Social Effects on Vocal Rate with Echoic Mimicry Using Prosody-only Voice

Noriko Suzuki†‡, Kazuhiko Kakehi‡, Yugo Takeuchi†, Michio Okada†

† ATR Media Integration & Communications Research Labs., ‡ Nagoya University
{noriko, yugo, okada}@mic.atr.co.jp, kakehi@info.human.nagoya-u.ac.jp

Abstract
We have been studying some of the essential factors that constitute interpersonal relations between humans and computers by focusing on social bonding in proto-communications (the interaction between adults and human infants or pets). From this viewpoint, this paper presents psychological experiments on the interaction between humans and animated characters that mimic the human voice at the prosodic level using prosody-only voice under different vocal rate of character’s voice: (a) faster than normal, (b) normal speed, and (c) slower than normal. We examine the subjects’ impression towards animated characters with the above conditions of their voice using post-questionnaire and analyse the change of the speech rate of subjects. The results indicate that most humans may prefer an animated character with a faster voice to that with a slower voice. Moreover, the speech rate of humans changes to opposite of vocal rate of animated characters’ voice. I.e. the speech rate of subjects becomes slower when the voice of character becomes faster, and vice versa.

1. Introduction
In daily life, humans talk with infants or pets that make utterances in prosody-only voice, such as the babbling of an infant or the call of an animal. This raises the following questions: Considering that little linguistic information is conveyed through these utterances, why do humans enjoy interacting with these primitive partners? Can this phenomenon be applied to human-computer interaction?

We tried to achieve social interaction with empathy between humans and computers by using an example of such primitive interaction between humans and infants or pets [2]. Specifically, we focused on echoic mimicry utterances with prosody-only voice, i.e., the babbling of a baby or the call of an animal, as a key to enhancing empathic relationships with partners, even though this form of communication cannot send much linguistic information [1, 3, 4]. We have examined the social effects echoic mimicry using prosody-only voice through psychological experiments under different mimicry output probability conditions [3] and under different output response timing conditions [1, 4].

We focused on the social effects that a computer can give a human under different time-scale modification of voice. The change of vocal rate of a person affects psychological impression of them, as has been pointed out [5, 6, etc]. It has been also considered that the change of vocal rate of a person affects that of other participants in a conversation, and such a coordination conducts empathic relationship. Our particular interests were in what kinds of impressions the human develops when the computer vocalizes faster or slower than natural vocal speed, and what kinds of effects actual vocal rate of participants works as a result of coordination through interaction.

2. Method

[Experimental environment] An experiment was performed to
The system expresses a mimicking voice after a subject’s utterance (upper), and a non-mimicking voice of three morae with a constant prosody after a subject’s utterance (lower).

The system works as follows: an audio signal from a microphone is sampled at 16 kHz (A/D Conversion). The time sequence of the power pattern, i.e., the loudness of the human voice, is calculated for each frame of the audio (Power Detection). The segmentation of the speech is determined by the threshold energy based on the result of the power calculation (Speech Segmentation). The time sequence of an $f_0$ pattern, i.e., the pitch of the human voice, is calculated for each segment ($f_0$ Detection). For the experiment, $f_0$ patterns are detected by the Average Magnitude Differential Function (AMDF) method. Prosody-only voice is synthesized by combining sine waves based on the power calculation and $f_0$ pattern (Speech Generation). For the purpose of experimental control, we prepared prosody-only voice and non-mimicry voice (20%). For the conditions, prosody-only voice is set constant under every condition: mimicry voice (80%) and non-mimicry voice (20%). For the glance and motion conditions, the character is set to perform the same expression for each under all voice rate conditions. For the glance condition, it always rolls from side to side by default. When it detects the voice of a human, it stops rolling. Then, when it starts to talk to the subject, it becomes larger than its default size. The character responds after an 500 msec delay for both mimicry voice and non-mimicry voice when the subject talks to it.

The experiment was carried out in the following sequence.

1. Subjects watch a sample interaction video for about two minutes.
2. They listen to the instructions illustrated below.
3. They interact with the character using infant building blocks (Fig. 4) for a period of four minutes.
4. They answer a questionnaire about their impression of interaction with the character.
5. They repeat steps 1 - 4 two more times under different conditions.

[Instructions] The experimenter explains that the character on the computer screen in front of subject has the same intelligence level as a one-year-old child and it is therefore not very smart. The subject is told to teach the character the names of objects while assembling the objects with building blocks. While giving the instructions, the experimenter shows the subject samples of building blocks (giraffe (left), parrot (middle), rabbit (right)).

**Figure 4:** Examples of building blocks used by a subject (giraffe (left), parrot (middle), rabbit (right))

1 However, the output probability of mimicry voice with prosody-only voice is set constant under every condition: mimicry voice (80%) and non-mimicry voice (20%). For the glance and motion conditions, the character is set to perform the same expression for each under all voice rate conditions. For the glance condition, it always rolls from side to side by default. When it detects the voice of a human, it stops rolling. Then, when it starts to talk to the subject, it becomes larger than its default size. The character responds after an 500 msec delay for both mimicry voice and non-mimicry voice when the subject talks to it.
of animals or cars that can be created using the building blocks. [Evaluation Measure] The subject can start to interact with the character easily because the subject has been given the above instructions. In a psychological evaluation after the interactive session, the experimenter asks the subject to answer a questionnaire including the following evaluation items which are prepared based on a pilot study on psychological evaluations of interaction between humans and robots [7], or anthropomorphic agents [8]. All of these items were evaluated based on a 7-point scale by the subjects.

1. Social Desirability
   Cooperation: Degree of cooperation that the subject feels towards the character when the subject interacts with the character
   Task achievement: Degree of task achievement that the subject feels towards the character when the subject interacts with the character
   Stress-reduction: Degree of stress that the subject feels towards the character when the subject interacts with the character

2. Individual Familiarity
   Friendliness: Degree of friendliness that the subject feels towards the character when the subject interacts with the character
   Likability: Degree of likability that the subject feels towards the character when the subject interacts with the character
   Sympathy: Degree of sympathy that the subject feels towards the character when the subject interacts with the character

On psychological evaluation, it will be likely that the subjects prefer the character with the faster voice rate in social desirability while they prefer the character with the slower voice rate in individual familiarity.

[Analysis measure] Our measurements of the vocal rate of subjects in the interactive sessions is based on the average mora durations of the voice unit that was bounded by pauses longer than 100 msec according to the method in [9]. In this method, the average mora duration (AMD) of a voice unit is the duration of the voice unit divided by the number of morae appearing in the voice unit. Therefore, a smaller AMD of a voice unit means a higher speech rate, and a larger AMD means a lower speech rate.

On analysis of conversational data, it will be likely that the vocal rate of the subjects similarly changes faster or slower according to that of the character as a result of coordination.

3. Results and Discussions

I. Psychological evaluation
   Full factorial ANOVAs were performed on all of the measures as within-subject factors by using MOS values (Fig. 5).
   There were significant differences in individual familiarity among the three vocal rate of the character: $F(2, 23) = 3.452, p < .05$ between neutral and slower conditions in social desirability (Fig. 5 (left)). And There were no significant differences in individual familiarity among the three vocal rates of the character: $F(2, 23) = 1.456, p = .240$ among three conditions in individual familiarity (Fig. 5 (right)).

   The total tendencies of the MOS values were almost all the same, and the subjects evaluated the animated character positively according to the following order of vocal rate: slower < faster < neutral. This result indicates that the impression humans form of social desirability depends on the impression of individual familiarity independent of the system design.

   The reasons why the subjects felt more positive about the character with faster vocal rate than slower are as follows:
   - The subjects might rhythmically interact with the character of faster vocal rate at their own utterance speed because the character returned shorter responses after their talk.
   - The subjects might interpret the longer responses of the character with slower vocal rate as the talk back to them.

II. Analysis of conversational data
   Figure 6 shows the average speech rates of six subjects using AMD (average mora duration).

   The total tendencies of the average mora duration were almost all the same as the following order of the vocal rate: slower < faster except subject E. This result means that the speech rate of the subject did not change to match the vocal rate of the character but exhibited the reverse tendency: the subjects talked at faster speech rate while the character responded at slower vocal rate, and vice versa.

   The reasons why the speech rate of subjects change the reverse tendency of vocal rate of the character are as follows:
The subjects react inversely to try to maintain a constant rhythm of interaction.
The character with slower voice rate gave negative effects on the speech rate of the subjects on comparing with the character of faster voice rate because the character disturb the subjects utterance rhythm as the result of longer response.

4. Conclusions
In this paper, we focused on social effects on the vocal rate of a computer’s response, as the key towards achieving social interaction with empathy between humans and computers. We examined how slower or faster vocal rate of a computer come to affect humans by using an interactive system that mimics prosodic features of the human voice echoically with prosody-only voice. From the results, we obtained the following findings:

- The subjects more favorably interpret the system with the faster vocal rate than the system with the slower vocal rate because it returned shorter responses to them.
- Actual speech rate of subjects changed inversely to that of the system because the subject coordinated to make a more rhythmical interaction without any intention.

We believe that these findings can be applied to the design of interactive systems such as voice interface system including communication robots and CG characters.

As future work, we will study the other features related to the a computer’s response with echoic mimicry using prosody-only voice on time-scale including rhythm pattern of utterances, as a key towards forming interpersonal relations between humans and computers. We will base this work on the social bonding observed in proto-communications.

5. References