Toward a mutual adaptive interface: An interface induces a user's adaptation and utilizes this induced adaptation, and *vice versa*

Takanori Komatsu (komatsu@fun.ac.jp)
Future University-Hakodate, 116-2 Kamedanakano, Hakodate 041-8655 JAPAN
Atsushi Utsunomiya (au@cs.c.u-tokyo.ac.jp), **Kentaro Suzuki (suzuki@cs.c.u-tokyo.ac.jp)**
Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp), **Kazuo Hiraki (khiraki@idea.c.u-tokyo.ac.jp)**
The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902 JAPAN
Natsuki Oka(oka@mrit.mei.co.jp)
Matsushita Electric Industrial Co., Ltd., 3-4 Hikaridai, Seika, Kyoto 619-0237 JAPAN

Abstract

The purpose of this paper is to construct a meaning acquisition model as a basic technology for a mutual adaptive speech interface which can communicate smoothly with an everyday user. We then constructed the meaning acquisition model in consideration of the following assumptions: (1). the model needs to induce a user's adaptation and to utilize this induced adaptation for the meaning acquisition processes, (2). the model focuses on the prosodic information, rather than phoneme information on which most past interface studies focused. As a result, we could confirm that this model could recognize the intentions/meanings of users' verbal commands by inducing users' adaptations and utilizing these for the meaning acquisition process when appropriate instructions were given to them from an experimenter. This result would complement the interface studies that focused only on phoneme information, and contribute to a customization or personalization technology for a speech interface system.

Introduction

The final purpose of this study is to construct an adaptive speech interface system which can communicate with a user through a **mutual adaptation process**. This mutual adaptation process discussed here is the process of repeating the following: a user adapts to an interface, then the interface adapts to the user by using her/his adaptation, i.e., both learn to respond and adapt appropriately to each other's behavior by using the partner's adaptation. Such a process would commonly be observed in a pair who can communicate smoothly, e.g., a child and her parents, a dog and its owner. Therefore, this mutual adaptation process should also be realized in a relationship between a user and a desired interface system.

This mutual adaptation process between a user and an interface consists of two adaptation processes: one is a user's adaptation to an interface, and the other is an interface's adaptation to a user. From the former point of view, some researchers have studied an adaptable interface system which is designed to induce a user to adapt to the system intuitively and naturally (for example, Ueda *et al.*, 2003). From the latter point of view, some have studied an adaptive learning interface system which provides a smooth operating environment for a user by learning and adapting to the user's operation patterns (for example, Sears & Shneiderman, 1994). It is generally believed that humans have certain cognitive features that they can use to smoothly adapt to their interaction partners, even if these partners are not human beings, such as computers or cars. However there is no adaptive interface research that concretely studies the human cognitive features used for adapting smoothly to an interaction partner. Moreover, there is also no past research into the formation of a mutual adaptive relationship between a user and an interface, i.e., not only does an interface adapt to the user, but the user also adapt to the interface by using the interface's adaptation.

The purpose of this paper is to construct a meaning acquisition model as a basic technology for a mutual adaptive speech interface system. This system can recognize the intentions/meanings of a user's verbal commands by inducing user adaptation based on human cognitive features, and utilizing this induced adaptation for the meaning acquisition process. Let us suppose that this model succeeds in inducing a user's adaptation and utilizing this for the meaning acquisition process, and the user then adapts to this model repeatedly. In this case, the user and the model would eventually form a mutual adaptation process. In this paper, as a first step toward such a mutual adaptive interface, we tried to realize a part of mutual adaptation between a user and a meaning acquisition model; i.e., the model induces the user's adaptation and utilizes this induced adaptation for the meaning acquisition process. Specifically, at first, a communication experiment was carried out to observe and analyze the human cognitive feature used for communicating. A meaning acquisition model was then proposed and constructed based on the results of the communication experiment. Finally, a testing experiment was carried out to clarify whether this proposed meaning acquisition model could recognize actual everyday users' speeches.

In this study we focus on prosodic information as an input for the meaning acquisition model. Prosodic information cannot be written as texts/characters but is expressed as stress or inflection, rather than phoneme information on which most past interface studies focused. To utilize phoneme information, the mapping between particular units of speech and specified actions, such as a dictionary-like database, needs to be prepared *a priori*. In some cases, however, it would be preferable not only to use the given command sets, but also to configure one's own commands through interaction as a result

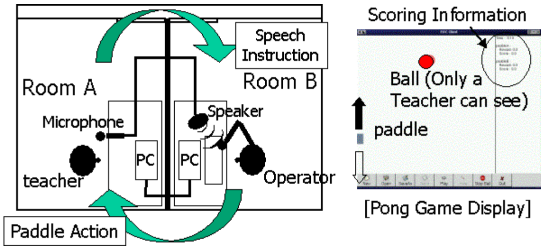

Figure 1: Game Environment

of the mutual adaptation processes. Recently, some researchers have started studying the roles of prosodic information in speech communication. They have found that the specific inflection patterns are universally interpreted with the same meanings regardless of language spoken, e.g., increasing intonation is interpreted as an interrogative (Scherer *et al.*, 1991) or a turn-taking signal (Pirrehumbert & Hirschberg, 1990). However, there is not yet any research which proposes a meaning acquisition model by focusing on these universal properties of prosodic information.

The meaning acquisition model proposed above, which focuses on prosodic information, would compliment the interface studies that focused only on phoneme information. Moreover, the result should contribute to achieving an interactive speech interface that would be practical for use by any user and which would provide insights for continued researches into Human-Agent Interaction (HAI) researches.

Communication Experiment

Purpose and Settings

In order to construct a desired meaning acquisition model described above, we need to clarify the human cognitive features used for communicating. To do so, at first, we carried out an experiment to observe how human subjects create a smooth communication by acquiring meaning for utterances in languages they do not understand. Pairs of subjects, one teacher and one operator in each pair, participated in this experiment: the teachers were placed in room A and the operators were placed in Room B (Figure 1). The goal of the subjects in each pair was to work together to get the highest possible score in “Pong”, a computer game rather like tennis or squash. Ten points were awarded to the subjects each time they hit the ball with a paddle, and ten points were deducted each time they missed it. The teachers’ role was to give instructions to the operators, and the operators’ was to move the paddle to hit the ball. The operators’ display did not show the ball (Figure 1), which was their target, so to operate the paddle they needed to understand the meanings/intentions of the teacher’s instructions, which were made linguistically incomprehensible. In this experiment, each pair played two consecutive 10 minutes game, with 3 minutes of rest between them.

Table 1: Average CDV and HV in Group 1

Category	(CDV, HV)
Category 1 (2pairs)	(0.5, 0.5), (0.3, 0.2)
Category 2 (5pairs)	(0.9, 0.3), (1.0, 0.2), (1.0, 0.5), (0.8, 0.6), (1.0, 0.6)
Category 3 (4pairs)	(1.0, 0.9), (1.0, 0.7), (1.0, 0.7), (0.9, 0.8)

Subjects

There were two groups of subjects. In each group, the player could not linguistically understand the instruction of the teacher.

- Group 1 (11 pairs, 22 Japanese, 22-28 years old, 18 men and 4 women): Each pair of subjects shared the same mother tongue. To make the teacher’s instructions linguistically incomprehensible for the operator, the teacher’s instructions were transmitted through a low-pass-filter (LPF). The LPF masked the teacher’s speech phonemes but did not affect the prosodic features of their speech.
- Group 2 (6 pairs, 23-26 years old, 10 men and 2 women): Each pair of subjects did not share the same mother tongue and the teachers were asked to speak their mother tongue: i.e., the operator could not linguistically understand what the teachers were saying even though no LPF was used.

The experimenter told the teachers to use as verbal instructions whatever words or sentences that they wanted.

Results

To evaluate the pairs’ performance, two values were assigned to each of the operators’ actions: moving the paddle and hitting the ball. For each action, if the operator moved the paddle in the teachers’ intended direction, the correct direction value (CDV) was awarded one point for each action; if they moved it in a different direction, the CDV was zero. If the operators hit the ball, the hit value (HV) was awarded one point for each action; if they missed it, the HV was zero. We used a statistical testing hypothesis formed by using binominal distribution to group the subjects, and then the pairs of subjects were divided into the three following categories:

Category 1 Average CDV less than 0.8.
Category 2 Average CDV more than 0.8; Average HV less than 0.7.
Category 3 Average CDV more than 0.8; Average HV more than 0.7.

Tables 1 and 2 show the average values of the last ten actions for the three categories. In Group 1, out of 11 pairs, two operators failed to understand any instructions (Category 1). Among the nine reminding pairs, five operators succeeded in moving the paddle in the direction

Table 2: Average CDV and HV in Group 2

Category	(CDV, HV, teacher–operator)
Category 1 (2pairs)	(0.5, 0.5, Chinese–Japanese), (0.4, 0.4, Chinese–Japanese)
Category 2 (2pairs)	(1.0, 0.6, Indonesian–American), (0.8, 0.6, Chinese–Japanese)
Category 3 (2pairs)	(1.0, 0.9, Spanish–Filipino), (1.0, 0.7, Korean–Chinese)

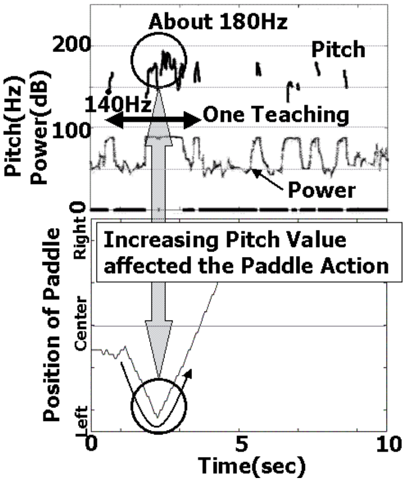

Figure 2: High-Pitch Element in Teacher's Instruction

the teachers intended but could not hit the ball well (Category 2). In the four remaining pairs the operators could move to the teachers' intended position and hit the ball well (Category 3). In Group 2, two out of six pairs were in Category 1, two in Category 2, and the remaining two were in Category 3.

Here, when the operator scored over 0.8 average CDV (i.e., pairs in Category 2 and 3), we recognized that the operator in this pair succeeded in understanding the teacher's instructions somehow. In these pairs, we observed mutual adaptation processes; i.e., not only did the player try to learn the intentions/meanings of the teacher's instructions, but also the teacher simultaneously revised the manner of giving instructions to fit the player's mode of learning. Concretely, in these pairs, we could observe the following specific behaviors that were regarded as **one of the human cognitive features used for communicating**.

- Teachers
  1. Decreasing the types of instructions.
  2. Making the operator focus on her/his actions by increasing voice pitch (See, Figure 2). We named this sound feature "Attention Prosody (AP)".
- Operators
  1. When an instruction was given, moving the paddle to indicate their comprehension of the given instruction.
  2. Moving the paddle differently according to different types of instruction.
  3. Correcting their paddle actions by using the given APs.

As a result, we can assume that the operators can recognize the intentions/meanings of teachers' verbal commands by inducing the teachers' adaptations, and utilizing these adaptations for the meaning acquisition processes. In addition, we can observe that the AP sound features were universally interpreted by the operators as "caution on their current action" and had a significant role in meaning acquisition process[1].

Overview of a Meaning Acquisition Model

To construct a meaning acquisition model which can recognize the meanings of the given verbal commands, this model needs to realize the operators' observed behaviors in the previous communication experiment. Therefore, the model needs to satisfy the following requirements.

1. Recognizing that given verbal command indicates certain paddle action.

2. Finding critical sound features in speeches to distinguish different types of instructions.

3. Extracting AP sound features from verbal commands to use for the meaning acquisition process.

To meet the above requirements, we made the following assumptions in creating the model:

1. When a paddling action is correct (hitting the ball), the model should recognize that the meaning of the given instruction indicated the current action. Conversely, when an action is incorrect (hearing an AP sound), the model should recognize that the meaning of the given instruction did not indicate the current action.

2. Certain probability distribution should be selected to explain an incoming instance, which is a paddle action paired with an instruction sound (eight-dimensional sound vector such as pitch, zero-cross number and so on, see Figure 3), from a mixture of normal distributions. Here, each distribution expresses each intention/meaning of an instruction.

To recognize the meanings of instructions, this model must learn to estimate the parameters (average and standard deviation, in each dimension) of probability distributions to explain the incoming instances. As a basic methodology, we used the EM algorithm (Dempster *et al.*, 1977) which can be used even for variables whose values have never been directly observed, provided the

[1] For a more detailed description of this communication experiment, refer to the article, Komatsu *et al.*,(2002).

```
1. Acquire the pitch, paddle position and reward data.
2. Correct the noisy and deviated pitch values (Ignoring such values and replace these with the interpolated ones).
3. Smooth the corrected pitch data (Calculating the moving average of this corrected pitch data).
4. Acquire the eight-dimensional sound vector.
     Calculating the (1) differential pitch value, (2) second-order differential pitch, (3) zero-cross number,
     (4) number of pitch's sudden transitions, (5) (1)'s sudden transitions, (6) (2)'s sudden transitions
     and (7) fully voice length with using (8) pitch value which is acquired in 3.
5. Check the onset and end point of the instruction.
     (When there are no sound over one second, presume that the instruction is finished.)
6. When a reward is given
          6-1. Calculate the paddle action value, which is weighted by the recent actions.
          6-2. Average the eight-dimensional sound vector during the instruction.
          6-3. Using the results of 6-1 and 6-2, start the extended EM algorithm to estimate the parameter values.
          6-4. Update the parameter values with the result of 6-3.
7. Return to 1.
```

Figure 3: Learning Procedures of Proposed Meaning Acquisition Model

general form of the probability distribution governing these variables is known. To estimate each distribution's parameters, the EM algorithm can use positive instances that are acquired when the positive reward is given; however, this algorithm cannot use the negative instances that are acquired when a negative reward is given, because this speech sound did not indicate this user's paddle action. Therefore, we developed an extended EM algorithm that could include negative instances for estimating the parameters. The detailed learning procedure is described in Figure 3.

To evaluate the basic competence of this meaning acquisition model, we carried out a testing experiment to confirm whether this model could learn to recognize the meanings of instructions through interaction with a human instructor. As a testing environment, this meaning acquisition model was incorporated into the paddle component of the software for the "Pong" game, that the human operator moved in the communication experiment. As already mentioned, this model does not focus on the phoneme information, so it must learn to recognize the intentions/meanings of instruction through prosodic sound features so as to distinguish the different instruction types, regardless of the actual language being spoken. It can do so if there are enough of these prosodic sound features. Therefore, to evaluate the performance of this model, an instructor used the five following types of instruction, and we observed whether this model could learn to recognize the meanings of these instructions.

(1). High-pitched utterances for upward and low-pitched ones for downward while saying "ahhh."

(2). A long voice for up and a choppy voice for down (while saying "ahhh").

(3). "UE" for up and "SHITA" for down, in Japanese.

(4). "UP" and "DOWN" in English.

(5). Inversion of (1).

As a result, we could observe that this model could learn to recognize the meanings of all five types of the given instructions. From this testing experiment, we thus confirmed that this model had sufficient ability to recognize the meanings of given instructions from an actual human instructor through interaction when that instructor was using the salient sound prosodic features. Moreover this model recognized the meanings without any *a priori* knowledge of instructions, e.g., a dictionary-like database.

Interaction with Actual Users

Purpose and Settings

In the previous section, we could confirm that the proposed meaning acquisition model could recognize the intentions/meanings of instructions given by a human instructor. However, this instructor was an ideal one because the instructor already knew and understood well this model. Therefore, to apply this model for an interface system which would work in actual everyday situations in the future, we were required to confirm whether this model could recognize the instruction of users who had no specific knowledge of this model. Another testing experiment was then carried out with subjects participating in the way everyday users would. The goal of each subject was to teach verbal commands to the constructed meaning acquisition model that was driving the "Pong" paddle, and to make the paddle move as desired. Recently, Harada (2002) reported that most people hesitate to talk naturally to agents that do not have an actual physical entity, e.g., a life-like agent in a computer or a computer itself. However, finding the conditions that will induce humans to speak to computers naturally is worthwhile. In this experiment, we focus on the effect of experimenter instructions as one of the conditions. For example, an experimenter gave an instruction to the subject such as, "Please talk to this model **as if talking to your friends**." We assumed that this kind of instruction would be equivalent to catchphrases that would help users intuitively understand how to use and interact with this model without reading thick manual documents.

Subjects

Two groups of subjects participated. The experimenter gave different instructions to each group.

- Group A (4 subjects, 3 Japanese and 1 Filipino, 23-29 years old, 3 men and 1 woman): the experimenter gave the subjects these instructions: "The game paddle is operated by a learning computer and your task is to teach this computer to make the paddle move as you want by using verbal instructions. So please **start teaching this computer**." The aim of this instruction was to specify to subjects that the teaching target was a computer agent.
- Group B (6 subjects, 6 Japanese, 22-26 years old, 4 men and 2 women); the experimenter gave the subjects these instructions: "...So please **start teaching as if talking to someone**." Unlike the instructions for Group A, the aim of this instruction was to specify to the subjects that the teaching target was something other than a simple computer.

Results

To judge whether subjects succeeded in teaching the instruction to the model or not, the same CDV and HV values were used as with the communication experiment, i.e., when the average CDV became more than 0.8, we recognized that the model had succeeded in recognizing the subject's intention. To distinguish whether subjects talked to the paddle naturally or not, an AP ratio was newly employed; an AP ratio is calculated by dividing the number of APs observed by the total number of instructions. In the previous communication experiment, we observed an AP ratio of about 5% for most pairs of subjects; therefore, we assumed that an AP ratio of around 5% meant that the subject talked to the agent naturally. In this experiment, each subject played the game for about 30 minutes and if s/he scored a CDV above an average of 0.8, the experiment was terminated. Table 3 shows the consumed time, AP ratio, types of final instructions, maximum–minimum varieties of instructions and (CDV, HV) values.

Group A All four subjects succeeded in teaching the verbal instructions to the model. However, these subjects did not change the types and varieties of instructions and consistently used only two types of instructions (corresponded to "up" and "down"). Therefore, the model had no opportunities for inducing the subjects' adaptations (e.g., inducing to decrease the types of subjects' instructions) and utilizing these for meaning acquisition processes. Thus, we could confirm that the relationship between them were different from ones between subjects in communication experiment. In addition, they did not use AP sound very frequently to train the meaning acquisition model. This means that they consistently gave the unemotional instruction to the model. And all subjects in Group A reported that they felt great stress while giving the instructions to this model. Additional studies were then required to investigate the relationships among an unemotional speech, a user's stress and an existence of mutual adaptation.

Group B Four out of six subjects succeeded in teaching the verbal instructions to the model. They decreased the varieties of instructions according to the learning modes of the meaning acquisition model, and utilized the AP sound features to make the model focus on current model's action. Moreover, they achieved an AP ration of about 3%, and no subjects reported that they felt any stress during the experiment. Therefore, these subjects used natural instructions compared to the subjects in Group A. Here, we could confirm that the model succeeded in recognizing the subjects' verbal commands by inducing the subjects' adaptations and utilizing these induced adaptations for the meaning acquisition processes. So it can be said that a part of mutual adaptation process existed between the subjects and the model.

As a result, we confirmed that the proposed meaning acquisition model had sufficient competence to recognize the intentions/meanings of everyday users instructions. The model could recognize the given verbal instructions by inducing the users' adaptations and utilizing these induced adaptations, with applying the given AP sound features as a negative reward for the meaning acquisition process. Here, it can be said that the model and the subjects could form a part of mutual adaptation process when appropriate instructions were given to them like Group B. In addition, we gained the insight that an appropriate condition would exist for inducing humans to talk naturally to computers. Although additional studies are required to investigate the effectiveness of this kind of instructions or other conditions (such as agent's appearance, physical entity and so on), the instructions provided to the subjects in Group B would be available as catchphrases for a speech interface system based on this meaning acquisition model that would help users understand intuitively how to use and interact with such an interface.

Discussion and Conclusions

As described in the previous section, the proposed meaning acquisition model and the subjects could form a part of mutual adaptation process when appropriate instructions were given to them. To realize a "true" mutual adaptation with using the acquired a part of mutual adaptation, we should resolve the following issues:

- **Avoiding the model's unnatural reactions** We could observe that subjects sometime felt that the model's reactions were unnatural: e.g., the model could not immediately respond by changing its action in accordance to a sudden strategy change by a teacher (while human subjects in the communication

Table 3: Result of testing experiment with subjects participating in the way everyday users would

Subject	Time (sec)	AP ratio (%)	Final Instructions (upward/downward and others)	max–min varieties of instruction	(CDV, HV)
(Group A)					
a	1425	0.8	Taas/Baba (Tagalog)	2–2	(1.0, 0.6)
b	1690	1.8	UE/SHITA (Japanese)	2–2	(0.9, 0.7)
c	485	0.0	Sony/Aiwa	2–2	(1.0, 0.8)
d	770	0.0	UEUE/SHITASHITA	2–2	(1.0, 0.7)
(Group B)					
e	1351	3.5	UEUE/SHITASHITA	10–2	(1.0, 0.7)
f	1574	0.3	UE/SHITA (toward, a bit, more)	8–5	(0.6, 0.4)
g	1670	3.2	UE/SHITAAA	3–2	(1.0, 0.7)
h	373	6.1	UE/SHITA	5–2	(0.9, 0.5)
i	1569	2.3	UEUE/SHITASHITA (a bit, passed)	9–4	(0.4, 0.4)
J	1166	4.1	UEUE/SHITA	4–2	(1.0, 0.5)

experiment could do). We assumed that the statistical learning algorithm, which is implemented in this model, caused the above phenomenon, and disrupted the subjects' adaptations to this model.

- **Scalability for real world** This model's competence was tested in our "Pong" game environment, so that it is suspicious whether this model can be applied for a multifunctional interface. It is expected that a user will use a wide variety of instructions for such an interface. However this model was constructed based on an assumption that the given instruction indicates certain action, so that the model cannot recognize the meanings of adverbial or evaluation instructions (such as "a bit", "good" and so on).

To resolve these issues by improving this model's capabilities, we expected that a true mutual adaptation process between users and this model could be achieved.

The purpose of this paper is to construct a meaning acquisition model as a basic technology for a mutual adaptive speech interface which can communicate smoothly with an everyday user. We then constructed the meaning acquisition model in consideration of the following assumptions: (1). the model needs to induce a user's adaptation and to utilize this induced adaptation for the meaning acquisition processes, (2). the model focuses on the prosodic information, rather than phoneme information on which most past interface studies focused. As a result, we could confirm that this model could recognize the intentions/meanings of users' verbal commands by inducing users' adaptations and utilizing these for the meaning acquisition process when appropriate instructions were given to them from an experimenter. In addition, AP sound features were utilized as a negative reward in this meaning acquisition process. We expect that this meaning acquisition model could contribute as basic technology to achieving an auto-customization of speech interface or an interface for a pet-robot which can create an intimate relationship with users.

References

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of Royal Statistical Society B*, *39*, 1–28.

Harada, T. E. (2002). Effects of agency and social contexts faced to verbal interface system. *Proceedings of the 19th Japanese Cognitive Science Society Annual Meeting (In Japanese)*, (pp. 14–15).

Komatsu, T., Suzuki, K., Ueda, K., Hiraki, K., & Oka, N. (2002). Mutual adaptive meaning acquisition by paralaguage information: Experimental analysis of communication establishing process. *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 548–553). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pirrehumbert, J. B., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, and M .E . Pollack (Eds.), *Intentions in Communication*, Cambridge, MA: MIT Press.

Scherer, K. R., Bense, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding, *Motivation and Emotion*, *15*, 123–148.

Sears, A. & Shneiderman, B. (1994). Split menus: Effectively using selection frequency to organize menus, *ACM Transactions on Computer-Human Interaction*,*1(1)*, 27–51.

Ueda, K., Endo, M., & Suzuki, H. (2003). Task decomposition: Why do some novice users have difficulties in manipulation the user-interfaces of dairy electronic applications?, *Proceedings of the 10th International Conference on Human-Computer Interaction (HCII2003)*, to appear.