HUMAN MOTION AND SPEECH ANALYSIS TO CONSTRUCT DECISION MODEL FOR A ROBOT TO END COMMUNICATING WITH A HUMAN

Naoki OTSUKA, Makoto MURAKAMI

Department of Information and Computer Sciences
Toyo University
Saitama, Japan
E-mail: ti050150@cc.eng.toyo.ac.jp, murakami_m@toyonet.toyo.ac.jp

ABSTRACT

The purpose of this paper is to develop a robot that moves independently, communicates with a human, and explicitly extracts information from the human mind that is rarely expressed verbally.

In a spoken dialog system for information collection, it is desirable to continue communicating with the user as long as possible, but not if the user does not wish to communicate. Therefore, the system should be able to terminate the communication before the user starts to object to using it.

In this paper, to enable the construction of a decision model for a system to decide when to stop communicating with a human, we acquired speech and motion data from individuals who were asked many questions by another person. We then analyze their speech and body motion when they do not mind answering the questions, and also when they wish the questioning to cease. From the results, we can identify differences in speech power, length of pauses, speech rate, and body motion.

Keywords: information retrieval, information collection, human interface, communication robot, terminate communication

1. INTRODUCTION

Recently, many researchers have attempted to reuse a variety of information in cyberspace. One of the most typical research fields is information retrieval [1]. Various forms of information that can be reused or retrieved are information represented on the World Wide Web, information acquired by sensors positioned in the real world [2], information captured by sensors on a wearable computer, which is mobile when worn by a human [3], and information collected by sensors on robots, that travel independently or assisted by remote control [4], etc. Impressions of a song, movie or sightseeing location are considered very valuable information that can be reused. If we could obtain individuals' impressions about various songs, movies or sightseeing locations, together with their situations, and personal details such as sex, birthplace, age and so on, it would be possible to guess the mood swings in a human when listening to a song/tune (or watching a movie or sightseeing). This would enable us to suggest an appropriate tune (or movie or sightseeing spot) to an individual whose personal details and situation were similar to those of the individual from whom the appropriate information was acquired. For example, we could recommend the movie "My Sassy Girl" to a Japanese male in his twenties, who was depressed due to a failed romance. These ideas are based on the collaborative filtering of various types of information represented on the World Wide Web or netnews [5].

Although some people voluntarily share their impressions on the World Wide Web or a Blog, most people rarely do so. Furthermore, such information cannot be acquired by passive sensors without communication between people. Our aim is to develop a robot that can move independently, communicate with a human, and explicitly extract valuable information from the human mind that is rarely expressed for reuse as verbal information. Such a system is called a spoken dialog system for information collection.

2. SPOKEN DIALOG SYSTEM FOR INFORMATION COLLECTION

We consider the differences between a spoken dialog system used for collecting information and a system that performs an interactive problem-solving task such as travel planning or traffic routing, which is the most conventional type of spoken dialog system [6]. The user of a system performing a task must initiate communication with the system. On the other hand, if the user of a system collecting information does not want to use it, the system itself must decide whether or not to start communicating with the user. The purpose of a system performing a task is to accomplish the task quickly and accurately, whereas a system collecting information must acquire a vast amount of information and must, therefore, employ a dialog strategy for commending the communication. Moreover, in a system performing a task, communication terminates once the task has been accomplished. In a system collecting information, it is desirable to continue communicating with the user as long as possible, but not if the user does not wish to communicate. Therefore, the system should terminate communication before the user starts to object to using it.

3. CONSTRUCTION OF DECISION MODEL TO TERMINATE COMMUNICATION

In this paper, we deal with the problems encountered in terminating communication. Humans tend to observe the other person to estimate his mood from body language, facial expressions and the tone of speech. A robot also needs to decide whether to end communication based on audio and visual information captured by microphones and video cameras. Therefore, we analyze the relationship between human behavior patterns while talking and the decision as to whether to end communication with the individual or not. Finally, we construct a decision model to enable the system to decide when to end communication with the human. In this section, we describe the method used in constructing the decision model.

First, we acquire speech and motion data from an individual both when he wishes to end communication and when he wishes to continue communicating. Then, we extract a feature vector from the acquired data, and construct a decision model for the system to end communication with a human as a probabilistic model as follows.

Let $\boldsymbol{x}$ be the n-dimensional feature vector extracted from the human speech and motion data; ω_0, the decision class to end communication; and ω_1, the decision class not to terminate communication. We calculate each class-conditional probability density function $P(\boldsymbol{x}|\omega_i)$ from the set of the feature vector $\boldsymbol{x}$ in the corresponding class ω_i. In this paper, we represent the class-conditional pdf $P(\boldsymbol{x}|\omega_i)$ as the Gaussian Mixture Model

$$P(\boldsymbol{x}|\omega_i) = \sum_{k=1}^{M} c_{ik} N(\boldsymbol{x}; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}), \qquad (1)$$

where M is the number of Gaussian components; c_k, the weighted coefficient of component k; $N(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, an n-variate Gaussian pdf; $\boldsymbol{\mu}_k$, the n-dimensional mean vector of the Gaussian component k; and $\boldsymbol{\Sigma}_k$, the $n \times n$ covariance matrix of the component k. The Gaussian Mixture Model is the most general pdf, the parameter estimation method of which has been formulated. We estimate the parameters $c_k, \boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ using the EM Algorithm.

In the recognition process, the likelihood of the decision to end communication $P(\boldsymbol{x}|\omega_0)$ and that of the decision not to end communication $P(\boldsymbol{x}|\omega_1)$ are calculated from the input vector $\boldsymbol{x}$ and Gaussian Mixture Model represented as Eq. (1). The result is then output according to the following decision rule:

$$\text{Decide } \omega_0 \text{ if } P(\boldsymbol{x}|\omega_0) > P(\boldsymbol{x}|\omega_1); \text{ otherwise decide } \omega_1.$$

4. EXPERIMENT

In this paper, we acquire human speech and motion data when an individual wishes both to terminate the communication and to continue communicating. We then analyze the prosody of each speech data, and observe the differences in the various motion data.

Fig. 1: The experimental setup

4.1 Data acquisition

We asked a male college subject many questions for an extended period of time to force him to stop answering the questions. During questioning, the subject and the questioner were not in the same room, and the questions were asked remotely using two PCs connected to a network and with the voice chat software 'Skype' installed. A flash animation of a speaking robot appeared on the subject's monitor, while the voice of the questioner was changed into a robot-like-voice using voice changing software 'MorphVox'. This was an attempt to simulate a situation in which a robot questions a human, because we appreciate that communication between a human and a system is somewhat different from that between humans. There were the usual things on the desk, such as a book, mobile phone, something to drink, etc. and the subject was allowed to use them freely, to create a normal situation. The subject's voice and body gestures were recorded with a headset microphone and video camera. Figure 1 shows the subject's situation while answering the questions.

On completion of the data recording, the subject was asked the following three questions. "Did you wish to stop answering the questions?", "At which point did you wish to stop answering the questions?" "Were there some questions that you did not wish to answer? If so, which were they?".

4.2 Speech analysis

We recorded 40 minutes of the college student's speech data. His answers to the three questions posed after the recording indicate that he wanted to stop answering questions after about 20 minutes, and that he did not wish to answer the questions about his family which were asked after about 28 minutes. Consequently, to illustrate the difference in the data, we used the first 20 minutes of speech data as data showing the subject's willingness to continue the communication, while the last 20 minutes was used as data showing that the system should terminate communication.

Depending on a person's mental state, the prosody of his voice such as loudness, height, intonation, and so on varies. We focus on three kinds of prosody, namely loudness, response time, and speed, to extract the following three prosodic

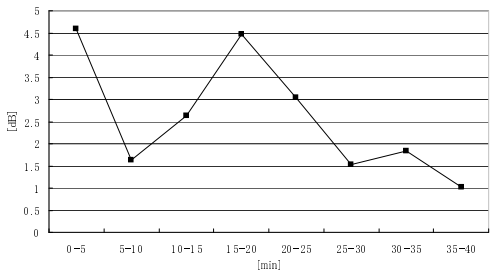

Fig. 2: Variation in the average of the speech power at 5 minute intervals

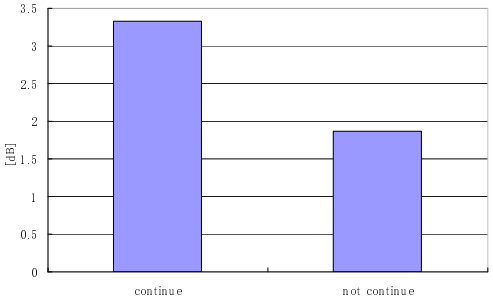

Fig. 3: Difference between the average power of the speech data in the situation in which the system should continue communicating and that in which the system should not.

features: speech power, the length of the pause between the end of the question and the start of the answer, and speech rate which is measured as the ratio of the number of mora to the time length.

Figure 2 shows the variation in the average of the speech power at 5 minute intervals. The average of power for the first 5 minutes and at 15-20 minutes is high, but after 20 minutes it decreases gradually. The power at about 10 minutes is small, perhaps due to the fact that some of the questions at this time were difficult to answer. Figure 3 clearly illustrates the difference in average power of the speech data in the situation in which the system should continue the communication and that in which the system should terminate communication. This indicates that speech power is higher when the user wishes to continue communicating than when he wishes to end the communication.

Figures 4 and 6 show the variation in the average of the pause length and the speech rate at 5 minute intervals, respectively. The average of the pause length first increases gradually, then remains almost constant, and finally increases again. The average of speech rate decreases gradually, then remains almost constant, and finally decreases again. Figures 5 and 7 show, respectively, the differences between the average pause length and speech rate of the speech data in the situation in which the system should continue communicating and that in which the system should not. These results indicate that the pause length is longer and the speech rate slower when the user wishes to continue the communication than when he wishes to terminate the communication.

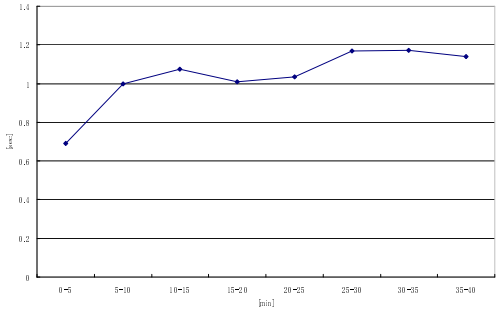

Fig. 4: Variation in the average of the pause length at 5 minute intervals

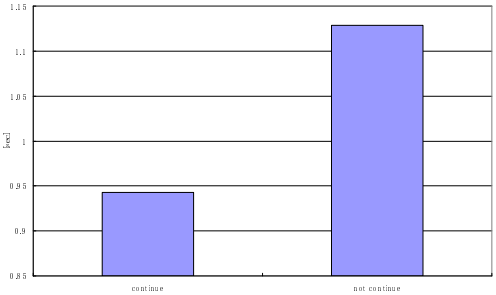

Fig. 5: Difference between the pause length of the speech data in the situation in which the system should continue the communication and that in which the system should not.

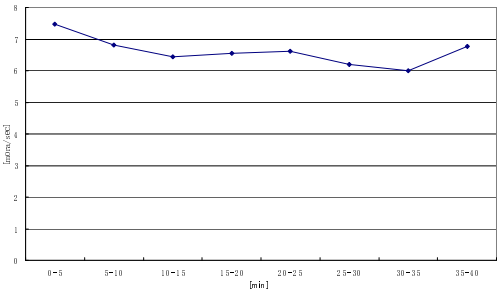

Fig. 6: Variation in the average of the speech rate at 5 minute intervals

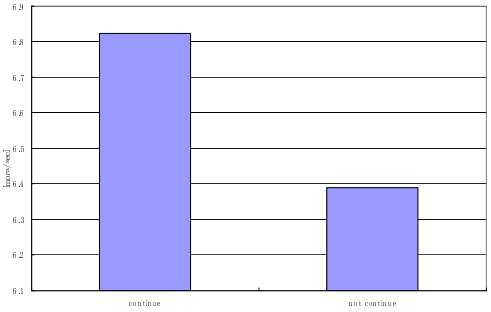

Fig. 7: Difference between the speech rate of the speech data in the situation in which the system should continue the communication and that in which the system should not.

Fig. 8: Example of body motion indicating that the system can continue the communication

Fig. 9: Example of body motion indicating that the system should terminate communication

In addition, we tested the significance of the differences in speech power, pause length, and speech rate. The results show that all the null hypotheses are rejected in a 5% rejection region. Therefore, the differences in these features are significant.

4.3 Motion analysis

Using two video cameras, one placed in front of him and another diagonally behind him, we recorded another college student to capture his facial expressions, pose, and body motion. The video data is 50 minutes long. The subject's answers to the questions asked after the recording show that he wanted to stop answering questions after about 30 minutes. Consequently, to illustrate the difference in the data, we use the first 30 minutes of video data as data showing the situation in which the system should continue the communication, and the data after 30 minutes as sample data showing that the system should end the communication.

Figure 8 depicts an example of motion showing that the system should continue the communication, while Figures 9 and 10 show two examples of motion confirming that the system should terminate the communication.

Minimal motion appears in the situation in which the system should continue communicating, while motion such as stretching, reading a book, or shoulder turns is indicative of the situation in which the system should terminate the communication.

5. CONCLUSION

To construct the decision model to enable a system to stop communicating with a human, we acquired the speech and motion data of individuals who were asked many questions by a third party. We then analyzed their speech and body motion when they were not averse to answering the questions and when they wanted to stop answering questions. The results show the differences in speech power, pause length, speech rate, and body motion.

We will analyze other prosodic features such as pitch, intonation, and so on. Then we will attach colored markers on the joints of a subject's body, record the subject with multiple video cameras, calculate the 3D-position of the colored markers, and analyse the acquired 3D motion data. Finally, we will construct the decision model for a system to decide whether or not to terminate communication with a human as a Gaussian Mixture Model.

Fig. 10: A further example of body motion indicating that the system should terminate communication

6. REFERENCES

[1] A. Singhal, "Modern information retrieval: a brief overview," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, vol. 24, no. 4, pp. 35–43, 2001.

[2] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D. Culler, and K. Pister, "System architecture directions for networked sensors," in *Proc. of ASPLOS*, 2000, pp. 93–104.

[3] B.J. Rhodes, "The wearable remembrance agent: a system for augmented memory," in *Proc. of ISWC97*, 1997, pp. 123–128.

[4] J. Meguro, J. Takiguchi, Y. Amano, and T. Hashizume, "3D reconstruction using multibaseline omnidirectional motion stereo based on GPS/dead-reckoning compound navigation system," *Int. J. of Robotics Research*, vol. 26, no. 6, pp. 625–636, 2007.

[5] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl, "GroupLens: an open architecture for collaborative filtering of netnews," in *Proc. of CSCW*, 1994, pp. 175–186.

[6] R. Cole, L. Hirschman, and L. Atlas, et al., "The challenge of spoken language systems: research directions for the nineties," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 1–21, 1995.