NEXT MAJOR APPLICATION SYSTEMS AND KEY TECHNIQUES IN SPEECH RECOGNITION TECHNOLOGY

Kazuyo TANAKA

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki 305, JAPAN
ktanaka@etl.go.jp

ABSTRACT

In this paper, we discuss several major speech recognition applications which will contribute to some human activities in a decade. At first, recent Japanese speech-related national projects directed toward future intelligent systems are briefly reviewed. Then we discuss three systems as the next major speech applications: *substantially robust systems, multimodal interaction systems* and *multilingual dialogue systems*. Evaluation of the performance of these systems is separately discussed in view of both total systems and specific technologies. We suggest that the degree of the difficulty of some kinds of specific tasks can be even more precisely measured, while the total system performance evaluation will become more difficult in future complex systems. Last, we take up *phrase spotting, distance calculation for phonetic symbol sequences, adaptation/learning, and software modularization/multi-agents* as the key techniques in constructing the above applications.

1. INTRODUCTION

In the past decade, we have achieved remarkable progress in some aereas of speech recognition technologies. The most significant factor that contribute to this progress is large scale speech databases, though several statistical techniques have also been important factors. The speech recognition performance therefore depends basically upon collected speech sample data and language data concerning their respective acoustic and/or language environments. This implies that the systems are insufficiently robust and will be one of the main factors preventing proliferation of speech application systems. This is my basic position to consider future application systems.

In this paper, we will first discuss several application systems which are considered to be coming into being in the next decade. We also briefly introduce recent Japanese Government fund-assisted speech-related projects. We will then discuss three systems as the next major speech recognition applications: *substantially robust systems, multimodal interaction systems* and *multilingual dialogue systems*, though it is likely that their performance will be limited in the near future. In chapter 3 we discuss the performance evaluations of these systems, separately evaluating total system performance and specific technologies. Lastly, we describe the key techniques putting together the above application systems, taking up the following four techniques: *phrase spotting, phonetic symbol distance calculation, adaptation/learning and software modularization/Multi-agent.*

The reference applications cited in the following discussions are mainly from Japanese systems.

2. SPEECH APPLICATION SYSTEMS INTO THE NEXT DECADE

2.1 Japanese Speech-Related Projects

Several speech-related projects have been carried out in Japan supported (in part) by national funds. From these, we selected two major projects: Real World Computing (RWC) Program conducted in ETL and the Real World Computing Partnership (RWCP), and a speech-to-speech translation project conducted in ATR Laboratories. These projects are basically directed toward intelligent application systems in the next decade.

(1) *RWC Program*:

The RWC Program started in 1992 and now 5 year later the research and development program has begun to establish fundamental technologies for the following five research topics: *Multimodal systems, autonomous learning systems, self-organizing information base systems, theory and algorithm bases, and real world adaptive devices.* The first two topics include spoken dialogue understanding, and are described as:

- Multimodal system: A personified agent type human interface that enables users to communicate with computers in integrated interaction functions such as speech and vision [Hayamizu et al, 1997; Mukai et al, 1997].

- Autonomous Learning System: An agent system that can move in real environments autonomously, collect and learn information related to the environment and people, through sensing and interactions with human beings [Asoh et al,1997].

The first prototype of the multimodal interaction system was developed in the first half of the RWC program, a block diagram of which is shown in Fig.1[Itou et al, 1995]. The system is characterized by the integration of the component functions of speech and image. It suggests one direction in future speech application systems.

(2) ATR speech-to-speech translation project:

The project originally started in 1987 and from 1993, has been conducted in the ATR Interpreting Telecommunications Research Laboratories (ATR-ITL). The aim is to establish component technologies of multilingual speech translation technologies for natural speech and to achieve cross-language global communications [Yamazaki, 1995]. The research subthemes are listed as: a) spontaneous speech recognition, b) prosody

processing, c) cooperative integrated translation, and d) integration of speech and language processing [Mima et al, 1997].

Real time natural speech translation is one of the main speech processing applications, especially in Japan where there exists a big demand. Of course its performance will be fairly limited in the near future.

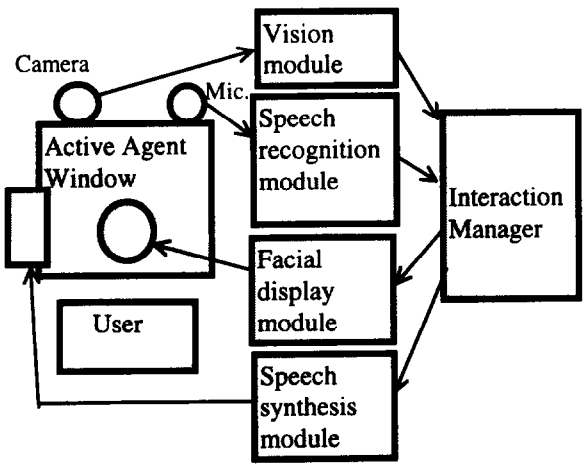

Fig. 1 System configuration of a multimodal system developed at ETL in 1995.

2.2 Next Major Application Systems

We list here speech recognition application systems which will come into valuable use within acceptable performance standards in the next decade.

(1) *Substantially Robust Systems*:

Current speech recognition systems do not have robustness in a wide range of speakers and environments. For example, recognition performance deteriorates due not only to acoustic-phonetic variations but also due to the conversational characteristics of the voices of children, handicapped and elderly people. Speech application machines installed in public spaces or linked to computer/telephone nets necessarily require recognition under such conditions in addition to proper responese during unexpected noise sources. Portable terminals and equipment for automobiles also require such robustness.

This type of systems will be available for tasks in the form of guidance, shopping, information retrieval, etc. The critical difference from the current systems is that the variations are basically unexpected, so that conventional training techniques will be less effective.

(2) *Multimodal Interaction Systems*:

The multimodal system will become main-stream in human-machine interface. There have already been proposed several prototype systems, in which speech is one of the main modalities [Ando et al, 1994; Takebayashi, 1994; Kamio et al, 1994; Watanuki et al, 1995] . Most of the systems employ humanized agents as metaphors of the systems. It encourages users to speak to computer systems.

One direction of the tasks proper for the multimodal system is to combine it with one of the expert systems which are already in some practical use, thereby improving the human-computer interface [Nitta et al,1997]. If visual input/output and audio input/output are employed in addition to speech, more sophisticated functions will be implemented in such areas as consultation systems, training/ education systems, etc.

The key point in system implementation is the integration of the modalities. It means a) using an individual mode to conform to its characteristics, b) time axis management of all modarities [Itou et al, 1996], and c) describing one modality function conditioned by other modalities. Usually, these systems are implemented by a multi-modular or multi-agent system. In relation to this point, further progress in programming techniques will be required. One technical aspect in multimodal systems is to verify or identify the speaker's location or individuality by using computer vision technique. It also should combine the system to hands-free speech input.

(3)*Multilingual Spoken Dialogue Systems*:

This category includes speech-to-speech translation systems such as the ATR-ITL project or Verbmobil project [Bub et al, 1997]. However, I think other types of systems such as speech recognition and synthesis systems supporting multi-language will also be substantially useful. In these systems, it is possible that parts of the software modules can be distributed and reusable in other language systems, while currently system implementation starts from gathering real speech samples. For example, HMM based recognition is applicable to another language recognition, but it needs the speech samples and their transcriptions of the corresponding language. It would be effective if we could adopt a common transcription unit set, as an IPA-like symbol set, and define a kind of distance measure between this set and the individual language transcription sets.

3. PERFORMANCE EVALUATION

3.1 Evaluation of Total Systems

In general, evaluating the performance of practical applications is difficult, particularly in the aspect of usability of the total system. The application systems discussed in section 2.2 are characterized by complexity in their system configurations. This implies that the system evaluation must become more complicated.

If we intend to evaluate the relative performance of systems, two types of the methodologies are considered (depending on the system category). One is a so-called black box evaluation by users (/subjects), in which a main evaluation factor is a number of times of user's actions to achieve a given task, although several kinds of evaluation factors have been considered. However, as it is well known and discussed in EAGLES Report (and also discussed in the JEIDA Speech I/O Committee in Japan), there is yet no definitive methodology.

The other is a kind of battle game such as *RoboCup* [Kitano, 1997], which is a soccer game by robots, or *DiaLeague* [Hashida

et al,1996], which is a kind of linguistic level dialogue game in a map task. Both are games by autonomous systems. In speech application systems, it will be acceptable that human instructors participate in the game by voice command. Those methods have of course some useful aspects that evaluate total system performance in an objective measure. But the tasks are necessarily distinct from practical use applications because the real use is too complicated to define in the experimental condition.

As we mentioned in chapter 2, the multimodal system connected with an expert system is a candidate for the future application systems. In relation to this, we have collected spoken dialogue data which contains two topics: dialogue between car dealers and customers, and that between travel agents and customers [Tanaka et al, 1996]. Professional dealers and agents were employed to produce reality in the conversations (This was human-to-human dialogue). If we substitute the agent with a computer system, images and gestures will be used in addition to speech, in both directions of the user-to-system and system-to-user, and the content of the dialogues are directed to more complex task compared with ATIS-like domain [Pallet et al, 1992]. I think the travel agent system may be an acceptable application system for the expansion of current speech dialogue systems. It is basically object-oriented conversations, so that conventional experimental knowledge of the system evaluation can be used.

3.2 Evaluation of Specific Techniques

Before evaluating specific techniques, we have to define the measurements of tasks to be tested. If we can define such measurements, it will be possible to evaluate the complexity of the tasks more precisely.

Let's examine *perplexity* as an example. Perplexity is a well known index for estimating the grammatical difficulty of sentence sets [Bahl, et al, 1978]. However, perplexity corresponds to vocabulary size in the case of word sets while the recognition difficulty does not always depend on the vocabulary size. We appropriately estimate the speech recognition difficulty of word set by applying between-word distance calculation [Tanaka et al, 1997]. Therefore, if we intend to estimate recognition difficulty of sentence sets, we should estimate between-sentence distances. In other words, the between-phonemic-sequence distance is another important factor for estimating the speech recognition difficulty of sentence sets, because possible branching words that affect the recognition difficulty should be limited using the between-word distance distribution of the words at that moment. Therefore, a more reasonable index for the recognition difficulty can be provided if we combine this factor with the perplexity.

4. KEY TECHNIQUES IN FUTURE APPLICATION SYSTEMS

In the speech recognition/Understanding domain, the following techniques will be important:

(1) *Phrase Spotting*

The phrase spotting technique will be crucial in various aspects of the systems cited in section 2.2. In those systems, it is obviously difficult to recognize all utterances of individual speakers, therefore systems recognize only the utterances focused on by using certain information, such as dialogue context or that of another modality (eg. visual image). The phrase spotting techniques can utilize such focusing techniques as those used in acoustic domain (eg. microphone-array processing [Nakamura et al 1996], blind source separation [Cichocki et al, 1997]) , multimodal domain (eg. interaction between speech and image), conversational domain, etc.

The phrase spotting from a given acoustic-phonetic stream or phonetic symbol sequence can be carried out by a kind of dynamic programming techniques.

(2) *Phonetic Symbol Distance Calculation in Symbolic Domain*

Handling the distance between phonetic symbol sequences in symbolic domain is more efficient when compared with that in the acoustic domain. It is effective for large vocabulary processing in predicting word candidates, estimating speech recognition difficulty of given vocabularies, and for phrase spotting.

We proposed the following method for this issue [Tanaka et al, 1997]. We employ subphonetic segment units to describe words. The distance calculation is composed of two steps: in the first step, two phoneme sequences are converted to the subphonetic unit sequences. Next, the distance between the two sequences is estimated by optimal matching using dynamic programming, where the distance values between subphonetic units are defined using each HMM in such a way that the distances are very closely related to the acoustic-phonetic domain distance.

(3) *Adaptation / Learning*

Against speaker variations and noise or recording environments, adaptation (/learning) is an effective way to achieve acceptable performance. Currently, the model structure of the object is basically fixed and categories of training samples are known. However, critical points lie in the issues of what are the necessary samples to be provided for training or learning and how best to collect them. In other words, the system requires the ability to estimate the training value of the samples, as well as estimating the model structure itself..

Some ideas have been tested to automatically acquire a kind of the phonemic category set from spoken word samples without individual transcriptions by the learning method [Kojima et al,1997]. In other words, it simulates the process of an infant acquiring the speech ability. This type of simulation can be used for evaluating learning techniques in the real world domain.

I think current techniques for learning bring about a limited effect in practical systems. It often proceeds into directions different from the user's thinking.

(4) *Software Modularization / Multi-agent System*

As we mentioned above, future systems will be composed of complex configurations, to the extent that advanced programming techniques such as controlling multi-agent systems will be required. Units of software modules and knowledge bases will be more compact for maintenance and reused in several environments.

5. CONCLUDING REMARKS

We have discussed the next major applications in speech recognition technologies. Proper application systems provide the driving force to create new or improved techniques. At this point, it should be noted that the most important thing is developing the key techniques that will be effective in a wide range of technological areas. It should also be noted that speech technologies have received significant knowledge from basic speech research .

In the last, I wish to thank Satoru Hayamizu, Hiroaki Kojima and all members of the Speech Processing group in ETL for their usual discussion.

REFERENCES

[Ando et al, 1994] H. Ando, Y. Kitahara, N. Hataoka, "Evaluation of multimodal interface using speech and pointing gesture on an *Interior Design System*," (In Japanese) IEICEJ Trans. D-II, J77-D-II No.8, pp.1465-1474 (Aug. 1994).

[Asoh et al, 1997] H. Asoh, S. Hayamizu, I.Hara, Y. Motomura, et al, "Socially embedded learning of the office-conversant mobile robot Jijo2", Proceedings of 15th IJCAI, pp.880-885 (Aug. 1997).

[Bahl et al, 1978] L.R. Bahl, J.K. Baker, et. al., "Automatic Recognition of continuously spoken sentences from a finite state grammar," Proc. IEEE ICASSP-78, pp.418-421(1978).

[Bub et al, 1997] T. Bub, W. Wahlster, A. Waibel, "Verbmobil: The combination of deep and shallow processing for spontaneous speech translation," Proc. ICASSP97, pp.71-74 (Apr. 1997).

[Cichocki et al, 1997] A. Cichocki, J. Cao, I. Sabala, "On-line adaptive algorithm for blind equalization of multi-channel systems," Proc. of International Conference on Neural Information Processing (ICONIP-97), pp.649-652 (Nov. 1997).

[Hasida et al, 1996] K. Hasida, Y. Den, K. Nagao, H. Kashioka, et al, " Report of DiaLeague'96 Spring Session and future plans," (In Japanese) Proc. Japan Soc. AI, SIG-SLUD-9601, pp. (June 1996).

[Hayamizu et al, 1997] S.Hayamizu, K.Sakaue, O. Hasegawa, K. Itou, et al, "Multimodal interaction system at the ETL," Proc. 1997 Real World Computing Symposium, pp.16-29 (Jan. 1997).

[Itou et al, 1995] K. Itou , O. Hasegawa, T.Kurita, S. Hayamizu, K. Tanaka, K.Yamamoto, N.Otsu, "An active multimodal interaction system", Proc. ESCA Workshop on Spoken Dialogue System, pp.169-172(1995).

[Itou et al, 1996] K. Itou, S.Hayamizu, K.Tanaka, "A timing management method in multimodal dialogue systems", Proc. Acoustical Society of America and Acoustical Society of Japan Third Joint Meeting 1996,4aSC13, J of ASA pp.2758 (Dec. 1996).

[Kamio et al, 1994] H. Kamio, H. Matsuura, Y. Masai, T. Nitta, "Multimodal dialogue system MultiksDial," (In Japanese) IEICEJ Trans. D-II, J77-D-II No.8, pp.1429-1437 (Aug. 1994).

[Kitano et al, 1997] H.Kitano, M. Tambe, P. Stone, M. Velloso, et al, "The RoboCup synthetic agent challenge 97," Proceedings of 15th IJCAI, pp.24-29 (Aug. 1997).

[Kojima et al, 1997] H. Kojima, K. Tanaka, "Organizing phone models based on piecewise linear segment lattice of speech samples", Proc. of European Conference on Speech Communication and Technology(EUROSPEECH'97), pp.1219-1222 (Sept.1997).

[Mima et al, 1997] H. Mima, O.Furuse, H. Iida, "Improving performance of transfer-driven machine translation with extra-linguistic information from context, situation and environment," Proceedings of 15th IJCAI, pp.983-988 (Aug. 1997).

[Mukai et al, 1997] T. Mukai, T. Nishimura, S. Nagaya, J. Kiyama, et al, "Multimodal and real time dialogue through gesture-speech interface on personal computer," Proc. 1997 Real World Computing Symposium, pp.1-7(Jan. 1997).

[Nakamura et al, 1996] S.Nakamura, T. Yamada, T. Takiguchi, K. Shikano, "Hands free speech recognition by a microphone array and HMM composition," Proc. Acoust. Soc. Am. and Acoust. Soc. Jap. Third Joint Meeting 1996 4pSC23, pp.1149 –1154 (Dec. 1996).

[Nitta et al, 1997] K. Nitta, O. Hasegawa et al, "An experimental multimodal disputation system; MrBengo," (In Japanese) IEICEJ Trans. D-II, J80-D-II, No.8, pp.2081-2087 (Aug. 1997).

[Pallet et al, 1992] D.S. Pallet, N.L.Dahlgren, J.G. Fiscus, W.M. Fisher, et al, "DARPA February 1992 ATIS benchmark test results," Proc. Speech and Natural Language Workshop, pp.15-27 (Feb. 1992).

[Singh et al, 1997] M.P. Singh, D.G. Bobrow, M.N. Huhns, M. King, et al, "The next big thing: Position statements," Proc. of 15th IJCAI, pp.1511-1520 (Aug. 1997).

[Takebayashi, 1994] Y. Takebayashi, "Spontaneous speech dialogue system *Tosburg-II*—Towards the user-centered multimodal interface," (In Japanese) IEICEJ Trans. D-II, J77-D-II No.8, pp.1417-1428 (Aug. 1994).

[Tanaka et al, 1996] K. Tanaka, S. Hayamizu, Y. Yamashita, K.Shikano, S.Itahashi, R. Oka, "Design and Data Collection for a Spoken Dialogue Database in the Real World Computing Program", Proc. Acoust. Soc. Am. and Acoust. Soc. Jap. Third Joint Meeting 1996 4aSC18,pp.1027 - 1030(Dec. 1996).

[Tanaka et al, 1997] K.Tanaka, H. Kojima, "A between-word distance calculationin a symbol domain and its applications to speech recognition", Proc. of International Conference on Neural Information Processing (ICONIP-97), pp.1107-1111 (Nov. 1997).

[Yamazaki,1995] Y.Yamazaki, "Toward cross-language global communi-cations- challenging research for spontaneous speech translation," Proccedings of Telecom 95 Forum, pp.3-7 (1995).

[Watanuki et al, 1995] K.Watanuki, K.Sakamoto, F.Togawa, "Multimodal interaction in human communication," IEICEJ Trans. Inf. & Syst. Vol. E78-D No.6, pp.609-615 (1995).