1993년도 한국음향학회 학술논문발표회 논문집(제 12권 1(s)호)

Trend on the Technical Development of Japanese Speech Recognition

Katsuhiko SHIRAI

Department of Information and Computer Science, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169 JAPAN
e-mail: shirai@shirai.info.waseda.ac.jp

1. History

If we remind the research history of speech science and engineering in Japan, we can find several projects as following:

1971 Pattern Recognition Project by MITI

1973— Technical Committee of Speech Research of Acoustical Society of Japan, now jointly managed with IEICEJ.

1986—1993 (1st stage) Automatic Telephone Interpretation Research Project of ATR (Advanced Telecommunication Research International)

1993 — (2nd stage) Interpreting Telecommunications Research Project

1986 — 1993 Auditory and Visual Perception Research Labs.

1992 — Human Information Processing Research Labs.

1987 — 1990 Advanced Man-Machine Interface through Spoken Language
Priority area, Grant in Aid for Scientific, Research supported by Ministry of Education, Science and Culture

1993 — Understanding and Generation of Dialogue by Integrated Processing of Speech, Language and Concept
Priority area, same as above.

2. Basic Trends of Research in in speech recognition

The most successful and widely employed approach is the statistical method to solve the basic problems of Speech Recognition, that include the vocabulary size, utterance mode (from discrete to continuous), speaker variability, ambient noise and so on, Stochastic training

methods such as HMM and Neural Network are effective based on large speech database that can cover the wide variety of voice and its environments.

Recently, efficient algorithms have been developed to make possible automatic training of speech recognizers using large amount of speech data which do not used any manual labeling or tagging.

That technology and the accumulation of speech database have improved so much the performance of speech recognition system. And speaker independent continuous speech recognition of large vocabulary size has been reaching the level of practical use.

Therefore, recently in Japan, many research laboratories developed experimental speech recognition systems almost of which operate in dialogue mode and in real time.

Accumulation of these experiments gave us many suggestions to make better human-machine interface including voice communication channel.

The main results are

(1) It is not realistic to require the human speaker to speak always only grammatical sentences with clean pronunciation.

(2) The machine should accept large vocabulary. And further unknown words must be properly handled.

(3) The machine can be adaptable for the speaker

(4) Various ambient noise can be reduced.

(5) Multimodality is useful in man-machine interface.

Practical speech dialogue system should have less restrictions as possible on the user. The major constraints for the speaker have been concerned with vocabulary, grammar, speaking style, knowledge of the system, vocalization, timing and so on. It may be impossible to put no restriction, but many restrictions could be applied unconsciously for the user by suitably arranging the man-machine interface.

Actually, in the conversation of human, by combining the situation and visual informations, the hearer can easily grasp the meaning. Therefore, it is very important in the speech dialogue system to use some media other than voice. Not only giving the voice output, but also showing users a graphical display or some gesticulative signals in an appropriate timing, the speech turns out to be restricted enough, and may be easier for the system to recognize, even if the user is permitted to speak in very natural manner. Concerning such multi-modal systems, there isn't an effective methodology designing how multiple communication channels should be related. The most suitable expression for some information takes not necessarily an invariable form but depends on the situation and the timing. This design method is considered to be the most important to put the dialogue system in practice.

The key point here, first of all, is that on each situation the range of the view of the system should be limited enough to decrease the search space of the recognition process, but it must be assured that the user can continue the conversation almost satisfactory and the aim of the conversation can be achieved after the several turns. Second issue is, on the contrary, how to secure the high flexibility of the speaker. To allow users that they can extend the topic freely contradicts with the first point to limit the search space in the recognition. This

fact means that the rapid change of the topic should be basically introduced by the system and after the clear guidance certain flexibility must be added by giving a selective menu.

Therefore, in the spoken dialogue system design, how to manage the dialogue by allowing the user's initiative and how to lead and expand the conversation is more important issue than how to recognize the speech.

3. Several examples

Telephone Directory Assistance(NTT)

Current typical speech recognizer based on HMM phoneme recognizer can work with 1000—2000 vocabulary. Shikano and others (NTT) developed a algorithm for very large vocabulary (about 80,000 words) continuous speech recognitions. They use a two-stage LR parser together with phoneme HMMs. The algorithm was applied to a telephone directory assistance containing 70,000 subscribers. New Noise Reduction scheme was introduced for HMM based recognizer. In this method, HMM composition combines a noise-source HMM and a phoneme HMM into on noise-added phoneme HMM and the large training of noisy-speech HMM can be solved.

A Voice-Activated Extension Telephone Exchange System (KDD)

It has been long recognized that speech recognition in telephone is very important application area of speech technology. However, the band limitation, wide variety of channel characteristics and nonlinear distortion cause much difficulty to speech recognition.

A prototype of an extension telephone exchange system was developed for a company size of 200 employees. Realtime speaker-independent continuous speech recognition was realized using a DSPs and is now tested to collect a large amount of spontaneous speech through telephone line.

ASURA & ATREUS (ATR)

ATR Interpreting Telephony Research Labs. (currently, ATR Interpreting Telecommunications Research Labs) was established in 1986 to conduct fundamental study to realize automatic translation Telephone system.

Automatic interpreting telephony is a new technology that enables speech communication between people speaking different languages.

ATREUS is the front-end to the speech translation system ASURA, that was demonstrated in the international speech translation experiment among ATR, CMU and Siemans AG + Karlsruhe Univ.

Task domain of ATREUS is the international conference registration where the dialogue is goal-oriented and the vocabulary size is about 1500 words.

The specially of ATREUS is the capability of spoken adaptation for the HMM, and the HMM phone model and LR parser is combined in very effective manner.

Spontaneous Speech Dialogue System TOSBURG (TOSHIBA)

TOSBURG is a unique system that has the features as

- speaker-independent spontaneous speech understanding based on keywords
- user-initiated dialogue management
- multimodal response generation

It has four components; word spotter, a keyword lattice parser, a dialogue manager and a response generator. The speciality in the recognition is its robustness for the noise. And it can work for the wide variety of ill-formed sentences.

4. Project of Spoken Dialogue

The last project "Generation and Understanding of Dialogue by Integrated Processing of Speech, Language and Concept" was organized by Prof. S.DOSHITA and started in 1993 as a three years project, that was planned as the successor of the project "Advance Man-Machine Interface Through Spoken Language" which was also the Area of Grant-in-Aid for Science Research on Priority Area.

The aim of the project is concentrated on the research of total process of speech communication. There have been many independent works concerning speech recognition, language processing and cognitive process of dialogue. However, they are mutually correlated and can never be completely studied, if each of them is confined itself only in each area.

Since the spoken language is very much different from the written language, traditional methods on language processing which have mainly treated the written language are insufficient to consider the spoken dialogue. The main reason is the plentifulness of ambiguities that can occur in the various stages of spoken dialogue.

However, the research of spoken dialogue is very important because as the progress of the computer technology, the higher ability of the human interface is required and the speech I/O is strong demand of non professional user of computers.

Four areas were adopted as the core groups of the project.

(A) Speech Understanding and Synthesis Technologies in Spoken Dialogue
Leader: Y.NIIMI (Kyoto Inst. of Tech.)

(B) Language Analysis and Generation in Spoken Dialogue
Leader: H.TANAKA (Tokyo Inst. of Tech.)

(C) Conceptual Level Understanding and Representation in Spoken Dialogue
Leader: R.MIZOGUCHI (ISIR Osaka Univ.)

(D) Modeling of Processes in Spoken Dialogue
Leader: K.SHIRAI (Waseda Univ.)

Each group has from 6 to 8 members. And under the guidance of these research members, a large member of collaborating members are joining in the project.

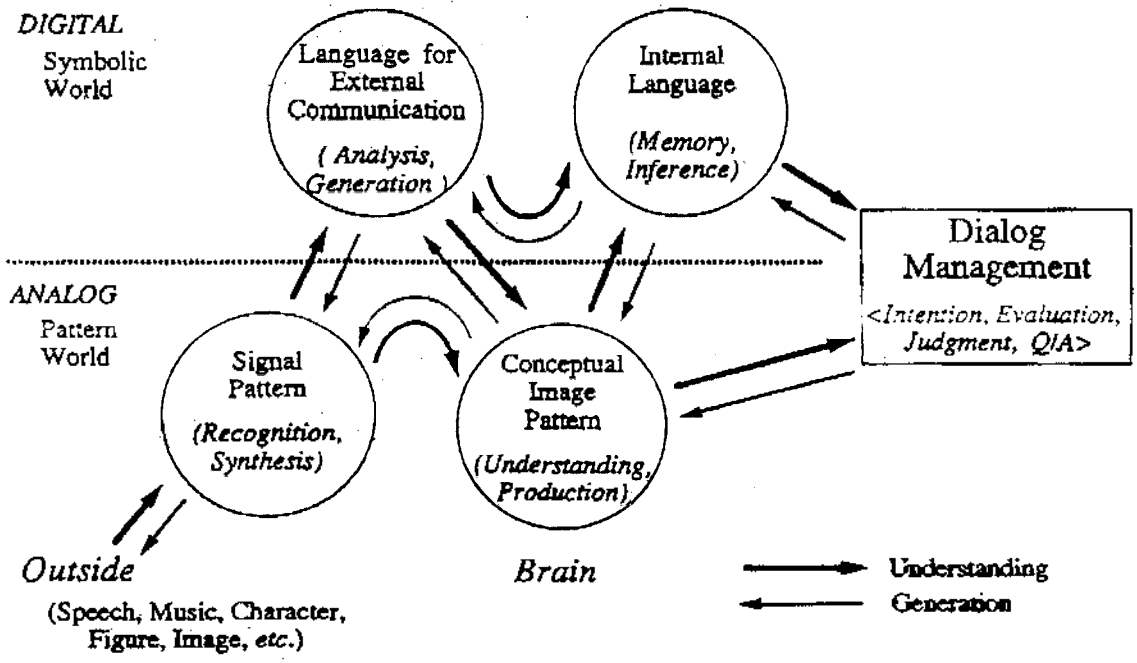

Fig. 1 Model for Dialog Understanding and Generation

5. Speech Corpora

Common speech database is indispensable for speech research to evaluate various speech analysis methods and speech recognition systems. Although the necessity of common speech database was discussed in the early stage of speech research, it tooks long time to realize such database in Japan.

These were initial efforts at ETC and Tohoku Univ. However, the first well organized speech database project was supported by JEIDA (Japan Electronic Industry Development Association). As the result of their efforts, now we can use the JEIDA Japanese Common Speech Data Corpus. The corpus is composed of 323 items uttered by 75 male and 75 female speakers. The speech database has been distributed among more than 45 organizations evolving in speech research.

ATR Interpreting Telephony Research Laboratories has been developing a large scale database since the establishment in 1986. The database is intended for the variety of use in speech research. Multiple transcriptions have been made in five different layers from phonemic description to fine acoustic-phonetic expressions.

In the research project of "Advanced Man-Machine interface through Spoken Language" which was carried out from 1987 to 1989, speech database is considered as one of the important themes. A preliminary work has been done to collect isolated utterances containing 109 syllables, 216 words and sentences from 20 speakers (10 males and 10 females, from 20 years to 60 years of age). Now, these database was distributed in the form of CD-ROM.

After that, the requirement for the large speech database including many speakers grew up, as the progress of the statistical method. Now, large scale speech database is also available through ASJ. The ASJ continuous speech corpora include ATR phoneme balanced 503 sentences and spontaneous task-oriented sentences. However, we need more spontaneous spoken dialogue corpora in the next stage.