AUTOMATIC SPEECH RECOGNITION AND ITS APPLICATION TO INFORMATION EXTRACTION

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 describes recent progress and the author's perspectives of speech recognition technology. Applications of speech recognition technology can be classified into two main areas, dictation and human-computer dialogue systems. In the dictation domain, the automatic broadcast news transcription is now actively investigated, especially under the DARPA project. The broadcast news dictation technology has recently been integrated with information extraction and retrieval technology and many application systems, such as automatic voice document indexing and retrieval systems, are under development. In the human-computer interaction domain, a variety of experimental systems for information retrieval through spoken dialogue are being investigated. In spite of the remarkable recent progress, we are still behind our ultimate goal of understanding free conversational speech uttered by any speaker under any environment. This paper also describes the most important research issues that we should attack in order to advance to our ultimate goal of fluent speech recognition.

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, 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 were actively investigated. 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 US, Europe and Japan stimulated by DARPA projects. Most of them can be classified into either dictation systems or human-computer dialogue systems.

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 ten 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, Spanish, Chinese and Japanese. Although there are several language-specific characteristics, similar recognition results have been obtained.

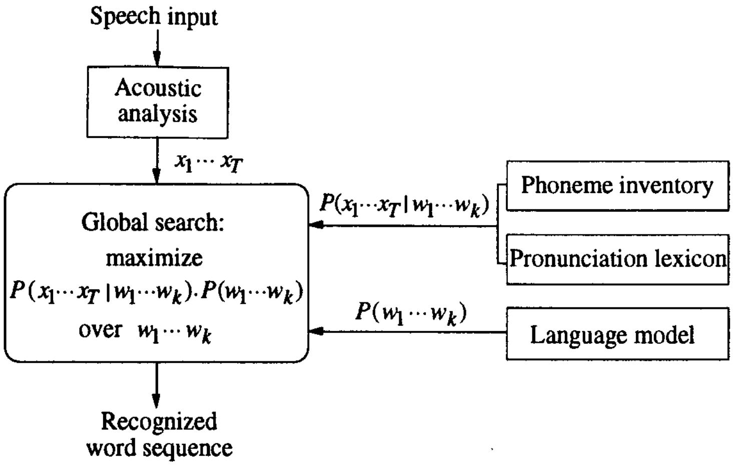

Fig. 1 - Mechanism of state-of-the-art speech recognizers.

The remainder of this paper is organized as follows. Section 2 describes recent progress in broadcast news dictation and its application to information extraction, and Section 3 describes human-computer dialogue systems. In spite of the remarkable recent progress, we are still far behind our ultimate goal of understanding free conversational speech uttered by any speaker under any environment. Section 4 describes how to increase the robustness of speech recognition, and Section 5 describes perspectives of linguistic modeling for spontaneous speech recognition/ understanding. Section 6 concludes the paper.

2. BROADCAST NEWS DICTATION AND INFORMATION EXTRACTION

2.1 DARPA Broadcast News Dictation Project

With the introduction of the broadcast news test bed to the DARPA project in 1995, the research effort took a profound step forward. Many of the deficiencies of the WSJ domain were resolved in the broadcast news domain [3]. Most importantly, the fact that broadcast news is a real-world domain of obvious value has lead to rapid technology transfer of speech recognition into other research areas and applications. Since the variations in speaking style and accent as well as in channel and environment conditions are totally unconstrained, broadcast news is a superb *stress test* that requires new algorithms to work across widely varying conditions. Algorithms need to solve a specific problem without degrading any other condition. Another advantage of this domain is that news is easy to collect and the supply of data is boundless. The data is *found speech*; it is completely uncontrived.

2.2 Japanese Broadcast News Dictation System

We have been developing a large-vocabulary continuous-speech recognition (LVCSR) system for Japanese broadcast-news speech transcription [4][5]. This is a part of a joint research with the NHK broadcast company whose goal is the closed-captioning of TV programs. The broadcast-news manuscripts that were used for constructing the language models were taken from the period between July 1992 and May 1996, and comprised roughly 500k sentences and 22M words. To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words by using a morphological analyzer since Japanese sentences are written without spaces between words. 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 covers about 98% of the words in the broadcast-news manuscripts. We calculated bigrams and trigrams and estimated unseen n-grams using Katz's back-off smoothing method.

Japanese text is written by a mixture of three kinds of characters: Chinese characters (Kanji) and two kinds of Japanese characters (Hira-gana and Kata-kana). Most Kanji have multiple readings, and correct readings can only be decided according to context. Conventional language models usually assign equal probability to all possible readings of each word. This causes recognition errors because the assigned probability is sometimes very different from the true probability. We therefore constructed a language model that depends on the readings of words in order to take into account the frequency and context-dependency of the readings. Broadcast news speech includes filled pauses at the beginning and in the middle of sentences, which cause recognition errors in our language models that use news manuscripts written prior to broadcasting. To cope with this problem, we introduced filled-pause modeling into the language model.

Table 1 - Experimental results of Japanese broadcast news dictation with various language models (word error rate [%])

Language model	Evaluation sets			
	m/c	m/n	f/c	f/n
LM1	17.6	37.2	14.3	41.2
LM2	16.8	35.9	13.6	39.3
LM3	14.2	33.1	12.9	38.1

News speech data, from TV broadcasts in July 1996, were divided into two parts, a clean part and a noisy part, and were separately evaluated. The clean part consisted of utterances with no background noise, and the noisy part consisted of utterances with background noise. The noisy part included spontaneous speech such as reports by correspondents. We extracted 50 male utterances and 50 female utterances for each part, yielding four evaluation sets; male-clean (m/c), male-noisy (m/n), female-clean (f/c), female-noisy (f/n). Each set included utterances by five or six speakers. All utterances were manually segmented into sentences. Table 1 shows the experimental results for the baseline language model (LM1) and the new language models. LM2 is the reading-dependent language model, and LM3 is a modification of LM2 by filled-pause modeling. For clean speech, LM2 reduced the word error rate by 4.7 % relative to LM1, and LM3 model reduced the word error rate by 10.9 % relative to LM2 on average.

2.3 Information Extraction in the DARPA Project

News is filled with events, people, and organizations and all manner of relations among them. The great richness of material and the naturally evolving content in broadcast news has leveraged its value into areas of research well beyond speech recognition. In the DARPA project, the Spoken Document Retrieval (SDR) of TREC and the Topic Detection and Tracking (TDT) program are supported by the same materials and systems that have been developed in the broadcast news dictation arena [3]. BBN'sRough'n'Reddy system extracts structural features of broadcast news. CMU's Informedia [6], MITRE's Broadcast Navigator, and SRI's Maestro have all exploited the multi-media features of news producing a wide range of capabilities for browsing news archives interactively. These systems integrate various diverse speech and language technologies including speech recognition, speaker change detection, speaker identification, name extaction, topic classification and information retrieval.

2.4 Information Extraction from Japanese Broadcast News

Summarizing transcribed news speech is useful for retrieving or indexing broadcast news. We investigated a method for extracting topic words from nouns in the speech recognition results on the basis of a significance measure [4][5]. The extracted topic words were compared with "true" topic words, which were given by three human subjects. The results are shown in Figure 2.

When the top five topic words were chosen (recall=13%), 87% of them were correct on average.

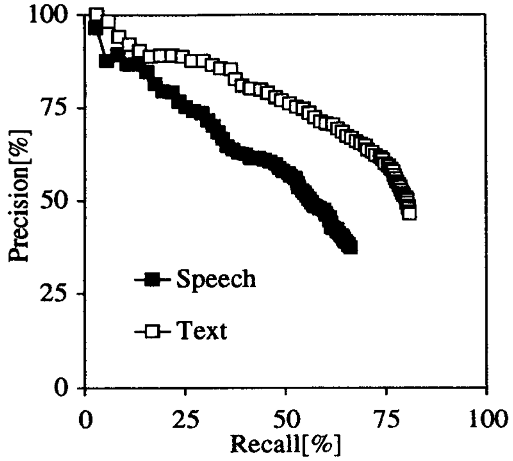

Fig. 2 - Topic word extraction results.

3. HUMAN-COMPUTER DIALOGUE SYSTEMS

3.1 Typical Systems in US and Europe

Recently a number of sites have been working on human-computer dialogue systems. The followings are typical examples.

(a) The View4You system at the University of Karksruhe

The University of Karlsruhe focuses its speech research on a content-addressable multimedia information retrieval system, under a multi-lingual environment, where queries and multimedia documents may appear in multiple languages [7]. The system is called "View4You" and their research is conducted in cooperation with the Informedia project at CMU [6]. In the View4You system, German and Servocroatian public newscasts are recorded daily. The newscasts are automatically segmented and an index is created for each of the segments by means of automatic speech recognition. The user can query the system in natural language by keyboard or through a speech utterance. The system returns a list of segments which is sorted by relevance with respect to the user query. By selecting a segment, the user can watch the corresponding part of the news show on his/her computer screen. The system overview is shown in Fig. 3.

(b) The SCAN- speech content based audio navigator at AT&T Labs

SCAN (Speech Content based Audio Navigator) is a spoken document retrieval system developed at AT&T Labs integrating speaker-independent, large-vocabulary speech recognition with information-retrieval to support query-based retrieval of information from speech archives [8]. Initial development focused on the application of SCAN to the broadcast news domain. An overview of the system architecture is provided in Fig. 4. The system consists of three components: (1) a speaker-independent large-vocabulary speech recognition engine which

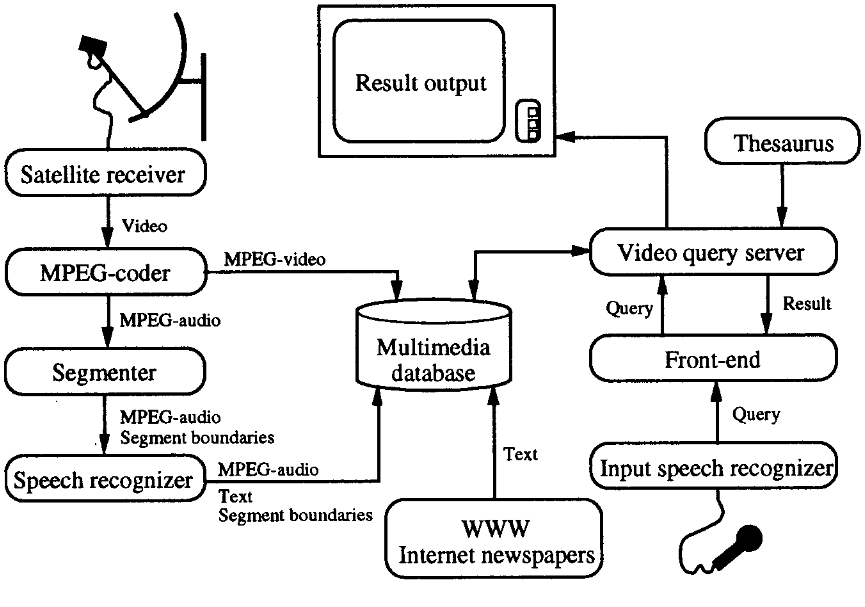

Fig. 3 - System overview of the View4You system.

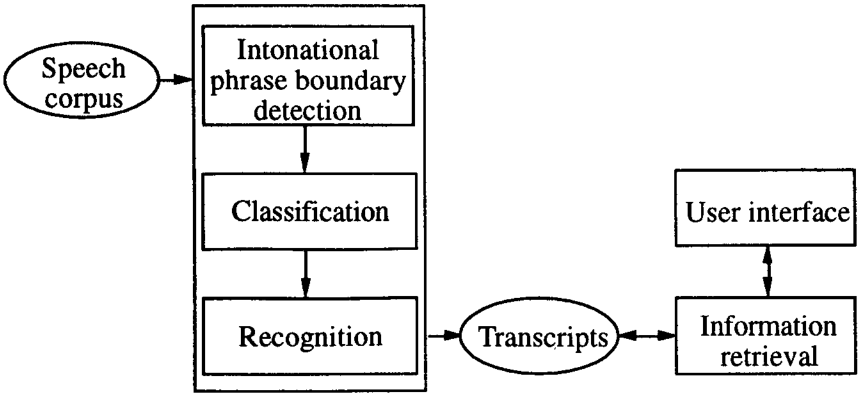

Fig. 4 - Overview of the SCAN spoken document system architecture.

segments the speech archive and generates transcripts, (2) an information-retrieval engine which indexes the transcriptions and formulates hypotheses regarding document relevance to user-submitted queries and (3) a graphical-user-interface which supports search and local contextual navigation based on the machine-generated transcripts and graphical representations of query-keyword distribution in the retrieved speech transcripts. The speech recognition component of SCAN includes an intonational phrase boundary detection module and a classification module, These subcomponents preprocess the speech data before passing the speech to the recognizer itself.

(c) The GALAXY-II conversational system at MIT

Galaxy is a client-server architecture developed at MIT for accessing on-line information using spoken dialogue [9]. It has served as the testbed for developing human language technology at MIT for several years. Recently, they have initiated a significant redesign of the GALAXY architecture to make it easier for researchers to develop their own applications, using either exclusively their own servers or intermixing them with servers developed by others. This redesign was done in part due to the fact that GALAXY has been designed as the first reference architecture for the new DARPA Communicator program. The resulting configuration of the GALAXY-II architecture is shown in Fig. 5. The boxes in this figure represent various human language technology servers as well as information and domain servers. The label in italics next to each box identifies the corresponding MIT system component. Interactions between servers are mediated by the hub and managed in the hub script. A particular dialogue session is initiated by a user either through interaction with a graphical interface at a Web site, through direct telephone dialup, or through a desktop agent.

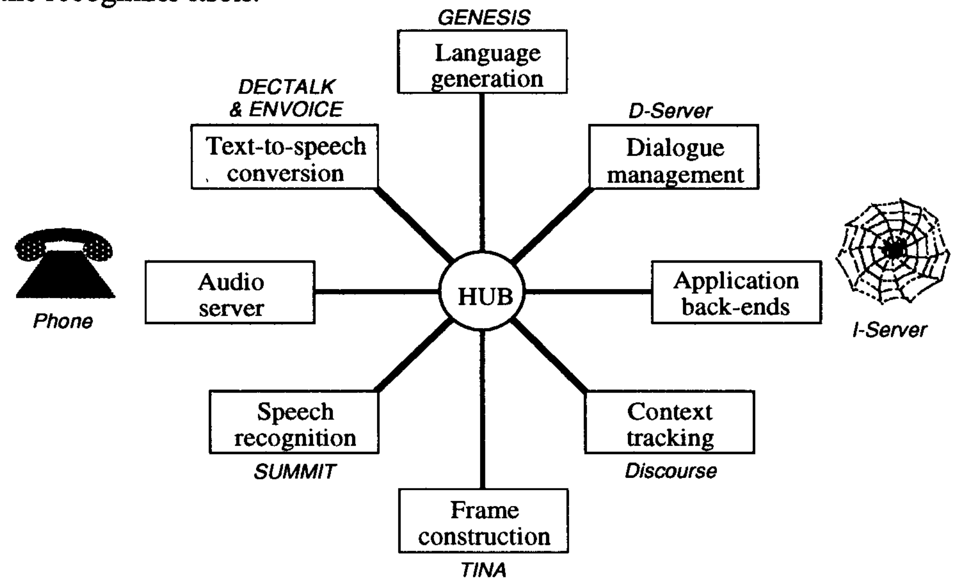

Fig. 5 - Architecture of GALAXY-II.

(d) The ARISE train travel information system at LIMSI

The ARISE (Automatic Railway Information Systems for Europe) projects aims developing prototype telephone information services for train travel information in several European countries [10]. In collaboration with the Vecsys company and with the SNCF (the French Railways), LIMSI has developed a prototype telephone service providing timetables, simulated fares and reservations, and information on reductions and services for the main French intercity connections. A prototype French/English service for the high speed trains between Paris and London is also under development. The system is based on the spoken language systems developed for the RailTel project [11] and the ESPRIT Mask project [12]. Compared to the RailTel system, the main advances in ARISE are in dialogue management, confidence measures, inclusion of optional spell mode for city/station names, and barge-in capability to allow more natural interaction between the user and the machine.

The speech recognizer uses n-gram backoff language models estimated on the transcriptions of spoken queries. Since the amount of language model training data is small, some grammatical classes, such as cities, days and months, are used to provide more robust estimates of the n-gram probabilities. A confidence score is associated with each hypothesized word, and if the score is below an empirically determined threshold, the hypothesized word is marked as uncertain. The uncertain words are ignored by the understanding component or used by the dialogue manager to start clarification subdialogues.

3.2 Designing a Multimodal Dialogue System for Information Retrieval

We have recently investigated a paradigm for designing multimodal dialogue systems [13]. 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. 6. 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.

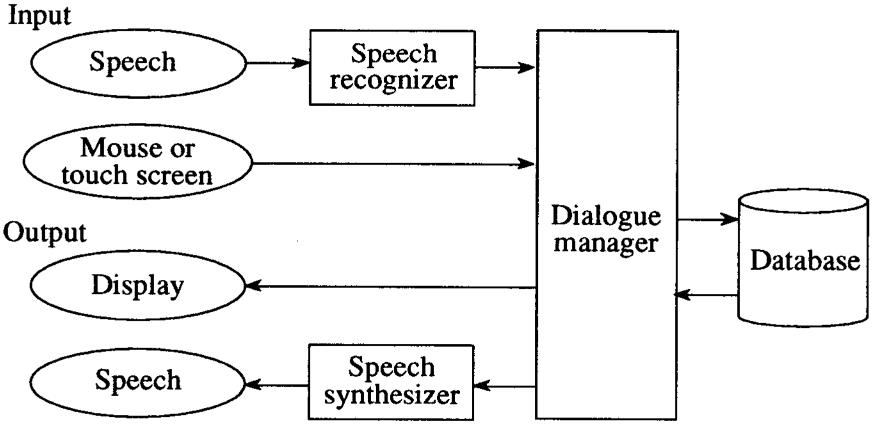

Fig. 6 - Multimodal dialogue system structure for information retrieval.

4. ROBUST SPEECH RECOGNITION

4.1 Automatic adaptation

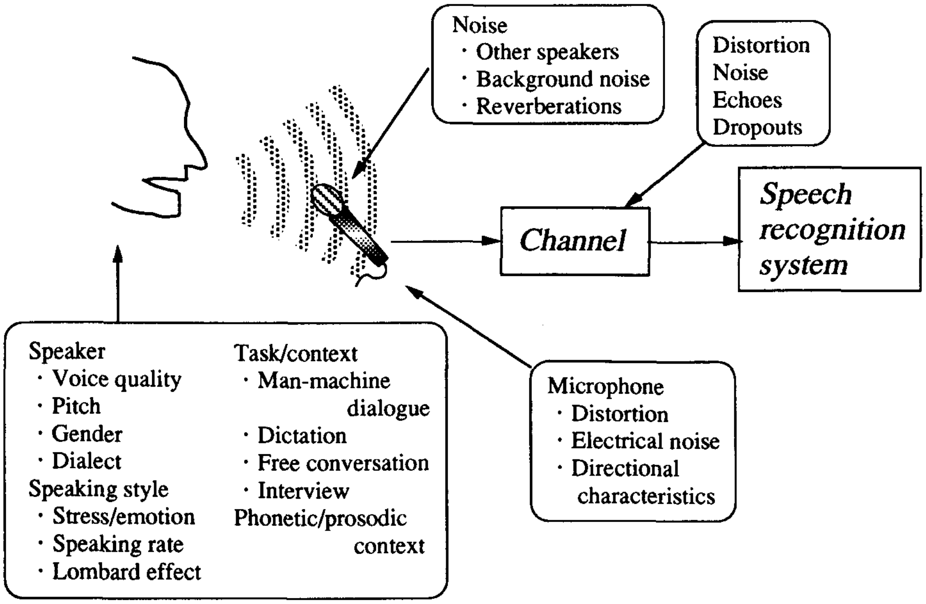

Fig. 7 - Main causes of acoustic variation in speech.

Ultimately, speech recognition systems should be capable of robust, speaker-independent or speaker-adaptive, continuous speech recognition. Figure 7 shows main causes of acoustic variation in speech [14]. It is crucial to establish methods that are robust against voice variation due to individuality, the physical and psychological condition of the speaker, telephone sets, microphones, network characteristics, additive background noise, speaking styles, and so on. Figure 8 shows main methods for making speech recognition systems robust against voice variation. It is also important for the systems to impose few restrictions on tasks and vocabulary. To solve these problems, it is essential to develop automatic adaptation techniques.

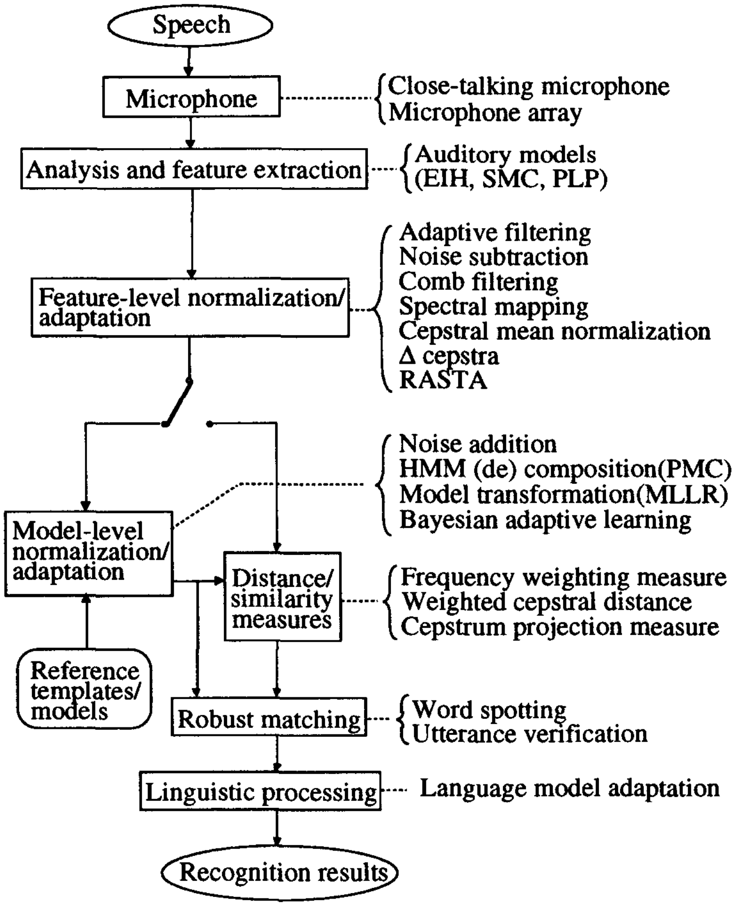
Speech
Microphone
Close-talking microphone
Microphone array
Analysis and feature extraction
Auditory models
(EIH, SMC, PLP)
Feature-level normalization/ adaptation
Adaptive filtering
Noise subtraction
Comb filtering
Spectral mapping
Cepstral mean normalization
Δ cepstra
RASTA
Noise addition
HMM (de) composition(PMC)
Model transformation(MLLR)
Bayesian adaptive learning
Model-level normalization/ adaptation
Distance/ similarity measures
Frequency weighting measure
Weighted cepstral distance
Cepstrum projection measure
Reference templates/ models
Robust matching
Word spotting
Utterance verification
Linguistic processing
Language model adaptation
Recognition results

Fig. 8 - Main methods to cope with voice variation in speech recognition.

Extraction and normalization of. (adaptation to) voice individuality is one of the most important issues [14]. 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 (text-dependent) and unsupervised (text-independent) methods. Unsupervised, on-line, 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 variablity or constraints. Modeling of the mechanism of speech production is expected to provide a useful modeling of speaker-to-speaker variability .

4.2 On-line speaker adaptation in broadcast news dictation

Since, in broadcast news, each speaker utters several sentences in succession, the recognition error rate can be reduced by adapting acoustic models incrementally within a segment that contains only one speaker. We applied on-line, unsupervised, instantaneous and incremental speaker adaptation combined with automatic detection of speaker changes [4]. The MLLR [15] -MAP [16] and VFS (vector-field smoothing) [17] methods were instantaneously and incrementally carried out for each utterance. The adaptation process is as follows. For the first input utterance, the speaker-independent model is used for both recognition and adaptation, and the first speaker-adapted model is created. For the second input utterance, the likelihood value of the utterance given the speaker-independent model and that given the speaker-adapted model 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 model is created. Otherwise, the existing speaker-adapted model is 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 model and some speaker-adapted models. If the speaker-independent model yields a larger likelihood than any of the speaker-adapted models, a speaker change is detected and a new speaker-adapted model is constructed. Experimental results show that the adaptation reduced the word error rate by 11.8 % relative to the speaker-independent models.

5. PRESPECTIVES OF LANGUAGE MODELING

5.1 Language modeling for spontaneous speech recognition

One of the most important issues 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 solve such problems [2]. 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 accomplish a task, 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.

5.2 Message-Driven Speech Recognition

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. 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. 9 [18], where M is the message (content) that a speaker intended to convey.

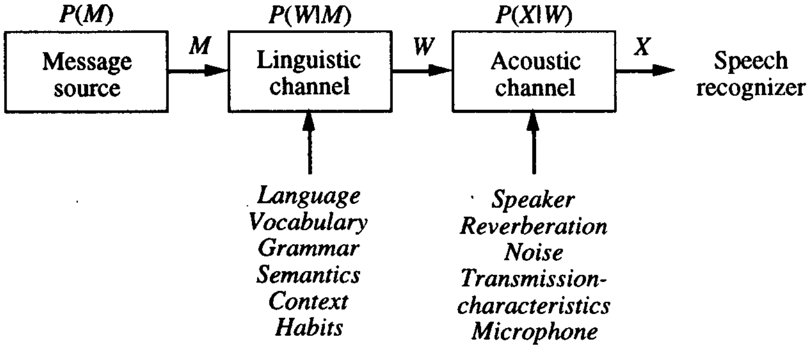

Fig. 9 - A communication - theoretic view of speech generation and recognition.

According to this model, the speech recognition process is represented as the maximization of the following a posteriori probability [4][5],

$$\max_{M} P(M|X) = \max_{M} \sum_{W} P(M|W)P(W|X). \qquad (1)$$

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

$$\max_{M} P(M|X) = \max_{M} \sum_{W} \frac{P(X|W)\,P(W|M)\,P(M)}{P(X)}. \qquad (2)$$

For simplicity, we can approximate the equation as

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

$P(X|W)$ 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.

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

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 [19]. 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 for the clean set were slightly reduced by this method.

6. CONCLUSIONS

Speech recognition technology has made a remarkable progress in the past 5 - 10 years. Based on the progress, various application systems have been developed using dictation and spoken dialogue technology. One of the most important applications is information extraction and retrieval. Using the 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 between humans, automatic speech recognition will continue to find applications, such as meeting/conference summarization, automatic closed captioning, and interpreting telephony. It is expected that speech recognizer will become the main input device of the "wearable" computers that are now actively investigated. In order to materialize these applications, we have to solve many problems. The most important issue is how to make the speech recognition systems robust against acoustic and lingustic 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] S. Furui: "Future directions in speech information processing", Proc. 16th ICA and 135th Meeting ASA, Seattle, pp. 1-4 (1998)

[3] F. Kubala: "Broadcast news is good news", DARPA Broadcast News Workshop, Virginia (1999)

[4] K. Ohtsuki, S. Furui, N. Sakurai, A. Iwasaki and Z.-P. Zhang: "Improvements in Japanese broadcast news transcription", DARPA Broadcast News Workshop, Virginia (1999)

[5] 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)

[6] M. Witbrock and A. G. Hauptmann: "Speech recognition and information retrieval: Experiments in retrieving spoken documents", Proc. DARPA Speech Recognition Workshop, Virginia, pp. 160-164 (1997). See also http://www.informedia.cs.cmu.edu/

[7] T. Kemp, P. Geutner, M. Schmidt, B. Tomaz, M. Weber, M. Westphal and A. Waibel: "The interactive systems labs View4You video indexing system", Proc. Int. Conf. Spoken Language Processing, Sydney, pp. 1639-1642 (1998)

[8] J. Choi, D. Hindle, J. Hirschberg, I. Magrin-Chagnolleau, C. Nakatani, F. Pereira, A. Singhal and S. Whittaker: "SCAN - speech content based audio navigator: a systems overview", Proc. Int. Conf. Spoken Language Processing, Sydney, pp. 2867-2870 (1998)

[9] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid and V. Zue: "GALAXY-II: a reference architecture for conversational system development", Proc. Int. Conf. Spoken Language Processing, Sydney, pp. 931-934 (1998)

[10] L. Lamel, S. Rosset, J. L. Gauvain and S. Bennacef: "The LIMSI ARISE system for train travel information", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Phoenix, pp. 501-504 (1999)

[11] L. F. Lamel, S. K. Bennacef, S. Rosset, L. Devillers, S. Foukia, J. J. Gangolf and J. L. Gauvain: "The LIMSI RailTel system: Field trial of a telephone service for rail travel information", Speech Communication, 23, pp. 67-82 (1997)

[12] J. L. Gauvain, J. J. Gangolf and L. Lamel: "Speech recognition for an information Kiosk", Proc. Int. Conf. Spoken Language Processing, Philadelphia, pp. 849-852 (1998)

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

[14] 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)

[15] 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).

[16] 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).

[17] 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)

[18] B.-H. Juang: "Automatic speech recognition: Problems, progress & prospects", IEEE Workshop on Neural Networks for Signal Processing (1996)

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