UBIQUITOUS SPEECH PROCESSING

Sadaoki Furui, Koji Iwano, Chiori Hori, Takahiro Shinozaki, Yohei Saito, and Satoshi Tamura

Tokyo Institute of Technology
Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
furui@cs.titech.ac.jp

ABSTRACT

In the ubiquitous (pervasive) computing era, it is expected that everybody will access information services anytime anywhere, and these services are expected to augment various human intelligent activities. Speech recognition technology can play an important role in this era by providing: (a) conversational systems for accessing information services and (b) systems for transcribing, understanding and summarizing ubiquitous speech documents such as meetings, lectures, presentations and voicemails. In the former systems, robust conversation using wireless handheld/hands-free devices in the real mobile computing environment will be crucial and as will multimodal speech recognition technology. To create the latter systems, the ability to understand and summarize speech documents is one of the key requirements. This paper presents technological perspectives and introduces several research activities being conducted from these standpoints in our research group.

1. UBIQUITOUS/WEARABLE COMPUTING ENVIRONMENT

Due principally to the technology of making computers smaller, more powerful and cheaper, the ubiquitous and wearable computing era is expected to come into being in the beginning of 21st century [1]. Making computers mobile and portable, exemplified by the present PDA (personal digital assistant) technology, is considered to be the transition phase to wearable computing [2]. Making computers more functional and smaller will generate not only quantitative changes but also qualitative changes in the way we use computers.

In the near future, a number of technical advances are anticipated: the transmission channel capacity of portable terminals will increase to several Mbps, the exchange of dynamic information will be possible in addition to that of simple characters and voice information, and computers, including portable equipment, will work together in autonomous collaboration [3]. These advances will give rise to sophisticated collaboration and coordination of human-machine systems based on autonomous protocol and information exchange between computers distributed everywhere. Indeed, the new characteristics of computing will greatly change the focus and approach of human computer interaction.

2. UBIQUITOUS SPEECH RECOGNITION

Speech is the primary, and the most convenient means of communication between people [4]. The conventional human-computer interface such as GUI, which assumes a keyboard, mouse, and bit-map display, is insufficient for the ubiquitous/wearable computing environment, especially for the wearables. Although handwritten character recognizers and keyboards that can be used with one hand have been developed as input devices for computers, speech recognition has recently received more interest. The main reason for this is that it permits both hands and eyes to be kept free and therefore is less restricted in its use and can achieve quicker communication. In addition, speech can convey not only linguistic information but also the emotion and identity of speakers.

In the history of speech recognition research, most speech recognizers and systems have been developed separately for each particular place and application, and each system has been shared by many people who come to the system or access it over the telephone. In speech transcription, common equipment has been adapted to each user's voice, but has never been considered as something to be carried around or controlled collaboratively with various other computers. In the ubiquitous/wearable computing environment, speech recognition will be, in most of cases, performed using a microphone and a networked computer worn and tailored to each person. Accordingly, future speech recognition systems will comprise very different structures than present systems. It is therefore important to investigate how to construct and use speech recognition systems in the new ubiquitous/wearable computing environment. Such an investigation will significantly change the foci of speech recognition research.

One of the important problems of state-of-the-art speech recognition technology is its robustness, i.e., the stability of its performance for the change of speakers, additive noise and distortions [5]. Some of the robustness problems can be easily solved if speech recognition by wearable computers is achieved. If everybody wears a device that recognizes his/her own voice, for example, the problem of speaker-to-speaker variability of voice can be eliminated. Models on the device can be incrementally adapted to the user using his/her daily utterances. Although it is possible to measure continuously the acoustic environmental conditions, such as noise, using the wearable itself, the process can be eliminated if computers are located everywhere and used to regularly measure the noise level and spectrum. By transmitting the noise information to the wearable recognizer, noise-adapted recognition can be easily performed.

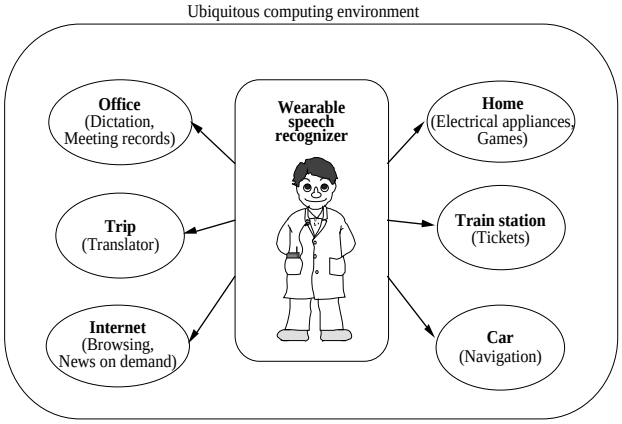

Fig. 1 Speech recognition in the ubiquitous/wearable computing environment.

If task-dependent information can be transmitted by radio to the wearable speech recognizer from the station of each service, such as train ticket machines, the recognizer can easily be adapted to the change of vocabulary and grammar according to the services as shown in Fig. 1. Take the case of speech recognition for car navigation as an example. Here, the vocabulary and grammar for car navigation are quickly transmitted to the wearable recognizer as soon as the user gets in the car, and input speech is recognized by the recognizer adapted to the user. In general, by storing general vocabulary and grammar in the speaker-adapted wearable

speech recognizers and quickly transmitting only a task-dependent part, task-adapted speech recognition can easily been performed.

3. CONVERSATIONAL SYSTEMS FOR ACCESSING INFORMATION SERVICES

3.1 Multimodal human/computer interface

Human beings favor the sensory dimensions of sight, sound, and touch as primary channels of communication. Machines that can accommodate these modes promise flexibilities and functionalities that transcend the traditional mouse and keyboard. Therefore, integration of multiple modalities in human/computer interfaces has long been viewed as a means for increasing ease of use. Recent examples include integration of modalities such as speech, gesture, gaze tracking and lip reading. Figure 2 shows the architecture of a multimodal task-oriented human/computer interaction system [6]. Each application defines a set of application programming interfaces (APIs) that can be invoked to cause different actions. The APIs determine the user's command vocabulary and grammar for both speech and gesture.

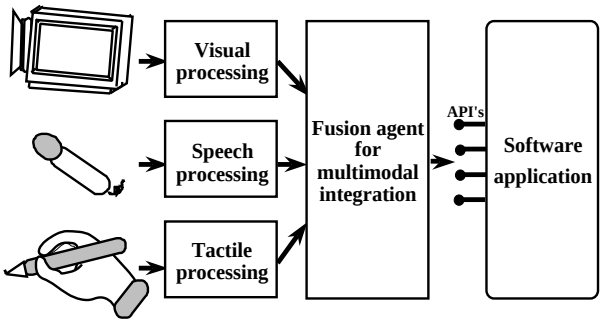

Fig. 2 Architecture of a multimodal human/computer interface.

Automatic lip reading has been investigated by many researchers as a complementary means to speech recognition. However it is still difficult to establish a robust and accurate approach for tracking the lip movements and extracting important features. In most of the systems, the shape of lips is measured based on pattern recognition techniques. Since the shape is sometimes difficult to measure correctly due to intensity and color variations, we are now trying to use statistical features based on optical flow, that is the measurements of image velocity vectors in the 2-D motion field calculated from spatio-temporal derivatives of image intensity or filtered versions of image [7] as features of lip movement.

We have tried several simple statistical features and found that the method using a combination of variances of vertical and horizontal components of flow vectors is effective. In order to increase the robustness, a gray-scale image covering the mouth area is passed through a smoothing process, and random noise is added before calculating the optical flow. The variance values are combined with cepstral coefficients and log-energy extracted from speech as well as their first and second order derivatives to construct a 41-dimensional feature vector for speech recognition. Connected digit utterances with the duration of 9minutes, spoken by a single male speaker, were used in the recognition experiment. The speech and images were recorded with frame rates of 100 frames/s and 15 frames/s, respectively, in an office environment. In order to cope with the frame rate difference, the optical flow was interpolated at every 10ms. Since the microphone was located approximately 1m away from the speaker, the speech SNR was approximately 11dB. Phoneme HMMs were trained using a training set and concatenated to build each digit model. Experimental results show that number of silence insertion errors in the connected digit utterances significantly decreased by incorporating the lip features to the acoustic features.

3.2 Designing a multimodal dialogue system for information retrieval

We have investigated a paradigm for designing multimodal dialogue systems [8]. An example task of the system was to retrieve particular information about different shops in the Tokyo Metropolitan area, such as their names, addresses and phone numbers. The system accepted speech and screen touching as input, and presented retrieved information on a screen display or by synthesized speech. The speech recognition part was modeled by the FSN (finite state network) consisting of keywords and fillers, both of which were implemented by the DAWG (directed acyclic word-graph) structure. The fillers accepted non-keywords/phrases occurring in spontaneous speech. A variety of dialogue strategies were designed and evaluated based on an objective cost function having a set of actions and states as parameters. Expected dialogue cost was calculated for each strategy, and the best strategy was selected according to the keyword recognition accuracy.

4. SYSTEMS FOR TRANSCRIBING, UNDERSTANDING AND SUMMARIZING UBIQUITOUS SPEECH DOCUMENTS

4.1 Transcription of presentations

Presentations at various conferences, such as the Acoustical Society of Japan (ASJ) meetings, and free presentations by voluntary subjects are recorded and transcribed in a project described in 5.1. Using these utterances, preliminary recognition experiments are being conducted at Tokyo Institute of Technology and Kyoto University. So far, experiments have been conducted using four presentations having 10 to 30 minutes each, all talking about speech [9].

In order to build a corpus appropriate for recognizing the lecture utterances, we collected text of transcribed lectures from the World Wide Web having roughly 76k sentences and 2M words. Spontaneous speech usually includes various filled pauses but they are not included in the lecture corpus. An effort was thus made to add filled pauses to the lecture corpus based on the statistical characteristics of the filled pauses, and then statistical language model was calculated for the most frequent 20k words. This language model is referred to LM1. Since all the presentations are talking about speech, a Japanese textbook titled "Speech Information Processing" authored by S. Furui was added to the transcribed lecture corpus to reduce OOV words, and language models were built in the same way as LM1. This language model is referred to LM2. Another language model referred to LM3 was built using the presentations transcribed in the project with the length of 18.7 hours consisting of 230k words. Two HMM acoustic models were employed: AM1 trained by using read speech uttered by 130 male speakers (40 hours) and AM2 trained using the presentations by 66 male speakers (12.4 hours).

Experimental results show that LM2 achieved almost the same OOV rate as LM3, and, although size of the presentation corpus is still very small, language and acoustic models based on the presentation corpus (LM3 and AM2) achieved much better recognition accuracy than any other models. This means that spontaneous speech corpus and task adaptation are crucial to build language and acoustic models for spontaneous speech recognition.

4.2 Transcription of discussions

Another recognition experiment has been performed using discussion speech in a broadcast TV program [9]. The following language models were built; LM1: constructed using broadcast news text over a 34-month period comprising 380k sentences [10], LM2: constructed using transcribed speech having 800k words collected from 159 days of the same program as the test set, and LM3: constructed by adding new words extracted

automatically from broadcast news articles with the topics similar to the test set to the LM2 using part-of-speech class N-gram language models.

Since cross-talk (double talk) frequently occurs during discussion and causes recognition errors, acoustic backing-off was applied to these periods, assuming that these periods can be automatically detected with some method. For these periods, acoustic scores were replaced by acoustic likelihood averaged over other periods and recognition was therefore performed merely based on linguistic scores.

Experimental results showed that test-set perplexity and OOV rate were both significantly reduced by using the transcription of the utterances in the same TV program, and that OOV rate was largely reduced by automatically adding new words. It was also found that the acoustic backing-off was effective to cope with the cross-talk problem.

4.3 Making minutes of meetings

We proposed a system for making minutes of meetings as one of the applications of speech recognition technology in the ubiquitous/wearable computing environment [1]. In this system, each attendant carries or wears a personalized computer to recognize his/her own utterances, and another computer connected by cable or radio to all the personalized computers is used as a meeting manager as shown in Fig. 3. Each recognizer is adapted to the attendant's voice, ways of speaking, and vocabulary using his/her utterances under various conditions. Therefore, relatively high recognition accuracy is expected even though speakers take turns during the meeting. The meeting manager computer collects structural information of the meeting, such as transition of topics and content and the speaker of each utterance, and transmits such information to each personalized recognizer. Thus, recognition accuracy can be improved and meanings of even fragmental utterances can be understood. Minutes of the meeting will be automatically made by collecting recognition results by each recognizer and other information, such as speaker change and the transition of topics, to the meeting manager. The minutes can also be summarized if necessary. In the future, if these results are displayed to all the participants' in real time, a computer-supported efficient meeting, a.k.a. CSCW (computer-supported cooperative work), will be realizable.

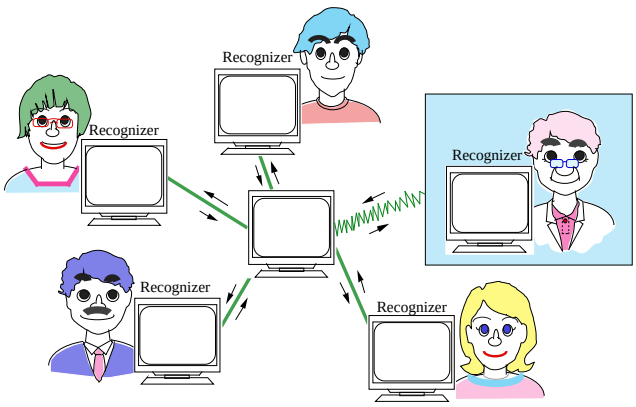

Fig. 3 Meeting synopsizing system using collaborative speech recognizers.

We have conducted a preliminary experiment using a broadcast TV discussion enjoined by politicians. Since recognition accuracy using present speech recognition technology is inadequate, we are now building an interactive system for making minutes [9]. The system first recognizes utterances and presents sentence hypotheses. When a user indicates recognition errors and provides corrections, the system modifies the acoustic and linguistic models and re-recognizes the input utterances. By iterating the process, minutes can be made with less labor than making them by hand. In order to speed up the iterative process, intermediate recognition results including speaker information are stored using a word graph.

4.4 Speech summarization

Transcribed spontaneous speech usually includes various kinds of redundant information. Therefore, a summarization technique to compress the information is expected to be useful in many applications, such as for closed captioning, indexing speech data for automatic retrieval, making abstracts of presentations, minutes of meetings and voicemails, and presenting information in news-on-demand systems.

We have proposed a method of automatically summarizing speech, sentence by sentence, by extracting a limited number of relatively important words from its automatic transcription according to a target compression ratio of the number of characters [11][12]. A word set that maximizes a summarization score consisting of a significance (topic) score and a confidence score of each word, a linguistic score (likelihood) of word concatenation, and a word modification score is extracted. The significance score is calculated for nouns based on the amount of information conveyed by each word, a flat score being given to words other than nouns. A posterior probability of each transcribed word, i.e. the ratio of the word hypothesis probability to that of all other hypotheses, is calculated using a word graph obtained by the decoder and used as a confidence measure. The confidence score is incorporated to give a penalty for acoustically as well as linguistically unreliable hypotheses. A trigram probability is used as a linguistic score. The word modification score is determined by a dependency structure obtained using Stochastic Dependency Context Free Grammar (SDCFG).

A set of words maximizing the summarization score is efficiently selected using a dynamic programming (DP) technique. In order to evaluate the summarization scores of various summarized sentences with different summarizing ratios obtained from the same original sentence, the summarization score is normalized according to the summarization ratio. Japanese broadcast news speech transcribed using our large vocabulary continuous speech recognition system was summarized. Experimental results show that all the scores consisting the summarization score are effective to make the summarized sentences readable and meaningful. The summarization method is now applied to presentation speech and extended to summarize articles consisting of multiple sentences.

5. JAPANESE NATIONAL PROJECTS

5.1 Project on spontaneous speech corpus and processing technology

The Science and Technology Agency Priority Program (Organized Research Combination System) entitled "Spontaneous Speech: Corpus and Processing Technology" was started in 1999 under the supervision of S. Furui [13]. The project will be conducted over a 5-year period in pursuit of the following three major themes (see Fig. 4):

1) Building a large-scale spontaneous speech corpus consisting of roughly 7M words with a total speech length of 800 hours. The majority of the recordings will be monologues such as lectures, presentations, and news commentaries. They will be manually given orthographic and phonetic transcription. One-tenth of the utterances (denoted as the "Core") will be tagged manually and used for constructing a morphological analysis program for automatically analyzing all of the 800-hour utterances. The Core will also be tagged with para-linguistic information including intonation.
2) Acoustic and linguistic modeling for spontaneous speech understanding and summarization using linguistic as well as para-linguistic information in speech.
3) Constructing a prototype of a spontaneous speech

summarization system.

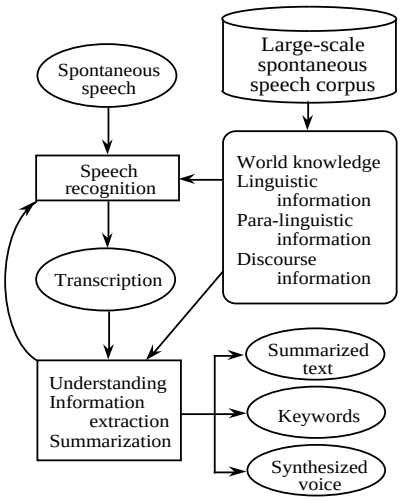

Fig. 4 Overview of the project on spontaneous speech corpus and processing technology

5.2 Project for diversified research on acoustic signals and sound fields

The Center for Integrated Acoustic Information Research (CIAIR) supervised by Prof. F. Itakura at Nagoya University was started in 1999 to create a center to carry out integrated research concerning the human-sound relation in terms of the following five questions [14]: 1) how can sound be spatially located? 2) how can the characteristics of sound be analyzed and synthesized? 3) how can speech and characters be transformed into each other? 4) how do humans communicate with speech? and 5) how do humans interpret sounds? Research organically integrating the fields of sound signal processing, spoken language processing and sound recognition, which have thus far been studied separately, is expected to break through the existing technical limits and open up a new research field concerning the human-sound interaction. Research activities will include various experiments and data collection with installations in actually running automobile systems to study three-dimensional reproduction of sounds and highly accurate speech recognition in an extremely noisy environment.

5.3 Project on language understanding and action control

Another national project related to speech understanding has recently started under supervision by Prof. H. Tanaka at Tokyo Institute of Technology. This project targets research on language understanding and its application to action control through building 3-D software robots in a virtual space in a computer and having them act based on natural spontaneous speech dialogue. Mechanisms of language understanding will be investigated from the viewpoint of actions of objects in response to the spontaneous speech dialogue. Multimodal dialogue will be one of the key research issues in this project. A preliminary investigation on a virtual actor system "Kairai", (Japanese for "puppet") , is reported in [15]. The research is concentrated on handling anaphora used to indicate objects or positions in the virtual world and on the study of ellipsis frequently used in command-style dialogues.

6. CONCLUSION

Speech recognition technology has made remarkable progress in the past 5-10 years. This progress has enabled various application systems to be developed using transcription and spoken dialogue technology. While we are still far from having a machine that converses with a human like a human, many important scientific advances have taken place, bringing us closer to the "Holy Grail" of automatic speech recognition and understanding by machine [4]. Speech recognition and understanding will become one of the key techniques for human computer interaction in the ubiquitous/wearable computing environment. To successfully use speech recognition in such an environment, every process such as start-stop control of recognition and adaptation to individuals and the surrounding environment must be performed without being noticed. Speech recognition should not be as it is in popular science fiction; instead it should be used unobtrusively, unconsciously and effortlessly [1]. It is also necessary to operate in a consistent manner no matter where the user goes.

The most important issue is how to make the speech recognition systems robust against acoustic and linguistic variation in spontaneous speech. In this context, a paradigm shift from speech recognition to understanding, where underlying messages of the speaker, that is, meaning/context that the speaker intended to convey, are extracted, instead of transcribing all the spoken words, will be indispensable. To reach such a goal, we need to have an efficient way of representing, storing, retrieving, and utilizing "world knowledge".

ACKNOWLEDGMENT

The authors wish to express their thanks to Professor Tatsuya Kawahara at Kyoto University for his contribution to the national projects and for several valuable comments and fruitful discussions. The authors would also like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast news and discussion database.

REFERENCES

[1] S. Furui: "Speech recognition technology in the ubiquitous/wearable computing environment", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Istanbul, pp. 3735-3738 (2000)
[2] A. Pentland: "Wearable intelligence", Scientific American, Vol. 276, No. 11 (Nov. 1998)
[3] M. Weiser: "The computer for the twenty-first century", Scientific American, pp. 94-104 (1991)
[4] B.-H. Juang and S. Furui: "Automatic recognition and understanding of spoken language – A first step towards natural human-machine communication", Proc. IEEE, 88, 8, pp. 1142-1165 (2000)
[5] S. Furui: "Steps toward natural human-machine communication in the 21st century," Proc. COST249 Workshop, "Voice Operated Telecom Services," Gent, Belgium, pp. 17-24 (2000).
[6] I. Marsic, A. Medl and J. Flanagan: "Natural communication with information systems", Proc. IEEE, 88, 8, pp. 1354-1366 (2000)
[7] J. L. Barron, D. J. Fleet and S. S. Beauchemin: "Systems and Experiment: Performance of Optical Flow Techniques", International Journal of Computer Vision, 12, 1, pp. 43-77 (1994)
[8] S. Furui and K. Yamaguchi: "Designing a multimodal dialogue system for information retrieval", Proc. Int. Conf. Spoken Language Processing, Sydney, pp. 1191-1194 (1998)
[9] T. Shinozaki, Y. Saito, C. Hori and S. Furui: "Toward spontaneous speech recognition", IEICE/ASJ Tech. Rep., SP2000-96 (2000) (in Japanese)
[10] K. Ohtsuki, S. Furui, N. Sakurai, A. Iwasaki and Z.-P. Zhang: "Improvements in Japanese broadcast news transcription", DARPA Broadcast News Workshop, Virginia, pp. 231-236 (1999)
[11] C. Hori and S. Furui; "Automatic speech summarization based on word significance and linguistic likelihood", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Istanbul, pp. 1579-1582 (2000)
[12] C. Hori and S. Furui: "Improvements in automatic speech summarization and evaluation methods", Proc. Int. Conf. Spoken Language Processing, Beijing, pp. IV-326-329 (2000)
[13] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki and T. Ohdaira: "Toward the realization of spontaneous speech recognition – Introduction of a Japanese priority program and preliminary results –", Proc. Int. Conf. Spoken Language Processing, Beijing, pp. III-518-521 (2000)
[14] http://www.ciair.coe.nagoya-u.ac.jp/index.html
[15] Y. Shinyama, T. Tokunaga and H. Tanaka: "'Kairai' – Software robots understanding natural language", Proc. 3rd Int. Workshop on Human-Computer Conversation, pp. 196-206 (2000)