

# A Real-Time System for Summarizing Human-Human Spontaneous Spoken Dialogues

Megumi Kameyama<sup>1</sup>, Goh Kawai<sup>1\*</sup>, Isao Arima<sup>2</sup>

<sup>1</sup>SRI International, <sup>2</sup>NTT Data Communications Systems Corp.

## Abstract

We have built a prototype Automatic Dialogue Summarizer (ADS) — a real-time system that automatically generates simple summaries of completely spontaneous human-human spoken dialogues without the machine interrupting the natural flow of conversation. Two dialogue participants (client and clerk) discuss conference room reservations (CRR) in Japanese, and the system dynamically updates summaries of what rooms were reserved or canceled for what times and by whom. This paper describes the system's architecture, its component technologies, and its performance. We discuss the robustness, efficiency, and effectiveness of the system, and the use of a spontaneous dialogue corpus for development and testing.

## 1 Introduction

We have built a system that generates summaries of spontaneous spoken dialogues. Dialogue participants discuss the reservations of conference rooms in Japanese. Summaries produced by the system indicate what rooms were reserved or canceled, for what times, and by whom. This paper describes the system's architecture, its component technologies, and its performance. We discuss how we achieved the system's robustness, efficiency, and effectiveness motivated by careful analyses of spontaneous dialogues. We conclude with a discussion of future prospects.

## 2 The Problem

The problem is to realize a system that automatically summarizes the key contents of completely spontaneous human-human spoken dialogues without the machine interrupting the natural flow of conversation. In the present case of CRR dialogues, the

system must constantly update and revise the current CRR summary state in the course of a dialogue. We define a CRR summary state to be a set of Reservation template objects with slots for who reserved or cancelled which room, for what time, on what date, and so on. A single dialogue may achieve multiple reservations or cancellations, corresponding to multiple Reservation template objects.

Such a system must overcome the known difficulties in discourse processing plus the complications of spontaneous spoken dialogues [5]. In discourse processing, as opposed to sentence processing, the system must deal with the context dependency of interpretation and dynamic updating of context. The complications of spontaneous spoken dialogues include overlapping utterances, misunderstandings, and miscommunications (coordination problems); denied requests, unstable agreements, and challenged assumptions (negotiation problems); false starts, filled pauses, repairs, ungrammaticality, and mispronunciation (disfluency problems). Moreover, the system speed must keep up with the natural dialogue speed.

By far the most outstanding difficulty in understanding task-oriented spontaneous spoken dialogues is the "topic determination" problem [6]. It is in general difficult to determine *how many* instances of the given task are being discussed in a dialogue, causing the system to recognize too many or too few task instances. When the dialogue concerns multiple task instances, it is also difficult to determine *which* task instance(s) a given utterance concerns. The overall accuracy of the system depends on the accuracy of this topic determination [7].

## 3 Architecture

The system is called the Automatic Dialogue Summarizer (ADS). Its user interface is shown in Figure 1. Two users (a clerk and a client), each with a microphone connected to a workstation, discuss

---

\*Presently at Tokyo University.

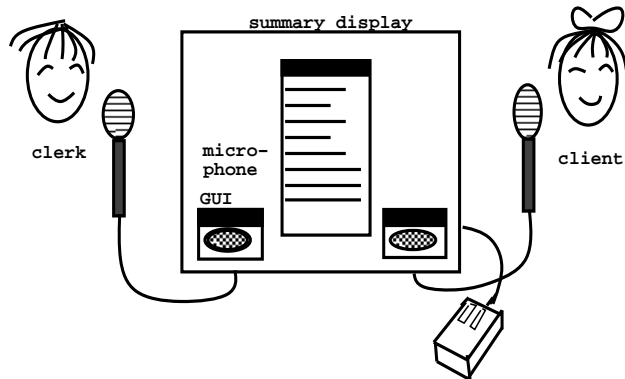


Figure 1: System User Interface

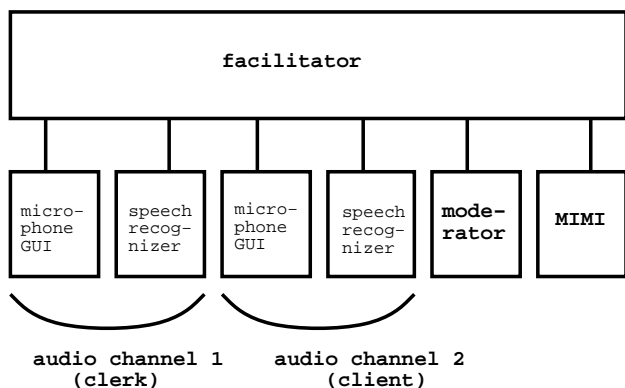


Figure 2: System Integration Using Open Agent Architecture

reservations or cancellations of conference rooms. A mouse-driven graphical user interface (GUI) is used to signal start of speech. Speech end detection is automatic. Both users can be speaking at the same time. A window displays speech recognition hypotheses and summarization results.

The ADS is integrated using SRI’s Open Agent Architecture (OAA) [1]. The OAA uses a distributed architecture in which a facilitator module is responsible for scheduling and maintaining the flow of communication among the client modules. Figure 2 shows the modules used in the system. Two channels for audio input allow for two users. Each user has a mouse-driven microphone GUI module to signal start of speech, which is used by the moderator module to sort the order of utterances. Speech is sent to a speech recognizer module, one for each user, and recognition hypotheses from both users are sent to the MIMI module (See subsection 3.2), which generates the summaries. The individual modules are described in more detail below.

### 3.1 Microphone GUI, Speech Recognizer, and Moderator

These three modules together accept, recognize, and order incoming speech.

The microphone GUI is a mouse-drive interface clicked by each user who starts talking. This signifies the start of a turn and is used to sort the order of utterances. Computer file timestamps cannot determine turn-taking behavior because they tell us only when an utterance ended.

The speech recognizer is a Japanese version of SRI’s Decipher<sup>TM</sup> continuous-speech, speaker-independent recognizer [2]. For real-time response, a four-feature discrete-density hidden Markov model (HMM) system with vocabulary fixed for the CRR task was used. The models reported here were trained on a read-speech database (2057-word vocabulary, 18355 utterances total, about 200 different speakers). A bigram language model [4] was used to approximate word connections expected in spontaneous speech.

Ideally, the recognizer should have been trained on the same spontaneous dialogue corpus used for developing the dialogue summarizer (See subsection 3.2). This was not feasible, however, because the dialogue transcription either ignored or unsystematically transcribed disfluencies and nonlinguistic articulations. We need to establish a systematic transcription method for these spontaneous speech characteristics.

The recognizer’s word error rate in a blind test of 5864 utterances (59076 words) of *read speech* was 16.86%. The test set perplexity was 17.5. Note that the recognizer was both trained and tested with read speech, which lacks a number of spontaneous-speech characteristics the ADS must deal with. Problems with this discrepancy are discussed in subsection 4.1.

The moderator synchronizes speech recognition hypotheses coming from both users, and translates it into a MIMI-readable form. Utterances are ordered based on which speaker started an utterance first, not on when speech terminated or when the speech recognizer finished processing the input. This solves the problem of overlapped utterances. Synchronized hypotheses are sent to MIMI by the facilitator.

### 3.2 MIMI: Dialogue Summarizer

MIMI (for “ears” in Japanese) automatically generates summaries from transcripts of CRR dialogues. MIMI is based on the FASTUS<sup>TM</sup> system for extracting key domain information from written texts [3]. FASTUS is a cascaded finite-state

transducer that processes each sentence through four phases that recognize words, basic phrases, complex phrases, and domain patterns. Each sentence loop ends with merging of the domain information. FASTUS recognizes only domain-relevant information, ignores unknown or irrelevant input, and merges redundant information — in a speed much faster than a human’s reading.

MIMI’s training and testing were based on the 150–dialogue CRR corpus described in Section 4. Its development guideline was minimalistic, whereby simpler approaches were made more complex only when necessitated by the dialogues [5, 6].

For real-time summarization of spontaneous spoken dialogues, MIMI made three major extensions to FASTUS: (1) flexible recognition of sentence/utterance units, (2) utterance lookahead, and (3) information override. (1) overcomes the lack of clear “sentence” boundaries in spontaneous speech. (2) is motivated by the fact that some linguistic units are made discontinuous by hesitations and the other speaker’s utterances. MIMI looks one utterance ahead before updating the summary with the current input, resulting in a one-utterance lag in the summary state display. (3) is motivated by the negotiation feature of spontaneous dialogues. This extends FASTUS’s incremental merging of incoming information.

As a result, in addition to recognizing only relevant information, being robust with errorful speech, and merging redundant information, MIMI flexibly adjusts to natural negotiative conversations with diverse utterances — in a speed much faster than a human’s speaking.<sup>1</sup>

## 4 Performance Evaluation

We collected 150 spontaneous human–human Japanese dialogues reserving or canceling conference rooms. The dialogues involved 50 different client speakers and one clerk speaker. Client and clerk’s utterances were recorded on separate audio channels and merged into a sequence of utterances based on turn taking or extended pauses. One-third of the dialogues involved multiple reservations or cancellations, while the rest dealt with a single reservation or cancellation. All dialogues were transcribed, and correct summaries were generated by hand. Half of the dialogues were used for developing and training MIMI, and the other half for blind testing.

---

<sup>1</sup>Kameyama [7] reports an in-depth analysis of MIMI’s system features and evaluation results.

The following results were obtained from a performance evaluation with the 75–dialogue blind test set.

### 4.1 Speech Recognizer

The speech recognizer was evaluated by processing the recorded spontaneous speech dialogues. All 150 dialogues were new to the recognizer, but we used only the 75 dialogues in MIMI’s blind test set, which consisted of 6250 utterances (41383 words) in total. The recognizer in this test faced a severe mismatch between the training and testing conditions, since the test data was spontaneous speech and all of the training data was read speech. The spontaneous dialogues differed from the read training corpus in number of disfluencies, sentence constructions, speaking rate, and word pronunciations. In addition, 29% (12013/41383) of the words were out of vocabulary. Eliminating all utterances from the test set with at least one out-of-vocabulary word resulted in a test set of 2114 utterances (3723 words). Due to a poor match to the language model which was trained on read speech (of which 34% were digits), the perplexity for the 2114 utterance test set was 384.<sup>2</sup>

Since there were inconsistencies in transcription conventions between read training and spontaneous test utterances, it was decided that percentage of words correct was a more appropriate measure of performance than word error rate. For the 2114 utterance test set, 50.2% of the words were correctly recognized.

It is known that state-of-the-art recognizers perform poorly with spontaneous speech, e.g., 50% word error rate for the spontaneous Switchboard conversations [8]. In addition, a study at SRI showed that performance improved from a 52.6% word error rate to 28.8% when subjects were instructed to read the verbatim transcripts from their spontaneous conversations. Modeling spontaneous speech phenomena such as casual speaking styles, words with omitted phones and syllables, and speech disfluencies is an active research area.

### 4.2 MIMI

The MIMI system was evaluated by directly processing the dialogue transcripts and scoring the summary output against human-generated summaries.

---

<sup>2</sup>This perplexity is similar to that observed when a language model was trained on read Wall Street Journal text and tested on spontaneous Switchboard conversations.

MIMI achieved 77.7% recall and 82.5% precision.<sup>3</sup> This performance achieves the level of a mature information extraction system.

An error-simulation experiment [7] showed that MIMI's performance degrades *linearly* with a steady increase in the input word error rate. We took the 75 dialogue transcriptions in the blind test set, and simulated word substitution errors at the rate ranging from 5% to 55%. We found that up through a 55% error rate, MIMI's performance decline linearly, with slopes ranging from -0.59 to -1.08. Since there is no error-rate threshold after which MIMI's performance suddenly drops, this result indicates MIMI's robustness against errorful input.

### 4.3 The Overall ADS

The ADS's overall performance was evaluated by feeding the recognizer output into MIMI, and scoring MIMI's summary output. The result was 52.0% recall and 54.8% precision. MIMI's performance thus dropped by 33.1% in recall and 33.6% in precision when at least about half of the input words was incorrectly recognized. This result confirms MIMI's robustness against errorful input. In fact, it seems that the information extraction approach to the natural language processing component of the ADS alleviates the full impact of word recognition errors. The content redundancy of typical spontaneous dialogues may be a factor. We would like to come to a full understanding of this observed fact in the future.

## 5 Discussion and Conclusion

A prototype Automatic Dialogue Summarizer demonstrates the benefit of integrating heterogeneous component technologies into a practical application, the effectiveness of data-driven engineering, and the robustness of the dialogue summarization component, MIMI, against errorful input.

How can we improve the overall performance of the ADS? One way is to improve recognition accuracy of content words while placing less emphasis on recognition of certain disfluencies and nonlinguistic articulations. In addition to the increase of training data, a more sophisticated language modeling in terms of grammar and dialogue models would be desirable. Another way is to enrich MIMI's contextual

representation and reasoning to overcome the topic determination problem.

One recurring question in this work is whether and how a domain-independent ADS is feasible. This must be investigated together with the pursuit for open-domain information extraction technologies in natural language processing.

## Acknowledgment

This work was sponsored by NTT DATA. The authors would like to thank Ray Perrault for his insightful guidance, and Adam Cheyer, Katsufumi Fukunishi, Nobuo Koizumi, Ken'ya Murakami, Patti Price, Otoyoshi Shirotsuka, and Kelsey Taussig for their assistance.

## References

- [1] Cohen, P.R., Cheyer, A., Wang, M., and Baeg, S.C., 1994. An Open Agent Architecture. In *Proc. AAAI 1994 Spring Symposium*. Stanford, CA, 1-8.
- [2] Cohen, M. et al., 1990. The Decipher Speech Recognition System. In *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP-90)*, Albuquerque, AZ, 77-79.
- [3] Hobbs, J., Appelt D., Bear J., Israel D., Kameyama M., Stickel M., and Tyson M., 1996. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In E. Roche and Y. Schabes, eds., *Finite State Devices for Natural Language Processing*, MIT Press, Cambridge, Massachusetts.
- [4] Jelinek, F. 1990. Self-organized Language Modeling for Speech Recognition. In Waibel, A. and K. Lee, eds., *Readings in Speech Recognition*, Morgan Kaufmann, San Mateo, CA, 450-506.
- [5] Kameyama, M. and Arima, I., 1993. A Minimalist Approach to Information Extraction from Spoken Dialogues. In *Proc. International Symposium on Spoken Dialogue (ISSD-93)*, Waseda University, Tokyo, Japan, 137-140.
- [6] Kameyama, M. and Arima, I., 1994. Coping with Aboutness Complexity in Information Extraction from Spoken Dialogues. In *Proc. International Conference on Spoken Language Processing (ICSLP-94)*, Yokohama, Japan, 87-90.
- [7] Kameyama, M. in preparation. Information Extraction from Spontaneous Spoken Dialogues. Manuscript. SRI International Artificial Intelligence Center.
- [8] Stolcke, A. and E. Shriberg. 1996. Statistical Language Modeling for Speech Disfluencies. In *Proc. IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP-96)*, Atlanta, GA.

<sup>3</sup> *Recall* is the number of answers the system got right divided by the number of possible right answers. It measures how comprehensive the system is. *Precision* is the number of answers the system got right divided by the number of answers the system gave. It measures the system's accuracy.