

Effects of Voice Prosody by Computers on Human Behaviors

Noriko Suzuki[†], Yohei Yabuta[‡], Yugo Takeuchi[‡], †, Yasuhiro Katagiri[†]

[†] ATR Media Information Science Labs., [‡] Shizuoka University

{noriko, katagiri}@atr.co.jp, {cs9090, takeuchi}@cs.inf.shizuoka.ac.jp

Abstract

This paper examines whether a human is aware of slight prosodic differences in a computer voice and changes his/her behaviors accordingly through interaction, when the prosodic difference carries informational significance. We conduct a route selection experiment, in which subjects were asked to find a route in a computer generated 3-D maze. The maze system occasionally provides a confirmation in response to the subject's choice of a route. The prosodic characteristics of confirmation utterances are made to marginally change according to whether the route selected is the right route for reaching the goal or a wrong route that ends up in a cul de sac. In this experiment, subjects are able to pick up the difference and successfully navigate through the maze. This result demonstrates that subjects are sensitive to even a slight change in the voice's prosodic characteristics and that computer voice prosody can affect the route selection behaviors of subjects.

1. Introduction

This paper presents an experimental result on the effects of prosodic difference in a computer generated voices on human behaviors in human-computer interaction.

People perceive internal states of conversational partners, including emotions, intentions and attitudes, not only in the content of their utterance, but also in the prosody of their voices. Over the past few years, a considerable number of studies have been made on the relationship between the internal states of humans and prosody in voice (e.g., [1, 2]). Therefore, designers of human-computer interaction focus on prosody as an important factor of expressing personality in computers. Some studies have tried to manifest personalities by controlling acoustic characteristics related to prosody in computer voices (e.g., [3, 4]). Although some studies have been made on the effects of prosody itself on human-computer interaction (e.g., [5]), there has been little focus on the effects of prosodic difference in computer voices.

This paper has the following two research aims: i) to examine whether a human can find slight prosodic differences in computer voices through interaction, and ii) to examine whether prosodic differences in computer voices affect both human impressions of computer voices and human behaviors. First, we conducted two experiments to examine sensitivity of a human to slight prosodic differences in a computer voice under both non-interactive and interactive environments. Second, we investigated the behavior change ratio of a human and impressions of computer voices according to prosodic differences in computer voices.

In this paper, we set up a route selection experiment as an interactive environment, in which subjects were asked to find a route in a computer generated 3-D maze. This maze system provides confirmation or collaborating voices, "ii?" (O.K.), when

subjects select a route at junctions. We mark the significance of route information, either normal alley or blind alley, by using different prosody in confirmation voices. We utilize pitch range as a parameter of prosody, since pitch range is considered to modulate expressions of emotions [1]. We discuss whether this difference in pitch range in confirmation voices can help subjects navigate successfully through the maze. In addition, we conduct psychological evaluation to investigate the subjects' impressions of the confirmation voices.

2. Related Works

The most closely related work to our approach is the expressive speech processing project [6]. This project is examining both factors and construction of expressive speech by using a corpus of natural speech. We focus on the effects of expressive speech by a computer on human behaviors. A system in the CASA project [4, 7] as well as the TVML system [3] try to produce a virtual personality by using prosodic differences in a computer voice. In contrast to their approach, we examine whether users perceive personality or emotion in prosodic difference through interaction according to the designer's intention.

3. Perception Experiment

Before we examined the effects of pitch range in a computer voice on a human-computer interactive session, we conducted a perception experiment under a non-interactive environment in order to compare how different or more natural subjects felt a voice sounded between the original pitch range and wider ranges.

3.1. Stimuli

We designed five kinds of source utterances by modifying a single utterance, "ii?" (O.K.), with different magnitudes of pitch range: 1.00, 1.25, 1.50, 1.75 and 2.00 (Figure 1). The utterance consists of a single vowel that forms a single word. The original utterance was read by a male Japanese speaker and recorded in the soundproof studio of our laboratory. It was digitized at a sampling rate of 48 kHz. Pitch range manipulation of the utterance was performed by the STRAIGHT technique [8]. We prepared nine utterance pairs (1 original pitch range and 5 pitch ranges pair x 2 orders - 1) as stimuli for the perception experiment.

3.2. Method

Nine Japanese native speakers who were undergraduates and graduate school students from 21 to 28 years old took part in the experiment. Subjects were instructed to compare impressions between first and second utterances in every utterance pair by using a pair of 7-point scales: one scale measured

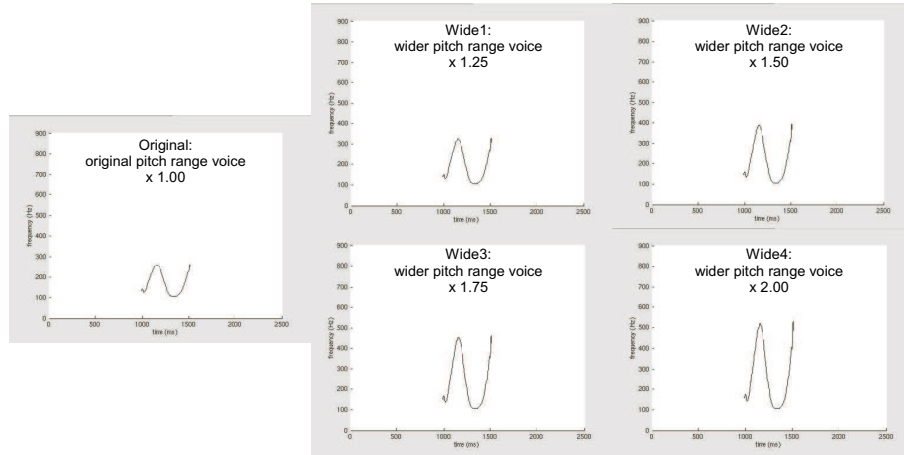


Figure 1: Perception experiment: stimuli voices “ii?” (O.K.?) at different pitch ranges

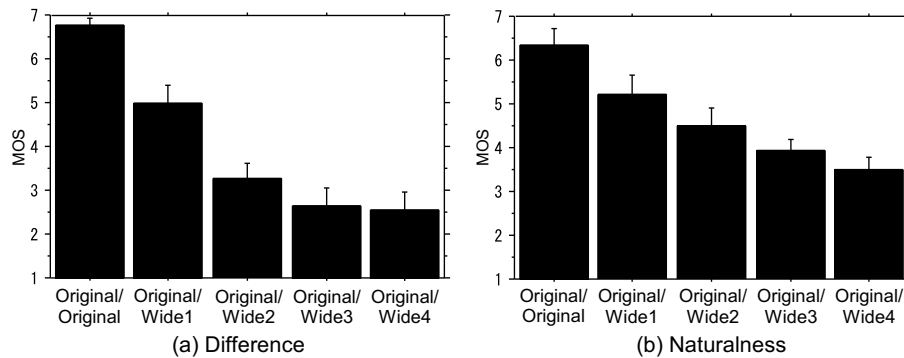


Figure 2: Perception experiment: results (a) Difference and (b) Naturalness

SAME/DIFFERENT perceptions and the other scale measured NATURAL/NOT NATURAL impressions. The more to the left a slash was marked, the more ‘DIFFERENT’ or ‘NATURAL’ was the impression, and the more to the right, the more ‘SAME’ or ‘NOT NATURAL’ the utterance sounded. The utterance pairs output at random order from a speaker.

3.3. Results

Figure 2 shows the average MOS (mean opinion score) value of the subjects for each evaluation item related to the subjects’ impressions of the nine stimuli utterance pairs: between original pitch range utterance (Original) and itself as well as between original pitch range utterance and wider pitch range utterances (Wide1–Wide4).

There were significant differences among five pairs for *difference* ($F(4, 40) = 25.15, p < .01$) as well as for *naturalness* ($F(4, 40) = 10.26, p < .01$) from the results of full factorial ANOVAs using within-subject factors.

Both evaluation items show that the total tendencies of the MOS values were almost completely the same and also that the subjects felt that the utterances with nearer pitch range to the original utterance were more natural and closer to the same. In other words, subjects perceived difference and naturalness between two utterances according to differences of pitch range that we designed.

4. Interactive Experiment

This experiment examines whether subjects are aware of slight prosodic differences in a computer voice and change their behaviors accordingly through interaction sessions, when the difference in pitch range carries informational significance. We used a route selection task, in which subjects were asked to find a route to the exit of a computer generated 3-D maze. まるい

4.1. Method

Subjects: 43 university students (from 18 to 25 years old).

Task: Subjects were given a route selection task by using a computer generated 3-D maze on a 50-inch plasma display with touch panel sensors. They were instructed to arrive at an exit within 500 steps from the entrance of the maze. They were shown how to select a route by touching buttons for four directions on the display. After they selected a route at a crossroads or T junction, the confirmation voice was produced from a hidden speaker behind the subjects and they have a chance to reselect the route. They were instructed that their partner sometimes confirms their route selection via voices. We set up the maze system for task that nobody arrived at an exit.

Confirmation voices: We selected two voices of the same content, “ii?” (O.K.?), with different pitch ranges, Original: x 1.00 and Wide2: x 1.50, as confirmation voices. In this experiment, we required that the difference between two confirmation voices is perceived by the subjects and that both voices sound natu-

ral. We selected the Wide2 voice, 1.50 times wider pitch range than Original, as a control confirmation voice relative to Original voice, as a result of the previous perception experiment.

4.2. Evaluation items

In this experiment, we used the following two items for evaluation:

I. Analysis of human selection behavior: Selection change ratio of subjects when their partner confirmed their route selection via voice.

II. Psychological evaluation: Average MOS values were obtained by giving a post-experimental questionnaire including the following items: five items on impressions of confirmation voices, four items on impressions toward the difference of the two confirmation voices, three items on the impressions toward subjects own behaviors, four items on the impressions of task itself.

4.3. Procedure

We used the following experimental procedure:

1. While receiving the instructions, the subjects played a short rehearsal session of the route selection task in a computer generated 3-D maze.
2. After the instructions and the rehearsal, the subjects started to perform the task:
 - (a) When the subjects selected \uparrow button, they moved forward. When the subjects selected \leftarrow , \rightarrow or \downarrow , they turn left, turn right or go backward.
 - (b) After they selected a route at a crossroads or T junction, the confirmation voice was output from a hidden speaker behind the subjects, and they had a chance to reselect a route. One trial consisted of 20 passage ways and 10 junctions. We set up the task so that the subjects always heard confirmation voices 10 times in every trial.
 - (c) They repeated (a) and (b) for five trials.
3. After subjects heard confirmation voices 50 times, the 3-D maze system showed a message of 'exceeds 500 steps' on the plasma display and the task was stopped.
4. They answered a post-experimental questionnaire.

4.4. Conditions

Table 1 shows three conditions of pitch range in confirmation voices and each condition's informational significance for the route.

Table 1: *Combinations between pitch range in confirmation voice and route information.*

Conditions	Confirmation voice: O	Confirmation voice: W
Rule1	normal alley	blind alley
Rule2	blind alley	normal alley
Random	both alleys	both alleys

O: Confirmation voice with original pitch range

W: Confirmation voice with 1.50 wider pitch range

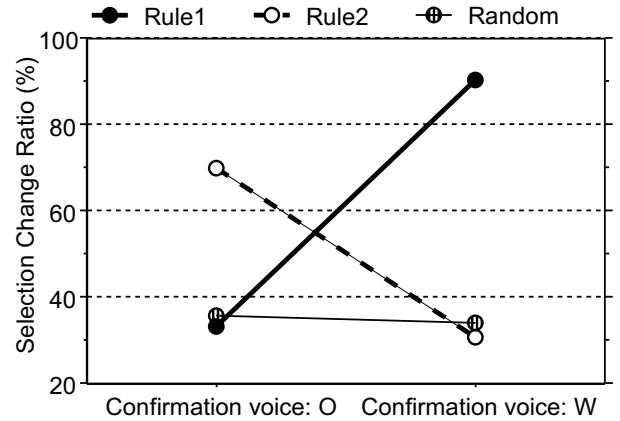


Figure 3: *Results of interactive experiment: selection change ratio*

4.5. Hypothesis

We assumed that the subjects would find that prosodic difference in confirmation voices indicates informational significance for a route. Furthermore, we have the following predictions:

Rule1: The subjects find that a confirmation voice with a wider pitch range prompts reselection of a route. After they hear a confirmation voice with a wider pitch range, they reselect the route at a higher ratio than the case of a confirmation voice with original pitch range. They have a positive impression of a confirmation voice from their partner.

Rule2: The subjects find that a confirmation voice with the original pitch range prompts reselection of a route. After they hear a confirmation voice with the original pitch range, they reselect the route at a higher ratio than the case of a confirmation voice with a wider pitch range. They have a positive impression of a confirmation voice from their partner.

Random: The subjects do not find any rule correspondence between prosodic difference in confirmation voices and the route. Even after they hear both types of confirmation voices, they reselect the route at a lower ratio than that under other conditions. They do not have a positive impression of a confirmation voice from their partner.

4.6. Results and Discussion

I. Human selection behavior: Figure 3 shows selection change ratio of subjects after they heard a confirmation or collaborating voice from their partner under three conditions. There were significant differences between selection change ratio of a confirmation voice with a wider pitch range and that of a confirmation voice with the original pitch range under both Rule1 ($\chi^2 = 201.63$) and Rule2 ($\chi^2 = 89.69$) conditions. However, there were no significant differences between selection change ratio of a confirmation voice with wider pitch range and that of a confirmation voice with original pitch range under Random condition ($\chi^2 = .08$). These results suggest our predictions are supported.

Moreover, the selection change ratio of a confirmation voice with a wider pitch range (Rule1) is higher than that of a confirmation voice with the original pitch range (Rule2) ($\chi^2 = 26.98$). This result suggests the following points:

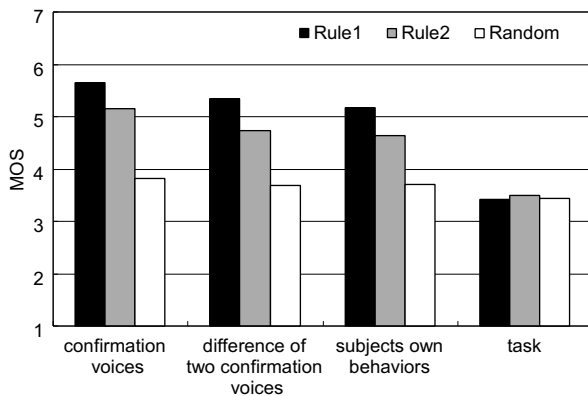


Figure 4: Results of interactive experiment: psychological evaluation

- The subjects found stronger prompting of route reselection from a confirmation voice with a wider pitch range than that with normal pitch range.
- The subjects perceived confirmation voices as social messages from their partner rather than as just attention sounds.

II. Psychological evaluation: Figure 4 shows the average MOS (mean opinion score) value of the subjects on each evaluation item under three conditions. There were significant differences between Rule1 and Random conditions as well as between Rule2 and Random conditions for all evaluation items except the impressions of the task itself as results of full factorial ANOVAs using between-subject factors: for *the impressions of confirmation voices* ($F(2, 40) = 9.44, p < .01$), for *the impressions toward difference of the two confirmation voices* ($F(2, 40) = 6.48, p < .01$) and for *the impressions toward subjects own behaviors* ($F(2, 40) = 15.35, p < .01$). This result supports the following conclusions:

- The subjects found that prosodic difference of confirmation voice indicates a different significance for route information under Rule1 and Rule2 conditions.
- The subjects felt that these meaningful confirmation voices were useful for reselecting the right route under Rule1 and Rule2 conditions.

5. Conclusions

In this paper, we focused on the effects of prosodic difference, i.e., original pitch range or wider pitch range, as a key toward expressing intentions or emotions by a computer. We indicated significance in route information by using prosodic difference in the route selection task. We examined the sensitivity of the subjects to slight prosodic differences in a computer voice under both non-interactive and interactive experiments. Moreover, we investigated the behavior change ratio of a human and impressions of computer voices according to the prosodic difference in computer voices. From the results, we obtained the following findings:

- The subjects are sensitive to even a slight difference in pitch range of confirmation or collaborating voices in both non-interactive and interactive environments.

- In interactive environment, the subjects find the link between the difference in pitch range of confirmation voices and route information.
 - They regard the confirmation voice as message from their partner and they use it to reselect a right route.
 - They also have positive impressions of the confirmation voice.

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

As future work, we will study whether a human perceive different intentions or emotions in computer voices with different prosody under different status of interaction. We will base this work on the power of prosody in human-computer interaction.

Acknowledgments

This research was supported in part by the Telecommunications Advancement Organization of Japan.

6. References

- [1] Scherer, K.R., Ladd, D.R. and Silverman, K.E.A., “Vocal cues to speaker affect: testing two models”, *J. Acoust. Soc. Amer.*, Vol. 76, No. 5, pp. 1346–1356, 1984.
- [2] Couper-Kuhlen, E. and Selting, M. (Eds.), *Prosody in conversation – interactional studies*, Cambridge University Press, 1996.
- [3] Hayashi, M., Ueda, H. and Kurihara, T., “TVML (TV program Making Language) - automatic TV program generation from text-based script -”, in *Proc. of Imagina’99*, 1999.
- [4] Nass, C., and Lee, K. M., “Does computer-generated speech manifest personality?”, in *Proc. of CHI2000*, pp. 329–336, 2000.
- [5] Suzuki, N., Takeuchi, Y., Ishii, K. and Okada, M., “Effects of echoic mimicry using hummed sounds on human-computer interaction”, *Speech Communication*, Vol. 40, No. 4, pp. 559–573, 2003.
- [6] Campbell, N., “Building a corpus of natural speech -and tools for the processing of expressive speech- the JST CREST ESP Project”, in *Proc. of EuroSpeech2001*, pp. 1525–1528, 2001.
- [7] Nass, C., Steuer, J. and Tauber, E., “Computers are social actors”, in *Proc. of CHI94*, pp. 72–78, 1994.
- [8] Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A., “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds”, *J. Speech Communication*, Vol. 27, No. 3–4, pp. 187–207, 1999.