/incongruent emotional expressions.

#### 4.1. Method

The bimodal stimuli were obtained by simultaneously presenting vocal and facial expressions. For better synchronization of lip movements and speech, the start point of the voice signals of the audio stimuli were adjusted along with the start point of the voice signals of the visual stimuli. The matching could be either “congruent,” in which the vocal and facial expressions portrayed the same emotion (i.e., angry face/angry voice), or “incongruent,” in which the vocal and facial expressions portrayed different emotions (i.e., angry face/sad voice). The bimodal stimuli of different intensity levels of emotional expressions (i.e., moderate angry face/weak sad voice, or moderate fearful face/strong fearful face) were not presented for either case (congruent or incongruent combinations). Figure 2 shows the number and combinations of bimodal stimuli per actor. The number of all bimodal stimuli expressed by all actors came to 292.

Twenty Japanese (12 males and 8 females, mean age = 33.7 (range = 24–63)) participated in the bimodal perception test. The test was organized into two blocks: congruent expression only and incongruent expression only (in a counter-balanced order). All participants took part in both blocks and evaluated all 292 bimodal stimuli. Prior to the experimental blocks, all participants judged 20 practice trials including both congruent and incongruent stimuli.

There were two tasks in each trial. One was to judge the naturalness of the emotional combination of the congruent/incongruent expressions for each stimulus. Some researchers [15, 3] have pointed out that incongruent information between modalities was difficult for observers to integrate at the perceptual level of processing information. This was true for our bimodal perception test of incongruent emotional information. The participants were asked to judge whether each stimulus was a “natural emotional expression” or an “unnatural emotional expression” with regard to the process of information integration. They selected “natural” if they thought the simultaneously presented vocal and facial emotional expressions could be integrated as one emotional expression. The other task was to rate the intensity of the six emotions perceived from the congruent/incongruent emotional expressions on a six-point scale. The method of presenting stimuli and rating the six emotions was the same as in the unimodal perception test described in Section 3.1. The mean scores of each of the six perceived emotions were calculated as the perceived emotional intensities for each stimulus.

#### 4.2. Results and discussion

##### 4.2.1. Naturalness of emotional combination of bimodal stimuli

The 182 stimuli that the majority of participants judged as natural were regarded as natural emotional expressions. Table 2 shows the number of bimodal stimuli across the vocal and facial emotional expressions that were judged natural. All stimuli of the four congruent emotional expressions (NEU, ANG, JOY, and FEA) were judged natural (shaded in gray). However, 11 out of the 12 stimuli of the congruent SAD expressions were judged natural. As for the incongruent emotional expressions, all stimuli of the SAD facial and FEA vocal expressions were also judged natural. More than half of the stimuli of incongruent expressions with the JOY facial expression were not judged natural (underlined and in bold face in Table 2). JOY was considered to be difficult to integrate with another negative emotional expression because it was the only positive emotion among the expressions in this experiment.

##### 4.2.2. Emotions perceived from congruent/incongruent emotional expressions

Mean scores of six perceived emotions were calculated for each stimulus and were then analyzed with $25 \times 6$ ANOVA with the factor of 25 combinations of congruent and incongruent emotional expressions and each of the six perceived emotions (ANG, JOY, FEA, SAD, DIS, and SUR). The results revealed significant main effects of perceived emotions ($F(5, 948) = N = 73$; 5 emotions $^2$ modalities x 3 levels of intensity – 2 duplications (NEU + NEU))

<table>
<thead>
<tr>
<th>Vocal expression</th>
<th>NEU</th>
<th>ANG</th>
<th>JOY</th>
<th>FEA</th>
<th>SAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEU</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>ANG</td>
<td>8</td>
<td>12</td>
<td>7</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>JOY</td>
<td>4</td>
<td>2</td>
<td>12</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>FEA</td>
<td>7</td>
<td>10</td>
<td>8</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>SAD</td>
<td>10</td>
<td>9</td>
<td>6</td>
<td>12</td>
<td>11</td>
</tr>
</tbody>
</table>

Figure 2: Number and combinations of bimodal stimuli per actor.

Table 2: Number of natural combinations of vocal and facial emotional expressions. Cells shaded in gray show combinations in which all stimuli were judged as natural. Underlined numbers shows combinations in which very few stimuli were judged as natural.
Table 3: Mean scores of perceived emotional intensity of bi-modal emotional expressions. Bold-face numbers indicate a significant difference between the bimodal expression and the control expression (Dunnett, \( p < 0.05 \)). NEU–NEU (neutral voice and neutral face) is the control expression.

<table>
<thead>
<tr>
<th>Expression (Voice–Face)</th>
<th>Perception</th>
<th>ANG</th>
<th>JOY</th>
<th>FEA</th>
<th>SAD</th>
<th>DIS</th>
<th>SUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEU–NEU</td>
<td>0.05</td>
<td>0</td>
<td>0.05</td>
<td>0.08</td>
<td>0.14</td>
<td>0.14</td>
<td>0.41</td>
</tr>
<tr>
<td>ANG–ANG</td>
<td>3</td>
<td>0.15</td>
<td>0.11</td>
<td>1.77</td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANG–FEA</td>
<td>1.16</td>
<td>0.01</td>
<td>0.82</td>
<td>0.08</td>
<td>1.82</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>ANG–JOY</td>
<td>0.4</td>
<td>2.75</td>
<td>0.02</td>
<td>0.18</td>
<td>0.48</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>ANG–NEU</td>
<td>0.68</td>
<td>0.04</td>
<td>0.1</td>
<td>0.1</td>
<td>0.84</td>
<td>1.34</td>
<td></td>
</tr>
<tr>
<td>ANG–SAD</td>
<td>1.55</td>
<td>0</td>
<td>0.29</td>
<td>0.72</td>
<td>1.84</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>FEA–ANG</td>
<td>1.12</td>
<td>0</td>
<td>1.19</td>
<td>1.3</td>
<td>1.18</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td>FEA–FEA</td>
<td>0.08</td>
<td>0</td>
<td>2.26</td>
<td>0.78</td>
<td>1.12</td>
<td>2.25</td>
<td></td>
</tr>
<tr>
<td>FEA–JOY</td>
<td>0.1</td>
<td>0.01</td>
<td>0.75</td>
<td>0.61</td>
<td>0.18</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>FEA–SAD</td>
<td>0.16</td>
<td>0.03</td>
<td>1.23</td>
<td>2.21</td>
<td>0.8</td>
<td>1.55</td>
<td></td>
</tr>
<tr>
<td>JOY–ANG</td>
<td>1.96</td>
<td>0.26</td>
<td>0.24</td>
<td>0.19</td>
<td>1.81</td>
<td>1.83</td>
<td></td>
</tr>
<tr>
<td>JOY–FEA</td>
<td>0.36</td>
<td>0.43</td>
<td>1.54</td>
<td>0.29</td>
<td>1.75</td>
<td>2.34</td>
<td></td>
</tr>
<tr>
<td>JOY–JOY</td>
<td>0.01</td>
<td>3.68</td>
<td>0.03</td>
<td>0.04</td>
<td>0.07</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>JOY–NEU</td>
<td>0.19</td>
<td>0.38</td>
<td>0.02</td>
<td>0.06</td>
<td>0.19</td>
<td>1.15</td>
<td></td>
</tr>
<tr>
<td>JOY–SAD</td>
<td>0.38</td>
<td>0.48</td>
<td>0.69</td>
<td>1.16</td>
<td>1.66</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td>NEU–ANG</td>
<td>1.22</td>
<td>0.04</td>
<td>0.18</td>
<td>0.11</td>
<td>1.65</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>NEU–FEA</td>
<td>0.3</td>
<td>0.03</td>
<td>0.8</td>
<td>0.19</td>
<td>1.44</td>
<td>1.54</td>
<td></td>
</tr>
<tr>
<td>NEU–JOY</td>
<td>0.05</td>
<td>2.52</td>
<td>0</td>
<td>0.06</td>
<td>15.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEU–SAD</td>
<td>0.34</td>
<td>0.02</td>
<td>0.54</td>
<td>0.84</td>
<td>1.61</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>SAD–ANG</td>
<td>1.20</td>
<td>0.06</td>
<td>0.56</td>
<td>1.43</td>
<td>1.21</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>SAD–FEA</td>
<td>0.13</td>
<td>0.02</td>
<td>2.06</td>
<td>1.28</td>
<td>1.24</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td>SAD–JOY</td>
<td>0</td>
<td>3.4</td>
<td>0.25</td>
<td>1.15</td>
<td>0</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>SAD–NEU</td>
<td>0.05</td>
<td>0.06</td>
<td>0.24</td>
<td>1.01</td>
<td>0.11</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>SAD–SAD</td>
<td>0.15</td>
<td>0</td>
<td>0.79</td>
<td>2.39</td>
<td>0.83</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>

80.71, \( p < 2.2e-16 \) and 25 combinations of emotional expressions for each perceived emotional category (\( F(23, 948) = 11.10, p < 2.2e-16 \)), as well as the interaction between the main effects (\( F(115, 948) = 20.73, p < 2.2e-16 \)). A post hoc test (Dunnett, \( p < 0.05 \)) was conducted for each perceived emotion with the 25 combinations of emotional expressions. Table 3 shows the results of the post hoc test for ANG, JOY, FEA, SAD, DIS, and SUR perception.

For the perception of ANG, JOY, and FEA (Table 3), any congruent and incongruent emotional expressions that showed a significant difference to the control expression, NEU–NEU (neutral voice with neutral face), were comprised of both or either vocal or facial expression of the corresponding emotion to the perceived emotion. These are valid results because the emotional expression of both or either of the two modalities affected the perception of emotions consistent with those expressions.

A more interesting phenomenon was that emotions that were not expressed either vocally and facially were perceived. The inconsistencies between the perception and expression of emotions for SAD, DIS, and SUR are shown in Table 3. For the SAD perception, six emotional expressions show significant difference to the control expression. Among the six emotional expressions, FEA–ANG (fearful voice with angry face) was not comprised of either the vocal or facial expression of the corresponding emotion to SAD perception. The same phenomenon appeared for DIS and SUR perception. As mentioned earlier, the audio and visual stimuli in this experiment did not include DIS and SUR expressions. According to Table 3, 13 and 10 bimodal expressions show a significantly higher intensity of DIS and SUR, respectively, compared to the control expression. Eight out of 10 bimodal expressions for SUR perception were comprised of both or either vocal or facial expressions of FEA.

We conclude that the SUR perception from those eight bimodal stimuli was caused by the perceptual characteristics of the unimodal expression that FEA expression could not discriminate from SUR because both vocal and facial FEA expressions did not yield significantly higher average ratings on the correspond-

\[ S = \frac{|x_1 y_2 - x_2 y_1|}{2}, \]  

where \((x_1, y_1)\) and \((x_2, y_2)\) are the coordinates of two vertices if

Figure 3: Pattern diagram of (a) line and (b) triangle representing the relationship between emotional perceptions of bimodal and unimodal stimuli.

5. Dimensional mapping to emotional space

To visualize and to investigate possibility of perception of the third emotion which is not expressed by both voice and face of a bimodal stimulus, a principal component analysis (PCA) was performed with all the results of unimodal and bimodal perception tests.

5.1. Method

The main idea of our method to analyze and visualize the perception of a third emotion is to calculate perceptual spatial differences between unimodal stimuli and a bimodal stimulus distributed on an emotional space. A PCA was performed on the mean scores of the six perceived emotions for each audio and visual unimodal stimuli to reduce the number of six perceived emotional categories into the dimension explaining 85% of the eigenvalues. Three principal components were regarded as the dimensions to create the perceptual emotional space. The principal component scores of the vocal and facial stimuli were regarded as coordinate points on the emotional space and those of the bimodal stimuli were predicted with the eigenvectors of the three principal components as coordinate points on the emotional space.

Two shapes are considered to represent the difference between two unimodal stimuli and a bimodal stimulus made up of vocal and facial expression (Fig. 3). One type (Type A) forms a line (a) in Fig. 3) that is drawn between two unimodal stimuli from a vocal expression to a facial expression. The bimodal stimulus comprising the same vocal and facial expression as the unimodal stimulus is placed just on or close to the line between the two unimodal stimuli. This type of bimodal stimuli perception is considered to be affected only by each of the unimodal emotional expression and not to cause the perception of a third emotion. The other type (Type B) forms a triangle ((b) in Fig. 3), and the bimodal stimulus is not placed on the line between two unimodal stimuli but rather is placed at some distance from the line. This type of bimodal stimuli perception is considered to cause the perception of the third emotion.

A triangular area formed by two unimodal stimuli and a bimodal stimulus was calculated to compare these two types of perception. An area \( S \) was calculated by

\[ S = \frac{|x_1 y_2 - x_2 y_1|}{2}, \]

where \((x_1, y_1)\) and \((x_2, y_2)\) are the coordinates of two vertices if
5.2. Result and discussions

5.2.1. Interpretation of three principal components creating emotional space

The perceptual emotional space was formed with the axes of the three principal components. To discuss fluctuations of bimodal stimuli from lines between unimodal stimuli, the three principal components were interpreted as the axes that represent the characteristics of an emotional space. Figures 4 (a), (b), and (c) show the eigenvectors of the principal components PC1, PC2, and PC3, respectively. JOY in Fig. 4 (a) obtained the highest score among the six perceived emotional categories while the scores of the remaining five emotions were all under 0. The PC1 is interpreted as a pleasure axis because JOY is the only positive emotion among the six. The PC2 in Fig. 4 (b) is interpreted as a dominance axis. ANG and DIS, which are characterized by dominant attitudes toward others, obtained a higher score than the other four emotions, while FEA and SAD, which are characterized by submissive attitudes toward others, obtained a lower score. The PC3 is interpreted as an arousal axis because SAD, which is considered a low-arousal emotion, obtained a low score and SUR, which is considered a high-arousal emotion, obtained a high score. The interpretations for PC1 and PC2 were considered to be valid because both the pleasure and dominant axes have previously been proposed by Mehrabian and Russel, and our relative scores for each emotion are consistent with their Pleasure-Arousal-Dominance (PAD) model [8]. However, the arousal axis of ANG, which is usually considered a high-arousal emotion, contradicts their PAD model: in our study, ANG obtained a middle score and was located close to DIS on the arousal axis. Disgust and surprise are often experienced with anger, as mentioned in Section 2. Some of our ANG stimuli might be close expressions of DIS, which is considered a low-arousal emotion but also an unpleasant and dominant emotion just the same as ANG, according to the PAD model. In fact, the participants of our perception test seemed to perceive DIS from ANG stimuli simultaneously. Results of the unimodal perception test (Table 1) suggest that although vocal and facial ANG expressions had significant differences in terms of ANG and DIS perception, the mean scores of DIS perception of ANG expressions were higher than the other four perceived emotions.

5.2.2. Triangle of perceptual difference among modalities

Three two-dimensional space, pleasure–dominance, pleasure–arousal, and dominance–arousal, were obtained as a result of the PCA. A one-way ANOVA was performed with the factor 25 combinations of congruent and incongruent bimodal emotional expressions for each dimension. Results showed a significant main effect of 25 types of congruent and incongruent emotional expression ($F(23, 158) = 5.57$, $p < 1.64e-11$, $F(23, 158) = 4.54$, $p < 4.45e-09$, $F(23, 158) = 4.45$, $p < 7.32e-09$, for each dimension respectively).

A post hoc test (Dunnett, $p < 0.05$) was conducted for each area of the triangle of the bimodal expression with 25 types of bimodal emotional expression. Compared to the control expression of NEU–NEU (neutral voice with neutral face), the bimodal expressions of ANG–FEA, ANG–SAD, FEA–ANG, JOY–ANG, JOY–SAD, NEU–FEA, and NEU–SAD had significantly wide triangles in the pleasure–dominance dimension. Triangular areas of JOY–ANG and JOY–SAD in the pleasure–arousal dimension and ANG–FEA, FEA–ANG, JOY–ANG, and JOY–SAD in the dominance–arousal dimension were also significantly wider than that of the control expression. All bimodal expressions that had significantly different areas from that of the control expressions also showed significantly different scores from that of the control expressions in the post hoc tests in the bimodal perception tests mentioned in Section 4.2.2.

Figures 5 and 6 show triangles representing significant perceptual differences between unimodal stimuli and bimodal stimuli on the pleasure (PC1)–dominance (PC2) dimension and on the dominance (PC2)–arousal (PC3) dimension. A vertex with a “V” mark on the triangle indicates a unimodal stimulus of a vocal emotional expression, and a vertex with an “F” mark indicates a stimulus of a facial emotional expression. The remaining vertex with a “·” mark indicates a bimodal stimuli of congruent/incongruent emotional expression. Figure 5 (b) and Fig. 6 (d) show the control expressions. These figures indicate not only the wider triangles of the bimodal expressions compared to that of the control expression but also in what direction the bimodal stimuli were distant from the lines between two unimodal stimuli. Many of the bimodal emotional expressions fluctuate toward the negative direction on the pleasure axis and toward the positive direction on the dominance axis. On the arousal axis, some of the bimodal expressions fluctuate in the high-arousal direction, but one bimodal expression fluctuates in the low-arousal direction even though both of its unimodal stimuli that were used for the bimodal stimuli had a higher arousal level than the bimodal stimuli, such as FEA–ANG in Fig. 6.

6. Conclusions

We investigated what types of incongruent vocal and facial emotional expressions are perceived as integrated emotional expressions through audio and visual stimuli, and what emotions are perceived from those incongruent emotional expressions. The results of a perception test with audiovisual dynamic stimuli showed that some incongruent emotional expressions were judged as natural emotional expressions and that emotions other than those expressed were perceived from incongruent stimuli. Moreover, the result of dimensional mapping to emotional space created with PCA indicates the possibility of perceiving a third emotion that is not expressed in either the voice or face used as audiovisual stimuli. For further work, we would investigate relationship between integrated emotional perception and physical measurement of vocal and facial expressions, such as prosodic information and dynamic information of action unit of facial expressions.

7. Acknowledgments

We would like to thank Ms. Asami Itoh, Ms. Mika Ohshima, Mr. Yuji Nakamura, and Mr. Tsutomu Haga for kindly providing their voices and facial expressions as our audiovisual stimuli. We also thank Mr. Tomokazu Yokoyama for arranging the actors’ recording schedule.
Figure 5: Triangles showing perceptual differences between unimodal stimuli and bimodal stimuli on pleasure (PC1)–dominance (PC2) dimension (a partial result of a post hoc test (Dunnett, p < 0.05)).

Figure 6: Triangles showing perceptual differences between unimodal stimuli and bimodal stimuli on dominance (PC2)–arousal (PC3) dimension (a partial result of a post hoc test (Dunnett, p < 0.05)).

8. References