Motion Analysis in Vocalized Surprise Expressions and Motion Generation in Android Robots

Carlos T. Ishi, Takashi Minato, Hiroshi Ishiguro

***Abstract*—Surprise expressions often occur in dialogue interactions, and they are often accompanied by verbal interjectional utterances. We are dealing with the challenge of generating natural human-like motions during speech in android robots that have a highly human-like appearance. In this study, we focus on the analysis and motion generation of vocalized surprise expression. We first analyze facial, head and body motions during vocalized surprise appearing in human-human dialogue interactions. Analysis results indicate differences in the motion types for different types of surprise expression as well as different degrees of surprise expression. Consequently, we propose motion-generation methods based on the analysis results and evaluate the different modalities (eyebrows/eyelids, head and body torso) and different motion control levels for the proposed method. This work is carried out through subjective experiments. Evaluation results indicate the importance of each modality in the perception of surprise degree, naturalness, and the spontaneous vs. intentional expression of surprise.**

***Index Terms*—Android robots, Emotion, Facial expressions, Motion generation, Surprise expression.**

I. Introduction

ANDROID robots have a highly human-like appearance, which gives them the ability to achieve natural communication with humans through several types of non-verbal information, such as facial expressions and gestures. Among the many studies related to facial expression in robots, most of them are related to symbolic (static) expression of the six traditional emotions (happy, sad, anger, disgust, fear and surprise) [1-9]. However, in real interactions, humans convey several types of emotions and attitudes by making subtle changes in facial expression.

Furthermore, when expressing an emotion, humans not only use facial expressions but also synchronize several other modalities, such as head and body movements as well as vocalic expressions. However, there has been very little research on emotion-expression methods incorporating a theory of the relationships among different modalities. Since androids have highly human-like appearance, this deficiency can cause a strongly negative impression (the “uncanny valley”) when an unnatural facial expression or motion is produced. From this viewpoint, it is important to clarify methodologies to generate motions that look natural.

For the expression of emotions, it is important to synchronize a variety of modalities, including facial movements, speech, and head/body movements, in order to clearly convey an emotion. For example, in the emotion-recognition field, it has been reported that the use of both audio and visual modalities provides higher recognition rates than using a single modality [10], [11]. Also, results of CG animation experiments clarified that using a combination of face and head modalities, in addition to the speech modality, improves the expression of an emotion, in comparison to using only the face modality [12].

Other studies investigated the synchronization of speech and facial expression. It has been reported that when there is mismatch between the emotions conveyed by the voice and by facial expressions, the emotion perceived from the facial expression is altered [13]. It has also been reported that when voice and facial expressions are presented, if the emotion expression of one of the modalities is ambiguous, the judgement of the perceived emotion is strongly influenced by the other modality [14]. For the facial parts, it has been reported that a systematic link exists between rapid upward-downward eyebrow movements and the voice’s fundamental frequency [15]. To achieve natural motion generation, the movements of the facial parts should also be synchronized with the changes in speech features. From this perspective, head and body movements should also be synchronized with speech. However, no previous research has tackled the challenge of developing suitable multimodal expression control in android robots.

Various studies have proposed control methods and systems for generating several types of facial expressions in android robots [3–9]. These are mostly based on FACS (Facial Action Coding System [16]) and methods for positioning and controlling the actuators to reproduce human-like facial expressions or, beyond this, modeling skin deformation based on mechanical deformation models. However, in all of these works, there has been no evaluation of the synchronization of

Manuscript received: Feb., 15th, 2017; Revised Apr., 25th, 2017; Accepted May, 2017. This paper was recommended for publication by Editor Dongheui Lee upon evaluation of the Associate Editor and Reviewers’ comments. This work was supported by JST, ERATO, Grant Number JPMJER1401.

C. T. Ishi is with the ATR Hiroshi Ishiguro Labs., Kyoto, 619-0288 Japan (phone: +81-774-95-2457; fax: +81-774-95-1408; e-mail: carlos@atr.jp).

T. Minato is with the ATR Hiroshi Ishiguro Labs., Kyoto, 619-0288 Japan (e-mail: minato@atr.jp).

H. Ishiguro is with the ATR Hiroshi Ishiguro Labs., Kyoto, 619-0288 Japan (e-mail: ishiguro@sys.es.osaka-u.ac.jp).

Digital Object Identifier (DOI) : see top of this page.

speech and facial expression and the face-body-head coordination. For expressing differences of nuance in emotion, it is important to evaluate the effects of multimodal expression rather than merely evaluating symbolic facial expressions.

From a multimodal perspective on motion control synchronized with voice, several methods for automatically generating lip and head motions from the speech signal of a tele-operator have been proposed [17–20]. Motion generation synchronized with laughter has also been recently proposed [21]. In the present study, we focus on motion analysis and generation during vocalized surprise expressions, which commonly occur in daily conversational interactions. Surprise expressions are not only simply related to emotional reactions but also used for expressing an attitude, such as showing interest in the dialogue partner. Such expressions have important social functions in human-human communication, so it is also important to clarify which types of modalities are closely related to such social meanings. Furthermore, surprise expressions are usually shorter in duration than other emotion expressions like happiness, sadness, anger and fear, and thus it is important to investigate the timing control between voice and movements of facial parts, head and body.

In the present study, we first analyzed face, head and body motions in vocalized surprise expressions appearing in natural human-human dialogue interactions. The dynamic properties of a motion in synchrony with speech (i.e., when a motion starts and ends relative to the vocalized surprise expression) were also investigated. Then, based on the analysis results, we proposed a method for motion generation in an android robot and investigated the effects of each modality through subjective experiments.

II. Motion analysis during vocalized surprise

A. Analysis data

We first conducted audio-visual analysis on surprise utterances appearing in human-human dialogue interactions.

For analysis, we use the multimodal conversational speech database recorded at ATR/IRC Labs. The database contains face-to-face dialogue interactions between several pairs of Japanese speakers, including audio, video and (head) motion capture data for each of the dialogue partners. Each dialogue has about 10–15 minutes of free conversations. The database contains segmentation and text transcriptions as well as dialogue act labels, including surprise labels for interjectional utterances.

We searched for all utterances containing either a surprise label or an exclamation mark (!) in the text transcription, which may indicate surprise expressions. In all, 636 utterances were extracted from the data of 28 adult speakers. Regarding the linguistic contents, interjections “e” were the most predominant (40%), followed by the interjections “a” (20%) and “he” (11%).

B. Annotation data: surprise degree and type

The perceived degree of surprise was first annotated for each of the surprise speech segments by listening only to the audio signals (i.e., based only on speech information). In order to account for contextual information, subjects were allowed to listen to audio from both dialogue partners, including 5 seconds before and 5 seconds after the surprise utterance.

A question arose about the criteria used to evaluate surprise degree. Consequently, we asked subjects to annotate the perceived degree of surprise expression regardless of whether the surprise was emotionally/spontaneously produced or socially/intentionally produced. Additionally, we asked subjects to annotate labels for intentionally produced surprise reactions and for quoted surprise expressions.

- e: emotional surprise (spontaneously produced reaction)
- s: social surprise expression (intentionally produced in order to smooth dialogue interaction; also includes acted-out surprise)
- q: quoted surprise expression (speaker expresses a past surprise utterance within the current dialogue)

Four native speakers of Japanese (research assistants) annotated the perceptual degree of surprise expression, and the above labels for all surprise utterances. Complete agreement among three or more annotators was found in 47% of the utterances, while agreement among two or more annotators was found in 97% of the utterances. Most of the disagreements among raters were found to have a difference of 1 point. For the surprise expression types, agreement among three or more annotators was reached in 82% of the utterances.

The perceptual scores were averaged across the annotators and normalized to a scale of 0 to 3 for the analysis of this study, resulting in 63 utterances for surprise degree “0,” 361 for “1,” 187 for “2,” and 25 for “3.” The majority of surprise expressions in daily conversations have a low degree of expressivity (“1”).

The 63 utterances for surprise degree “0” were excluded from subsequent analysis, resulting in a total of 573 surprise utterances. From those, 365 (64%) were classified as emotional/spontaneous, 166 (29%) were classified as social/intentional, and 42 (7%) were classified as quoted.

C. Annotation data: motion type

The following label sets were used to annotate the visual features related to motions and facial expressions during surprise utterances.

- eyelids: {normally opened, slightly widened, widened}
- eyebrows: {neutral, slightly raised, clearly raised}
- head: {no motion, up, down, left or right, up-down, tilted, nod, others (including motions synchronized with other motions like body)}
- upper body: {no motion, front, back, up, down, left or right, tilted, turn, others (including motions synchronized with other motions like head and arms)}

For each surprise utterance, the labels related to motion and facial expressions were annotated by one research assistant, who monitored the video and the motion data displays. In cases where multiple items were perceived, multiple label selection was permitted. The annotations were later checked and refined by another research assistant.

The most predominant motion type was eyebrow raise (usually accompanied by eyelid widening), found in 20% of the utterances, followed by an upward or up-down head motion (15%), body backward or upward motion (10%), and head nodding (5%). No motion was observed in about 30% of the utterances. This means that in daily interactions, speakers do not change the facial expression or make gestures each time a surprise is expressed. Moreover, the appearance of a motion can depend on the degree of expressed surprise, as will be discussed in the next sub-section.

In order to analyze the dynamic features of a motion, the intervals where facial, head or body parts are moving to target positions, or moving back to their neutral positions, were also segmented. The segmentation was conducted by one research assistant and later checked and refined by another research assistant.

D. *Analysis results of motion during surprise events*

Fig. 1 shows the overall distributions of motion occurrence for different degrees of perceived surprise expression. (The motion types are not distinguished in these results.)

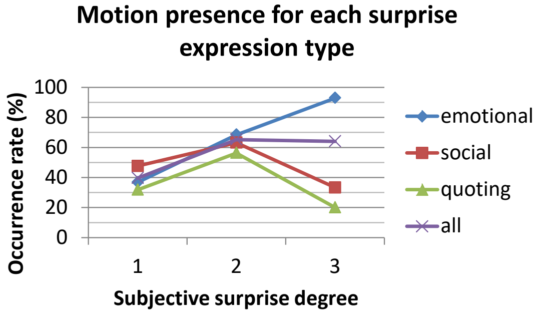

Fig. 1. Distributions of motion occurrence rates during surprise utterances, according to perceived surprise degree categories.

As an overall trend, these results show that the occurrence rate of a motion increases as the degree of surprise expression increases ("all"). Moreover, this trend becomes clearer for emotional/spontaneous surprise expression, where the occurrence rate of a motion approaches 100% for the highest degree of surprise expression. The results also show that social/intentional and quoted surprise utterances may or may not be accompanied by a motion, regardless of the degree of surprise expression.

Fig. 2 shows the distributions of motion occurrence rates for the most predominant motion types (eyebrow raise, upper body backward motion and head upward motion) for each surprise degree category. Here, the occurrence of eyebrow raise motion is higher for the middle and high degrees of surprise expression (levels 2 and 3), but the occurrence rate of body motions is much higher for the high degree of surprise expression (level 3).

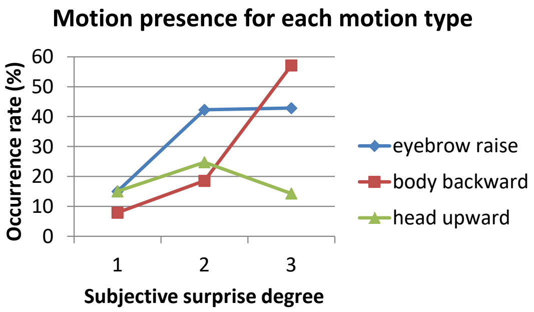

Fig. 2. Motion occurrence rate of motion types by each surprise degree category.

E. *Analysis of motion timing during surprise events*

For the generation of motions during emotion expressions, it is important to control the timings of onset and offset of different motions in synchrony with the speech utterances. In this section, we present analysis results on the timings of different modalities around the surprise utterance.

Onset and offset durations for eyebrow and body motion were measured for the interjections "e" and "a," which are the ones that appeared with the highest frequencies. Average and standard deviations were estimated for two motion levels (level 1 and level 2).

For eyebrow raise, the onset duration was faster than the offset duration for both levels, with averages around 200 to 300 ms for onset and 400 to 500 ms for offset. A slightly longer duration was found for level 2, since the amount of movement is bigger. The standard deviations for offset duration were much larger, since the eyebrows sometimes took 1 to 2 seconds to return to the neutral position.

For the upper body, onset and offset durations were both around 0.8 seconds for level 1 but around 1.2 seconds and 1.5 seconds for level 2.

Fig. 3 shows the distributions of the differences between motion and surprise utterances for each motion type. Here, "start" indicates the difference between the start time of a motion and the start time of the utterance, while "end" indicates the difference between the end time of a motion and the end time of the utterance.

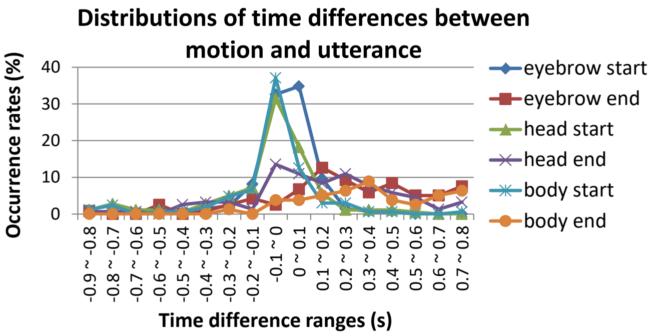

Fig. 3. Distributions of time differences between motion and utterances in surprise expression.

Results show that the start times of all eyebrow, head and body motions are mostly in the range of -0.1 to 0.1 seconds, which means that the motions are usually synchronized with the surprise utterances. The distributions of the end times are more spread but are concentrated on positive values for the time differences. This means that the motions go back to the neutral positions after the surprise utterance finishes.

III. Motion generation in android robots

A. *Proposed motion-generation method and adaptation to the android ERICA*

Based on the analysis results, we proposed a motion-generation method accounting for the following four factors: facial expression control (eyebrow raise accompanied by eyelid widening), head motion control (head pitch direction), and upper-body motion control (torso pitch direction). The block diagram in Fig. 4 illustrates how the proposed motion generation method works during surprise utterances.

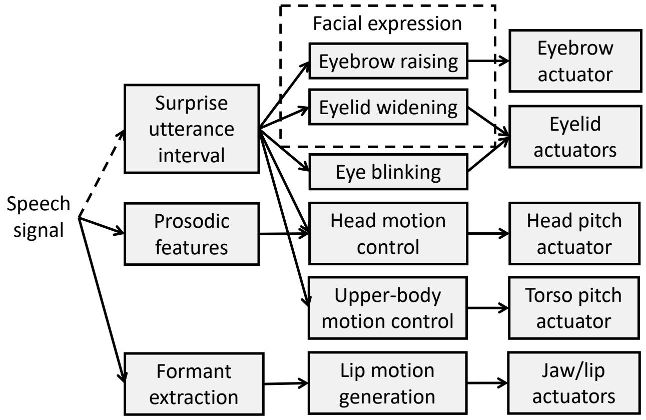

Fig. 4. Proposed method for motion generation in surprise utterances.

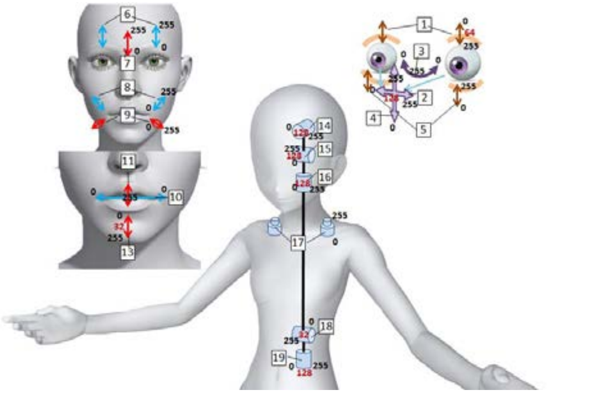

Fig. 5. Actuators of female-type android ERICA.

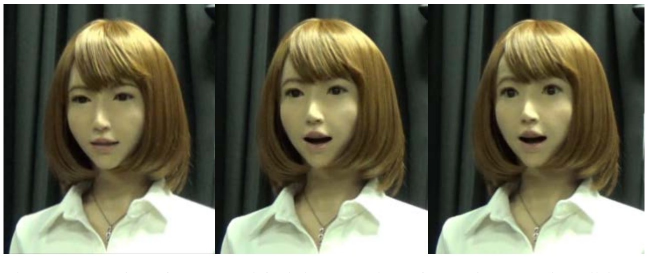

Fig. 6. Examples of generated facial expressions for eyebrow and eyelid control at level 0 (neutral idle face, left), level 1 (slight surprise face, middle) and level 2 (clear surprise face, right)

A female-type android robot, called ERICA, was used to evaluate the proposed motion-generation method. However, the method can be applied to any robot having similar degrees of freedom (DOFs). Fig. 5 shows the robot's actuators, and Fig. 6 shows the external appearance of the robot's face.

The current version of the android robot has 13 DOFs for the face, 3 DOFs for the head motion, and 2 DOFs for the upper-body motion. Among these, the following were controlled in the present work: 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 control (actuator 13), head pitch control (actuator 15), and upper-body pitch control (actuator 18). All actuator commands range from 0 to 255.

For the facial expression during surprise, the eyebrow raise and eyelid widening are coordinated and controlled at two levels of expression. The target actuator values were 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 our robot, the target eyebrow actuators were set to act[6] = 127 for level 1 and act[6] = 255 for level 2, and the upper and lower eyelid actuators were 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}. Fig. 6 shows an example 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 slight change in the facial expression.

Based on the analysis results, 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 seconds after the end of the utterance. A half cosine function is used to smoothly move the actuators back to the neutral position.

Also from the analysis results, an eye blinking usually occurs as the facial expression turns back to the neutral face. This was realized in our android by closing the eyes (act[1] = 255; act[5] = 255) during a period of 100 ms and then opening the eyes back to the neutral idle face (act[1] = 90; act[5] = 80).

For the upper-body motion control, we proposed a method for moving the upper body in the backward direction and then moving back to the neutral position based on the timing analysis results in the previous section. Two levels were controlled corresponding to about 2 degrees for level 1 and 4 degrees for level 2, which is the maximum angle achieved by our android. From 0.3 seconds after the end point of the surprise utterance interval, the upper body is moved back to the neutral position. Based on the analysis results, the onset duration to achieve the maximum angle is set to 0.8 seconds, while the offset duration to move back to the neutral idle position is set to 1.5 seconds. As in the eyebrow control, half cosine functions were used to smooth motion velocity changes in the current and target positions.

For head motion control, we used a method for controlling the head pitch (vertical movements) from the voice pitch (fundamental frequencies, F0). Although this control strategy is not exactly what humans do during speech, natural head motion was observed in a previous work on laughing-speech motion generation [21]. Furthermore, we can expect natural motion to also be generated during surprise utterances because humans tend to raise the head for high F0s, resulting in a face-up or up-down motion in surprise intervals. The following expression is used to convert F0 values to the head pitch actuator:

$$headpitch\_F0[\mathrm{t}] = (F0[\mathrm{t}] - center\_F0) * F0\_scale \qquad (1)$$

where center_F0 is the speaker's average F0 value (around 120 Hz for male and around 240 Hz for female speakers) converted to semitone units, F0[t] is the current F0 value (in semitones), and F0_scale is a scale factor for mapping the F0 (voice pitch) changes to head pitch movements. In the present experiment, F0_scale factor was set in such a way that a 1 semitone change in voice pitch corresponds to an approximately 1 degree change in head pitch rotation.

From the preliminary experiments, we felt it unnatural that the head was facing the upward direction while the body moved in the backward direction during a surprise expression. In fact, in our analysis, we observed that the speaker usually looks at the dialogue partner when expressing surprise. Thus, we provided an additional control in the head pitch actuator to move in the inverse direction to the body pitch movement:

$$headpitch[\mathrm{t}] = headpitch\_F0[\mathrm{t}] - torsopitch[\mathrm{t}] \qquad (2)$$

For non-vocalized surprise expressions, the jaw should be dropped. In vocalized surprise expressions, lip motion and thus the jaw motion should also be synchronized with the speech contents. In our method, lip motion is controlled based on a previously proposed formant-based lip motion control method [17]. The jaw actuator is controlled using the estimated lip heights. Appropriate lip shapes can be generated in vocalized surprise segments with different vowel qualities (such as in "eh!" and "ah!"), since the method is based on the vowel formants.

B. Evaluation of proposed motion-generation method

In the present work, we assume that the surprise utterance intervals are given in order to investigate the effects of controlling different modalities as a way to express different degrees of surprise.

We evaluated the proposed motion generation method with the interjectional utterances "e" and "a," which are the ones that most frequently occur in the database for expressing surprise. We extracted from our database 16 dialogue passages of about 10 seconds including interjectional utterances "e" or "a" expressing different degrees of surprise. Then we generated motion in our android, based on the speech signal and the surprise utterance interval information.

In order to evaluate the effects of different modalities and different degrees of motion control, motions were generated in the android according to the six conditions listed in Table I.

The motion types in Table I 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 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.

TABLE I
MODALITIES CONTROLLED FOR GENERATING SIX MOTION TYPES

Motion	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)

Video clips were recorded, for each motion type and each dialogue passage, to use in the subjective experiments. We conducted video-based evaluation instead of face-to-face evaluation since the participants do not interact with the robot.

For short interjectional utterances (around 200 ms), the range for body motion was small. Thus, 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 other 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)x3 + (1+4)x6) were used for evaluation.

In the experiment, participants were asked to watch all 63 videos and to grade each video with a perceptual subjective score. The order of the videos was randomized, and participants were allowed to play them at most two times each.

The perceptual subjective scores were graded according to the scales shown below. The numbers within parentheses were used to quantify the perceptual scores.

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 (human-like)? 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).

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

C. Evaluation results

In order to account for the effects of the voice modality, the utterances used in the experiment were separated into three

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2017.2700941, IEEE Robotics and Automation Letters

groups according to their perceptual degrees of surprise graded only from the voice (Section II.B). The resulting number of utterances were 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).

Fig. 7 shows the average subjective scores of the three factors (surprise expression degree, motion naturalness, and intentional-emotional impression) for each motion type, arranged by the voice groups (surprise expression degrees by voice only). It should be emphasized 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 were conducted to investigate the effects of presence/absence or degree of motion control, and statistical significance tests were conducted through t-tests. First, a comparison of the results for e1+h1+b0 and e2+h1+b0 motion types provided the effects of controlling the motion degrees of eyebrow and eyelids. As shown in the upper panel of Fig. 7, the average perceptual score for surprise expression degree increases 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 shows that a slight change in the eyebrow/eyelid control is effective for changing the perceived degree of surprise.

A comparison of the results for e2+h0+b0 and e2+h1+b0 motion types provided the effects of controlling the head motion modality. The difference 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., $p = 0.09$), a bit smaller than the effects of eyebrow/eyelid control.

The naturalness scores in the middle panel of Fig. 7 show slightly natural to natural scores in almost all motion types. A comparison between e2+h0+b0 and e2+h1+b0 indicates that head motion has important effects on the naturalness (human-like) perception when the body does not move (b0). On average, the naturalness scores are increased by about 0.5 points ($p < 0.01$). However, the naturalness scores approach 0 in motion types with fewer motion numbers (e1+h1+b0 and e2+h0+b0) in voice group 3 (high surprise expression degree by voice-only). This is thought to be due to the mismatch between surprise expressions by voice and motion modalities.

Comparisons of the results for e2+h0+b0 and e2+h0+b2 motion types or between e2+h1+b0 and e2+h1+b2 motion types provide the effects of controlling the body motion modality. Consequently, when head motion is not controlled (h0), the effects of controlling or not the body motion (b0 vs. b2) increases the surprise degrees by about 0.4 to 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; $p < 0.05$), since 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 e2+h1 condition).

Regarding the effects of the voice modality, results in the upper panel of Fig. 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). On the other hand, 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).

Regarding the subjective spontaneity degree, the bottom panel of Fig. 7 shows that the average scores in e1+h1+b0 and e2+h0+b0 motion types are negative in voice group 3, indicating that if the amount of motion decreases, the surprise utterances might be perceived as intentional rather than emotional ($p < 0.01$).

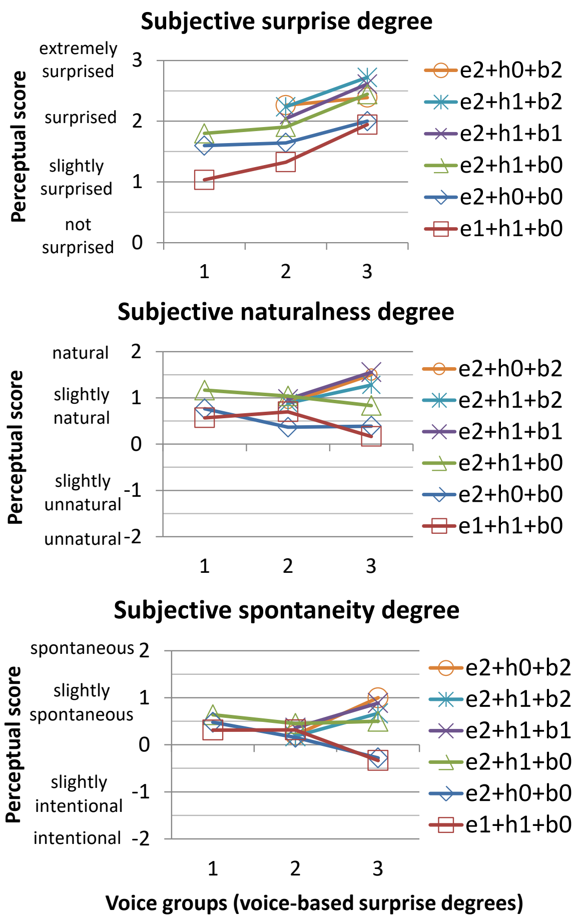

Fig. 7. Subjective perceptual scores of surprise expression degree (top), naturalness degree (mid), and emotional/intentional impression degree (bottom) for each motion type, arranged by voice groups (horizontal axis: voice-based surprise expression degrees).

IV. Discussion

In this study, we focused on analysis and evaluation of short and long interjectional utterances. Although the occurrence rates were smaller, short and long sentences were also found in

15% of the surprise utterances. Further evaluation of motion control in longer sentences is necessary for investigating the generalization of the results in the present study.

We conducted video-based evaluation instead of face-to-face evaluation, since the participants do not interact with the robot, and consequently this avoids the influence of other factors such as eye gazing during the experiments. We intend to conduct face-to-face evaluation after solving eye gazing control issues. Nonetheless, we can expect similar results to the video-based evaluation in the present study, considering that in a previous study it was shown that subjective experiment results did not change between video-based and face-to-face interaction during head motion control experiments [19].

V. Conclusion and final remarks

In the present study, we first analyzed facial, head and body motions during vocalized surprise expressions appearing in human-human dialogue interactions. The analysis results provided the following findings: 1) 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. 2) Different motion types have different occurrence rates according to the surprise expression degree. In particular, body backward motion appears at higher frequency when expressing high surprise degrees. 3) 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.

Consequently, we proposed motion-generation methods based on the analysis results and evaluated the effects of different modalities (eyebrows/eyelids, head and upper body) and different motion control levels. This work was done through subjective experiments using an android robot. Evaluation results indicate the following: 1) Eyebrow/eyelid motion control was most effective in changing the expression degrees of surprise. 2) Upper-body motion control was effective for increasing the expression degrees of surprise and naturalness. 3) Head motion was more effective for increasing perceptual naturalness. 4) The surprise expression degrees for different motion types are biased by the surprise degrees expressed by the voice-only modality. 5) Utterances with high surprise degrees may be interpreted as “intentional” surprise expressions if they are not accompanied by upper-body motion. Although the proposed method was evaluated in the android ERICA, it can be adapted to any robot having similar degrees of freedom (DOFs).

Future works include prediction of surprise expression degrees from acoustic features, with the aim of automating the generation of surprise motion during teleoperation.

Acknowledgment

We would like to thank Mika Morita, Megumi Taniguchi, Kyoko Nakanishi, and Tomo Funayama for their contributions to the experiment setup and data analysis.

References

[1] C. Breazeal, Emotion and sociable humanoid robots, *International Journal of Human-Computer Studies*, 59, pp. 119-155, 2003.

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

[3] Y. Wu, N. M. Thalmann, D. Thalmann, A Dynamic Wrinkle Model in Facial Animation and Skin Aging, *Journal of Visualization and Computer Animation*, Vol.6, No.4, pp.195-205, 1995.

[4] T. Hashimoto, S. Hiramatsu, T. Tsuji, H. Kobayashi, Development of the Face Robot {SAYA} for Rich Facial Expressions, Proc. of the *SICE-ICASE International Joint Conference*, pp.5423-5428, 2006.

[5] D. Lee, T. Lee, B. So, M. Choi, E. Shin, K. Yang, M. Baek, H. Kim, H. Lee, Development of an Android for Emotional Expression and Human Interaction, Proc. of *the 17th World Congress The International Federation of Automatic Control*, pp.4336-4337, 2008.

[6] D. Mazzei, N. Lazzeri, D. Hanson, D. de Rossi, HEFES an Hybrid Engine for Facial Expressions Synthesis to control human-like androids and avatars, Proc. the *4th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*, pp.95-200, 2012.

[7] Y. Tadesse, S. Priya, Graphical Facial Expression Analysis and Design Method An Approach to Determine Humanoid Skin Deformation, *Journal of Mechanisms and Robotics*, Vol.4, No.2, pp.021010, 2012.

[8] H.Ahn, D. Lee, D. Choi, D. Lee, M. Hur, H. Lee, T. Kanda, Designing of Android Head System by Applying Facial Muscle Mechanism of Humans Proc. of *IEEE-RAS International Conference on Humanoid Robots*, pp.799-804, 2012.

[9] D. Loza, S. Marcos, E. Zalama, J. G. Garcia-Bermejo, J. L. Gonzalez, Application of the FACS in the Design and Construction of a Mechatronic Head with Realistic Appearance, *Journal of Physicla Agents*, Vol.7, No.1, pp.31-38, 2013.

[10] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Min Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, Analysis of emotion recognition using facial expressions, speech and multimodal information, Proc. of *the 6th international conference on Multimodal interfaces*, pp.205-211, 2004.

[11] F. Alonso-Mart, M. Malfaz, J. Sequeira, J. F. Gorostiza, M. A. Salichs, A Multimodal Emotion Detection System during Human-Robot Interaction, *Sensors*, Vol.13, No.11, pp.15549-15581, 2013.

[12] B. Uz, U. Gudukbay, B. Ozguc, Realistic Speech Animation of Synthetic Faces, Proc. of *the Computer Animation*, pp.111-118, 1998.

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

[14] D. W. Massaro, P. B. Egan, Perceiving affect from the voice and the face, *Psychonomic Bulletin & Review*, Vol.3, No.2, pp.215-221, 1996.

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

[16] P. Ekman and W. V. Friesen, Head and body cues in the judgment of emotion: A reformulation, *Perceptual and motor skills*, vol.24, no.3, pp.711-724, 1967.

[17] C. Ishi, C. Liu, H. Ishiguro, N. Hagita. “Evaluation of formant-based lip motion generation in tele-operated humanoid robots,” Proc. *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS 2012*), Vilamoura, Portugal, pp. 2377-2382, October, 2012.

[18] C.T. Ishi, C. Liu, H. Ishiguro, and N. Hagita. “Head motion during dialogue speech and nod timing con-trol in humanoid robots,” Proc. of *5th ACM/IEEE International Conference on Human-Robot Interaction* (*HRI 2010*), pp. 293-300, 2010.

[19] C. Liu, C. Ishi, H. Ishiguro, and N. Hagita. Generation of nodding, head tilting and gazing for human-robot speech interaction. *International Journal of Humanoid Robotics* (*IJHR*), vol. 10, no. 1, January 2013.

[20] S. Kurima, C. Ishi, T. Minato, and H. Ishiguro. Online Speech-Driven Head Motion Generating System and Evaluation on a Tele-Operated Robot, *IEEE International Symposium on Robot and Human Interactive Communication* (*ROMAN 2015*), pp. 529-534, 2015.

[21] C. Ishi, T. Funayama, T. Minato, and H. Ishiguro (2016). “Motion generation in android robots during laughing speech,” *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS 2016*), pp. 3327-3332, Oct., 2016.