MULTIMODAL EXPRESSION FOR HUMANOID ROBOTS BY INTEGRATION OF HUMAN SPEECH MIMICKING AND FACIAL COLOR

Tokitomo Ariyoshi,†, Kazuhiro Nakadai†, and Hiroshi Tsujino†

†Honda Research Institute Japan, Co., Ltd.
8-1 Honcho, Wako-shi, Saitama, 351-0114, Japan
{toki.ariyoshi, nakadai, tsujino}@jp.honda-ri.com

ABSTRACT
Multimodal expression is essential for humanoid robots to communicate with people naturally and intelligibly. This paper describes multimodal expression for humanoid robots by mimicking human speech with the ability of expression through “facial colors”. Currently the robot is able to express joy (by turning yellow in the face), anger (red), sadness (blue), and relaxation (green). These colors have been selected according to color psychology. The human speech mimicking is based on prosody extraction of pitch, loudness and temporal information with speech synthesis based on the extracted prosody. The multimodal expression system implemented on Honda ASIMO shows that facial colors improve affective speech recognition by over 15%. In addition, qualitative observations that use speech and facial color with conflicting affective meanings producing complex affection have been reported.

1. INTRODUCTION
Recently, automatic speech recognition (ASR) and speech synthesis have been studied extensively and applied to a wide variety of systems such as telephone service systems, ticket machines, car navigation systems, and so on. An application of these technologies to a humanoid is also one of the most challenging and interesting topics from both an engineering and a scientific viewpoint. In this case, robustness in real-world, real-time processing as well as human-like performance is necessary to communicate with people. To achieve this, we need to focus not only on a single speech technology but also on various technologies such as multimodal recognition and expression. The total performance of the system depends on the integration of these technologies. We have previously reported robot audition such as sound source localization, sound source separation and speech recognition in the real world [1]. To enhance human-humanoid communication in the real world, we have in this paper focused on the humanoid’s expression. To form an expression, not only verbal information (speech) but non-verbal information such as prosody and paralinguistic information are essential. We have studied considerable work in the field of speech synthesis [2]. These researchers have reported that pitch, loudness, and temporal information are essential to produce affective and natural speech[2, 3]. However, the quality of the synthesized speech is not very good, that is, the recognition rate of affective meaning in the synthesized speech is not very high. To improve the recognition performance, work has been extended to research on multimodal expression by using a “talking head” that has a human-like face, with motion of the lips and tongue [4]. The talking head acts as a software agent, and is effective in exploring the mechanisms of human perception and expression. However, in the case of robots, human-like complex facial expression is difficult. In robotics, there are a few papers on multimodal expression[5, 6]. Most of them assume human-like facial expression and gestures for multimodal expression. However, such humanoids are not reasonable from the point of view of cost. In addition, it is difficult to achieve natural expression because of the complexity of the problem, such as controlling a lot of action units in Ekman’s FACS[7]. This problem is described as “uncanny valley” [8]. We have introduced color psychology [9] to enhance multimodal expression. In color psychology, it is known that each color has affective meaning. We achieve facial expression by changing colors. For speech synthesis, we use speech mimicking based on pitch, loudness and temporal information because this prosody information is known to be the essential cues for making affective speech. We developed a multimodal expression system for a humanoid by integration of human speech mimicking and facial color expression. In this paper, we report the effect of facial colors on the multimodal expression system. We also report how facial colors can be used effectively and synergistically to effect multimodal expression such as the production of more complex affection by using speech and facial color with conflicting affective meanings.
2. MULTIMODAL EXPRESSION SYSTEM

We developed the multimodal expression system shown in Fig. 1. The system consists of two sub systems - a human speech mimicking subsystem and a facial color subsystem.

2.1. Human Speech Mimicking Subsystem

This subsystem is implemented as a multi-agent system. There are four agents — “Speech Recognizer”, “Pitch Estimator”, “Power Estimator”, and “Mimic Engine.” Speech signals captured by a microphone are sent to three agents – Speech Recognizer, Pitch Estimator and Power Estimator in parallel. These three agents extract name and duration of phoneme, power and pitch. The extracted feature is converted into a feature event including its extracted time, that is, all types of events are structured in the same fashion. This architecture gives the system flexibility in the sense that another agent is easily added to the system and an agent is easy to be replaced with an alternative agent. The feature events are sent to the Mimic Engine that generates a prosodic phoneme event by synchronization and association of feature events. The phoneme event is stored in a database, modified according to instructions given by the user in advance, and sent to the speech synthesizer. Thus, speech is mimicked by resynthesizing the prosodic phoneme events.

2.1.1. Speech Recognizer

Speech Recognizer which is based on the “Julius” developed at Kyoto University [10], extracts a phoneme that includes its name, and duration obtained from alignment of phonemes. A phoneme event en includes the start time, the name of the phoneme, and the duration.

2.1.2. Pitch Estimator

Pitch Estimator continuously estimates pitch of the input speech stream, by using harmonic relation. The input speech is digitized at a sampling rate of 16 kHz. The spectrum is calculated from the input speech by using FFT of 512 points with window length of 25 ms and window shift of 10 ms. From the spectrum, harmonics having the same fundamental frequency are clustered. A set of the clustered harmonics is assumed to be a single sound. When multiple sounds are extracted, a harmonics set that satisfies the following conditions is selected as the target speech:

1. fundamental frequency of the harmonics set is from 80 Hz to 300 Hz.
2. power summation of the harmonics is maximum.

The pitch is regarded as fundamental frequency of the harmonics set. Thus, the pitch event ef which includes pitch and its extracted time is extracted at the interval of 10 ms.

2.1.3. Power Estimator

Power Estimator extracts the power of the input speech stream as a summation of speech signal for every 10 ms. Power event ep includes power value and its time stamp.

2.1.4. Mimic Engine

Mimic Engine receives feature events. It generates a prosodic phoneme event by synchronization and association. It modifies the prosodic phoneme event, and resynthesizes speech according to the prosodic phoneme event. To integrate feature events, a phoneme event en is selected, and sequences of ev and ef within duration of en are selected. The prosodic phoneme event Ep, that includes en, and the sequences of ev and ef is generated. Every prosodic phoneme event is stored to a phoneme event database for future use such as correction of feature extraction errors and mimic morphing. Feature parameters of the phoneme event can be modified to Ep, in mimic modifier. This module is controlled manually so that users can specify various mimic ways, such as pitch-only and combination of multiple mimic ways. Then, Ep or Ep sends to speech synthesizer. The speech synthesizer is based on “FineVoice” developed by NTT-IT, which uses FinalFLUET speech synthesis engine based on pitch synchronous overlap-add (PSOLA)[11].

2.2. Facial Color Subsystem

The facial color subsystem changes the humanoid’s face color in synchronization with the human speech mimicking subsystem. The subsystem is also implemented as a multi-agent system. It consists of three agents – Facial Color Selector, Color Intensity Modulator, and Facial Color Engine.

2.2.1. Facial Color Selector

Facial Color Selector selects one of four colors such as yellow, red, blue, and green corresponding to joy, anger, sadness, and relaxation, respectively. This correspondence is
based on the affective meaning of colors in color psychology [9]. The input of Facial Color Selector is user’s instruction such as a type of affections or colors.

2.2.2. Color Intensity Modulator
Color Intensity Modulator sends “on” and “off” when the first and the last prosodic phoneme event are received. Future work will enable color intensity change according to information in prosodic phoneme events.

2.2.3. Facial Color Engine
Facial Color Engine controls facial color expression. We use Honda ASIMO for a test-bed, the color of a part of the humanoid’s head as shown in Fig. 1 is controlled. The colored face and the mimicked speech are presented at the same time by output from Color Intensity Modulator.

3. EVALUATION
We evaluated the multimodal expression system for a humanoid. We used the Honda ASIMO humanoid as the test-bed. The performance of the system was evaluated on the following two points:

Exp. 1: performance of the human speech mimicking system in retaining affective meaning, and

Exp. 2: effects of facial color in multimodal expression.

Five Japanese short sentences which are common in daily conversation were used – ‘Doumo Kon-nichiwa (Hello)’, ‘Hontouni Yokatta (It was really good.)’, ‘Nande Sounaruno (Why is it so?)’, ‘Kono Asobi Tanoshii (This play is pleasant.)’, and ‘Ha Souiukoto (Oh, it is such a kind of things.)’. The affective meaning of these sentences varies with change in intonation. A male speaks the sentences with four different intonations, joy, anger, sadness and relaxation. Thus, an original speech data set of 20 sentences is generated by collecting the sentences spoken by the male. A mimicked speech data set is generated by the human speech mimicking subsystem from the original data set.

In Exp. 1, eight subjects were used. They listened to the original data set and a mimicked speech, and filled out questionnaires to answer one of five choices that included the four affections and other. This way the recognition rates of affection included in affective speeches by subjects were measured.

For Exp. 2, an AV data set of 160 audio-visual examples was prepared from the original and mimicked data sets by combination with four facial colors, i.e., yellow, red, blue and green. We had six subjects in this experiment. They listened to the speeches and looked at the colored face at the same time, and answered the same questionnaires as Exp. 1.

The results are shown in Figs. 2 and 3. Figs. 2a) and 3a) show the results in the case of real human speeches, while Figs. 2b) and 3b) are the results in the case of mimicked speeches. Each figure shows qualitative results without statistical test.

3.1. Performance of Mimicking Human Speech
In the case of real speeches, each affective speech is recognized well. This proves the validity of the recorded data set. From Fig. 2b), recognition of mimicked speeches has deteriorated. The affective meanings of “Joy” and “Anger” are still retained. The result shows a tendency for “Sadness” and “Relaxation” to be recognized as “Anger” and “Joy”, respectively. This indicates that the human speech mimicking system has enough performance to keep relatively strong affections such as “Joy” and “Anger”. When we consider that “Joy” and “Relaxation” are positive affections and “Anger” and “Sadness” are negative ones, we can say that the system holds that positive and negative affections.

3.2. Improvement by Facial Colors
From Figs. 2a) and 3a), reinforce tendencies shown in Tabs. 1 and 2 is observed. Likewise, Tab. 2 is consistent with information in Figs. 2b) and 3b). In Tabs. 1 and 2, “++”, “−−” and “○” mean excitement, inhabitation and neutral effects with respect to affective meaning, respectively. “++” and “−−” mean greater influence than single “+” and “−”. The observations show that red and blue have affirmative meanings of anger and sadness respectively, which is consistent with our understanding of color psychology. In particular, because blue has a strong effect of sadness, The use of blue facial color is effective, and should be used carefully in human-robot interaction because of negative affective meaning.

It is accepted that yellow means joy in color psychology. This affective meaning of yellow holds in the case of real speech, while only red excites an affective meaning of joy in mimicked speech. We hypothesize that yellow has little influence on affective meaning. In mimicked speech, about 60% of mimicked joy speech is misrecognized as other affections, that is, mimicking itself does not convey enough information. Therefore, to emphasize affective meaning of joy, red with strong affective effect is necessary.

In the case of green, a competitive phenomenon is observed between real and mimicked speeches. This is caused by two different faces in green. It is generally accepted that green means relaxation in color psychology, while it has been reported that green implies weakness, fatigue and sadness. In mimicked speech, the former affective meaning affects human recognition. The latter affective meaning affects the real speech. It is for future work to analyze how to select one of the two different faces of green in human recognition.

3.3. Other Observations
Several subjects reported that they sometimes feel more complex affections such as pretense of joy to hide anger or sadness. This happens when speech and facial color with conflicting affective meanings are shown. For example, when joyful speech and red face are shown, the subjects feel that the robot pretends to be joyful while covering its real affection of anger.
Generally, the use of facial color helps human recognition of affective meaning expressed by the robot. In mimicked speech, improvement of recognition rate reaches 15% on average when red for joy and anger, blue for sadness and green for relaxation are used as facial color. In real speech, the recognition rate also improves 16% on average when red for anger, yellow for joy and blue for sadness are used, improvement of recognition rate reaches 15% on average.

4. FUTURE WORK

We have shown preliminary performance for multimodal expression. To achieve humanoids with natural communication functions, we have to do more work. More complex multi-modal expression such as temporary intensity change of facial color should be investigated. The generality of how a multimodal expression should be assessed including whether such a general way exists should be addressed. We should consider how to integrate multimodal expression with internal emotional states of the humanoid. As reported by the subjects, conflicting affective meaning between speech and facial color can produce more complex and new affections. To use multimodal expression practically, context sharing in conversation among people and a humanoid should be considered; otherwise the expression could be received as other unpleasant meanings. This means it is important to consider dialog and the relationship between dialog and expression.

5. CONCLUSION

To realize a humanoid system that communicates with people naturally, we proposed multimodal expression by integration of human speech mimicking and facial colors. The proposed multimodal expression shows that the effective use of facial color improves the performance in recognition of affection, and in addition produces complex affective meaning when affective meanings of speech and a facial color are in conflict.

6. REFERENCES