INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Motion analysis in vocalized surprise expressions

Carlos Ishi[1], *Takashi Minato*[1], *Hiroshi Ishiguro*[1,2]

[1]ATR/HIL, Japan
[2]Osaka University, Japan

carlos@atr.jp, minato@atr.jp, ishiguro@irl.sys.es.osaka-u.ac.jp

Abstract

The background of our research is the generation of natural human-like motions during speech in android robots that have a highly human-like appearance. Mismatches in speech and motion are sources of unnaturalness, especially when emotion expressions are involved. Surprise expressions often occur in dialogue interactions, and they are often accompanied by verbal interjectional utterances. In this study, we analyze facial, head and body motions during several types of vocalized surprise expressions appearing in human-human dialogue interactions. The analysis results indicate an inter-dependence between motion types and different types of surprise expression (such as emotional, social or quoted) as well as different degrees of surprise expression. The synchronization between motion and surprise utterances is also analyzed.

Index Terms: surprise expression, facial expressions, motion analysis, paralinguistic information, multimodal analysis

1. Introduction

The background of our research is the development of android robots that can interact and behave as humans [1]. 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 based on FACS (Facial Action Coding System [2]) and intend to reproduce symbolic (static) facial expressions of the six traditional emotions (happy, sad, anger, disgust, fear and surprise) [3–9]. 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, along with vocalic expressions. However, there has been very little research on emotion-expression methods incorporating 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 perspective, it is important to clarify the coordination among different modalities to generate motions that look natural and clearly convey an emotion.

There are several studies regarding the multimodal analysis during emotion expression. 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 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 [16–20]. Relationships between laughter types and laughter motions were analyzed [21], and motion generation synchronized with laughter speech has also been proposed [22]. However, motion generation methods during other emotional speech events including surprise have not been proposed so far.

In the present study, we focus on motion analysis 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's talk. 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.

We analyzed facial, head and body motions during 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.

2. Motion analysis during vocalized surprise

The multimodal data used for motion analysis during vocalized surprise expressions are described in Section 2.1.

http://dx.doi.org/10.21437/Interspeech.2017-631

Annotation procedures are explained in Section 2.2 for surprise degrees and types, and in Section 2.3 for motion types.

2.1. 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 [17], [21]. 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%), as shown in Fig. 1. Along with surprise, "e" is often used for expressing unexpectedness, "a" for noticing, and "he" for admiration or sympathy.

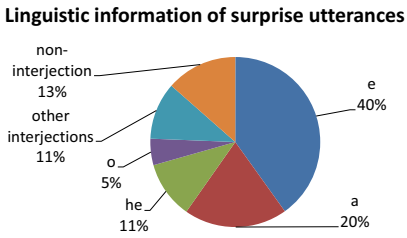

Figure 1: *Overall distributions of the morpheme types in surprise utterances.*

2.2. Annotation data: surprise degree and type

The perceived degree of surprise was first annotated in a 4-point scale (from 0 for not surprised to 3 for very surprised) 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 10% for surprise degree "0," 57% for "1," 29% for "2," and 4% for "3." The 63 utterances for surprise degree "0" were excluded from subsequent analysis, resulting in a total of 573 surprise utterances. From those, 60% were classified as emotional/spontaneous, 33% as social/intentional, and 7% as quoted.

2.3. 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.

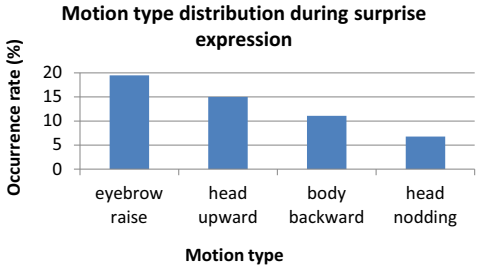

Figure 3: *Overall distributions of the motion types in surprise utterances.*

Fig. 3 shows the distributions for the predominant motion types. 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 Section 3.1.

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

3. Data analysis

Section 3.1 presents analysis results of the relationship between motion occurrences and different degrees and types of surprise utterances. Analysis of the dynamic features of motions and synchronization with surprise utterances is presented in Section 3.2. In Section 3.3 presents preliminary analysis on the differences between surprise types.

3.1. Analysis results of motion during surprise events

Fig. 4 shows the overall distributions of motion occurrence for different degrees of perceived surprise expression. (The motion types are not distinguished in these results.)

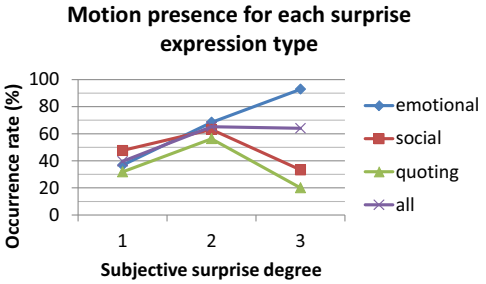

Figure 4: *Distributions of motion occurrence rates during surprise utterances, according to perceived surprise degree categories.*

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 high degree of surprise expression (level 3) and is significantly higher than social/intentional and quoted surprise utterances ($p < 0.05$ by qui-square test). The results indicate that social and quoted surprise utterances may or may not be accompanied by a motion, regardless of the degree of surprise expression.

Fig. 5 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. The results in Fig. 5 show that 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) ($p < 0.05$ by chi-square tests).

Fig. 6 shows the distributions of motion types for the interjections "e", "a" and "he", which are the most predominant morphemes, appearing for verbally expressing surprise. The results in Fig. 6 show that the different morphemes have different distributions of motion types. It can be observed that the predominant motion is eyebrow raise in the interjection "e" ($p < 0.05$), upward head motion in interjection "a" ($p < 0.05$), and nodding in "he" ($p < 0.01$ by chi-square tests). These differences are thought to be because these interjections also convey other paralinguistic information, along with the surprise expression (unexpectedness in "e", noticing in "a", and admiration in "he").

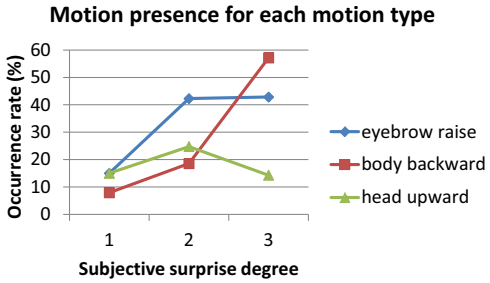

Figure 5: *Motion occurrence rate of motion types by each surprise degree category.*

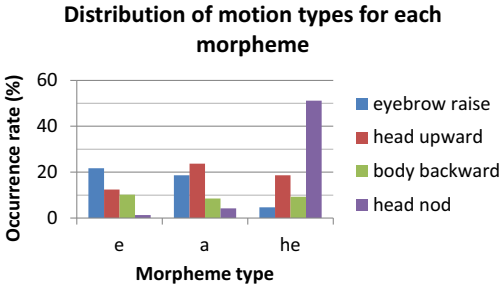

Figure 6: *Distributions of motion types for each morpheme type.*

3.2. 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 for small and level 2 for large movements). Figs. 7 and 8 show averages and standard deviations of the onset and offset durations for eyebrow raise and upper body backward motions, respectively.

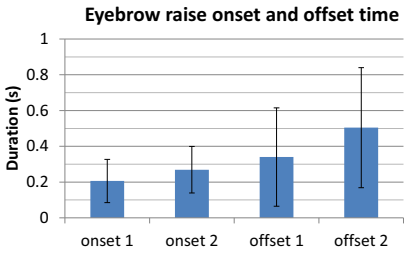

Figure 7: *Distributions of the onset and offset times for eyebrow raise motion, for levels 1 and 2.*

For eyebrow raise, the onset duration was faster than the offset duration for both levels ($p < 0.01$ by t-tests), 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 ($p < 0.01$ for onset and $p < 0.05$ for offset, by t-tests), 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.

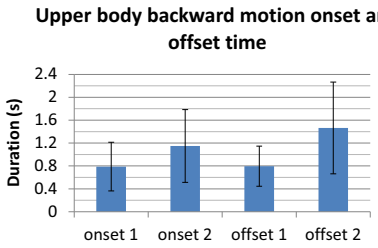

Figure 8: *Distributions of the onset and offset times for upper body backward motion, for levels 1 and 2.*

For the upper body, onset and offset durations were both longer for level 2 ($p < 0.01$ by t-tests). The differences between mean onset and offset durations were not significant, being around 0.8 seconds for level 1, but around 1.2 seconds and 1.5 seconds for level 2.

Fig. 9 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.

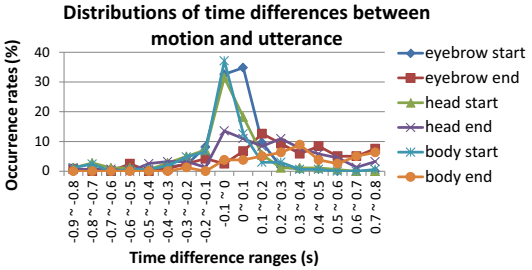

Figure 9: *Distributions of time differences between motion and utterances in surprise expression.*

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. These trends are similar with motion synchronization during laughter speech found in a previous study [21].

3.3. Analysis of differences between surprise types

It was shown in Section 3.1 that the occurrence rates of a motion depend on the surprise type. In this section we present some analysis results on differences between surprise types.

Among several prosodic features, we have observed a trend of social/intentional surprise being usually longer in duration compared to emotional/spontaneous surprise utterances. Fig. 10 shows the distributions of three duration categories {D1: shorter than 400ms; D2: between 400 ms and 800 ms; D3: longer than 800ms} and degrees of emotional to social surprise expression (the levels 0 to 3 correspond to the number of annotators that judged the surprise utterances as "social").

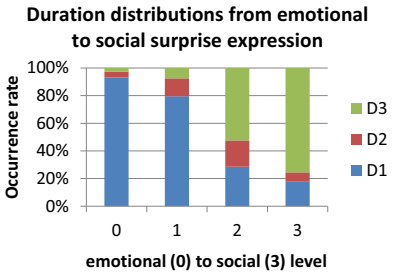

Figure 10: *Distributions of duration categories of monosyllabic interjections, from emotional to social surprise expression.*

It is clear that in emotional/spontaneous surprise expression (level 0 in Fig. 7) short interjections are predominant (category D1), while in social/intentional surprise expression (level 3 in Fig. 7) very long interjections are predominant (category D3) ($p < 0.01$ by chi-square tests). It can also be observed that the levels 1 and 2 show intermediate distributions for the duration categories. These results indicate that the duration of the interjectional utterances may be an important cue for expressing either an emotional or a social surprise, besides motion control. The effects of surprise expression duration and control of different motion modalities are subjects for future investigation.

4. Conclusions

In the present study, we 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 high surprise degrees are expressed. 3) Different interjection types have different distributions for motion types. 4) 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.

Future works include automatic surprise interval detection and application of the analysis results in the present study to motion generation of android robots. Finally, it would be interesting to verify if the findings of the present study is invariant across different languages and cultures.

5. Acknowledgements

This work was supported by JST/ERATO (Grant Number JPMJER1401). We would like to thank Mika Morita, Megumi Taniguchi, Kyoko Nakanishi, and Tomo Funayama for their contributions to the data annotation and data analysis.

6. References

[1] D.F. Glas, T. Minato, C.T. Ishi, T. Kawahara, and H. Ishiguro, "ERICA: The ERATO Intelligent Conversational Android," Proc. of 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2016), pp. 22-29, Aug. 2016.

[2] 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.

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

[4] 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.

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

[6] 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, Proceedings of the 17th World Congress The International Federation of Automatic Control, pp.4336-4337, 2008.

[7] 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.

[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 Proceedings 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, Proceedings 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, Proceedings 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, Proceedings 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, Proceedings of the 4th International Conference on Spoken Language Processing, pp.2175-2179, 1996.

[16] C. Ishi, C. Liu, H. Ishiguro, N. Hagita. (2012). "Evaluation of a formant-based speech-driven lip motion generation," In 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), Portland, Oregon, pp. P1a.04, September, 2012.

[17] C. Ishi, H. Ishiguro, N. Hagita (2013). Analysis of relationship between head motion events and speech in dialogue conversations. Speech Communication 57 (2014) 233–243, June 2013.

[18] C.T. Ishi, C. Liu, H. Ishiguro, and N. Hagita. "Head motion during dialogue speech and nod timing control 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, H. Hatano, H. Ishiguro (2016). "Audiovisual analysis of relations between laughter types and laughter motions," Proc. of the 8th international conference on Speech Prosody (Speech Prosody 2016), pp. 806-810, May, 2016.

[22] 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.