論文 / 著書情報
Article / Book Information

Title	Steps toward natural human-machine communication in the 21st Century
Authors	SADAOKI FURUI
Citation	ISCA Workshop on Voice Operated Telecom Service, Vol. , No. , pp. 17-24
Pub. date	2000, 5

STEPS TOWARD NATURAL HUMAN-MACHINE COMMUNICATION IN THE 21ST CENTURY

Sadaoki Furui

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

ABSTRACT

This paper first overviews the state of the art in automatic speech recognition and understanding technology including the introduction of research being undertaken at Tokyo Institute of Technology. It then discusses the limitations of the current technology and points out challenges that lie ahead of us, such as how to create flexible and robust acoustic/linguistic modeling and decoding techniques. For processing spontaneous speech, indispensable will be a paradigm shift from speech recognition to understanding where underlying messages of the speaker are extracted, instead of transcribing all the spoken words. Building a large corpus of spontaneous speech to construct reliable acoustic and linguistic models is also crucial. 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 initial years of the 21st century. In such an environment, speech recognition will be widely used as one of the principal methods of human-computer interaction. This paper thus also addresses how to construct and use speech recognition systems in the ubiquitous/wearable computing environment.

1. INTRODUCTION

The field of automatic speech recognition has witnessed a number of significant advances in the past 5-10 years, spurred on by advances in signal processing, algorithms, computational architectures, and hardware. These advances include the widespread adoption of a statistical pattern recognition paradigm, a data-driven approach which makes use of a rich set of speech utterances from a large population of speakers, the use of stochastic acoustic and language modeling, and the use of dynamic programming-based search methods.

A series of (D)ARPA projects have been a major driving force of the recent progress in research on large-vocabulary, continuous-speech recognition. Specifically, actively investigated were dictation of speech reading newspapers, such as north America business newspapers including the Wall Street Journal (WSJ), and conversational speech recognition using an Air Travel Information System (ATIS) task. More recent DARPA programs are the broadcast news dictation and natural conversational speech recognition using Switchboard and Call Home tasks. Research on human-computer dialogue systems, the Communicator program, has also started [1]. Various other systems have been actively investigated in the US, Europe and Japan stimulated by these DARPA projects. Most of these can be classified into either dictation systems or human-computer dialogue systems.

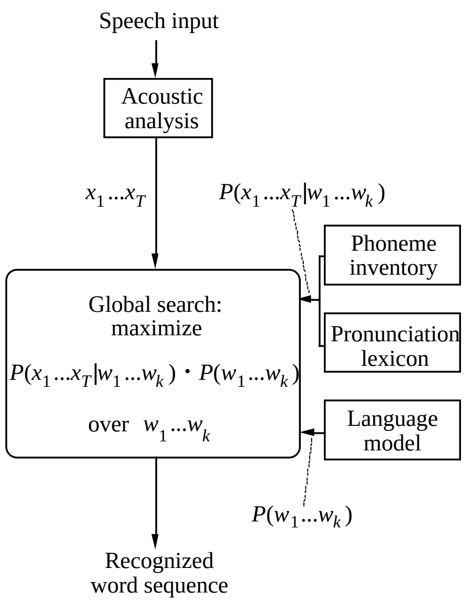

Fig. 1 - Mechanism of state-of-the-art speech recognizers.

Figure 1 shows a mechanism of state-of-the-art speech recognizers [2]. Common features of these systems are the use of cepstral parameters and their regression coefficients as speech features, triphone HMMs as acoustic models, vocabularies of several thousand or several tens of thousand entries, and stochastic language models such as bigrams and trigrams. Such methods have been applied not only to English but also to French, German, Italian and Japanese. Although there are several language-specific characteristics, similar recognition results have been obtained.

2. SPEECH RECOGNITION RESEARCH AT TOKYO INSTITUTE OF TECHNOLOGY

2.1 Broadcast News Dictation

We have been developing a Japanese broadcast-news speech transcription system [3] as part of a joint research project with NHK broadcasting. The goal is the closed-captioning of TV programs. The language models were constructed using broadcast-news manuscripts taken from NHK TV news broadcasts made between July 1992 and May 1996. The broadcasts comprised roughly 500k sentences and 22M words (morphemes). To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words using a morphological analyzer because Japanese sentences are written without spaces between words. Since many Japanese words have multiple readings and correct readings can only be decided according to context, we constructed a language model in which a word with multiple readings is split into different language model entries according to those readings. We also introduced filled-pause modeling into the language model. A word-frequency list was derived for the news manuscripts, and the 20k most frequently used words were selected as vocabulary words. This 20k vocabulary covered approximately 98% of the words in the manuscripts. We calculated bigrams and trigrams and estimated unseen n-grams using Katz's back-off smoothing method.

The feature vector consisted of 16 cepstral coefficients, normalized logarithmic power, and their delta features (regression coefficients). The total number of parameters in each vector was 34. Cepstral coefficients were normalized by the cepstral mean subtraction (CMS) method. The acoustic models were gender-dependent shared-state triphone HMMs and were designed using tree-based clustering. They were trained using phonetically-balanced sentences and dialogues read by 53 male speakers and 56 female speakers. The contents were completely different from the broadcast-news task. The total number of training utterances was 13,270 for the males and 13,367 for the females, and the total length of the training data was approximately 20 hours for each gender. The total number of HMM states was approximately 2,000 for each gender, and the number of Gaussian mixture components per state was four.

Clean speech data consisting of 50 male and 50 female utterances with no background noise were extracted from TV news broadcast in July 1996 and used as evaluation utterances. The male and female sets included utterances by five or six speakers, respectively. All utterances were manually segmented into sentences. The mean word error rates for male and female utterances were 14.2% and 12.9%, respectively.

2.2 Designing a Multimodal Dialogue System for Information Retrieval

We have recently investigated a paradigm for designing multimodal dialogue systems [4]. 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 as shown in Fig. 2. 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 number of keywords was 306, consisting of district names and business names. The fillers accepted roughly 100,000 non-keywords/phrases occuring 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.

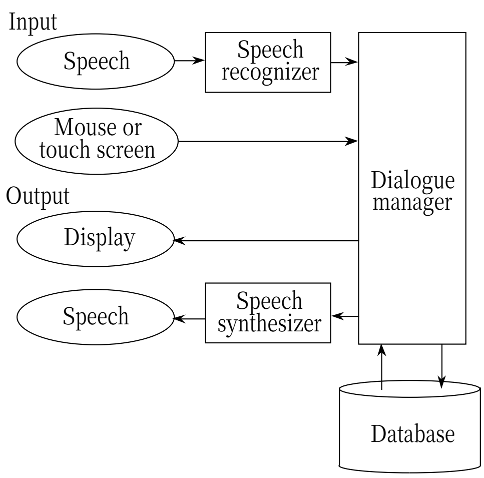

Fig. 2 - Multimodal dialogue system structure for information retrieval.

3. TOWARD ROBUST SPEECH RECOGNITION

3.1 Automatic Adaptation

Ultimately, speech recognition systems should be capable of robust, speaker-independent or speaker-adaptive, continuous speech recognition. Figure 3 shows the main causes of acoustic variation in speech [5]. 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 other aspects. Figure 4 presents the main methods for making speech recognition systems robust against voice variation. Also important is for the systems to impose few restrictions on tasks and vocabulary. Developing automatic adaptation techniques is essential to resolve these problems.

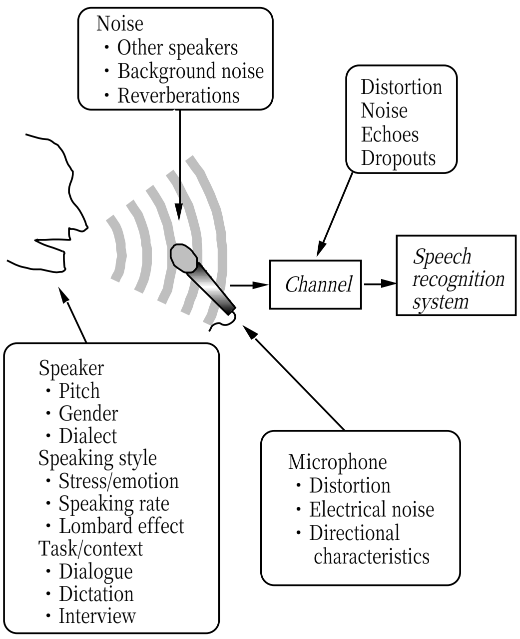

Fig. 3 - Main causes of acoustic variation in speech.

Extraction and normalization of (adaptation to) voice individuality is one of the most important issues in this regard. A small percentage of people occasionally cause systems to produce exceptionally low recognition rates. This is an example of the "sheep and goats" phenomenon. Speaker adaptation (normalization) methods can usually be classified into supervised and unsupervised methods. Unsupervised, online, instantaneous/incremental adaptation is ideal, since the system works as if it were a speaker-independent system, and it performs increasingly better as it is used. However, since we have to adapt many phonemes using a limited size of utterances including only a limited number of phonemes, it is crucial to use reasonable modeling of speaker-to-speaker variability or constraints. Modeling of the mechanism of speech production is expected to provide a useful modeling of speaker-to-speaker variability [6] .

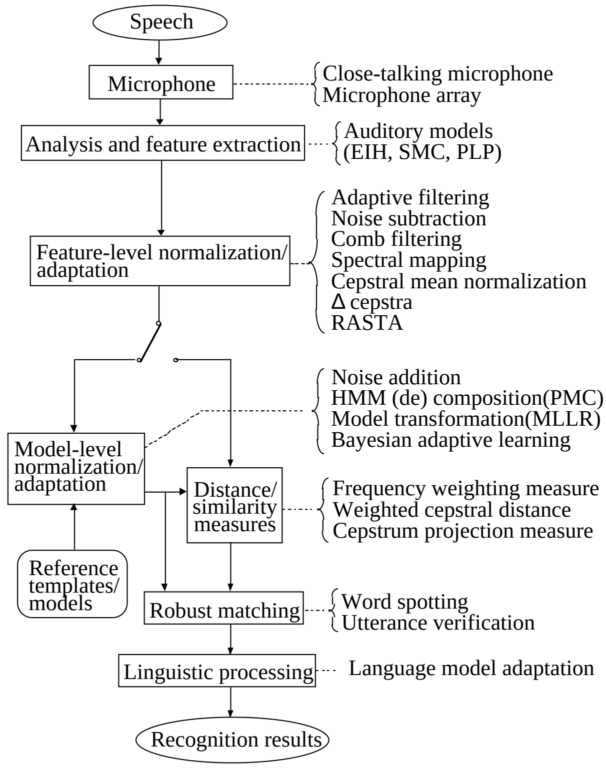

Fig. 4 - Main methods to contend with voice variation in speech recognition.

3.2 Online Incremental Speaker Adaptation

In many applications of speech recognition, speakers change frequently, new speakers appear, and each of them utters a series of several sentences. In such a situation, an unsupervised and online adaptation method, which uses the unknown utterance itself for adaptation, is expected to be effective. The adaptation should also work incrementally within a segment in which one speaker utters several sentences. To create such a system, we must ensure that speaker change is detected automatically and correctly.

We have proposed an online, unsupervised, instantaneous and incremental speaker adaptation method combined with automatic detection of speaker changes for broadcast news dictation[7]. The MLLR [8] -MAP [9] and VFS (vector-field smoothing) [10] methods were instantaneously and incrementally carried out for each utterance. The speaker change is detected by comparing likelihoods using speaker-independent and speaker-adaptive GMMs (Gaussian mixture models). A single-state speaker-independent GMM with 64 Gaussian distributions is constructed in the training stage using the same utterances that were used to construct the speaker-independent phone HMM.

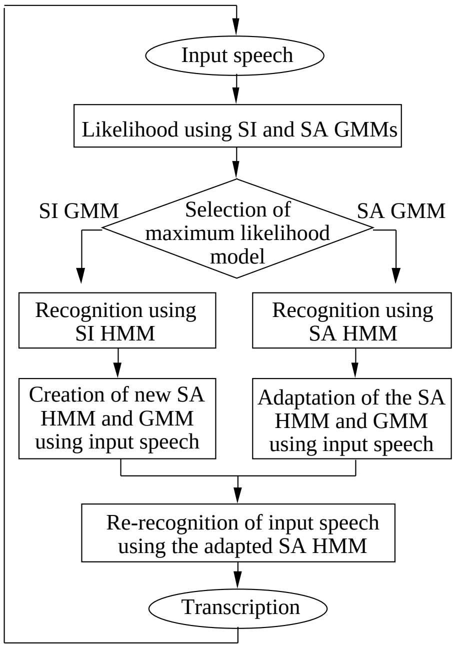

Fig. 5 - Speaker adaptation process
(SI: speaker independent; SA: speaker adapted).

The adaptation process is shown in Fig. 5. For the first input utterance, the speaker-independent model is used for both recognition and adaptation, and the first speaker-adapted HMM and GMM are created. For the second input utterance, the likelihood value of the utterance given the speaker-independent GMM and that given the speaker-adapted GMM are calculated and compared. If the former value is larger, the utterance is considered to be the beginning of a new speaker, and another speaker-adapted HMM as well as GMM are created. Otherwise, the existing speaker-adapted HMM and GMM are incrementally adapted. For the succeeding input utterances, speaker changes are detected in the same way by comparing the acoustic likelihood values of each utterance obtained from the speaker-independent GMM and speaker-adapted GMMs. If the speaker-independent GMM yields a larger likelihood than any of the speaker-adapted GMMs, a speaker change is detected and new speaker-adapted HMM and GMM are constructed. Since GMM has only one state, it is adapted using a transformation matrix obtained by a single-cluster MLLR adaptation procedure applied to the phone HMM although HMM is actually adapted using seven phonemic clusters.

The same utterances used in the experiments described in Subsection 2.2 were used in the evaluation. Experimental results show that the adaptation reduced the word error rate by 10.0% relative to the speaker-independent models.

4. PERSPECTIVES OF LANGUAGE MODELING

4.1 Language Modeling for Spontaneous Speech Recognition

Another important issue for speech recognition is how to create language models (rules) for spontaneous speech. 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 words, ungrammatical sentences, disfluency, partial words, repairs, hesitations, and repetitions. It is crucial to develop robust and flexible parsing algorithms that match the characteristics of spontaneous speech. A paradigm shift from the present transcription-based approach to a detection-based approach will be important to resolve such problems [11]. How to extract contextual information, predict users' responses, and focus on key words are very important issues.

Stochastic 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. It would also be useful to integrate unification grammars and context-free grammars for efficient word prediction. 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 do the latter, however, use a different linguistic style. Adaptation of linguistic models according to tasks, topics and speaking styles is a very important issue since collecting a large linguistic database for every new task is difficult and costly.

4.2 Detection-Based Paradigm for Speech Understanding

A detection-based system consists of detectors, each of which aims at detecting the presence of a prescribed event, such as a phoneme, a word, a phrase, and a linguistic notion such as an expression of travel destination. The detector uses a model for the event and an anti-model that provides contrast to the event. It follows the Neymann-Pearson lemma in that the likelihood ratio is used as the test statistic against a threshold.

The following issues need to be addressed in this formulation [11]:

- How to train the models and anti-models? The idea of discriminative training can be applied using the verification error as the optimization criterion.
- How to choose detection units? Reasonable choices are words and key phrases.
- How to include language models and event context/constraints which can help raise the system performance in the integrated search after the detectors propose individual decisions?

Several simple implementations of this paradigm have shown promise in dealing with natural utterances containing many spontaneous speech phenomena [12].

4.3 Message-Driven Speech Recognition and Understanding

State-of-the-art automatic speech recognition systems employ the criterion of maximizing $P(W|X)$, where W is a word sequence, and X is an acoustic observation sequence. This criterion is reasonable for dictating read speech.

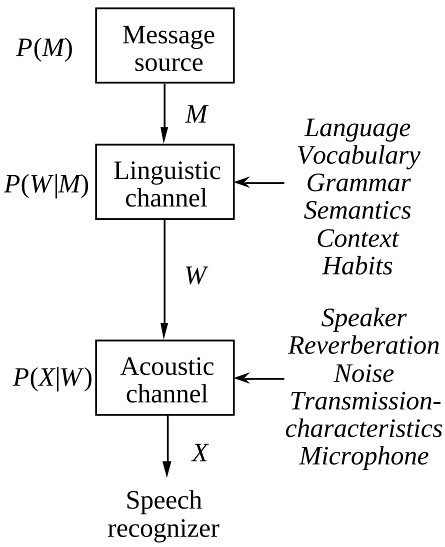

Fig. 6 - A communication-theoretic view of speech generation and recognition.

However, the ultimate goal of automatic speech recognition is to extract the underlying messages of the speaker from the speech signals. Hence we need to model the process of speech generation and recognition as shown in Fig. 6 [13], where M is the message (content) that a speaker intended to convey. There is also a possibility to give feedback from the "understanding module" to the speech recognition module such that decoding hypotheses can be properly adjusted and, hopefully, converge to the most correct word sequence as well as the most correct understanding of the utterance.

According to this model, the speech recognition process is represented as the maximization of the following a posteriori probability [3, 13]:

$$\max_M P(M \mid X) = \max_M \sum_W P(M \mid W) P(W \mid X). \qquad (1)$$

Using Bayes' rule, Eq. (1) can be expressed as

$$\max_M P(M \mid X) = \max_M \sum_W \frac{P(X \mid W) P(W \mid M) P(M)}{P(X)}. \qquad (2)$$

For simplicity, we can approximate the equation as

$$\max_M P(M \mid X) = \max_{M,W} \frac{P(X \mid W) P(W \mid M) P(M)}{P(X)}. \qquad (3)$$

$P(X|W)$ is calculated using hidden Markov models in the same way as in usual recognition processes. We assume that $P(M)$ has a uniform probability for all M. Therefore, we only need to consider further the term $P(W|M)$. We assume that $P(W|M)$ can be expressed as follows:

$$P(W \mid M) \approx P(W)^{1-\lambda} P'(W \mid M)^{\lambda}, \qquad (4)$$

where λ, $0 \le \lambda \le 1$, is a weighting factor. $P(W)$, the first term of the right-hand side, represents a part of $P(W|M)$ that is independent of M and can be given by a general statistical language model. $P'(W|M)$, the second term of the right-hand side, represents the part of $P(W|M)$ that depends on M. We consider that M is represented by a co-occurrence of words based on the distributional hypothesis by Harris [14]. Since this approach formulates $P'(W|M)$ without explicitly representing M, it can use information about the speaker's message M without being affected by the quantization problem of topic classes. This new formulation of speech recognition was applied to the Japanese broadcast news dictation, and it was found that word error rates were slightly reduced by this method.

4.4 Information Extraction from Speech

Summarizing transcribed speech is useful for retrieving or indexing speech data. We investigated a method for extracting topic words from broadcast news. Since we found in our preliminary experiments using human subjects that most of the topic words were nouns, we extracted only nouns in the speech recognition results on the basis of a significance measure [3, 13]. The extracted topic-words were compared with "true" topic-words, which were given by three human subjects. The results are shown in Figure 7. When the top five topic-words were chosen (recall = 13%), 87% of them were correct on average.

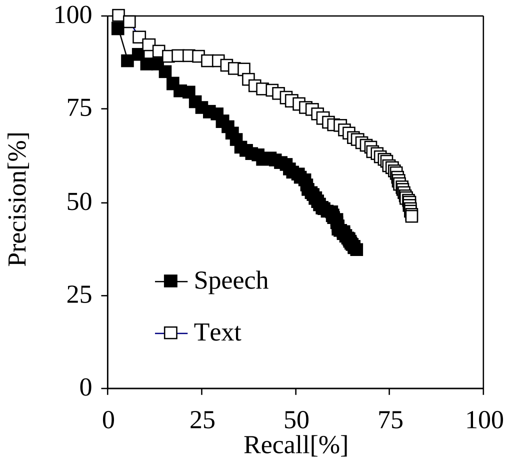

Fig. 7 – Topic-word extraction results.

4.5 Speech Summarization

In closed captioning, the number of words presented on the TV screen for a professional announcers' broadcast news speech sometimes exceeds the number of words that people can read and understand. In addition, transcribed speech usually includes some redundant information. Therefore, a summarization technique is desired to be applied to the closed captioning system to compress the information of speech. Furthermore, summarization for transcribed speech is also expected to be useful for indexing speech data for automatic retrieval and for making abstracts of presentations and minutes of meetings.

We have proposed a new method of automatically summarizing speech by extracting a limited number of relatively important words from its automatic transcription according to a target compression ratio for the number of characters [15]. To determine a word set to be extracted, we define a summarization score consisting of a topic score (significance measure) of words and a linguistic score (likelihood) of the word concatenation. The topic score is only calculated for nouns using the significance measure employed in the experiment in Subsection 4.4. A flat score is given to words other than nouns. A set of words maximizing the score is efficiently selected using a dynamic programming (DP) technique.

Japanese broadcast news speech transcribed using our large vocabulary continuous speech recognition system was summarized. A trigram language model for summarization was built using text from Mainichi newspapers published in 1996, comprising approximately 1.7M sentences consisting of 29M words. This was done because we consider newspaper text to usually be more compact and simpler than broadcast news text and therefore more appropriate for building language models for summarization than the latter text.

As a result, 86% of important words in the original speech were correctly included in the summarizing sentences and 72% of the summarizing sentences could retain the meanings of the original speech under the 60-70% summarization condition.

5. SPEECH RECOGNITION IN THE UBIQUITOUS/WEARABLE COMPUTING ENVIRONMENT

5.1 Ubicomp/Wearable Environment

The continuing progress in hardware and software development technologies have lead to the augmentation of computer performance at such a rapid pace that it improves several hundred times in every 10-year period. Resultingly, computers are getting smaller, more powerful and cheaper. Regardless of whether and to what degree they are noticed by users, computers will proliferate into every facet of our daily lives. People will actually walk through their day-to-day lives wearing several computers at a time. Thus, making computers mobile and portable, exemplified by the present PDA (personal digital assistant) technology, is considered to be the transition phase to wearable computing [16]. 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, various computers including portable equipment existing everywhere will work together in autonomous collaboration [17]. Indeed, the new characteristics of computing will greatly change the focus and approach of human computer interaction.

The transmission channel capacity of portable terminals will be easily expanded to the level of several Mbps, taking advantage of the technological progress available through, for example, MMAC (Multimedia Mobile Access Communication) systems. The exchange of dynamic information will be possible in addition to that of simple characters and voice information. In turn, this will give rise to sophisticated collaboration and coordination of human-machine systems based on autonomous protocol and information exchange between computers distributed everywhere.

5.2 Features of Speech Recognition in the Ubicomp/Wearable Environment

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 assumed to be shared by many people who come to the system or access it over the telephone. In speech dictation, 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. 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.

Some of the robustness problems described in Subsection 3.1 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. Moreover, if everybody always wears a personal microphone near his/her mouth, the signal-to-noise ratio can be largely improved, compared with using a distant microphone. A system is proposed for making minutes of meetings as one of the applications of speech recognition technology in the Ubicomp/Wearable environment [18]. 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.

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. Take the case of speech recognition for car navigation as an example. Here, input speech can be recognized by the recognizer adapted to the user more easily if the vocabulary and grammar for car navigation are quickly transmitted to the wearable recognizer as soon as the user gets in the car. 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.

6. JAPANESE NATIONAL PROJECT ON SPONTANEOUS SPEECH CORPUS AND PROCESSING TECHNOLOGY

As described in Subsection 4.1, one of the most difficult issues for speech recognition is how to build language models for spontaneous speech. For this purpose, we need to have large spontaneous speech corpora. In the United States, "Switchboard" and "Call Home" spontaneous speech corpora have been built since several years ago. In Japan, we have recently started a Science and Technology Agency Priority Program (Organized Research Combination System) entitled "Spontaneous Speech: Corpus and Processing Technology" supervised by S. Furui. The principal organizations working together to conduct this project are the National Language Research Institute under the Ministry of Education, the Communication Research Laboratory under the Ministry of Posts and Telecommunications, and Tokyo Institute of Technology.

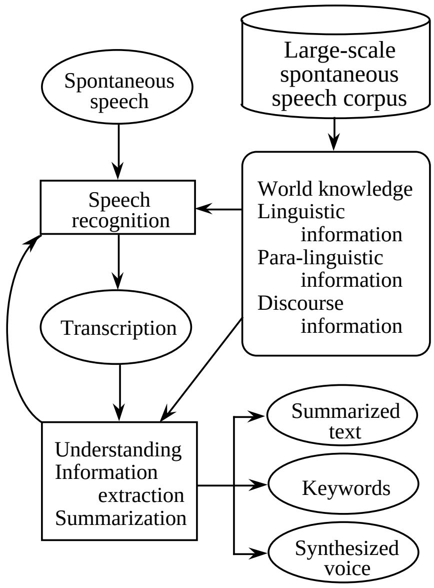

Fig. 8 - Overview of the national project.

The project will be conducted over a 5-year period under the following three major themes as in Fig. 8

1) Building a large-scale spontaneous speech corpus consisting of roughly 7M words with the total speech length of 1000 hours. Mainly recorded will be monologues such as lectures, presentations, and news commentaries. They will be manually given orthographic and phonetic transcription. One-tenth of the utterances ("Core") will be tagged manually and used for making a morphological analysis program for automatically analyzing all of the 1000-hour utterances. The Core will also be tagged with para-

linguistic information including intonation.
2) Acoustic and linguistic modeling for spontaneous speech understanding using linguistic as well as para-linguistic information in speech.
3) Investigating spontaneous speech summarization technology.

The technology created in this project is expected to be applicable to wide areas such as indexing of speech data (broadcast news, etc.) for information extraction and retrieval, transcription of lectures, preparing minutes of meetings, closed captioning, and aids for the handicapped.

7. CONCLUSION: NEED FOR A PARADIGM SHIFT

Speech recognition technology has made remarkable progress in the past 5-10 years. This progress has enabled various application systems to have been developed using dictation and spoken dialogue technology. One of the most important applications is information extraction and retrieval. Using speech recognition technology, broadcast news can be automatically indexed, producing a wide range of capabilities for browsing news archives interactively. Since speech is the most natural and efficient communication method for humans, automatic speech recognition will continue to find applications such as meeting/conference summarization, automatic closed captioning, human computer interfaces and interpreting telephony. It is expected that the speech recognizer will become the main input device of the "wearable" computers that are now actively investigated. In order to make these applications a reality, many problems must be solved. The most important issue is how to make the speech recognition systems robust against acoustic and linguistic variation in 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.

REFERENCES

[1] http://fofoca.mitre.org

[2] H. Ney: "Corpus-based statistical methods in speech and language processing", in Corpus-based Methods in Language and Speech Processing (S. Young and G. Bloothooft, Ed.), pp. 1-26 (1997)

[3] 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)

[4] S. Furui and K. Yamaguchi: "Designing a multimodal dialogue system for information retrieval", Proc. Int. Conf. Spoken Language Processing, Sydney, pp. 1191-1194 (1998)

[5] S. Furui: "Recent advances in robust speech recognition", Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, pp. 11-20 (1997)

[6] S. Furui: "Future directions in speech information processing", Proc. 16th ICA and 135th Meeting ASA, Seattle, pp. 1-4 (1998)

[7], S. Furui, Z.-P. Zhang and K. Ohtsuki: "On-line incremental speaker adaptation for broadcast news transcription, Proc. IEEE Automatic Speech Recognition and Understanding Workshop, Keystone, pp. 165-168 (1999)

[8] C. J. Leggetter and P. C. Woodland: "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, pp. 171-185 (1995)

[9] J. -L. Gauvain and C.-H. Lee: "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains" IEEE Trans. on Speech and Audio Processing, 2, 2, pp. 291-298 (1994)

[10] K. Ohkura, M. Sugiyama and S. Sagayama: "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs", Proc. Int. Conf. Spoken Language Processing, Banff, pp. 369-372 (1992)

[11] B.-H. Juang and S. Furui: "Automatic recognition and understanding of spoken language – A first step towards natural human-machine communication", Proc. IEEE (to be published)

[12] T. Kawahara, C. H. Lee and B. H. Juang: "Combining key-phrase detection and subword-based verification for flexible speech understanding", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Munich, pp. 1159-1162 (1997)

[13] K. Ohtsuki, S. Furui, A. Iwasaki and N. Sakurai: "Message-driven speech recognition and topic-word extraction", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Phoenix, pp. 625-628 (1999)

[14] Z. S. Harris: "Co-occurrence and transformation in linguistic structure", Language, 33, pp. 283-340 (1957)

[15] C. Hori and S. Furui; "Automatic speech summarization based on word significance and linguistic likelihood", Technical Report of IEICE, SP99-110 (1999)

[16] A. Pentland: "Wearable intelligence", Scientific American, Vol. 276, No. 11 (Nov. 1998)

[17] M. Weiser: "The computer for the twenty-first century", Scientific American, pp. 94-104 (1991)

[18] S. Furui: "Speech recognition under ubiquitous/wearable computing environment", Technical Report of IEICE, SP99-114 (1999)