Chapter

Motion Generation during Vocalized Emotional Expressions and Evaluation in Android Robots

Carlos T. Ishi

Abstract

Vocalized emotional expressions such as laughter and surprise often occur in natural dialogue interactions and are important factors to be considered in order to achieve smooth robot-mediated communication. Miscommunication may be caused if there is a mismatch between audio and visual modalities, especially in android robots, which have a highly humanlike appearance. In this chapter, motion generation methods are introduced for laughter and vocalized surprise events, based on analysis results of human behaviors during dialogue interactions. The effectiveness of controlling different modalities of the face, head, and upper body (eyebrow raising, eyelid widening/narrowing, lip corner/cheek raising, eye blinking, head motion, and torso motion control) and different motion control levels are evaluated using an android robot. Subjective experiments indicate the importance of each modality in the perception of motion naturalness (humanlikeness) and the degree of emotional expression.

Keywords: emotion expression, laughter, surprise, motion generation, human-robot interaction, nonverbal information

1. Introduction

Vocalized emotional expressions such as laughter and surprise (usually accompanied by verbal interjectional utterances) often occur in daily dialogue interactions, having important social functions in human-human communication. Laughter and surprise utterances are not only simply related to funny or emotional reactions but also can express an attitude (like friendliness or interest) [1, 2].

Therefore, it is important to account for such vocalized emotional/attitudinal expressions in robot-mediated communication as well. Since android robots have a highly humanlike appearance, natural communication with humans can be achieved through several types of nonverbal information, such as facial expressions and head/body gestures. There are numerous studies regarding facial expression generation in robots [3–11]. Most of these are related to symbolic (static) facial expression of the six traditional emotions (happy, sad, anger, disgust, fear, and surprise). However, in real daily interactions, humans can express several types of emotions and attitudes by making subtle changes in facial expression and head/body motion.

When expressing an emotion, humans not only use facial expressions but also synchronize other modalities, such as head and body movements as well as vocalic expressions. Due to a high humanlike appearance in androids, the lack of a modality or of a suitable synchronization among different modalities can cause a strongly negative impression (the "uncanny valley"), when an unnatural facial expression or motion is produced. Therefore, it is important to clarify methodologies to generate motions that look natural, through appropriate timing control.

The author's research group has been working on improving human-robot communication, by implementing humanlike motions in several types of humanoid robots. So far, several methods for automatically generating lip and head motions of a humanoid robot in synchrony with the speech signal have been proposed and evaluated [12–15]. Throughout the evaluation experiments, it has been observed that more natural (humanlike) behaviors by a robot are expected, as the appearance of the robot approaches the one of a human, such as in android robots. Furthermore, it has been observed that unnaturalness occurs when there is a mismatch between voice and motion, especially during short-term emotional expressions, like in laughter and surprise. To achieve a smooth human-robot interaction, it is essential that natural (humanlike) behaviors are expressed by the robot.

In this chapter, motion generation for two vocalized emotional expressions, laughter and surprise, is being focused on. These are usually shorter in duration in comparison to other emotion expressions like happiness, sadness, anger, and fear, and thus it is important to account for a suitable timing control between voice and movements of facial parts, head, and body. The control of different modalities is investigated for achieving natural motion generation during laughter and surprise events of humanoid robots (i.e., when the robot produces a laughter or a vocalized surprise reaction).

In Section 2, related works on motion analysis and generation during emotion expression are presented. In Section 3, the motion generation methods for laughter and surprise expressions are described, along with the motion control methods of an android robot. The motion generation methods are based on analysis results of human behaviors during dialogue interactions [16, 17]. Sections 4 and 5 present evaluation results on the effectiveness of controlling different modalities of the face, head, and upper body (eyebrow raising, eyelid widening/narrowing, lip corner/cheek raising, eye blinking, head motion and torso motion control) and different motion control levels for laughter and surprise expressions. The effects of each modality are investigated through subjective experiments using an android robot as test bed. Section 6 concludes the chapter and presents future work topics. The contents of this chapter are partially included in the author's previously published studies [18, 19]. Readers are invited to refer to those studies, for more details on the motion analysis results.

2. Related work

As stated in the introduction, it is important to synchronize a variety of modalities, including facial movements, speech, and head/body movements, in order to suitably express an emotion.

It has been reported in the emotion-recognition field that the use of both audio and visual modalities provides higher recognition rates than using a single modality [20, 21]. It is also reported that using face and head modalities in combination to the speech modality improves the expression of an emotion in CG (computer graphics) animation, in comparison to using only the face modality [22].

The synchronization of speech and facial expression has also been investigated.

It has been reported that the emotion perceived from the facial expression is altered, when there is a mismatch between the emotions conveyed by the voice and by facial expressions [23]. It has also been reported that when both voice and facial expressions are presented, the judgment of the perceived emotion is strongly influenced by one of the modalities, if the emotion expression of the other modality is ambiguous [24]. It has also been reported that there is a systematic link between eyebrow movements and the fundamental frequency of the voice [25].

Various methods have been proposed for generating several types of facial expressions in android robots [5–11]. However, most of these methods are based on FACS (facial action coding system [26]) for positioning and controlling the actuators to reproduce humanlike facial expressions or for modeling skin deformation based on mechanical deformation models. Furthermore, there has been no evaluation of the synchronization of speech and facial expression and the face-body-head coordination, in all of these works. It is important to evaluate the effects of multimodal expression, for expressing differences of nuance in emotion rather than merely evaluating symbolic facial expressions. Previous studies indicate that the facial parts should also be moved in synchrony with the changes in speech features, in order to achieve natural motion generation. From the same perspective, head and body modalities should also be controlled in synchrony with speech.

However, no previous studies have tackled the challenge of developing suitable multimodal expression control in android robots.

Regarding laughter motion generation particularly, several studies have been reported in the CG animation field. Most of them are related to the ILHAIRE project [27]. For example, a model which generates facial motion position only from laughter intensity is proposed, based on the relation between laughter intensity and facial motion [28]. In [29], the laughter synthesis model above is extended by adding laughter duration as input and selecting recorded facial motion sequences from human motion data. A multimodal laughter animation synthesis method is proposed in [30], by generating lip and jaw motions from speech and pseudo-phoneme features, head and eyebrow motions from pseudo-phoneme and duration features, and torso and shoulder motions from head pitch rotation. In [31], methods to generate rhythmic body movements (torso leaning and shoulder vibration) during laughter are proposed. The torso leaning and shoulder vibrations are reconstructed from human-captured data through synthesis of two harmonics.

Another issue regarding robotics application is that android robots have limitations in the motion DOF (degrees of freedom) and motion range, different from CG agents. Those studies on CG agents have assumed rich 3D models for facial motions, which cannot be directly applied to the android robot control. Therefore, it is important to clarify the effectiveness of different motion generation strategies for providing natural impressions during emotional expressions, under limited DOFs. Some studies have implemented facial expression of smiling or laughing in robots for human-robot interaction [3, 4]. However, these dealt with symbolic facial expressions, so that dynamic features and other modalities during laughter are not taken into account.

In this study, the motion coordination and the effects of several modalities are taken into account for the motion generation in laughter and vocalized surprise expressions.

3. Motion generation in laughter and surprise expressions

The motion generation methods during laughter and surprise utterances are based on analysis results on human-human dialogue interaction data [16–19]. The motion generation methods account for dynamic properties of a motion in synchrony with speech (i.e., when a motion starts and ends relative to the laughter/surprise

expression). The main results of motion timing analyses are summarized in Section 3.1. The motion generation approaches for laughter and surprise expressions and the motion control methods in an android robot are described in Section 3.2.

3.1 Motion timing analysis results

Analysis on laughter motion data indicates that the start time of the smiling facial expression (eye narrowing and lip corner raising) usually matches with the start time of the laughing speech, while the end time of the smiling face (i.e., the instant the face turns back to the normal face) is usually delayed relatively to the end time of the laughing speech by 1.2 ± 0.5 s. An eye blinking is usually accompanied at the instant the face turns back from the smiling face to the normal face. This was observed in 70% of the laughter events. Regarding lip corner raising, it was observed that the lip corners are clearly raised at the laughter segments by expressing a smiling face, while they are slightly raised over a longer period in non-laughing intervals by expressing a slightly smiling face. The percentage in time of smiling faces was 20%, while by including slight smiling faces the percentage in time was 81% on average, ranging from 65 to 100% (i.e., one of the speakers showed slight smiling facial expressions over the whole dialogue). Obviously, these percentages are dependent on the person and the dialogue context. In the analyzed data, most of the conversations were in joyful context. Regarding the upper-body motion, both forward and backward motions are observed. The pitch angle rotation velocities for upper-body motion were 10 ± 5°/s for forward and −10 ± 4°/s for backward directions.

The main findings for the analysis on surprise motion are as follows. First, the occurrence rate of a motion during surprise utterances varies depending on whether the surprise expression is emotional/spontaneous, intentional/social, or quoted, and this rate is highly correlated to the degree of expression in emotional/spontaneous surprise. Second, different motion types have different occurrence rates according to the surprise expression degree. In particular, body backward motion appears with higher frequency when expressing high surprise degrees. Regarding motion time issues, the onset instants of face, head, and body motion are most of the time synchronized with the start time of the surprise utterances, while offset instants are usually later than the end time of the utterances, similarly to the observations in laughter motion analysis. However, the offset times were different. For eyebrow raise, the onset duration was faster than the offset duration, with averages around 200–300 ms for onset and 400–500 ms for offset. For the upper body, onset and offset durations were both around 0.8 s for small movements, and around 1.2 and 1.5 s for large movements.

More details on the motion analysis results can be found in [16–19], including different types and functionalities of laughter and surprise in natural dialogue interactions.

3.2 Description of motion generation in laughter and surprise expressions and control methods in an android robot

Based on the motion timing analysis results presented in Section 3.1, motion generation methods during laughter and surprise utterances are proposed, by accounting for the following modalities: facial expression control (eyelid narrowing and lip corner raising for laughter, eyelid widening and eyebrow raising for surprise), head motion control (head pitch direction), eye blinking control at the transition between smiling/surprising face to the neutral face, and body motion control (torso pitch direction).

Figure 1 shows a block diagram of the motion generation method for laughter and surprise utterances in an android robot. The method requires the speech signal and the laughing/surprise intervals as input. In autonomous robots, the laughing speech intervals

and surprise utterance intervals are given a priori, while in tele-operated robots, these have to be automatically detected from the speech signal of the tele-operator.

A female-type android robot, called ERICA, was used to evaluate the effects of different modalities for motion generation. However, the methodology can be applied to any robot having equivalent degrees of freedom (DOFs). **Figure 2** shows the external appearance and the actuators of the android robot.

As shown in **Figure 2**, the android ERICA has 13 degrees of freedom for the face, 3 degrees of freedom for the head motion, and 2 degrees of freedom for the upper-body motion. Among these, the following ones were controlled for laughter and surprise expressions: upper eyelid control (actuator 1), lower eyelid control (actuator 5), eyebrow raise control (actuator 6), lip corner raise control (actuator 8, cheek is also raised), lip corner stretch control (actuator 10), jaw lowering (mouth opening) control (actuator 13), head pitch control (actuator 15), and upper-body pitch control (actuator 18). All actuator commands range from 0 to 255. The numbers in red in **Figure 2** indicate default actuator values for the neutral position.

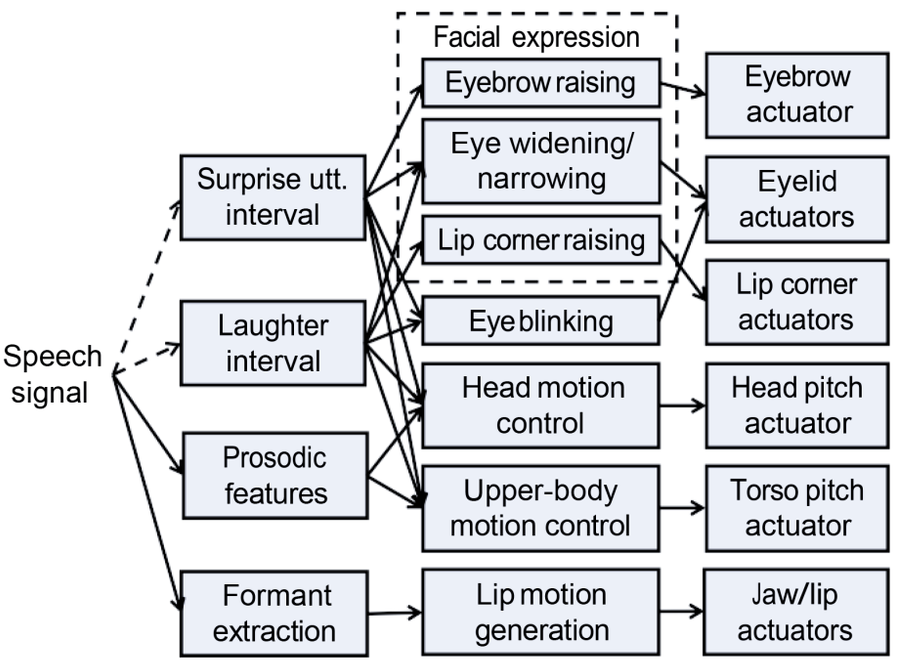

Figure 1.
Block diagram of the motion generation during laughing speech and surprise utterances.

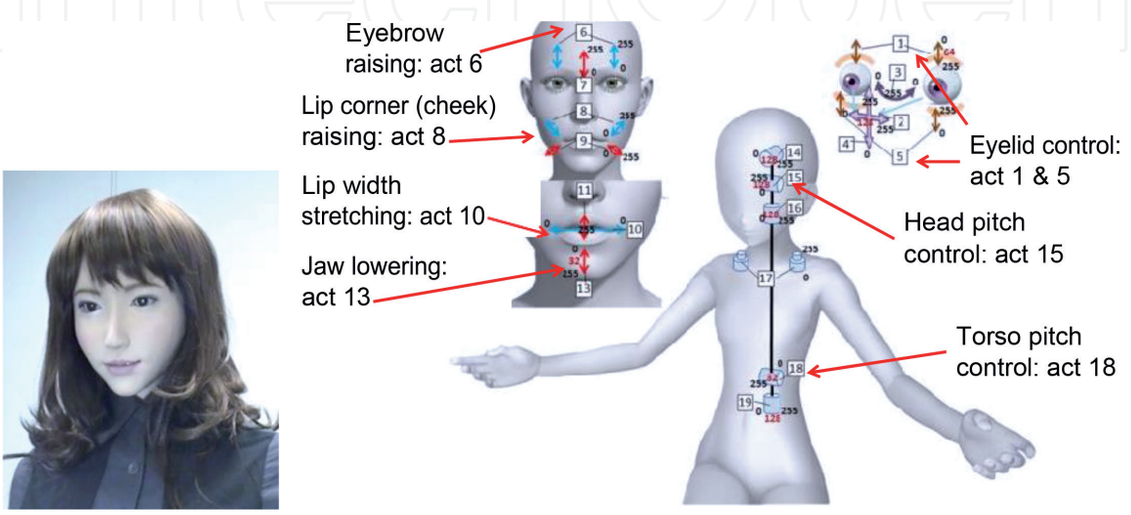

Figure 2.
External appearance of the female-type android robot ERICA and corresponding actuators.

3.2.1 Facial motion control

Before explaining how the facial motion is controlled for different facial expressions, it is worth to clarify that the actuator values presented in this section were manually adjusted for the android ERICA, in order to achieve a desired facial expression. Thus, the actuator values are included for reference, but for other robots having different actuation ranges, these values have to be adjusted by looking at the resulting facial expressions.

For the facial expression during laughter, the lip corner is raised (act[8] = 200), and the eyelids are narrowed (act[1] = 128, act[5] = 128). These values were set so that a smiling face can be clearly identified, as shown in the right panel of **Figure 3** (compare the generated smiling face in the right panel with the neutral face in the left panel). The mouth aperture depends on the vocalic contents of the laughing voice, as will be explained later. The timing of the facial motion control is based on the analysis results, so that the eyelid and lip corner actuator commands are sent at the instant the laughing speech interval starts, and the actuator commands are set back to the neutral position 1–1.5 s after the end of the laughing speech interval.

During preliminary analysis on motion generation, it has been observed that the facial expression of the neutral face (i.e., in non-laughter intervals) looked scary for the context of a joyful conversation. In fact, the lip corners were slightly or clearly raised in 80% of the dialogue intervals. A slight smile face was kept during non-laughter intervals, by controlling the eyelids and lip corner actuators to have intermediate values between the laughter smiling face and the neutral (non-expression) face. For the facial expression during the idle slight smile face, the lip corner is partially raised (act[8] = 100), and the eyelids are partially narrowed (act[1] = 90, act[5] = 80), to obtain the impression of a slight smiling face, as shown in the middle panel of **Figure 3**.

For the facial expression in surprise utterances, the eyebrow raise and eyelid widening are coordinated and controlled at two levels of expression. The target actuator values are set by looking at the facial expressions of the android robot, in order to provide an appearance of a slight surprise face for level 1 and a clear surprise face for level 2. For the android ERICA, the target eyebrow actuators are set to act[6] = 127 for level 1 and act[6] = 255 for level 2, and the upper and lower eyelid actuators are set to {act[1] = 80; act[5] = 60} for level 1 and {act[1] = 40; act[5] = 30} for level 2. For the neutral idle face (corresponding to level 0), these actuators are set to {act[6] = 0; act[1] = 90; act[5] = 80}. As stated before, these values have to be manually adjusted for different robots, in a way to obtain the desired

Figure 3.
Examples of generated facial expressions by eyelid and lip corner control: neutral face (left), idle slight smile face in non-laughter intervals (middle), and smile face during laughter intervals (right).

facial expression. **Figure 4** shows examples of the produced facial expressions for each of these levels. The facial expression at level 1 may not appear to be a surprised facial expression by only looking at the static picture. However, when looking at the facial movements from the neutral face, it is possible to perceive a change in the facial expression.

Regarding the timing of motion control, the eyelid and eyebrow actuator commands are sent at the instant the surprise utterance interval starts, and the actuator commands are set to move back to the neutral position within 0.5 s after the end of the utterance.

For both laughter and surprise expressions, an eye blinking motion is added, considering that an eye blinking is usually accompanied when the facial expression turns back to the neutral face. An eye blink is implemented in the android, by closing the eyes (act[1] = 255 and act[5] = 255) during a brief period of 100 ms and opening the eyes back to the neutral face (act[1] = 64, act[5] = 0) or to an idle smiling face (act[1] = 90; act[5] = 80), as shown in the left and middle panels of **Figure 3**.

3.2.2 Upper-body motion control

For laughing speech, the upper body is moved to the forward and backward directions. In order to achieve smooth movements, the upper-body actuator is controlled according to half cosine functions, as defined in the following expressions:

$$torsopitch[\mathrm{t}] = upbody_{target} \times \frac{1 - \cos\left(\pi \frac{t}{T_{max}}\right)}{2} \tag{1}$$

The upper body is moved from the start point of a laughing speech interval, in order to achieve a maximum target angle corresponding to the actuation value $upbody_{target}$, in a time interval of T_{max}.

From the end point of the laughing speech interval, the upper body is moved back to the neutral position according to an inverse cosine function as shown in the following expression.

$$torsopitch[\mathrm{t}] = \left(upbody_{end} - upbody_{neutral}\right) \times \frac{1 - \cos\left(\pi + \pi \frac{t - t_{end}}{T_{max}}\right)}{2} \tag{2}$$

$upbody_{end}$ and t_{end} are the actuator value and the time at the end point of the laughter speech interval, and $upbody_{neutral}$ corresponds to the actuator value for the

Figure 4.
Examples of generated facial expressions for eyebrow and eyelid control at level 0 (neutral idle face, left), level 1 (slight surprise face, middle), and level 2 (clear surprise face, right).

android's neutral pose. Thus, if the laughter interval is shorter than T_{max}, the upper body does not achieve the maximum angle.

The $upbody_{target}$ was adjusted to −10 degrees (which is the mean body pitch angle range in human data), and the time interval T_{max} to achieve the maximum angle was adjusted to 1.5 s (a bit longer than the human average time, to avoid jerky motion in the android).

For surprise utterances, the upper body is moved in the backward direction at the start point of the surprise utterance and then moved back to the neutral position. Two levels are controlled corresponding to about 2 degrees for level 1 and 4 degrees for level 2 (which was the maximum angle achieved by the android).

Regarding the timing control, the upper body is moved back to the neutral position from 0.3 s after the end point of the surprise utterance interval. The onset duration to achieve the maximum angle is set to 0.8 s, while the offset duration to move back to the neutral idle position is set to 1.5 s. Half cosine functions are used to smooth motion velocity changes in the current and target positions, as in the expressions (1) and (2) for laughter motion control.

The torso pitch actuator in the android (actuator 18) is then controlled around the neutral pose actuator value ($upbody_{neutral}$), according to the following expression:

$$\mathrm{act}[18] = upbody_{neutral} + torsopitch[\mathrm{t}] \tag{3}$$

3.2.3 Head motion control

For the head motion control, a method for controlling the head pitch (vertical movements) from the voice pitch (fundamental frequencies, *F*0) is employed. This is based on the fact that there is some correlation between head motion and voice pitch [15, 32]. Although this correlation is not very high (i.e., this control strategy is not exactly what humans do during speech), natural head motions are expected to be generated during laughing and surprise expressions, since humans usually tend to raise the head for high *F*0s especially in inhaling laughter intervals, and high-pitched surprise utterances. The following expression is used to convert *F*0 values to the head pitch actuator:

$$headpitch\_F0[\mathrm{t}] = (F0[\mathrm{t}] - center\_F0) \times F0\_scale \tag{4}$$

where *center_F*0 is the speaker's average *F*0 value (around 120 Hz for male and around 240 Hz for female speakers) converted to semitone units and *F*0 is the current.

*F*0 value (in semitones) and *F*0*_scale* is a scale factor for mapping the *F*0 (voice pitch) changes to head pitch movements. For the experiments, *F*0*_scale* is set in a way that a 1-semitone change in voice pitch corresponds to ~1-degree change in head pitch rotation.

Preliminary evaluation has shown that the robot motion looked unnatural during a surprise expression, when the head was facing the upward direction, while the body moved in the backward direction. In fact, it has been observed from the human motion data that the speaker is usually looking at the dialogue partner during a surprise expression. The following additional control in the head pitch actuator deals with this issue, by moving the head in the inverse direction to the body pitch movement:

$$headpitch[\mathrm{t}] = headpitch\_F0[\mathrm{t}] - torsopitch[\mathrm{t}] \tag{5}$$

The head pitch actuator in the android (actuator 15) is then controlled around the neutral pose actuator value ($headpitch_{neutral}$), according to the following expression:

$$act[15] = headpitch_{neutral} + headpitch[t] \quad (6)$$

3.2.4 Lip motion control

The lip motion is controlled based on a formant-based lip motion control method [12]. The method is based on the fact the first and second formants (resonance frequencies in the vocal tract) can be associated to the lip height and lip width, respectively, after some speaker normalization procedure. The jaw actuator (actuator 13) is controlled using the estimated lip heights, and the lip stretch actuator (actuator 10) is controlled using the estimated lip widths.

In this way, appropriate lip shapes can be generated in laughter segments with different vowel qualities (such as in "hahaha" and "huhuhu") as well as in vocalized surprise segments with different vowel qualities (such as in "eh!" and "ah!"), since the method is based on the vowel formants.

4. Evaluation of the laughter motion generation

This section presents evaluation results on the laughter motion generation method, by controlling different modalities of the face, head, and body. The experimental setup is described in Section 4.1; the evaluation results and the interpretation of the results are presented in Section 4.2.

4.1 Experimental setup

Two conversation passages of about 30 s including multiple laughter events were extracted from a dialogue database, and the corresponding motion data was generated in the android ERICA, based on the method described in the Section 3.2. The speech signal and the laughter speech interval information are provided as input. The two conversation passages were extracted from different speakers and will be named "voice 1" and "voice 2." "voice 1" includes social and embarrassed laughter, while "voice 2" includes emotional and funny laughter.

Table 1 shows the five motion types (named "A"–"E") generated in the android, taking the effects of different modalities into account.

"Eyelids" and "lip corners" are controlled to express a smiling facial expression (corresponding to Duchenne smile faces [33]) during laughter. These are present in all conditions. "Lip corners" corresponds to a lip corner raising motion, which is also accompanied by a cheek raising motion in the android, while "eyelids" corresponds to an eye narrowing motion.

"Eye blink" corresponds to an eye blinking motion, when the face expression is turned back to the neutral (idle) face, from a smiling face. "Head" corresponds to the motion control of the head pitch (vertical head movements) from the voice pitch. "Idle smile face" corresponds to a slight smiling face during non-laughter intervals. "Upper body" corresponds to the motion control of the torso pitch (front-back upper-body movements) in long laughter events.

Video clips are recorded for each motion type and used in the subjective evaluation experiments. Video-based evaluation is conducted instead of face-to-face evaluation since the participants do not interact with the robot. Pairwise comparisons are conducted in order to investigate the effects of the different motion controls. The evaluated motion pairs are described in **Table 2**.

Motion	Controlled modalities
A	Face (eyelids + lip corners) + eye blink + head
B	Face (eyelids + lip corners) + head
C	Face (eyelids + lip corners) + eye blink
D	Face (eyelids + lip corners) + eye blink + head + idle smiling face
E	Face (eyelids + lip corners) + eye blink + head + idle smiling face + upper body

Table 1.
The controlled modalities for generating five motion types during laughter events.

Motion pair	Differences in the controlled modalities
A vs. B	Presence/absence of "eye blink" control ("eyelids," "lip corners," and "head" are in common)
A vs. C	Presence/absence of "head" control ("eyelids," "lip corners," and "eye blink" are in common)
A vs. D	Absence/presence of "idle smiling face" control ("eyelids," "lip corners," "eye blink" and "head" are in common)
D vs. E	Absence/presence of "upper-body" control ("eyelids," "lip corners," "eye blink," "head," and "slightly smiling face" are in common)

Table 2.
Motion pairs for comparison of the effects of different modalities in laughter.

In the evaluation experiments, pairs of videos are presented for the participants. The order of the videos for each pair is randomized. The videos are allowed to be replayed at most two times each.

After watching each pair of videos, participants are asked to grade the preference scores for pairwise comparison, and the overall naturalness scores for the individual motions, in 7-point scales, according to the questionnaire below. The numbers within parenthesis are used to quantify the perceptual scores.

Q1. Which motion looked more natural (humanlike)? Motion A is clearly more natural (−3), Motion A is more natural (−2), Motion A is slightly more natural (−1), Difficult to decide (0), Motion B is slightly more natural (1), Motion B is more natural (2), Motion B is clearly more natural (3).

Q2. Is the motion natural (humanlike)? very unnatural (−3), unnatural (−2), slightly unnatural (−1), difficult to decide (0), slightly natural (1), natural (2), very natural (3).

The first question was answered for each video pair, while the second question was answered for each of the individual videos. For the motion types A and D, which appear multiple times, individual scores are graded only once, at the first time the videos are seen. Besides the perceptual scores, participants are also asked to write the reason of their judgments, if a motion is perceived as unnatural.

The sequence of motion pairs above was evaluated for each of the conversation passages ("voice 1" and "voice 2"). Twelve remunerated subjects (male and female, aged from 20 to 40 s) participated in the evaluation experiments.

4.2 Evaluation results

Figure 5 shows the evaluation results for pairwise comparisons. Statistical analyses are conducted by t-tests (* for $p < 0.05$ and ** for $p < 0.01$ confidences). For the preference scores in the pairwise comparison, significance tests are conducted in comparison to 0 scores, which correspond to unperceivable differences.

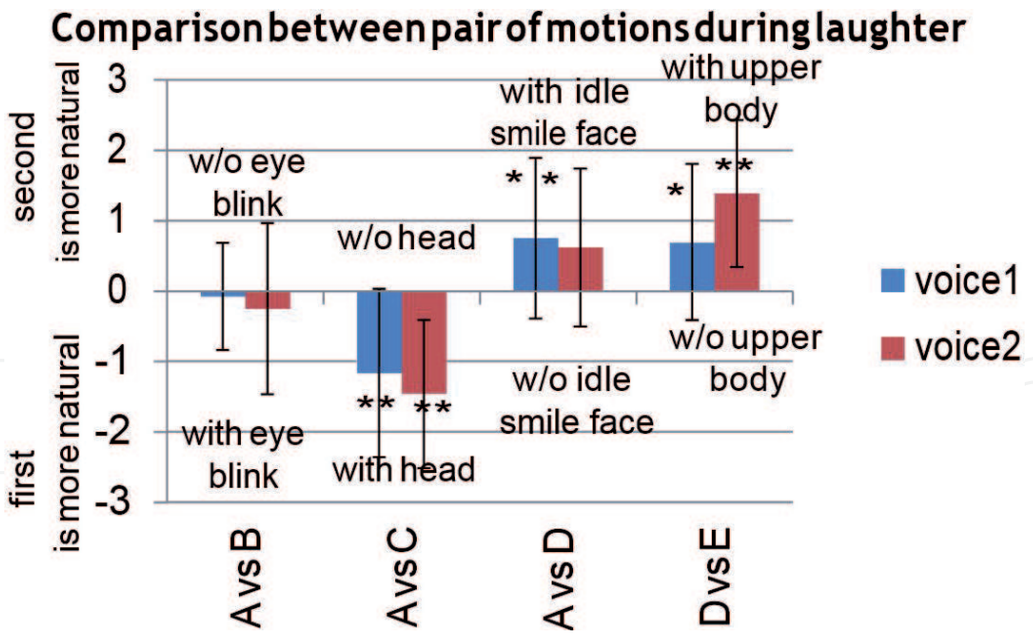

Figure 5.
Subjective preference scores between motion pairs in laughter motion generation (average scores and standard deviations). (Negative average scores indicate the first condition was preferred, while positive average scores indicate that the second condition was preferred).

The differences between the motion types A and B (with and without eye blinking control) are subtle, so that most of the participants could not perceive differences. However, subjective scores showed that the inclusion of eye blinking control was evaluated to look more natural for both conversation passages ("voice 1" and "voice 2").

The comparison between the motion types A and C (with and without head motion control) indicates that the inclusion of head motion control clearly increases the motion naturalness ($p < 0.01$) for both "voice 1" and "voice 2." The participants' judgments were remarkable (the differences in the motion videos were clear).

The comparison between the motion types A and D (without or with idle smile face) indicates that keeping a slight smiling face in the intervals other than laughing speech was also effective to increase motion naturalness ($p < 0.01$).

Finally, the comparison between the motion types D and E (with and without upper-body motion) indicates that the inclusion of upper-body motion also increases motion naturalness ($p < 0.05$ for "voice 1," $p < 0.01$ for "voice 2"). The differences are more evident in "voice 2" (in comparison to "voice 1") since "voice 2" contained longer duration for the laughter events within the conversation passage, and consequently the upper-body movements were more clear.

Figure 6 shows the results for perceived naturalness graded for each motion type. The results of subjective scores shown in **Figure 6** indicate that, overall, slightly natural to natural motions could be achieved by the laughter motion generation method including all motion control types.

The motion type C is the only one that received negative average scores, meaning that if the head does not move, the laughter motions will look unnatural. This indicates that the *F*0-based method for head pitch control is effective for increasing motion naturalness during laughter. However, some of the participants pointed out that the motions would look more natural, if other axes of the head also move. This is a topic for future work.

Regarding the insertion of eye blinking, at the instant the facial expression turns back to the neutral face (motion type B), although the comparisons between motion types A and B were not statistically significant (since the visual difference

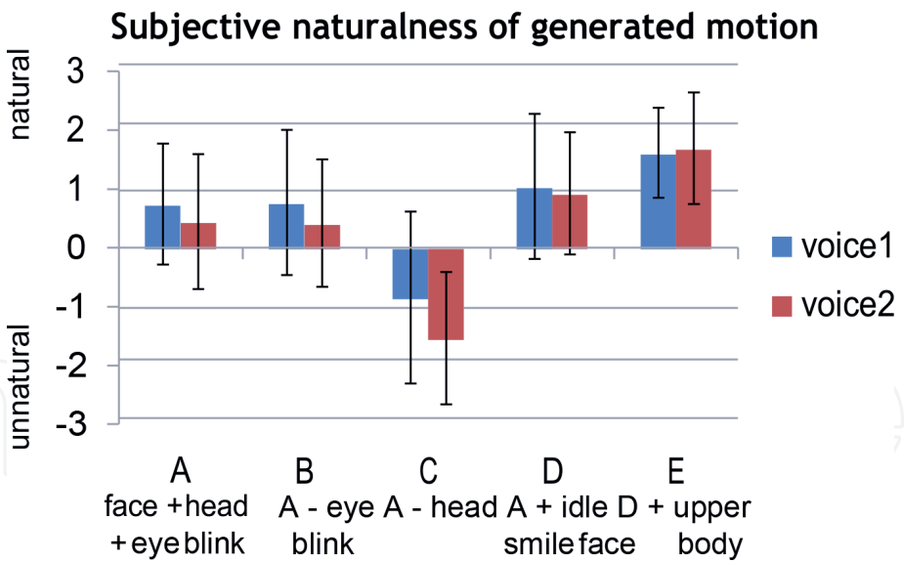

Figure 6.
Subjective naturalness scores for each motion type in laughter motion generation (average scores and standard deviations).

is subtle). However, the participants who perceived the difference judged the presence of eye blinking to be more natural. The eye blinking control is thought to work as a cushion to alleviate the unnaturalness caused by sudden changes in facial expression. The insertion of such a small motion could possibly be used as a general method for other facial expressions.

The control of idle slight smile face in non-laughter intervals (motion type D) was shown to be effective to improve the naturalness, since the conversation context was in joyful situations. However, for a more appropriate control of slight smile face, detection of the situation might be important.

The reason why motion type E (with upper-body motion) was clearly judged as more natural than motion type D (without upper-body motion) for "voice 2" is that it looks unnatural if the upper body does not move during long and strong emotional laughter. The proposed upper-body motion control was effective to relieve such unnaturalness. Regarding intensity of the laughter, although it was implicitly accounted in the present work, by assuming high correlation between pitch and duration with intensity, it could also be explicitly modeled on the generated motions.

5. Evaluation of the surprise motion generation

This section presents evaluation results on the surprise motion generation method, by controlling different modalities and different control levels of the face, head, and body. The experimental setup is described in Section 5.1, and the evaluation results are presented and discussed in Section 5.2.

5.1 Experimental setup

The surprise motion generation method was evaluated for the interjectional utterances "e" and "a," which are the ones that most frequently occurred in dialogue interactions for expressing surprise. Sixteen dialogue passages of about 10 s including interjectional utterances "e" or "a" expressing different degrees of surprise were extracted from the dialogue database. Then motion was generated in the android,

based on the method described in Section 3.2. The speech signal and the surprise utterance interval information are provided as input for motion generation.

Table 3 lists the six motion types generated in the android, for evaluating the effects of different modalities and different degrees of motion control, during surprise expressions. The motion types in **Table 3** are named according to the modality and control levels: "e" stands for eyebrow and eyelids, "h" for head, and "b" for body. The numbers following these letters indicate the control levels. Level "0" indicates no control, level "1" indicates small movements, and level "2" indicates large movements. The facial expressions of levels "1" and "2" for "eyebrows + eyelids" are shown in the middle and right panels in **Figure 4**. The levels "1" and "2" for body motion indicate maximum range of 2 and 4 degrees, respectively, as explained in Section 3.2. The head movements are controlled from the voice, so that 1-semitone change in voice pitch corresponds to ~1 degree for head pitch (Section 3.2). The six motion types were chosen in order to reduce the efforts of the annotators while allowing the comparison of pairs between presence/absence and degree of a motion.

Video clips were recorded, for each motion type and each dialogue passage, to be used in the subjective experiments.

Considering that the range and amount of body movements will be small in short interjectional utterances (around 200 ms), only the three motion types without body control (e2 + h0 + b0, e1 + h1 + b0, and e2 + h1 + b0) were evaluated for short interjectional utterances. For the long interjectional utterances, all six motion types were evaluated. From the eight "a" utterances, seven were short, while from the eight "e" utterances, four were short. Thus, a total of 63 videos ((7 + 4) × 3 short utterances + (1 + 4) × 6 long utterances) were used for evaluation.

In the experiments, the participants are asked to watch all 63 videos and to grade each video with perceptual subjective scores, according to the questionnaire below. The numbers within the parentheses were used to quantify the perceptual scores. The order of the videos is randomized, and the participants are allowed to watch at most two times each.

Q1. What is the perceived degree of surprise expression (regardless of whether an expression is emotional/spontaneous or social/intentional)? No expression (0), slight expression (1), clear expression (2), strong expression (3).

Q2. Is the motion natural (humanlike)? Very unnatural (−3), unnatural (−2), slightly unnatural (−1), difficult to decide (0), slightly natural (1), natural (2), very natural (3).

Q3. Do you feel that the surprise expression is emotional/spontaneous or social/intentional? Intentional (−2), slightly intentional (−1), difficult to decide (0), slightly emotional (1), emotional (2).

Motion type	Controlled modalities
e2 + h0 + b0	Eyebrows + eyelids (level 2)
e2 + h0 + b2	Eyebrows + eyelids (level 2) + body (level 2)
e1 + h1 + b0	Eyebrows + eyelids (level 1) + head
e2 + h1 + b0	Eyebrows + eyelids (level 2) + head
e2 + h1 + b1	Eyebrows + eyelids (level 2) + head + body (level 1)
e2 + h1 + b2	Eyebrows + eyelids (level 2) + head + body (level 2)

Table 3.
Modalities controlled for generating six motion types in surprise utterances.

Eighteen remunerated subjects (male and female, aged from 20s to 40s) participated in the evaluation experiments.

5.2 Evaluation results

We consider that the degree of surprise expression is affected by both audio and visual modalities. In order to account for the effects of the voice modality, the utterances used in the experiment were categorized into three levels, according to their perceptual degrees of surprise graded only from the voice. The resulting number of utterances was 8 for voice group 1 (all short interjections), 7 for voice group 2 (3 short and 4 long interjections), and 1 for voice group 3 (long interjection).

Figure 7 shows the average subjective scores (vertical axes) for surprise expression degree, motion naturalness, and emotional/intentional impression, according to the voice groups (horizontal axes: surprise expression degrees by voice only), for each of the six motion types. Note that the different levels in the horizontal axis are based on voice only, while the subjective scores in the vertical axes are based on voice plus motion modalities.

Pairwise comparisons are conducted to investigate the effects of presence/absence or degree of motion control, and statistical significance tests are conducted through t-tests. Firstly, the effects of controlling the motion degrees of eyebrow and eyelids are analyzed by comparing motion types e1 + h1 + b0 and e2 + h1 + b0. It can be observed in the upper panel of **Figure 7** that the average perceptual scores for surprise expression degree increase by about 0.7 points (on a 0–3-point scale) for voice group 1 ($p < 0.01$) and by about 0.5 points in voice group 3 ($p < 0.01$). This indicates that a slight change in the eyebrow/eyelid control is effective for changing the perceived degree of surprise.

Next, the effects of controlling the head motion modality are analyzed by comparing the results for the motion types e2 + h0 + b0 and e2 + h1 + b0. The differences in surprise expression degree between these two motion types are about 0.2 points for voice group 1 ($p < 0.01$) and about 0.4 points for voice group 3 (n.s., $p = 0.09$), which are slightly smaller than the effects of eyebrow/eyelid control.

The effects of controlling the body motion modality are analyzed by comparing the results between the motion types e2 + h0 + b0 and e2 + h0 + b2 (when head motion is not controlled) or between the motion types e2 + h1 + b0 and e2 + h1 + b2 (when head motion is controlled). It is observed that, when head motion is not controlled (h0), the effects of controlling or not the body motion (b0 vs. b2) increase the surprise degrees by about 0.4–0.5 points for voice groups 2 and 3 ($p < 0.01$). When head motion is controlled (h1), the increase in the perceptual surprise degree is smaller by about 0.3 points (h1 + b0 vs. h1 + b2; $p < 0.05$), probably because the contribution of head motion is superimposed. Although the differences were not statistically significant, a gradual increase can be observed for the gradual control of body motion (b0 vs. b1 vs. b2, for the motion type e2 + h1).

Regarding the naturalness scores, the results in the middle panel of **Figure 7** indicate slightly natural to natural scores in almost all motion types. By comparing the motion types e2 + h0 + b0 and e2 + h1 + b0, it can be inferred that head motion has important effects on the naturalness (humanlike) perception when the body does not move (b0). The naturalness scores are increased by about 0.5 points on average ($p < 0.01$), by inclusion of head motion.

Regarding the subjective spontaneity degree, the results in the bottom panel of **Figure 7** show that the average scores in motion types e1 + h1 + b0 and e2 + h0 + b0 are negative in voice group 3, indicating that if the amount of motion decreases,

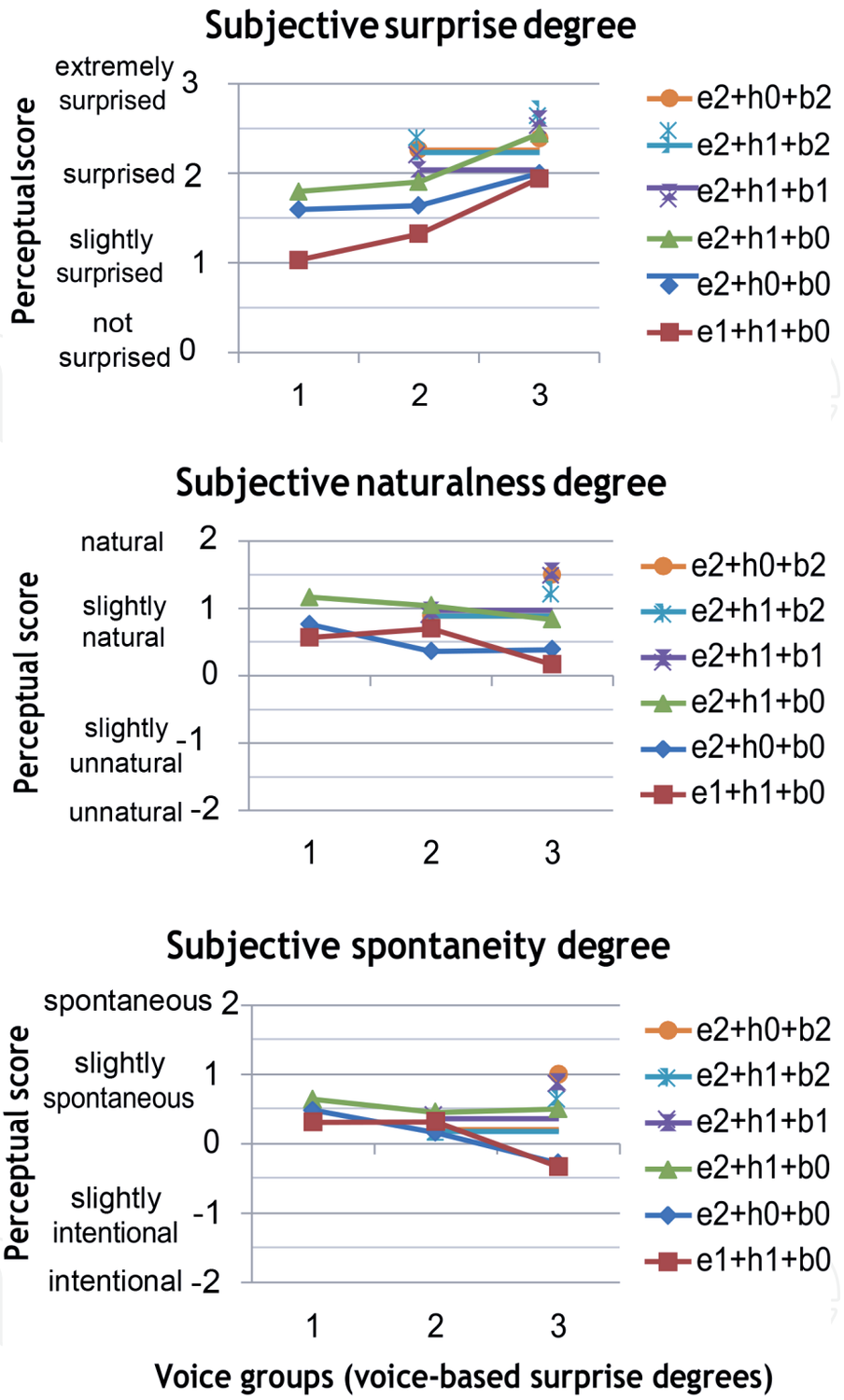

Figure 7.
Subjective perceptual scores of surprise expression degree (top), naturalness degree (mid), and emotional/intentional impression degree (bottom) for each motion type, according to the voice groups (horizontal axis: voice-based surprise expression degrees).

the surprise utterances might be perceived as intentional rather than emotional ($p < 0.01$). On the other hand, the naturalness scores decrease in motion types with fewer motion (e1 + h1 + b0 and e2 + h0 + b0) in voice group 3 (high surprise expression degree by voice only), as shown in the middle panel of **Figure 7**. This is thought to be due to the mismatch between surprise expressions by voice and motion modalities. Another interpretation is that these motion types are perceived as being unnatural, because from the dialogue context, an emotional/spontaneous expression was expected.

Finally, regarding the effects of the voice modality, the results in the upper panel of **Figure 7** for the subjective surprise degrees clearly show that within a motion type, the subjective surprise degrees increase according to the voice groups (voice-based surprise degrees). This means that the perception of surprise degree is dependent on the surprise expression from the voice and, moreover, that by controlling the motion degrees of different modalities, the degree of surprise expression transmitted from the combination of voice and motion can be biased by a certain amount. For example, for the utterances in voice group 1, the subjective surprise degree can be raised to 1.8 on average by controlling the head and eyebrows (e2 + h1 + b0), while the utterances in voice group 3 can have their subjective surprise degree reduced to around 2 if the head and body are not controlled (e2 + h0 + b0).

6. Conclusion and final remarks

Methods for motion generation synchronized with laughter speech and vocalized surprise expressions were described, based on analysis results of human behaviors on facial, head, and body motions during dialogue interactions.

The effectiveness of controlling different modalities of the face, head, and upper body (eyebrow raising, eyelid widening/narrowing, lip corner/cheek raising, eye blinking, head pitch, and torso pitch motion control) and different motion control levels were evaluated using an android robot. The evaluation was conducted through subjective experiments, by comparing motions generated with different modalities and different motion control levels.

Evaluation results for laughter motion generation indicated that motion is perceived as unnatural, if only the facial expression (lip corner raising and eyelid narrowing) is controlled (without head and body motion control). The motion naturalness scores increased when head pitch, eye blinking (at the instant the facial expression turns back to neutral face), idle smile face (during non-laughter intervals), and upper-body motion are also controlled. The best naturalness scores are achieved when all modalities are controlled.

Evaluation results for surprise motion generation indicated that (1) eyebrow/eyelid motion control is effective in changing the perceptual degrees of surprise expression, (2) upper-body motion control is effective for increasing the degrees of surprise expression and naturalness, (3) head motion is more effective for increasing naturalness (rather than surprise degree), (4) the degrees of surprise expression for different motion types are biased by the surprise degrees expressed by the voice-only modality, and (5) utterances with high surprise degrees may be interpreted as intentional (rather than emotional or spontaneous) if they are not accompanied by upper-body motion.

In the present study, it was shown that with a limited number of DOFs (lip corner, eyelids, eyebrows, head pitch, torso pitch), natural motion could be generated for laughter and surprise expressions. Although the android robot ERICA is used as a test bed for evaluation, the described motion generation approach can be generalized for any robot having equivalent DOFs.

Remaining topics for future work include automatic detection of laughing speech intervals and surprise utterance intervals from acoustic features, in order to automate the motion generation process from the input speech signal. Prediction of surprise expression degrees from acoustic features and explicit modeling of laughter intensity are also remaining tasks for motion generation automation. The control strategy of head tilt and shake axes, the investigation of eye blinking insertion for alleviating unnaturalness caused by sudden changes in other facial expressions,

and the detection of situation for slight smile face control are remaining topics for improving motion naturalness.

Acknowledgements

This research was supported by JST/ERATO, Grant Number JPMJER1401. Special thanks go to Takashi Minato for the contributions in the android motion control and discussions on motion generation. The author also thanks Mika Morita, Megumi Taniguchi, Kyoko Nakanishi, and Tomo Funayama for their contributions in the data analysis and experimental setup.

Author details

Carlos T. Ishi
ATR Hiroshi Ishiguro Labs., Kyoto, Japan

*Address all correspondence to: carlos@atr.jp

References

[1] Devillers L, Vidrascu L. Positive and negative emotional states behind the laughs in spontaneous spoken dialogs. In: Proc. of Interdisciplinary Workshop on the Phonetics of Laughter; 2007. pp. 37-40

[2] Campbell N. Whom we laugh with affects how we laugh. In: Proc. of Interdisciplinary Workshop on The Phonetics of Laughter; 2007. pp. 61-65

[3] Breazeal C. Emotion and sociable humanoid robots. International Journal of Human-Computer Studies. 2003;**59**:119-155

[4] Zecca M, Endo N, Momoki S, Itoh K, Takanishi A. Design of the humanoid robot KOBIAN-preliminary analysis of facial and whole body emotion expression capabilities. In: Proc. of the 8th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2008); 2008. pp. 487-492

[5] Wu Y, Thalmann NM, Thalmann D, Dynamic Wrinkle A. Model in facial animation and skin aging. Journal of Visualization and Computer Animation. 1995;**6**(4):195-205

[6] Hashimoto T, Hiramatsu S, Tsuji T, Kobayashi H. Development of the face robot {SAYA} for rich facial expressions. In: Proceedings of the SICE-ICASE International Joint Conference; 2006. pp. 5423-5428

[7] Lee D, Lee T, So B, Choi M, Shin E, Yang K, et al. Development of an android for emotional expression and human interaction. In: Proceedings of the 17th World Congress the International Federation of Automatic Control; 2008. pp. 4336-4337

[8] Mazzei D, Lazzeri N, Hanson D, de Rossi D. HEFES an hybrid engine for facial expressions synthesis to control human-like androids and avatars. In: Proc. the 4th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics; 2012. pp. 95-200

[9] Tadesse Y, Priya S. Graphical facial expression analysis and design method an approach to determine humanoid skin deformation. Journal of Mechanisms and Robotics. 2012;**4**(2):021010

[10] Ahn H, Lee D, Choi D, Lee D, Hur M, Lee H, et al. Designing of android head system by applying facial muscle mechanism of humans. In: Proceedings of IEEE-RAS International Conference on Humanoid Robots; 2012. pp. 799-804

[11] Loza D, Marcos S, Zalama E, Garcia-Bermejo JG, Gonzalez JL. Application of the FACS in the design and construction of a mechatronic head with realistic appearance. Journal of Physicla Agents. 2013;7(1):31-38

[12] Ishi C, Liu C, Ishiguro H, Hagita N. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012); 2012. pp. 2377-2382

[13] Ishi CT, Liu C, Ishiguro H, Hagita N. Head motion during dialogue speech and nod timing control in humanoid robots. In: Proc. of 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010); 2010. pp. 293-300

[14] Liu C, Ishi C, Ishiguro H, Hagita N. Generation of nodding, head tilting and gazing for human-robot speech interaction. International Journal of Humanoid Robotics (IJHR). 2013;**10**(1):1-19

[15] Sakai K, Minato T, Ishi CT, Ishiguro H. Novel speech motion

generation by modelling dynamics of human speech production. Frontiers in Robotics and AI. 2017;**4**(49):14

[16] Ishi C, Hatano H, Ishiguro H. Audiovisual analysis of relations between laughter types and laughter motions. In: Proc. of the 8th International Conference on Speech Prosody (Speech Prosody 2016); 2016. pp. 806-810

[17] Ishi C, Minato T, Ishiguro H. Motion analysis in vocalized surprise expressions. In: Proc. Interspeech 2017; 2017. pp. 874-878

[18] Ishi CT, Minato T, Ishiguro H. Motion analysis in vocalized surprise expressions and motion generation in android robots. IEEE Robotics and Automation Letters. 2017;**2**(3):1748-1754

[19] Ishi CT, Minato T, Ishiguro H. Analysis and generation of laughter motions, and evaluation in an android robot. APSIPA Transactions on Signal and Information Processing. 2019;**8**(e6):1-10

[20] Busso C, Deng Z, Yildirim S, Bulut M, Min Lee C, Kazemzadeh A, et al. Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th International Conference on Multimodal Interfaces; 2004. pp. 205-211

[21] Alonso-Mart F, Malfaz M, Sequeira J, Gorostiza JF, Salichs MA. A multimodal emotion detection system during human-robot interaction. Sensors. 2013;**13**(11):15549-15581

[22] Uz B, Gudukbay U, Ozguc B. Realistic speech animation of synthetic faces. In: Proceedings of the Computer Animation; 1998. pp. 111-118

[23] Adams A, Mahmoud M, Baltrusaitis T, Robinso P. Decoupling facial expressions and head motions in complex emotions. In: Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction; 2015. pp. 274-280

[24] Massaro DW, Egan PB. Perceiving affect from the voice and the face. Psychonomic Bulletin & Review. 1996;**3**(2):215-221

[25] Cave C, Guaitella I, Bertrand R, Santi S, Harlay F, Espesser R. About the relationship between eyebrow movements and *F0* variations. In: Proceedings of the 4th International Conference on Spoken Language Processing; 1996. pp. 2175-2179

[26] Ekman P, Friesen WV. Head and body cues in the judgment of emotion: A reformulation. Perceptual and Motor Skills. 1967;**24**(3):711-724

[27] The ILHAIRE project (Incorporating laughter into human avatar interactions: Research and experiments). Available from: http://www.ilhaire.eu/

[28] Niewiadomski R, Pelachaud C. Towards multimodal expression of laughter. In: IVA; 2012. pp. 231-244

[29] Niewiadomski R, Hofmann J, Urbain J, Platt T, Wagner J, Piot B, et al. Laugh-aware virtual agent and its impact on user amusement. In: AAMAS; 2013. pp. 619-626

[30] Ding Y, Prepin K, Huang J, Pelachaud C, Artieres T. Laughter animation synthesis. In: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems; 2014. pp. 773-780

[31] Niewiadomski R, Mancini M, Ding Y, Pelachaud C, Volpe G. Rhythmic body movements of laughter. In: Proc. of 16th International Conference on Multimodal Interaction; 2014. pp. 299-306

[32] Yehia HC, Kuratate T, Vatikiotis-Bateson E. Linking facial animation, head motion and speech acoustics. Journal of Phonetics. 2002;**30**:555-568

[33] Ekman P, Davidson RJ, Friesen WV. The Duchenne smile: Emotional expression and brain physiology II. Journal of Personality and Social Psychology. 1990;**58**(2):342-353