論文 / 著書情報
Article / Book Information

Title	Prospects for Spoken Dialogue Systems in a Multimedia Enviroment
Authors	Sadaoki Furui
Citation	ESCA Workshop, , , pp. 9-16
Pub. date	1995, 7

Prospects for Spoken Dialogue Systems in a Multimedia Environment

Sadaoki Furui

NTT Human Interface Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo, 180 Japan
Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152 Japan
furui@speech-sun.ntt.jp

ABSTRACT

This paper describes prospects of spoken dialogue systems focusing on their integration with a multimedia environment in future application areas and services. It tries to forecast where progress will be achieved in the near future and what applications will become commonplace as a result of the increased capabilities. It also describes the most important research problems to be solved in order to achieve useful systems. The problems for speech synthesis include natural and intelligible voice production, prosody control based on meaning, the control of synthesized voice quality and choice of individual speaking styles, multi-lingual synthesis, choice of application-oriented speaking styles, and synthesis from concepts. The problems for speech recognition include robust recognition against speech variations, adaptation/normalization to variations due to environmental conditions and speakers, automatic knowledge acquisition for acoustic and linguistic modeling, spontaneous speech recognition, and naturalness and ease of human/machine interaction. It is also important to establish evaluation methods for measuring the quality of technology and systems in a multimedia environment.

1. INTRODUCTION

For the majority of mankind, speech production and understanding are quite natural and unconsciously acquired processes performed quickly and effectively throughout our daily lives. Speech recognition and synthesis systems are expected to play important roles in an advanced multi-media society with user-friendly human-machine interfaces. Services using these systems will include voice dialing, database access and management, guidance and transactions, automated reservations, various order-made services, dictation and editing, electronic secretarial assistance, robots, automated interpreting (translating) telephony, security control, and aids for the handicapped (e.g., reading aids for the blind and speaking aids for the vocally handicapped) [11].
Speech recognition systems include not only those that recognize message content but also those that recognize the identity of the speaker. Speaker recognition techniques are expected to be widely used in the future to verity the claimed identity of users in telephone banking and shopping services, information retrieval services, remote access to computers, credit-card calls, etc.

Spoken dialogue (voice interaction) with computers is one of the most promising areas for applying advanced speech technology. Figure 1 shows a typical structure for task-specific voice control and dialogue systems [5]. Although the speech recognizer, which converts spoken input into text, and the language analyzer, which extracts meaning from text, are separated into two boxes in the figure, it is desirable that they work closely together, since it is necessary to use semantic information efficiently in the recognizer to obtain correct texts. How to combine these two functions is a most important issue, especially in conversational speech recognition (understanding).

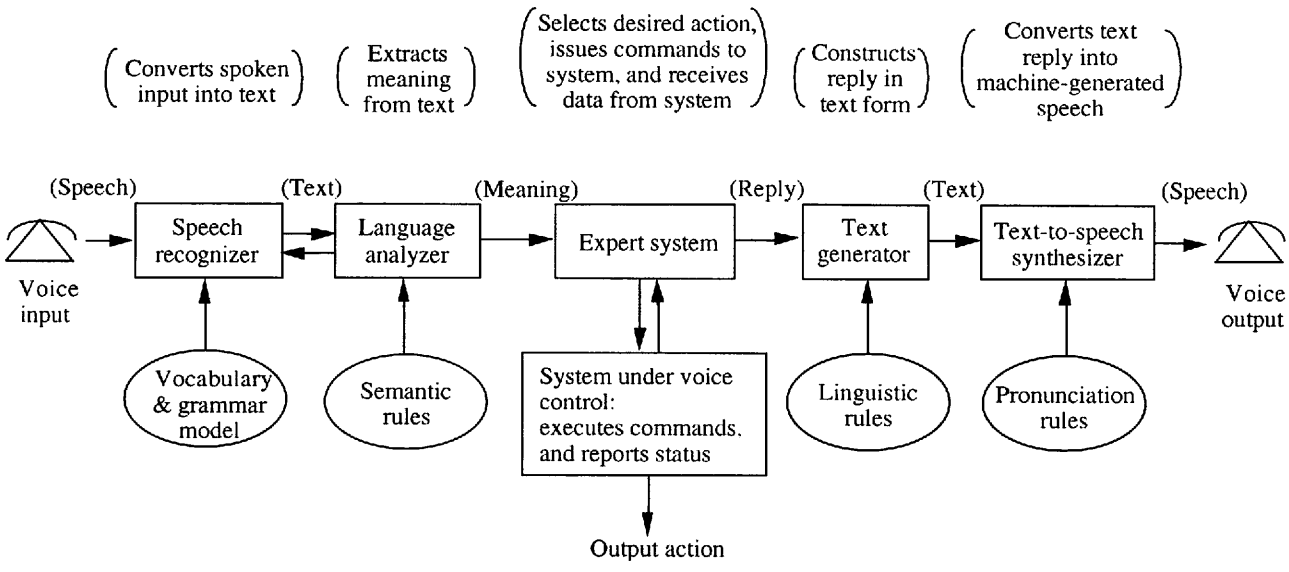

Fig. 1. Typical structure for task-specific voice control and dialogue systems.

The meanings extracted by the language analyzer are used to drive an expert system to select the desired action, to issue commands to various systems, and to receive data from these systems. Replies from the expert system are transferred to a text generator which constructs reply texts. Finally the text replies are converted into speech by a text-to-speech synthesizer. "Synthesis from concepts" is performed by the combination of the text generator and the text-to-speech synthesizer.

Dictation is the process of converting speech into word sequences, where it is not necessarily essential to extract the meaning from the text. On the other hand, in human-machine dialogue, the meaning of the input speech is most important and it is unnecessary to convert it into a correct sequence of words. Analysis and automatic processing of spoken dialogue in human-human and human-machine communication have recently been studied by many researchers in the field of speech communication from both scientific and application points of view [10][13]. Research on spoken dialogue is highly interdisciplinary, being related to many research areas. Therefore, there are many aspects to the studies of spoken dialogue.

2. ISSUES FOR SPOKEN DIALOGUE SYSTEMS IN A MULTIMEDIA ENVIRONMENT

2.1 Spoken dialogue systems

Ultimate speech recognition systems should be capable of robust, speaker-independent or speaker-adaptive, continuous speech recognition. Ultimate speech synthesis and recognition systems that are really useful and comfortable for users should match or exceed human capabilities. That is, they should be faster, more accurate, more intelligent, more knowledgeable, less expensive, and easier to communicate with than human beings. For this purpose, the ultimate systems must be able to handle conceptual information.

It is, however, neither necessary nor useful to try to use speech for every kind of input and output in computerized systems. Although speech is the fastest and easiest input and output means for simple exchange of information with computers, it is inferior to other means in conveying complex information. A variety of information can be displayed at once by images and text. It is important to have an optimal division of roles and cooperation in a multimedia environment that includes images, text, tactile signals, handwriting, etc. Input by speech (recognition) and output by images and text is considered to be an ideal combination for most interactive systems with computers. Displaying recognition results is very

useful to enable users to detect recognition errors and to reduce redundancy in conversation with computers.

The most promising application area for speech technology is telecommunications [12]. The fields of communications, computing, and networking are now converging in the form of personal information/communication terminals. In the near future, personal communications services (known as PCS) will become popular, and everybody will have their own portable telephone. Several technologies will play major roles in this communications revolution, but one of the key ones will be speech processing technology. Advances in speech synthesis/recognition technology will make telephone sets useful personal terminals for communicating with computer systems.

Ultimate seamless telecommunication systems are expected to use "virtual reality" technology. Teleconferencing and information retrieval systems that employ virtual reality will become possible. It is interesting to consider the roles of speech synthesis and recognition technology in these systems.

2.2 Dialogue modeling

Dialogue modeling in a multimedia environment is a very new, important, and interesting research topic. In typical spoken dialogue systems, a dialogue manager keeps the dialogue history and resolves anaphoric and elliptical sentences. How to extract contextual information, predict users' responses, and focus on key words in spoken utterances are very difficult and important issues. Automatic knowledge acquisition is very important in achieving systems that can automatically follow variations in tasks, including the topics of a conversation.

From the human interface point of view, future computerized systems should be able to automatically acquire new knowledge about the thinking process of individual users, automatically correct user errors, and understand the intention of users by accepting rough instructions and inferring details. A hierarchical multimedia interface that initially uses diagrams and images (including icons) to express global information, and then uses linguistic expression, such as spoken and written language for details, would be a good interface that matches the human thinking process.

2.3 Speech synthesis

A typical structure of a text-to-speech synthesizer consists of text-preprocessing, text-to-phoneme conversion (letters-to-sound), prosodic contour generation (pitch and duration), and speech signal reconstruction (synthesis). Although a large number of text-to-speech synthesis systems are commercially available and they can generally provide intelligible speech, they still lack naturalness. The quality of sound attained has recently become much better than before, as the result of using waveform concatenation methods for speech signal reconstruction, but it is still a long way from that of a naturally articulated voice.

There are several serious problems in the first two stages. First the reading of a written text is often error-prone, such as inaccurate pronunciation of proper names, foreign words, abbreviations, and acronyms. Second we do not have accurate rules for prosody control. Prosody is the major source of unnaturalness in present synthesized speech, which tends to be perceived as monotonous or jerky and fails to transmit meaning [14].

Other problems in speech synthesis include control of the synthesized voice quality and choice of individual speaking styles, multi-lingual and multi-dialectal synthesis, choice of application-oriented speaking styles (e.g., announcements, database access, newspaper reading, spoken e-mail, conversation), addition of emotion, and synthesis from concepts.

One crucial issue for smooth and successful conversation between computerized systems and people by means of speech recognizers and synthesizers is how to prompt the users by synthesized commands or questions. It has been reported that when users were asked to respond to the machine with isolated words, the percentage of isolated word responses depended strongly on the prompts made by the machine [1]. It was also reported that the intonation of prompted speech as well as the content could strongly influence the user responses.

2.4 Robust speech recognition

It is crucial to establish methods that are robust against voice variation due to individuality, the physical and psychological condition of the speaker, telephone sets, microphones, network characteristics, additive background noise, speaking styles, and so on [4][6]. It is also important for the systems to

have few restrictions on tasks and vocabulary. To solve these problems, it is essential to develop automatic adaptation techniques.

To cope with the problem of background noise, we have proposed an HMM composition technique in which a noise-source HMM and clean phoneme HMMs are combined into noise-added phoneme HMMs [7]. Experimental results showed that this technique gave similar recognition rates to those of HMMs trained by using a large noise-added speech database. We have recently extended the HMM composition technique to simultaneously cope with additive noise and multiplicative distortion [8]. Recognition experiments confirmed that this method greatly improves recognition rates for noisy and distorted speech. The efficiency and flexibility of the algorithm and its adaptability to new noises make it suitable as a basis for a speech recognizer resistant to noise.

Extraction and normalization of (adaptation to) voice individuality is one of the most important issues [3]. A small fraction of people occasionally produce exceptionally low recognition rates (the so-called "sheep and goats" phenomenon). Speaker normalization (adaptation) methods can usually be classified into supervised (text-dependent) and unsupervised (text-independent) methods. Experiments have shown that people can adapt to a new speaker's voice after hearing just a few syllables, irrespective of the phonetic content of the syllables.

2.5 Language modeling for spontaneous speech

One of the most important issues for speech recognition is how to create language models (rules) for spoken language. When recognizing spontaneous speech in dialogues, it is necessary to deal with variations that are not encountered when recognizing speech that is read from texts. These variations include extraneous words, out-of-vocabulary (OOV) words, ungrammatical sentences, botched utterances, restarts, repetitions, and style shifts.

It is crucial to develop robust and flexible parsing algorithms that match the characteristics of spontaneous speech. It is necessary to clarify why these expressions are used in spontaneous speech and how they are related to the processes of generating and understanding spoken dialogue. The parsing algorithms should rely on semantic information more than syntactic information. It will be necessary to appropriately adjust the methods of combining syntactic and semantic knowledge with acoustic knowledge according to the situation.

Statistical language modeling, such as bigrams and trigrams, has been a very powerful tool, so it would be very effective to extend its utility by incorporating semantic knowledge. We have recently investigated a grammar model based on a Markov model for the ATIS (Air Travel Information Service) task. An ergodic Markov model with states corresponding to word classes was trained and employed to rescore/reorder the N-best hypotheses. A small but significant improvement in selection of acceptable sentences was obtained with this method [8]. It will also be useful to integrate unification grammars and context-free grammars for efficient word prediction.

Since semantic and contextual information is heavily task dependent and collecting a large linguistic database for every new task is difficult and costly, portability or adaptation of linguistic models to new tasks and topics is also a very important issue. Style shifting is also an important problem in spontaneous speech recognition. In typical laboratory experiments, speakers are reading lists of words rather than trying to accomplish a real task. Users actually trying to accomplish task, however, use a different linguistic style. Not only the linguistic structures but also the acoustic characteristics of speech vary according to the task. Recognizers should be able to automatically acquire new knowledge about these features and trace these changes.

In spontaneous speech recognition, instability in the detection of end-points of speech is frequently observed. Additionally, the system should respond to the utterance as quickly as possible. To solve these problems, it is necessary to establish a method for determining when sufficient information has been acquired instead of detecting the end of input speech.

2.6 Objective evaluation methods

It is important to establish methods for measuring the quality of speech synthesis/recognition systems. Objective evaluation methods that ensure quantitative comparison of a broad range of techniques are essential to technological development in the speech processing field. Evaluation methods can be classified into the following two categories [2].

- Task evaluation: creating a measure capable of

evaluating the complexity and difficulty of tasks
- Technique evaluation: formulating both subjective and objective methods for evaluating techniques and algorithms for speech processing

Task evaluation is very important for speech recognition, since the performances of recognition techniques can be compared only when they are properly compensated for the difficulty of the task. Although several measures for task evaluation have already been proposed, such as word and phoneme perplexity, none of them is good enough at evaluating the difficulty in understanding the meanings of sentences. It may be very difficult to achieve a reliable measure for such purposes, since it involves quantifying all sources of linguistic variability.

Technique evaluation must take the viewpoint of improving the human-machine interface. Since human-machine interaction is very different from human-human interaction, it is crucial to evaluate multimedia systems under actual conditions. Ease of human-machine interaction must be properly measured. Recognition systems having minimum recognition errors are not always the best. There are various tradeoffs among the categories of errors, such as substitution, insertion, and deletion. Even if the error rate is relatively high, systems are acceptable if the error tendency is natural and matches the principles of human hearing and perception. It is crucial that recognition errors are easy to correct and that the system does not repeat the same errors.

To correctly evaluate technologies and to achieve steady progress, it is important to comprehensively evaluate a technique under actual field conditions instead of under a single set of controlled laboratory conditions. Even if recognition systems perform extremely well in laboratory evaluations and during demonstrations to prospective clients, they often do not perform nearly as well in the "real world". This is mainly because the speech that actually has to be recognized varies for many reasons and therefore usually differs from training speech. Systems designed for naive users are sometimes very boring for experienced users. How to automatically adjust the systems according to the skill of users is very important for making the systems comfortable for every user.

The following are also important evaluation items from the viewpoint of applications: incentive for customers to use the systems, low cost, creation of new revenues for suppliers, cooperation on standards and regulation, and quick prototyping and development.

3. MULTIMEDIA TELEPHONE DIRECTORY ASSISTANCE SYSTEM

3.1. System structure

We designed a multimedia spoken dialogue system for telephone directory assistance with three input devices (microphone, keyboard and mouse) and two output devices (speaker and display), based on our very-large-vocabulary continuous speech recognition and HMM composition techniques [9]. Providing access to directory information via spoken names and addresses is an interesting and useful application of large-vocabulary continuous speech recognition technology in telecommunication networks.

Figure 2 shows the structure of our continuous speech recognition system for telephone directory assistance. We have developed a two-stage generalized LR parser that uses two classes of LR tables: a main grammar table and several sub-grammar tables. These grammar tables are separately compiled from a context-free grammar. The sub-grammar tables deal with semantically classified items, such as city names, town names, block numbers, and subscriber names. The main grammar table controls

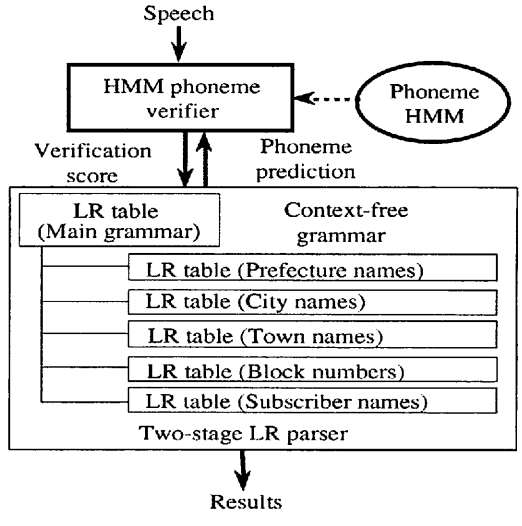

Fig. 2. Continuous speech recognition system for telephone directory assistance.

the relationships between these semantic items. Dividing the grammar into two classes has two advantages: since each grammar can be compiled separately, the time needed for compiling the LR table is reduced, and the system can easily be adapted to many types of utterances by changing the main grammar rules.

To reduce the search space without pruning the correct candidate, we use forward and backward trellis likelihoods, an adjusting window for choosing only the probable part of the trellis for each predicted phoneme, and an algorithm for merging candidates that have the same allophonic phoneme sequences and the same context-free grammar states. Candidates are also merged at the meaning level. This algorithm was tested on a telephone directory assistance system that recognizes spontaneous speech containing the names and addresses of more than 70,000 subscribers (vocabulary size is about 80,000). The experimental results show that the system performs well in spite of the large perplexity.

Figure 3 shows the basic structure of the dialogue system [9]. Since the interaction time, that is, the recognition speed is crucial in testing dialogue systems, we reduced the number of subscribers to 2,300 in this experimental system. The vocabulary size was roughly 4,000. This recognition system uses context-independent HMMs and does not use merging at the meaning level. Implemented on an HP-9000-735, the recognition currently takes about 20 seconds per sentence.

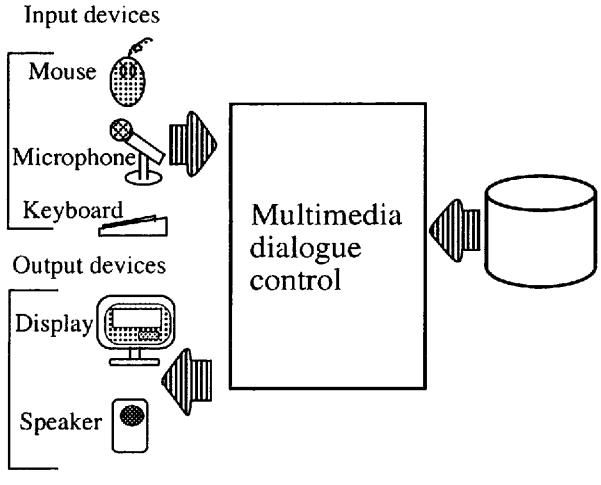

Fig. 3. Basic structure of multimedia dialogue system for telephone directory assistance.

Figure 4 shows an output window example in our system. The numbers on the left side of the window show the order of candidates. This window displays five potential subscriber candidates. For each candidate, the system displays five slots: city, town, block number, subscriber name, and telephone number. A simple example of how this dialogue system is used is as follows:

(1) After clicking the speech button, a user states an address and subscriber name.
(2) The system recognizes the input speech and displays five candidates.
(3) If one of the candidates is correct, the user obtains the telephone number by clicking the telephone number slot of the correct candidate.

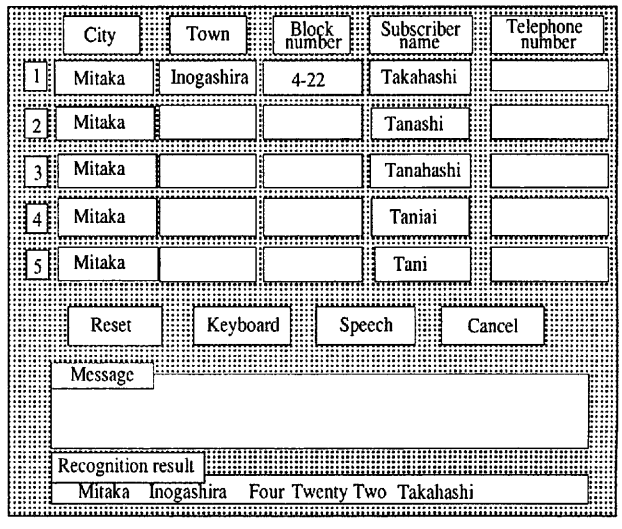

Fig. 4. Example of a window in the multimedia dialogue system for telephone directory assistance.

3.2. Dialogue controller

The main functions of the dialogue controller are as follows.

(1) *Display candidates*

After speech recognition, four potential candidates are displayed in order of their likelihood scores. The telephone directory assistance database constraint is not usually used in selecting these candidates. However, the fifth candidate is the candidate that satisfies the constraint in the telephone directory assistance database, because there is a high possibility that the candidate that satisfies the constraint is correct, even if it has a low likelihood score.

(2) *Error correction*

If there is no correct candidate among the five candidates, the user corrects the input error by choosing the candidate closest to the correct subscriber address and name, clicking the wrong keyword slot, and uttering a sentence with that specific semantic item. In error correction mode, the system switches the main grammar to the grammar in which the clicked item must be uttered. For example, if a user clicks the subscriber name slot, the system switches the main grammar to the grammar for utterances that need to include a subscriber name. The user can include some new information in the sentence, in addition to the specific item. The beam width is also increased to raise the recognition accuracy.

3.3. Evaluation

This system was evaluated from the human-machine-interface point of view. We asked 20 researchers in our laboratory to try to use this system. Dialogue experiments were performed to evaluate the following issues: (a) system performance (task completion rate, sentence understanding rate, task completion time, etc.), (b) user evaluation of the system, (c) content and manner of user utterances, and (d) problems encountered with the system.

(1) *Training*

The users were first requested to practice operating this system by themselves using a tutorial system, which was an interactive system implemented on a workstation. The tutorial system was designed to control and unify the guidance as well as knowledge given to each user. One practice sequence, including examples of correct recognition and incorrect recognition, takes roughly 10 minutes, in which users operate the system following instructions displayed on the screen. A typical request format is also shown and practiced in this stage. Pauses and speaking rates are not controlled.

(2) *Testing*

Twenty sheets of paper indicating the tasks using sketch maps were given to each user. Each task was indicated by the name and location of the person whose telephone number had to be requested on the map. The amount of information indicated on the sheet varied; for example, the first name or the town name of the person was sometimes not given. The users were requested to make inquiries based on the information given in each sheet. We used maps for indicating the tasks to avoid imposing any particular structure on the spoken sentences.

When the user could obtain the desired telephone number, he/she wrote down the number on the answer sheet, and proceeded to the next task. Even if the user could not get the telephone number after considerable effort, he/she was requested to proceed to the next task.

(3) *Questionnaires*

After testing, each user was asked to answer several questions, and the information obtained was compared with various logs recorded during the test.

(4) *Results*

The results of the experiment gave the task completion rate as 99%, which means that, in most of the trials, the users could get the correct telephone numbers. The average number of utterances for each task was 1.4, and the average sentence understanding rate was 57.8%. The average rate for the correct recognition result being indicated in the top five candidates was 77.5%. We found that the higher the top five recognition rate was, the lower the average number of utterances became.

The average time needed to complete each task was 57.2 seconds, and it decreased as the users became more experienced. About 75% of the users said that they preferred using the computer-based dialogue system to a telephone directory. About 55% of the users said that the system was easy to use. The main reason for negative answers to this question was closely related to the feeling that the response time of the system was too slow.

We have collected a speech database through these experiments for future analysis and experiments.

4. CONCLUSION

Spoken dialogue systems using speech recognition and synthesis technology are expected to play important roles in an advanced multimedia society with user-friendly human-machine interfaces. It is important to have an optimal division of roles and cooperation in a multimedia environment that includes images, text, tactile signals, handwriting, etc.

Input by speech recognition and output by images and text is an ideal combination in most interactive systems with computers. Displaying recognition results is also useful to enable users to detect recognition errors and to reduce redundancy in conversation with computers.

Dialogue modeling in a multimedia environment is a very new, important, and interesting research topic. Other important research issues include synthesizing a naturally articulated voice, appropriate text prompting, robust speech recognition, language modeling and flexible parsing for spontaneous speech, automatic knowledge acquisition and adaptation, and establishing objective evaluation methods. This paper also introduced a multimedia speech dialogue system for telephone directory assistance.

REFERENCES

[1] S. Basson: "Prompting the user in ASR applications," COST 232 Workshop on Speech Recognition over the Telephone Line, Rome, Italy (1992)

[2] S. Furui: "Evaluation methods for speech recognition systems and technology," Proc. Workshop on Int. Cooperation and Standardization of Speech Databases and Speech I/O Assessment Methods, Chiavari, Italy (1991)

[3] S. Furui: "Speaker-independent and speaker-adaptive recognition techniques," in Advances in Speech Signal Processing, ed. by S. Furui and M. M. Sondhi, Marcel Dekker, Inc., New York, pp. 597-622 (1992)

[4] S. Furui: "Toward robust speech recognition under adverse conditions," ESCA Workshop on Speech Processing in Adverse Conditions, Cannes-Mandeliew, pp. 31-42 (1992)

[5] S. Furui: "Towards the Ultimate Synthesis/Recognition System," *Voice Communication between Humans and Machines*, ed. by D. B. Roe & J. G. Wilpon, National Academy Press, Washington D. C., pp. 450-466 (1994)

[6] B. H. Juang: "Speech recognition in adverse environments," Computer Speech and Language, 5, pp. 275-294 (1991)

[7] F. Martin, K. Shikano and Y. Minami, "Recognition of Noisy Speech by Composition of Hidden Markov Models," Proc. Eurospeech '93, pp. 1031-1034 (1993)

[8] Y. Minami, M. Barlow, T. Matsuoka and S. Furui: "Recent topics in speech recognition research at NTT Laboratories," Proc. ARPA Spoken Language Systems Workshop, Austin, TX (1995)

[9] Y. Minami, K. Shikano, O. Yoshioka, S. Takahashi, T. Yamada and S. Furui: "A Large-Vocabulary Continuous Speech Recognition Algorithm and Its Application to a Multi-Modal Telephone Directory Assistance System," ARPA Workshop on Human Language Technology, Princeton, pp. 362-367 (1994)

[10] Proc. International Symposium on Spoken Dialogue - New Directions in Human and Man-Machine Communication, Waseda University, Tokyo, Japan (1993)

[11] L. R. Rabiner and B. H. Juang: "Fundamentals of Speech Recognition," Prentice-Hall, Inc., New Jersey (1993)

[12] L. R. Rabiner: "The role of voice processing in telecommunications," Proc. 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Kyoto, Japan, pp. 1-8 (1994)

[13] K. Shirai and S. Furui (ed.): *Special Issue on Spoken Dialogue*, Speech Communication, 15, 3-4, pp. 189-365 (1994)

[14] C. Sorin: "Speech synthesis applications and research: Present and future," Proc. 13th Int. Cong. on Acoustics (1989)