Implementation of Coordinative Nodding Behavior on Spoken Dialogue Systems

Jun-ichi HIRASAWA *Noboru MIYAZAKI* *Mikio NAKANO* *Takeshi KAWABATA*
NTT Basic Research Laboratories
3-1 Morinosato-Wakamiya, Atsugi, Kanagawa 243-0198, JAPAN
http://www.brl.ntt.co.jp/info/dug/
e-mail: jun@idea.brl.ntt.co.jp

ABSTRACT

This paper proposes a mechanism that contributes to the implementation of a spoken dialogue system with which a user can communicate effortlessly. In a dialogue, exchanges between participants promote the establishment of shared information and this leads to effortless communication. This is called "dialogue coordination". In particular, revealing the respondent's internal state, such as through nodding and back-channel feedback, promotes the establishment of shared information. This is called "manifestation", which is one aspect of coordinative behavior, and a mechanism for handling manifestation is introduced. In a human-human dialogue, the listener's manifestative behavior often occurs during a speaker's utterance. However, systems using conventional speech recognition technologies cannot respond during the speaker's utterance. In order to solve this problem, the proposed mechanism, ISTAR protocol transmission, utilizes the intermediate speech recognition results without waiting for the end of the speaker's utterance. This realizes a system with flexible manifestative behavior.

1. INTRODUCTION

A spoken dialogue system is one of the most desirable human-machine communication interfaces. First-time users of such a system may be able to manipulate it without any training, provided they could use it effortlessly as they converse with other humans in everyday life. The key to designing a spoken dialogue system is *effortless*ness of communication.

To date, many spoken dialogue systems have been proposed [1, 7, 11]. They are not necessarily effortless enough in spite of their ability to respond intelligently to some extent. One way of achieving effortlessness is to give the system an animated face. With such a face, one has a feeling of communicating with a human face to face [8, 9]. However, an animated face alone is not quite enough. One of the most important factors in achieving effortlessness is *dialogue coordination* [4]. A dialogue system with coordinative behavior[5, 10] not only provides a humanoid-animated face but also coordinates the communication between the user and machine.

Dialogue coordination [4] is behavior that promotes the establishment of shared information between conversants. In a dialogue, the conversants exchange information and they promote the establishment of shared information through such devices as confirmation and clarification while coordinating the dialogue flow. This promotion of the establishment of shared information is called "dialogue coordination" and it leads to effortless communication.

The final goal of our research is a system with dialogue coordination. The system, unlike most conventional spoken dialogue systems, should be able to flexibly perform coordinative behavior, i.e., should be designed so that a user and the system can work together to establish shared information.

When dialogue participants want to exchange information and promote the establishment of shared information, it is particularly important that the respondent reveals his/her internal state toward the speaker. We call this behavior *manifestation*, which is one aspect of coordinative behavior. Respondent's manifestative behavior expresses not only the content of the speaker's utterance, but also that the recipient is following the dialogue successfully through such devices as nodding and back-channel feedback [12]. This acknowledgment [6] contributes to the establishment of shared information. Our initial goal is to introduce a mechanism for handling manifestation.

In human-machine speech communication, manifestation that reveals the system's internal state can be said to be rather more important than in human-human communication, because errors in speech recognition are unavoidable. In order for a user to communicate with a system effortlessly, the system has to manifest its internal state for successful communication.

In human-human dialogues, the manifestative behavior of a respondent frequently occurs during a speaker's utterance. Since the users might talk to the system as they talk to a human, the manifestative behavior of the system should also be produced during the speaker's utterance. In order for the system to manifest its internal state before the end of the speaker's utterance, it must be able to use intermediate speech recognition results without waiting for

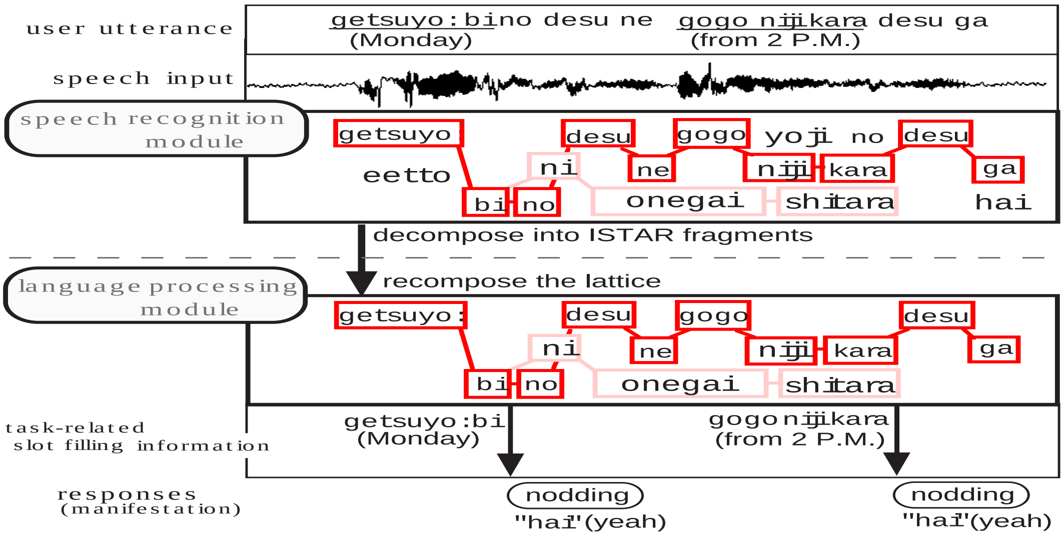

Figure 1: ISTAR protocol transmission. The word lattice is decomposed into a series of ISTAR fragments in the speech recognition module, and is recomposed in the language processing module.

the end of the user's utterance.

In this paper, we propose a mechanism, ISTAR protocol transmission, that enables the utilization of the intermediate speech recognition results, and describe its implementation in a system that achieves manifestative behavior using eye direction and pre-recorded voices and a subsumption architecture.

2. A PROBLEM

In human-human dialogues, responses, such as back-channel feedback or nodding, are given at any appropriate time even during a speaker's utterance. If a system could respond in a similar manner, users would be more comfortable talking to it. The realization of such a system requires an ability to utilize intermediate speech recognition results for manifestation as soon as possible without waiting for the end of the speech interval, i.e., incrementally along the time axis (left to right) [3].

However, all conventional speech recognition technologies except the one in [3] never produce recognition results before detecting the end of the speech interval. This is because conventional methods are designed to produce the best score pass in a word lattice. Though the study by Görz [3] can indeed utilize the intermediate speech recognition results, it is not designed for coordinative manifestation in dialogues.

3. ISTAR PROTOCOL

In order to utilize intermediate speech recognition results before the end of the speaker's utterance for the manifestative behavior, we propose a mechanism using the ISTAR (Incremental Structure Transmitter And Receiver) protocol (Fig.1). ISTAR is a protocol for incrementally transmitting structured information, i.e. a word lattice, from the speech recognizer to the language processor. The functions of the ISTAR protocol are (1) decomposition of the word lattice in the speech recognition module, (2) transmission of the decomposed fragments to another module, and (3) recomposition of the word lattice from the decomposed fragments.

The advantage in using ISTAR is that the subsequent module does not need to wait to begin processing and can utilize the speech recognition results right away, which meets the requirement for the implementation of the manifestative behavior. In addition, with ISTAR, no information on the word lattice structure in the speech recognition module is lost, since the word lattice can be thoroughly recomposed in the language processing module.

The actual process (cf. fig.1) is as follows: The speech recognition module receives speech input and begins the recognition process with the Viterbi algorithm using only frame-synchronous forward search. Whenever the recognizer arrives at the end of a word hypothesis, the recognition module produces an ISTAR fragment. The fragment includes a time code for the word hypothesis, a label for the word hypothesis, its score, and a left-context back pointer to the preceding word fragment. At this time, the word lattice is being decomposed into a series of ISTAR fragments. Because each fragment has its left-context pointer to the preceding fragment, the ISTAR transmission into the other module preserves the structured information in the word lattice.

An example of how the transmitted ISTAR fragments are

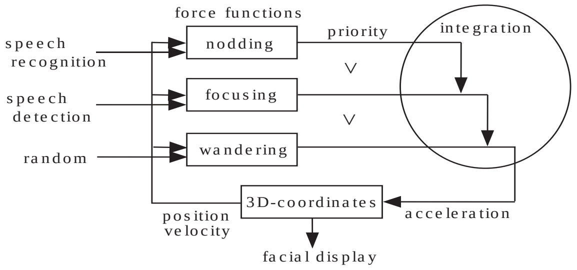

Figure 2: Subsumption architecture. Several force functions are integrated according to their priority value, and the eye direction at the next time is calculated.

used is as follows: The language processing module receives the consecutive ISTAR fragments one after another from the speech recognition module. The latest word hypothesis can calculate its score according to the recomposed word lattice. If the score is high enough for the system to accept what the user has said, the system manifests the successful communication by uttering "yeah" and nodding. When a recognition score is not high enough, the system indicates that communication has failed by uttering "what?".

At present, the ISTAR protocol transmission is applied to the production of word hypotheses [3]. However, the concept of the ISTAR protocol is not limited to the word hypothesis transmission; it is also applicable to phoneme-level transmission. If a speech recognition module produces ISTAR fragments whenever a phoneme hypothesis changes, the system can predict and supplement what a user is going to utter even if the user stops an utterance in the middle of a word.

4. IMPLEMENTED SYSTEM

We implemented a spoken dialogue system with ISTAR protocol transmission in order to obtain manifestative behavior. The system transmits intermediate speech recognition results using the ISTAR protocol, and manifests acceptance or incredulity of information relevant to execution of a given task by its eye direction and pre-recorded voices.

4.1. Subsumption Architecture

The eye motion, such as turning the eyes toward a user (*focusing*), *nodding* or *inclination* of the head, is a way to manifest the internal state of the system. The eye motion is controlled using dynamics-based subsumption architecture [2] (Fig.2) for module maintenance and realtime interaction among modules corresponding to the coordinative behavior. A priority-based control mechanism is adopted for this subsumptive action.

The system has 3D-coordinates describing position θ, velocity $\dot{\theta}$, and acceleration $\ddot{\theta}$, and has multiple force functions f_i. A force function is activated by each event trigger, and the acceleration at the next time frame, $\ddot{\theta}_{t+1}$, is calculated by f_i with current position θ_t and velocity $\dot{\theta}_t$, which determines the eye direction at the next time $t+1$. Each force function has a priority value and when several force functions are activated simultaneously, the accelerations are integrated according to these values such that

$$\ddot{\theta}_{t+1} = \frac{\displaystyle\sum_{active\ i} 2^{-p_i} \cdot f_i(\theta_t, \dot{\theta}_t, t)}{\displaystyle\sum_{active\ i} 2^{-p_i}}$$

where p_i is the force function priority value, and $\sum$ is the summation of the active force functions.

Below are some examples of force functions:

wandering: This force function is always activated with the minimum priority. This behavior means that the system is unaware of or is paying no attention to the user as the eyes wander aimlessly with low critical damping.

focusing: This force function is activated once the system detects speech input from the user. The focusing behavior (the system's face turning toward the user) utilizes rapid critical damping and signifies that the system is paying attention to the user.

nodding: This force function is activated momentarily when the system receives and accepts the information relevant to the execution of a given task.

4.2. System Specifications

Modality. Our system has limited modalities for its input/output channel. The input to the system from a user is only speech. The speech recognition module has a vocabulary set including about 50 words and network grammar. The output to the user has two modalities: eye direction and speech (pre-recorded voices). The eye direction is controlled subsumptively [2], and expresses four kinds of behavior: wandering, focusing, acknowledging nodding, and incredulous head inclination as its manifestation (Fig.3). A couple of dozen words or phrases are pre-recorded, such as "yeah"(used with acknowledging nodding) and "what?"(used with incredulous inclination). In addition, the system can produce some confirmation utterances using the pre-recorded voices.

Manifestation (behavior design). This system was applied to a meeting-room reservation task. This task is a kind of *slot-value filling task*, and content words uttered by a user can imply the user's intention to request a reservation of a meeting room.

In designing manifestative behavior, we have to specify three factors for each manifestation. The first is what to manifest: What is the internal state of the system to be manifested? The second is when to manifest: Manifestation at inappropriate times has a different un-intentional effect in dialogues. When should manifestation be produced?. The last is how to manifest: There are several

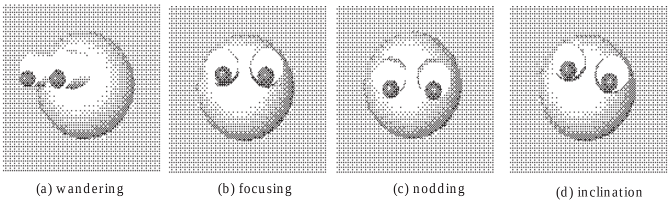

Figure 3: Repertoire of the manifestation. (a) Wandering manifests unawareness of speech input. (b) focusing detection of speech input, (c) nodding while uttering "yeah" acceptance of information related to task execution, and (d) inclining the head while uttering "what?" incredulity of the input from a user.

available forms; for example, back-channel feedback or partial repetition of a speaker's utterance. Which is the most suitable one?

We give one example of the manifestation behavior design. The system manifests (a) its acceptance of the task-related information as its internal state (b) immediately after the recognition of the word representing the information and (c) by nodding and back-channel feedback uttering "yeah". This design of the manifestative behavior was implemented, and the dialogue for this implementation is in `[MOVIE 0158.MOV]` on CD-ROM.

5. CONCLUSION

Aiming for an effortless spoken dialogue system and for implementation of coordinative behavior on the system, we created a mechanism for handling coordinative behavior, particularly for handling the manifestation, such as nodding and back-channel feedback. We showed that ISTAR protocol transmission, the utilization of intermediate results of speech recognition without waiting for the end of a speaker's utterance, makes flexible manifestative behavior possible.

Now that we have created a mechanism by which the system can respond at any time even during the speaker's utterance, the next challenge is an appropriate design for system coordinative behavior in dialogues. The design of system behavior has to answer three questions: (a) What is the system internal state to be manifested? (b) When should the manifestation be produced? (c) What is the most suitable form of the manifestation?

ACKNOWLEDGMENT

We would like to thank Dr. Ken'ichiro ISHII, the executive manager of the Information Science Research Laboratory for his encouragements and comments. We also thank the members of the Dialogue Understanding Research Group for valuable discussions.

References

[1] Allen,J.F., Miller,B.W., Ringger,E.K., and Sikorski,T. "A Robust System for Natural Spoken Dialogue", *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics*, pp.62-70, 1996.

[2] Brooks,R.A. "The Whole Iguana", Brady,M.(ed), *Robotics Science*, MIT Press. pp.432–456, 1989.

[3] Görz,G., Kesseler,M., Spilker,J., and Weber,H. "Research on Architectures for Integrated Speech/ Language Systems in Verbmobil", *Proc. of the 16th International Conf. on Computational Linguistics*, pp.484-489, 1996.

[4] Katagiri,Y. "Dialogue Coordination Functions of Japanese Sentence-Final Particles", *International Symposium on Spoken Dialogue*, pp.145-148, 1993.

[5] Nagao,K. and Takeuchi,A. "Speech Dialogue with Facial Displays: Multi-modal Human-Computer Conversation", *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp.102-109, 1994.

[6] Novick,D.G. and Sutton,G. "An Empirical Model of Acknowledgment for Spoken-Language Systems", *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp.96-101, 1994.

[7] Smith,R.W. and Hipp,D.R. *Spoken Natural Language Dialog Systems*, Oxford Univ. Press, 1994.

[8] Sundblad,O. and Sundblad,Y. "OLGA–a Multimodal Interactive Information Assistant", *Summary of the ACM Conf. on Human Factors in Computing Systems (CHI'98)*, pp.183-184, 1998.

[9] Takebayashi,Y., Tsuboi,H., Kanazawa,H., Sadamoto, Y., Hashimoto,H., and Sinichi,H. "A Real-Time Speech Dialogue System Using Spontaneous Speech Understanding", *IEICE Transactions on Information and Systems*, E76-D, No.11, pp.112–119, 1993.

[10] Thórisson,K.R. "Layered Modular Action Control for Communicative Humanoids", *Computer Animation* , 1997.

[11] Young,S.R., Hauptmann,W.H., Ward,W.H., Smith, E.T., and Werner,P. "High Level Knowledge Sources in Usable Speech Recognition Systems", *Communications of the ACM*, pp.183-194, 1989.

[12] Ward,N. "Using Prosodic Clues to Decide When to Produce Back-channel Utterances", *Proc. of International Conf. on Spoken Language Processing*, pp.1728-1731, 1996.