JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 18, 627-636 (2002)

Short Paper

A Portable Auto Attendant System With Sophisticated Dialog Structure

JHING-FA WANG[*+] AND HSIEN-CHANG WANG[+]
[+]*Department of Information Engineering*
[*]*Department of Electrical Engineering*
National Cheng Kung University
Tainan, 701 Taiwan
E-mail: {wangjf, wangsj}@csie.ncku.edu.tw

An attendant system connects the caller to the party he/she wants to talk to. Traditional systems require the caller to know the full name of the party. If the caller forgets the name, the system fails to provide service for the caller.

In this paper we propose a portable Auto Attendant System (AAS) with sophisticated dialog structure that gives a caller more flexibility while calling. The caller may interact with the system to request a phone number by providing just a work area, specialty, surname, or title, etc. If the party is absent, the system may provide extra information such as where he went, when he will be back, and what he is doing. The system is built modularly, with components such as speech recognizer, language model, dialog manager and text-to-speech that can be replaced if necessary. By simply changing the personnel record database, the system can easily be ported to other companies. The sophisticated dialog manager applies many strategies to allow natural interaction between user and system. Functions such as fuzzy request, user repairing, and extra information query, which are not provided by other systems, are integrated into our system.

Experimental results and comparisons to other systems show that our approach provides a more user friendly and natural interaction for auto attendant system.

Keywords: auto attendant system, fuzzy request, clear request, semantic frame, dialog manager

1. INTRODUCTION

In this decade, spoken dialog systems have been widely researched. The goal is to let a machine interact with a user to provide service. Applications such as railway ticket reservation [5], train timetable information inquiry [11], automobile purchasing guide [4], restaurant information system [6], weather information system [8], and many other applications [12-18] have been presented to the public and demonstrate an ability to serve people.

Our system is defined as an auto attendant system that can interact with a caller using spoken language. When using an attendant system, the caller first inquires the

Received May 2, 2000; revised February 5 & March 26 & July 6, 2001; accepted August 22, 2001.
Communicated by Wen-Lian Hsu.

party's phone number, and then the system builds a connection between the caller and the party or provides more information about the party (such as where he goes, when he will be back, etc.) if he is absent. From observation of real dialogs between the operator and caller, we may classify the inquiry into two types. One is *Clear request*, which means that the caller provides enough information for database access. The other is *fuzzy request* which means the caller provides only partial or vague information. An example of each is shown below.

Clear request: "Please tell me the phone number of Chang San in the department of computer science and information engineering".

Fuzzy request: "I want to know the phone number of the professor who specializes in speech coding".

Traditional AAS can deal with only clear inquiry. For example, Herman and Alex [1] reported on their ASR system which requires a caller to spell out the name of the party. Seide and Keller [7] proposed an automated directory information system which limits the caller to say the name of the person he wants. Neubert et al. also reported their result in building a directory name retrieval system [17] which lets a user say the name of the desire party. Many other AAS systems [12, 13, 15, 16, 18] also use a clear inquiry strategy to reduce their complexity. However, in many cases, if the caller forgets the full name of the party or can only remembered his specialty or title, an AAS adopting clear inquiry strategy will fail to complete the task. In this paper we propose a new system that takes into account special information to handle fuzzy request problems. Information such as specialty, title, and sex of the party is built in our system database. If a caller performs a fuzzy request, our system will ask him for extra information until the party is determined. Providing a fuzzy request function in our system makes it more user friendly.

The proposed system is designed modularly with five modules as described below.

Language modeling: to build a statistical language model needed for speech recognition.
Telephone speech recognition: to transfer input speech into keyword candidates.
Key-feature matching: to determine the features and generate the semantic frame.
Dialog manager: to decide how and what to respond.
Text-to-speech output: to output speech to the caller.

Each module can be tested, updated, or replaced independently, which makes our system more testable and configuriable. Our system aims to be practically used in many companies or organizations for office telephone-transfer automation. In order to easily port our system from one company/organization to another, we have collected many dialog structures from different companies to establish a general dialog-structure database. Armed with this special information, a new company only needs to provide the personnel database to port the system.

These five modules are introduced in the following sections. Section 2 describes recognition of telephone speech. Section 3 explains how to determine the semantics of the input. Section 4 is about the sophisticated dialog manager and strategies applied in this system. The experiments and results are described in section 5. Finally, we give the conclusion in section 6.

2. TELEPHONE SPEECH RECOGNIZER

2.1 Telephone Speech Recognition With Keyword Verification

User speech conveyed by telephone network is first recognized by the keyword spotting system developed in our laboratory [10]. As the previous results show, a 74.8% recognition rate for top-1 keyword candidate using a 2,583 words lexicon without applying any language model was achieved. The spotter uses an HMM model with three consonant-states and five vowel-states. The features of each frame are 26 dimensional vectors with 12-th order MFCC cepstrum, 12-th order delta MFCC cepstrum, delta energy, and delta-delta energy coefficients. The training speech is derived from the MAT telephone speech database, which contains telephone speech of 400 people [9]. The speech input is recognized as a keyword lattice by the keyword spotter, and a verification step is performed to find the most likely keywords in order to reject meaningless parts of the input such as "Mm", "Ah", cough, hesitation, etc. The result of the keyword spotter then forms a top-N keyword lattice which has the highest acoustic scores corresponding to the input speech.

Since a user tends to obey the system prompt, what the user will say in different rounds of the dialog is mostly predictable. Thus, we can apply different keyword sets for different dialog rounds; this is called the *dynamic keyword set* approach. Based on this approach, the keyword spotting system can reduce the searching space and improve the recognition rate. The recognition rate with and without a dynamic keyword set is compared in section 5.

3. KEY FEATURES AND SEMANTICS

3.1 Key Features

To form a semantic frame, the generated keyword sequence is then matched with the personnel record database using the key-features. Each personnel record in our system contains ten *key-features*, they are: 1) surname, 2) first name, 3) title, 4) sex, 5) department, 6) specialty, 7) phone number, 8) place, 9) time, and 10) event about the person. Note that the 8-th, 9-th, and 10-th key features describe where the party went, when he will be back, and what he is doing, respectively. These key features are stored in a definable personnel database. Three particular fields of the personnel database (place, time, and event) are defined for representing the status of the person if he is away from the phone. A caller can inquire about this information when the party is away. Providing this kind of extra information greatly increases the friendliness of our system.

3.2 Synonymous Key Features Matching

From all of the personnel records, those which most closely match the queried features are chosen as candidates. We propose the key feature matching approach which takes into account the synonyms of the input. For example, if the input contains a string "teacher Chang", then the personnel records having surname "Chang" and title "professor" or "doctor" are also matched to the input. Note that *professor* and *associate profes-*

sor are synonyms since people usually confuse them. Several synonyms defined in our system are listed below.

Title: teacher, doctor, professor, associate professor, …
Sex: female, Ms., Mrs., lady, …
Specialty: speech coding, speech compression, audio compression, …
Title: manager, administrator, …

With these synonyms defined, our system is able to deal with a more natural inquiry.

3.3 Forming the Semantic Frame

Every input has a corresponding speech act type (SAT) that represents the meaning or intention of an input [3]. In order to determine what action should be performed for the input, the result of the key-feature matching and the speech act type must be decided first. We use the semantic frame to keep such information. A semantic frame contains a speech act type plus ten semantic slots corresponding to the 10 key-features in the personnel record. The matched features are entered into the semantic slots.

To determine the SAT of the input, we use the same technique as proposed in our previous work [2], which is that we check if the input contains a certain pattern of SATs. Six SATs are defined in this domain, they are: 1) Inquiry for information, 2) Refinement, 3) Repair, 4) Reconfirm, 5) Greeting, and 6) Un-decidable. The first SAT, inquiry for information, is further divided into four sub-SATs which are: i) Ask for phone number, ii) ask location of the party, iii) ask comeback time of the party, and ix) ask current event of the party.

Fig. 1 shows the semantic frame of an example request. In our proposed system, the last three features (Place, Time, Event) in Fig. 1 are not always *NULL*. The caller may ask about where the party went, or when will he come back, etc., thus we also include these features in the semantic frame. For example, if the caller were informed that the party is away, he may ask, “Where is he off to?” In this case the semantic frame “*place*” is filled. The purpose is to provide more information about the party if the caller wants to know. In the following sections, we describe the sophisticated dialog manager and the strategies of response generation.

4. DIALOG MANAGER

Because of the uncertainty of speech recognition and the multiple results of database access, it is necessary to interact with the caller to find the phone number. We propose a dialog manager to control the interaction between user and system. In order to make the dialog go smoothly, our system has several strategies for interacting with the caller.

4.1 Dialog Strategies

The dialog strategies for the proposed dialog manager are listed below. Note that in our examples, “S” stands for system and “U” for user.

Input = *I would like to talk to Professor Wang in the Department of Electrical Engineering*.
SAT = *Inquiry for information (ask for phone number)* Surname: *Wang* Name: *NULL* Title: *Professor* Sex: *NULL* Department: *Electrical Engineering* Specialty: *NULL* Phone #: *NULL* Place: *NULL* Time: *NULL* Event: *NULL*

Fig. 1. Example of inquiry and its semantic frame.

***Mixed-initiated strategy*:** Differs from the system-initiated strategy. We use the mixed-initiated dialog approach. In a system-initiated system, the caller is limited to answering the question asked by the system. In a mixed-initiated system, the caller is allowed to ask any question spontaneously.

***Reconfirming strategy*:** To make sure that all the generated results are correct, system asks the caller to reconfirm them in the final round of the dialog.

***Repair strategy*:** If the caller finds that the system does not recognize his intention correctly, he may repair it at the next round of the dialog.

***Simple strategy*:** To simplify the dialog, we allow the caller to inquire only one phone number at a time. In addition, we do not ask the user to answer more than two questions in each round.

***Distinguishing caller and party*:** When calling an attendant, the user usually introduces himself to the operator. This kind of input should be carefully handled, or else the caller's intention may be misinterpreted. In the example "*I am Wang Hsien-Chang, I want to speak to Lee Chian*", two names are recognized. We know Wang Hsien-Chang is the caller since it is attached with the phrase "I am". Thus our system can provide a correct response to the user. If a system cannot distinguish the caller and the party, the response may look like "*System recognized two persons, Mr. Wang Hsien-Chang and professor Lee Chian, which one do you want?*"

4.2 Dialog Response Generation

The response of dialog manager depends on the semantics frame derived from input speech. The proper rule for each situation is applied to generate the relative responding sentence. We have six kinds of responses as described below.

Greeting response: In the beginning and end of a dialog, a greeting response such as *"How may I help you", "Good bye"* are used.

Refinement response: If there are more than three matched records, our system requires the caller to provide more information.

Reconfirming response: When the system has produced the phone numbers for the inquiry, it asks the caller to confirm it, such as *"Professor Wang Jhing-Fa, Phone number 62111, is that correct?"*

Repair response: If the caller corrects the response of the system, the old semantic slots are modified and new key-features are entered into the semantic frame. The dialog continues with the new semantics.

Pardon response: If the system cannot make a decision about a word sequence, or if the information provided by the caller conflicts with the database records, it will prompt to the caller to reply again.

Failed response: If the system cannot properly proceed with the input for more than three tries, we declare that it is a failed dialog, and the caller is switched to a human operator.

5. EXPERIMENTS AND RESULTS

Our experimental environment consists of a Pentium-II PC with a 4-port Dialogic card running the Microsoft NT operating system. The personnel database contains 200 records and the response time for each input is about the same as the input duration. The experiment is carried out with 100 callers (65 males and 35 females in our school.) They are asked to make more than one phone call during the experiment. In this way we have 272 dialogs for testing our system. The average age of the callers is 24 years old. Most callers are cooperative, which means they usually obey request made by the system.

The experiments contain two parts, the semantics accuracy and the dialog accuracy. With the aid of the dialog manager, although the speech recognition accuracy is not high enough, the performance of this auto attendant system is not bad. Most testers indicate that they were satisfied with the operation of our research.

5.1 Experiment on Semantic Accuracy

The speech recognizer is a keyword spotter, which has been reported by our colleagues in [10]. It showed a 74.8% recognition rate for top-1 candidate using 2,583 words as its lexicon without applying any language mode. In this work the speech recognition accuracy was tested for each semantic slot and SAT. The results are shown in the upper part of Table 1. The correct rates are measured for each semantic slot with and without applying the dynamic keyword set approach to test whether it works. Without applying the dynamic keyword set, the total lexicon size is 823 in this work. For the case with dynamic keyword set, since keyword size is reduced each time a semantic slot is determined, the average lexicon size is about 614. The average rates of correct speech recognition over semantic slots with and without applying dynamic keyword set approach are 88.8% and 93.2 %, respectively. The dynamic keyword set approach improves

the accuracy by 4.4%. Note that since many Chinese surnames are similar sounding, the speech recognition system often confused one name with another, causing a significant number of name recognition errors.

For each input, if the semantic slots and the SAT are correctly entered into the semantics frame, then it is an accurate processing of semantics. There were 652 inputs in the 272 test dialogs. 594 of their semantics were determined correctly, giving a semantic accuracy of 91.1% in this experiment, as shown in the bottom row of Table 1.

Table 1. Recognition rate of semantic slots and semantics frame.

Type		Correct rate	
		Without dynamic keyword set	**With dynamic keyword set**
Semantic slots	Surname	85.2%	88.5%
	Name	91.6%	96.7 %
	Title	91.3%	95.5 %
	Sex	84.8%	90.2 %
	Department	90.2%	93.3 %
	Specialty	84.9%	89.2 %
	Where	92.5%	98.6 %
	When	88.6%	92.5 %
	What	90.2%	94.6 %
Average of semantic slots		88.81%	93.2 %
SAT: Speech Act Type		85.7%	92.8 %
Semantics Frame		84.0%	91.1 %

5.2 Experiment on Dialog Accuracy

If the caller can finally know the desired phone number from our system within three repairs, then it is a successful dialog. Among 272 test dialogs, 185 are clear inquiries and 87 are fuzzy. Among the clear request dialogs, 172 were successful and 13 failed. For fuzzy requests, 78 are successful and nine failed. The average dialog success rate is 91.3%. The dialog accuracies are shown in Table 2.

Table 2. Results of the dialog success rate.

	Clear requests	**Fuzzy requests**	**Total requests**
Number of dialogs	185	87	272
Number of successful dialogs	172	78	250
Success rate	92.9 %	89.6 %	91.3%

5.3 Comparison of Different Auto Attendant Systems

Table 3 briefly compares our system with similar systems. Although telephone-based spoken dialog systems have wide-spread applications, this comparison fo-

cuses on auto attendant systems and voice dialing systems. Note that vocabulary size is small if the lexicon size is under 200 words, and medium if it is between 200 to 5,000 words. As illustrated in Table 3, our system provides intelligent ability for a caller to acquire extra information about a party when he is away, such as where he goes, when he will be back, and what he is doing.

Table 3. Comparison of different auto attendant systems.

Country, Institute	Vocabulary size	Accuracy	Portability	Mixed initiated	Dialog Repair strategy	Handling Fuzzy request	Providing extra info.	Distinguish-ing caller or party
USA, NYNEX [15]	Small	90.0 %	✓	×	N/A	×	×	×
UK, Vocalis [18]	Small	96.4 %	✓	×	N/A	×	×	×
USA, AT&T [16]	Small	96.1 %	×	×	×	×	×	×
Europe, EUROSCOM [12]	Medium	N/A	✓	✓	N/A	×	×	×
Japan, NTT [13]	Medium	N/A	✓	✓	✓	×	×	×
France, ENST [17]	Medium	82.0 %	✓	×	✓	×	×	×
Germany, Philips [14]	Medium	81.0 %	✓	✓	✓	✓	×	×
Taiwan, NCKU (This work)	Medium	91.3 %	✓	✓	✓	✓	✓	✓

6. CONCLUSIONS

In this paper we have presented a natural and flexible auto attendant system with sophisticated dialog structure. This system is capable of dealing with telephone number inquiry by either full-name or related information such as title, specialty, etc. Fuzzy request and dynamic keyword set techniques, which were not addressed in other systems, are proposed in this paper. In addition, our system is designed modularly and embedded with a general common dialog structure, making our system more testable, configurable and portable. Comparison to other auto attendant systems also shows that our system is more intelligent and useable. Although our approach demonstrates a feasible solution for building auto attendant systems, for a telephone-based spoken dialog system to work ideally, further effort should be spent on issues such as robust speech recognition over telephone networks and out-of-vocabulary problems.

REFERENCES

1. L. F. Lamel and S. K. Bennacef et al., "The LIMSI RailTel System: Field trial of a telephone service rail travel information," in *Proceedings of Workshop on Interactive Voic Technology for Telecommunications Applications*, 1997, pp. 67-82.
2. G. Bouwman, J. Sturm, and L. Boves, "Incorporating confidence measures in the Dutch train timetable information system developed in the arise project," in *Proceedings of EuroSpeech*, 1999, pp. 493-396.

3. M. Helen, B. Senis, and V. Zue et al., "WHEELS: A conversational system in the automobile classification domain," in *Proceedings of International Conference on Spoken Language Processing*, Vol. 1, 1996, pp. 542-545.
4. H. Tsuboi and Y. Takebayashi, "A real-time task-oriented speech understanding system using keyword spotting," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 1992, pp. 197-200.
5. V. Zue and S. Sneff, et al., "Jupiter: A telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing*, 2000, Vol. 8, pp. 85-96.
6. D. Johnston, et al., "Current and experimental applications of speech technology for telecom services in Europe," *Speech Communication*, Vol. 23, 1997, pp. 5-16.
7. M. Kitai, et al. "ASR and TTS telecommunications applications in Japan," *Speech Communication*, Vol. 23, 1997, pp. 17-30.
8. A. Kellner, et al. "PADIS – An automatic telephone switchboard and directory information system," *Speech Communication*, Vol. 23, 1997, pp. 95-111.
9. G. Cerf, et al., "Progress in deployment and further development of the NYNEX VoiceDialing service," *Speech Communication*, Vol. 23, 1997, pp. 41-50.
10. A. L. Gorin, G. Riccardi, and J. H. Wrighf, "How may I help you," *Speech Communication*, Vol. 23, 1997, pp. 113-117.
11. F. Neubert and G. Gravier, et al., "Directory name retrieval over the telephone in the picasso project," in *Proceedings of IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, 1998, pp. 31-36.
12. B. T. Tan, Y. Gu, and T. Thomas, "Implementation and evaluation of a voice-activated dialing system," in *Proceedings of IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, 1998, pp. 83-86.
13. H. Hild and A. Waibel, "Recognition of spelled names over the telephone," in *Proceedings of International Conference on Spoken Language Processing*, Vol. 1, 1996, pp. 346-349.
14. F. Seide and A. Kellner, "Toward an automated directory information system," in *Proceedings of EuroSpeech*, 1997, pp. 1327-1330.
15. C. H. Wu, Y. J. Chen, and Y. C. Hung, "Telephone speech multi-keyword spotting using fuzzy search algorithm and prosodic verification," in *Proceedings of International Conference on Spoken Language Processing*, Vol. 3, 1998, pp. 835-838.
16. H. C. Wang, "MAT – A project to collect mandarin speech data through telephone networks in Taiwan," *International Journal of Computational Linguistic Chinese Language Processing*, Vol. 2, 1997, pp. 73-89.
17. Y. J. Yang, L. F. Chien, and L. S. Lee, "Speaker intention modeling for large vocabulary mandarin spoken dialogs," in *Proceedings of International Conference on Spoken Language Processing*, Vol. 2, 1996, pp. 713-716.
18. H. C. Wang, J. F. Wang, and Y. N. Liu, "A conversational agent for food_ordering dialog based on venus-dictate," in *Proceedings of ROCLING X International Conference*, 1997, pp. 325-334.

Jhing-Fa Wang (王駿發) received his M.S. degree in Electrical Engineering from National Cheng-Kung University in 1979, and Ph.D. degree in Computer Science from

Stevens Institute of Technology in 1983. He has served as a professor in National Cheng-Kung University since 1984. He was the Director of the Department of Computer Science and Information Engineering from 1990 to 1996. His research interests include CAD/VLSI design, computer speech processing, neural nets for image processing, graph theory and bio-information engineering.

Hsien-Chang Wang (王獻章) received his B.S. degree in Computer Science and Information Engineering from Feng-Chia University in 1991, and M.S. degree from National Chung-Cheng University in 1993. He is now a Ph.D. candidate in the Institute of Computer Science and Information Engineering, National Cheng-Kung University. His current research interests include spoken language processing, spoken dialog systems, natural language processing, and mobile information systems.