論文 / 著書情報
Article / Book Information

Title	Recent progress in corpus-based spontaneous speech recognition
Authors	Sadaoki Furui
出典 / Citation	IEICE Transactions on Information and Systems, Vol. E.88-D, No. 3, pp. 366-375
発行日 / Pub. date	2005, 3
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2005 Institute of Electronics, Information and Communication Engineers.

INVITED PAPER *Special Section on Corpus-Based Speech Technologies*

Recent Progress in Corpus-Based Spontaneous Speech Recognition

Sadaoki FURUI[†a)], *Fellow*

SUMMARY This paper overviews recent progress in the development of corpus-based spontaneous speech recognition technology. Although speech is in almost any situation spontaneous, recognition of spontaneous speech is an area which has only recently emerged in the field of automatic speech recognition. Broadening the application of speech recognition depends crucially on raising recognition performance for spontaneous speech. For this purpose, it is necessary to build large spontaneous speech corpora for constructing acoustic and language models. This paper focuses on various achievements of a Japanese 5-year national project "Spontaneous Speech: Corpus and Processing Technology" that has recently been completed. Because of various spontaneous-speech specific phenomena, such as filled pauses, repairs, hesitations, repetitions and disfluencies, recognition of spontaneous speech requires various new techniques. These new techniques include flexible acoustic modeling, sentence boundary detection, pronunciation modeling, acoustic as well as language model adaptation, and automatic summarization. Particularly automatic summarization including indexing, a process which extracts important and reliable parts of the automatic transcription, is expected to play an important role in building various speech archives, speech-based information retrieval systems, and human-computer dialogue systems.

key words: *spontaneous speech recognition, corpus, model adaptation, indexing, summarization*

1. Introduction

Read speech and similar types of speech, e.g. news broadcasts reading a text, can be recognized with accuracy higher than 95%, using the state-of-the-art speech recognition technology. However, recognition accuracy drastically decreases for spontaneous speech [1]. One of the major reasons for this decrease is that acoustic and language models used up until now have generally been built using written language or speech read from a text. Spontaneous speech and speech from written language are very different, both acoustically and linguistically. Spontaneous speech includes filled pauses, repairs, hesitations, repetitions, partial words, and disfluencies. It is quite interesting to note that, although speech is almost always spontaneous, spontaneous speech recognition is a special area that emerged only about 10 years ago within the wider field of automatic speech recognition. Broadening the application of speech recognition depends crucially on raising the recognition performance for spontaneous speech.

In order to increase recognition performance for spontaneous speech, it is necessary to build acoustic and language models for spontaneous speech. Current methods applying statistical language modeling such as bigrams and trigrams of words or morphemes to spontaneous speech corpus may prove to be inadequate. Our knowledge of the structure of spontaneous speech is currently insufficient to achieve the necessary breakthroughs. Although spontaneous speech effects are quite common in human communication and may increase in human machine discourse as people become more comfortable conversing with machines, modeling of speech disfluencies is only in the initial stage. Since spontaneous speech includes various redundant expressions, recognition of spontaneous speech will require a paradigm shift from simply recognizing speech, where transcribing the spoken words is the primary focus, to understanding where underlying messages of the speaker are extracted.

Until now, our focus on speech recognition has been on its use as an interface in human-machine interactions, mostly for information access and extraction. With increases in cellular phone use and dependence on networked information resources, and as rapid access to information becomes an increasingly important economic factor, voice access to data and voice transactions will no doubt rise dramatically. However, there is also a growing interest in viewing speech not just as a means to access information, but as, itself, a source of information.

Since speech is the most natural and effective method of communication between human beings, various important speech documents, including lectures, presentations, meeting records and broadcast news, are produced everyday. However, it is not easy to quickly review, retrieve, selectively disseminate, and reuse these speech documents, if they are simply recorded as audio signal. Therefore, automatically transcribing speech using speech recognition technology is a crucial aspect of creating knowledge resources from speech.

We can envision a great information revolution on par with the development of writing systems, if we can successfully meet the challenges of speech, both as a medium for information access and as itself a source of information. Speech is still the means of communication used first and foremost by humans, and only a small percentage of human communication is written. Automatic speech recognition can add many of the advantages normally associated only with text (random access, sorting, and access at different times and places) to the many benefits of speech. Making this vision a reality requires significant advances.

Manuscript received August 2, 2004.

[†]The author is with Tokyo Institute of Technology, Tokyo, 152–8552 Japan.

a) E-mail: furui@cs.titech.ac.jp

DOI: 10.1093/ietisy/e88–d.3.366

2. Four Categories of Speech Recognition Tasks and Major Research Activities

Speech recognition tasks can be classified into four categories, as shown in Table 1, according to two criteria: whether it is targeting utterances from human to human or human to computer, and whether the utterances have a dialogue or monologue style [2]. The table lists typical tasks for each category.

Most of the practical application systems widely used now are classified as Category III, recognizing the utterances in human-computer dialogues, such as in airline information services tasks. DARPA-sponsored projects including ATIS and Communicator have laid the foundations of these systems. Unlike other categories, the systems in the Category III are usually designed and developed after clearly defining the application/task. The machines that we have attempted to design so far are, almost without exception, limited to the simple task of converting a speech signal into a word sequence and then determining, from the word sequence, a meaning that is "understandable". Here, the set of understandable messages is finite in number, each being associated with a particular action (e. g., route a call to a proper destination or issue a buy order for a particular stock). In this limited sense of speech communication, the focus is detection and recognition rather than inference and generation.

Category I targets human-to-human dialogues and includes DARPA-sponsored Switchboard and Call Home (Hub 5) tasks. Speech recognition research in this category aiming to make minutes of meetings (e.g. [3]) has recently started. Waibel et al. have been investigating a meeting browser that observes and tracks meetings for later review and summarization [4]. Akita et al. have investigated techniques for archiving discussions [5]. In their method, speakers are automatically indexed in an unsupervised way, and speech recognition is performed using the results of indexing. Processing human-human conversational speech under unpredictable recording conditions and vocabularies presents new challenges for spoken language processing.

A relatively new task classified into this category is the MALACH (Multilingual Access to Large spoken ArCHives) project [6]. Its goal is to advance the state-of-the-art technology for access to large multilingual collections of spontaneous conversational speech by exploiting an unmatched collection assembled by the Survivors of the Shoah Visual History Foundation (VHF). The VHF collection contains 116,000 hours of interviews that were conducted in 32 languages with nearly 52,000 survivors of the Holocaust. By the end of 2005, the entire collection will have been digitized (to 180 TB of MPEG-1 video) and manually indexed using an extensive controlled vocabulary and within-interview name authority control. Automatic transcription has been added for nearly 1,000 hours of English and Czech using speech recognition systems with word error rates below 40%. This collection is indeed a challenging task because of heavily accented, emotional and elderly spontaneous characteristics. Named entity tagging, topic segmentation, and unsupervised topic classification are also being investigated.

Table 1 Categorization of speech recognition tasks.

	Dialogue	Monologue
Human to human	(Category I) Switchboard, Call Home (Hub 5), meeting, interview	(Category II) Broadcast news (Hub 4), other programs, lecture, presentation, voice mail
Human to machine	(Category III) ATIS, Communicator, information retrieval, reservation	(Category IV) Dictation

Tasks belonging to Category II, which targets recognizing human-to-human monologues, include transcription of broadcast news (Hub 4), news programs, lectures, presentations, and voice mails (e.g. [7]). Speech recognition research in this category has recently become very active. Since the utterances in the Category II are made with the expectation that the audience can correctly understand what is spoken in the one-way communication, they are relatively easier to recognize than the utterances in Category I. If high recognition performance is achieved, a wide range of applications, such as making lecture notes, records of presentations and closed captions, archiving and retrieving these records, and retrieving voice mails, will be realized.

Various research has made clear that the utterances spoken by people talking to computers, such as those in Categories III and IV, especially when the speaker is conscious, are acoustically as well as linguistically very different from those spoken to other people, such as those in Categories I and II.

One of the typical tasks belonging to Category IV, which targets the recognition of monologues performed when people are talking to a computer, is dictation. Various commercial softwares for such purposes have been developed. Since the utterances in Category IV are made with the expectation that the utterances will be converted exactly into texts with correct characters, their spontaneity is much lower that that in Category III. In the four categories, spontaneity is considered to be the highest in Category I and the lowest in Category IV.

A survey on activities in spontaneous speech recognition in Europe over the last 10 years, including various research projects, can be found in [8]. A brief historical review on the development of this topic in Europe is presented, and various technical issues are addressed, distinguishing research projects on spontaneous speech recognition from other research activities in speech.

3. Corpora

The appetite of today's statistical speech processing techniques for training material are well described by the aphorism: "There's no data like more data." Large structured collections of speech and text are essential for progress in speech recognition research. Unlike the traditional approach, in which knowledge of speech behavior is "discovered" and "documented" by human experts, statistical methods provide an automatic procedure to directly "learn" regularities in the speech data. The need for a large set of good training data is, thus, more critical than ever. However, establishing a good speech database for the computer to uncover the characteristics of the signal is not a straightforward process. There are basically two broad issues to be carefully considered: one being the content and its annotation, and the other the collecting mechanism.

For natural dialog applications such as the ATIS program, a wizard setup is often used to collect the data. A wizard in this case is a human mimicking the machine interacting with the user. Through the interaction, natural queries in sentential forms are collected. While a wizard setup can produce a useful set of data, it lacks diversity, particularly in situations where the real machine may fail. A human wizard cannot intentionally simulate all types of machine error and thus the recorded data may fail to provide complete information of real human-machine interactions.

The recorded data needs to be verified, labeled, and annotated by people whose knowledge is introduced into the design of the system through its learning process (i.e. via supervised training of the system after the data has been labeled). Labeling and annotation for spontaneous speech can easily become unmanageable. For example: how do we annotate speech repairs and partial words? how do the phonetic transcribers reach a consensus in acoustic-phonetic labels when there is ambiguity? and how do we represent a semantic notion? Errors in labeling and annotation will result in system performance degradation. How to ensure the quality of the annotated results is thus a major concern. Research limited only to automating or creating tools to assist the verification procedure is in itself an interesting subject.

4. "Spontaneous Speech: Corpus and Processing Technology" Project

4.1 Overview of the Project

In the above-mentioned context, a 5-year Science and Technology Agency Priority Program entitled "Spontaneous Speech: Corpus and Processing Technology" was conducted in Japan from 1999 to 2004 [1], and the following three major results were obtained.

1) A large-scale spontaneous speech corpus, Corpus of Spontaneous Japanese (CSJ), consisting of roughly 7 M words with a total speech length of 650 hours has been built [9], [10].
2) Acoustic and language modeling for spontaneous speech recognition and understanding using linguistic as well as para-linguistic information in speech was investigated.
3) Spontaneous speech recognition and summarization technology was investigated.

4.2 Corpus of Spontaneous Japanese (CSJ)

Mainly recorded are monologues such as academic presentations (AP) and extemporaneous presentations (EP) as shown in Table 2. AP is live recordings of academic presentations in nine different academic societies covering the fields of engineering, social science and humanities. EP is studio recording of paid layman speakers' speech on everyday topics like "the most delightful memory of my life" presented in front of a small audience and in a relatively relaxed atmosphere. The age and gender of EP speakers are more balanced than that of AP speakers. The CSJ also includes some dialogue speech for the purpose of comparison with monologue speech. The recordings were manually given orthographic and phonetic transcription. Spontaneous speech-specific phenomena, such as filled pauses, word fragments, reduced articulation and mispronunciation, as well as non-speech events like laughter and coughing were also carefully tagged.

One-tenth of the utterances, hereafter referred to as the Core, were tagged manually and used for training a morphological analysis and part-of-speech (POS) tagging program [11] for automatically analyzing all of the 650-hour utterances. The Core consists of 70 APs, 107 EPs, 18 dialogues and 6 read speech files. They were also tagged with para- linguistic/intonation information, dependency-structure, discourse structure, and summarization (see Fig. 1). For intonation labeling of spontaneous speech, the traditional J_ToBI [12] was extended to X_JToBI [13], in which inventories of tonal events as well as break indices were considerably enriched.

In the Netherlands, the Spoken Dutch Corpus (in Dutch: Corpus Gesproken Nederlands, or CGN) project started in 1998 [14]. The project aims at the compilation and annotation of a corpus of 1,000 hours of spoken Dutch.

Table 2 Contents of the CSJ [10].

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Academic presentations (AP)	838	1006	Monolog	Spont.	299.5
Extemporaneous presentations (EP)	580	1715	Monolog	Spont.	327.5
Interview on AP	*(10)	10	Dialog	Spont.	2.1
Interview on EP	*(16)	16	Dialog	Spont.	3.4
Task oriented dialogue	*(16)	16	Dialog	Spont.	3.1
Free dialogue	*(16)	16	Dialog	Spont.	3.6
Reading text	*(244)	491	Dialog	Read	14.1
Reproduction	*(16)	16	Monolog	Read	5.5
*Counted as the speakers of AP or EP				Total hours	658.8

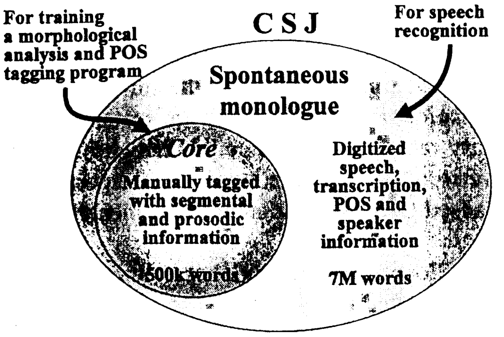

Fig. 1 Overall design of the Corpus of Spontaneous Japanese (CSJ).

Upon completion, the corpus is expected to constitute a valuable resource for research in the field of language and speech technology. Although the corpus will contain a fair amount of read speech, a major part of the data will consist of spontaneous speech, ranging from lectures to conversations recorded unobtrusively. All speech recordings will be made available together with several levels of annotations, from orthographic transcription to syntactic analyses and prosodic labeling, similar to the CSJ.

Detailed results of 2) and 3) obtained using the CSJ in the framework of the "Spontaneous Speech" project will be described in the following sections.

5. Progress Made and Difficulties Encountered in Spontaneous Speech Recognition

5.1 Effectiveness of Corpora

By constructing acoustic and language models using the CSJ, recognition errors for spontaneous presentation were reduced to roughly half compared to models constructed using read speech and written text [1], [15]. Increasing the size of training data for acoustic and language models has decreased the recognition error rate (WER: word error rate) as shown in Figs. 2 and 3 [16]. They show the results averaged over the three test sets which will be described in Sect. 5.5. Figure 2 indicates WER, adjusted test-set perplexity [17] and out-of-vocabulary (OOV) rate, as a function of the size of the language model training data with the condition that the acoustic model is constructed using the whole training data set (510 hours). The adjusted perplexity was used to normalize the effect of the increase of vocabulary size on the perplexity according to the increase of training data size.

On the other hand, Fig. 3 shows WER as a function of the size of acoustic model training data, when the language model is made using the whole training data set (6.84 M words). By increasing the language model training data size from 1/8 (0.86 M words) to 8/8 (6.84 M words), the WER, the perplexity and the OOV are relatively reduced by 17%, 19%, and 62%, respectively. By increasing the acoustic model training data from 1/8 (68 hours) to 8/8 (510 hours), the WER is reduced by 6.3%. The best WER of 25.3%, obtained by using the whole training data set for both acoustic and language modeling, shown at the extreme right condition in Fig. 3, is 2.9% lower in the absolute value than that shown in Fig. 2. This is because the experiment of Fig. 3 combined ΔΔCepstrum and ΔΔpower with the three features of Cepstrum, ΔCepstrum and Δpower which were used in the experiment of Fig. 2. All these results show that WER is significantly reduced by an increase of the training data and almost saturated by using the whole data set. This strongly confirms that the size of the CSJ is meaningful in modeling spontaneous presentation speech.

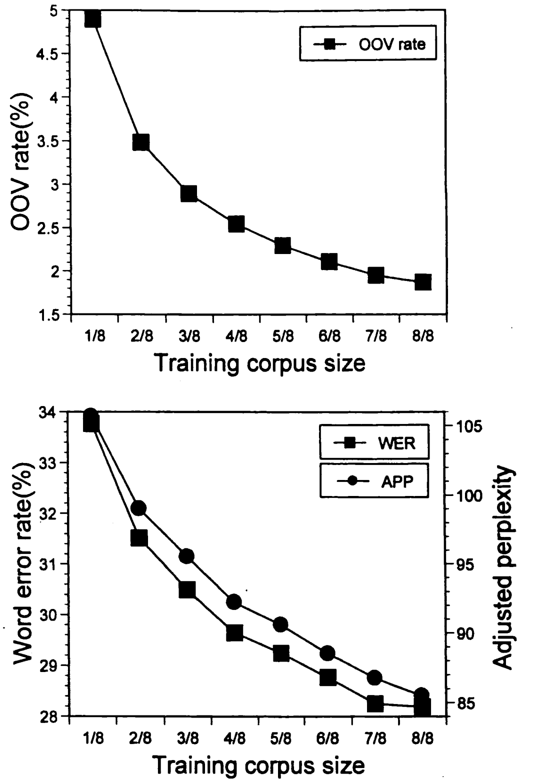

Fig. 2 WER, adjusted test-set perplexity and out-of-vocabulary (OOV) rate as a function of the size of language model training data.

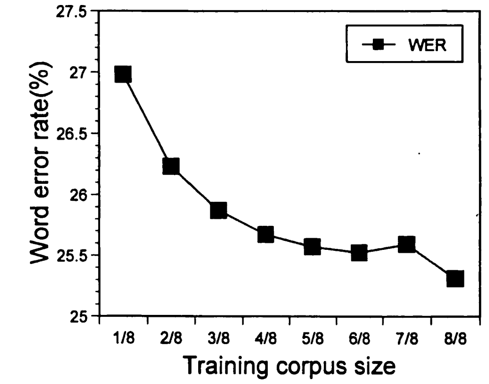

Fig. 3 WER as a function of the size of acoustic model training data.

5.2 Pronunciation Variation

In spontaneous speech, pronunciation variation is so diverse that multiple surface form entities are needed for many lexical items. Kawahara et al. [18] have found that statistical modeling of pronunciation variations integrated with language modeling is effective in suppressing false matching of less frequent entries. They have adopted a trigram model of word-pronunciation entries. Since both orthographic and phonetic transcriptions of the CSJ were made manually for each unit of utterance (sentence), word-based automatic alignment between them was performed to obtain the pronunciation entries for each word. This was incorporated as a post-processor of the morphological analyzer. Heuristic thresholding was applied to eliminate erroneous patterns, in which pronunciation entries whose occurrence probability in each lexical item is lower than a threshold were eliminated. As a result, 30,820 word-pronunciation entries (24437 distinct words) were obtained, on which a trigram model was trained. Experimental results show that the word-pronunciation trigram model is effective.

5.3 Sentence Boundary Detection

Another difficulty of spontaneous speech recognition is that generally no explicit sentence boundary is given. Therefore, it is impossible to recognize spontaneous speech sentence by sentence. Kawahara et al. developed a decoder in which no sentence boundaries are required [19]. The decoder can handle very long speech with no prior sentence segmentation. Experimental results show that the new decoder performed better than the previous version using sentence boundaries. Based on transcription results and pause lengths, sentence boundaries are automatically determined and punctuation marks are given. Specifically, a linguistic likelihood ratio between a model including sentence boundary and a model without boundary is compared with a threshold and then a decision is made.

5.4 Analysis of Spontaneous Speech Recognition Errors

Individual differences in spontaneous presentation speech recognition performances have been analyzed using 10 minutes from each presentation given by 51 male speakers, for a total of 510 minutes [20]. Seven kinds of speaker attributes were considered in the analysis. They were word accuracy (Acc), averaged acoustic frame likelihood (AL), speaking rate (SR), word perplexity (PP), out of vocabulary rate (OR), filled pause rate (FR) and repair rate (RR). The speaking rate, defined as the number of phonemes per second, and the averaged acoustic frame likelihood were calculated using the results of forced alignment of the reference tri-phone labels after removing pause periods. The word perplexity was calculated using trigrams, in which prediction of out of vocabulary words was not included. The filled pause rate and the repair rate were the number of filled pauses and repairs divided by the number of words, respectively.

Analysis results indicate that the attributes exhibiting a real correlation with the accuracy are speaking rate, out of vocabulary rate, and repair rate. Although other attributes also have correlation with the accuracy, the correlation is actually caused through these more fundamentally influential attributes.

The following equation has been obtained as a result of a linear regression model of the word accuracy with the six presentation attributes.

$$\begin{aligned} Acc = {} & 0.12AL - 0.88SR - 0.020PP - 2.2OR \\ & + 0.32FR - 3.0RR + 95 \end{aligned} \tag{1}$$

In the equation, the regression coefficient for the repair rate is −3.0 and the coefficient for the out of vocabulary rate is −2.2. This means that a 1% increase of the repair rate or the out of vocabulary rate corresponds respectively to a 3.0% or 2.2% decrease of the word accuracy. This is probably because a single recognition error caused by a repair or an out of vocabulary word triggers secondary errors due to linguistic constraints. The determination coefficient of the multiple linear regression is 0.48, which is significant at a 1% level. This means that roughly half of the variance of the word accuracy can be explained by the model.

Normalized representation of the regression analysis, in which the variables are normalized in terms of the mean and variance before the analysis in order to show the effects of explaining variables on the word accuracy, indicates that coefficients of the speaking rate, the out of vocabulary rate and the repair rate are relatively large.

5.5 Test Sets for Technology Evaluation

In order to evaluate the progress of spontaneous speech recognition technology, three test sets of presentations have been constructed from the CSJ so that they well represent the whole corpus with respect to various factors of spontaneous speech [21]. The above-mentioned analysis by Shinozaki et al. concluded that speaking rate (SR), out-of-vocabulary rate (OR) and repair rate (RR) are directly correlated with accuracy. Other factors mainly depend on one or more of these three. For example, word perplexity (PP) is correlated with the accuracy, but if its correlation with the OR is removed, we find actually that PP is not so correlated with the accuracy. However, OR is intrinsically dependent on vocabulary and is thus variable when the lexicon is modified. On the other hand, PP's difference among speech samples is generally more stable even when the language model is revised. Therefore, we decided to take into account PP instead of OR, in combination with SR and RR, in the test-set selection.

Since the speaking styles and vocabularies of academic and extemporaneous presentations are significantly different, we set up respective test sets. In addition, considering the fact that most of the academic presentations were given by male speakers, we set up two sets for the academic category: a male-only set and a gender-balanced set. Thus, we have three test sets, each of which consists of 10 speakers. Benchmark results of speech recognition using the three test sets are also presented in the above paper.

5.6 Model Complexity Control

Since there is always a limit of the size of data that we can collect for acoustic modeling, especially for spontaneous speech with large variations, it is very important to properly control the model complexity. The Bayesian approach has advantages in that it enables an appropriate model selection for given speech data. However, the Bayesian approach is difficult to apply to large-scale tasks such as spontaneous speech recognition. Watanabe et al. have tried to apply a practical Bayesian framework, Variational Bayesian Estimation and Clustering (VBEC), to spontaneous speech recognition [22]. In the Variational Bayesian (VB) approach, approximate posteriors (VB posteriors) are effectively obtained by using EM-like iterative calculations. The VBEC framework includes estimation of acoustic model posteriors, selection of appropriate acoustic models, and acoustic score calculations using the Bayesian prediction. Experimental results for the CSJ task indicate that VBEC model selection is effective in clustering triphone HMM states and determining the number of Gaussians per state.

6. Model Adaptation

It is very important to create a modeling methodology and associated data collection scheme that can be applied generally to many different tasks. To maximize performance, one should always strive for data that truly reflects the actual environment. This calls for a database collection plan that is consistent with the task. This data collection effort soon becomes unmanageable if the system designer has to redo data collection for each and every application that is being developed. It is therefore desirable to design a task-independent data set and a modeling method that delivers a reasonable performance upon first use and can quickly allow in-field trials for further revision as soon as task-dependent data become available. Research results in this area can offer the benefit of reduced application development cost.

6.1 Acoustic Model Adaptation

Word accuracy varies largely from speaker to speaker. There exist many factors that affect the accuracy of spontaneous speech recognition. They include individual voice characteristics, speaking manners, styles of language (grammar), vocabularies, topics, and noise, such as coughs. Even if utterances are recorded using the same close-talking microphones, acoustic conditions still vary according to the recording environment. A batch-type unsupervised adaptation method has been incorporated to cope with speech variation in the CSJ utterances [15]. The MLLR method using a binary regression class tree to transform Gaussian mean vectors was employed. The regression class tree was made using a centroid-splitting algorithm. The actual classes used for transformation were determined at run time according to the amount of data assigned to each class. By applying the adaptation, the error rate was reduced by 15% relative to the speaker independent condition.

Kawahara et al. proposed a speaking rate dependent decoding strategy [18]. The proposed method applies the most appropriate acoustic analysis, phone models, and decoding parameters according to the speaking rate. This method gained improvement of 1.9% in absolute word error rate.

6.2 Dynamic Bayesian Network Approach

As described in Sect. 5.4, one of the most important issues in spontaneous speech recognition is how to cope with the degradation of recognition accuracy due to speaking rate fluctuation within an utterance. Shinozaki et al. proposed an acoustic model for adjusting mixture weights and transition probabilities of the HMM for each frame according to the local speaking rate [23]. The proposed model was implemented along with variants and conventional models using the dynamic Bayesian network framework. The proposed model has a hidden variable representing variation of the "mode" of the speaking rate and its value controls the parameters of the underlying HMM. Model training and maximum probability assignment of the variables were conducted using the EM/GEM and inference algorithms for Bayesian networks. Utterances from meetings and lectures were used for evaluation where Bayesian network-based acoustic models rescored the utterance hypotheses obtained from a first-pass N-best list. In the experiments, the proposed method showed consistently higher performance than conventional models.

6.3 Language Model Adaptation

In order to cover a wide variation of tasks, mixture-based language models (LMs) are commonly used, in which several LMs specific to a particular topic or style of language are generated and stored. In general, an application domain can be characterized by the subtasks, and each task can be characterized by a topic or a set of topics. A given document collection like that of newspaper texts can be categorized into specific text clusters according to a given set of topics. Usually, the newspaper articles are manually classified into different genres like news, sports, movies, etc. Based on this information, topic specific text clusters can be derived and further, for each cluster, a topic specific LM can be generated. Automatic text clustering for topic assignment can also be used. The probability estimates from these component LMs are interpolated to produce an overall probability, where the interpolation weights are chosen to reflect the topic or style of language currently being recognized. A difficult problem is posed by the fact that new topics are always created and different representations are frequently used for the same topic. Therefore, how to dynamically model the set of topics is a crucial issue.

Lussier et al. investigated combinations of unsupervised language model adaptation methods for CSJ utterances [24]. Data sparsity is a common problem shared by all speech recognition tasks but it is especially acute in the case of spontaneous speech recognition. The method proposed combines information from two readily available sources, clusters of presentations from the training corpus and the transcription hypothesis, to create word-class n-gram models that are then interpolated with a general language model. The interpolation coefficient is estimated based on EM algorithm using a development set. Since this method performs in an offline manner using whole recognition results to suppress influences of local recognition errors, it is more robust against recognition errors than online adaptation methods. Experimental results show that a relative reduction in word error rate of 5–10% is obtained on the three CSJ test sets used.

Nanjo et al. also proposed offline, unsupervised methods of language model adaptation to a specific speaker and topic [25]. In their methods, texts similar to a test-set presentation (a set of transcription hypotheses) are selected from the training corpus based on the word perplexity and tf-idf measure, and a language model based on the selected texts is interpolated with a general language model. The transcription hypotheses are also directly used to generate a language model tuned to the presentation. After discarding bigram and trigram entries observed only once in the presentation, the language model is interpolated with the general language model using the complementary backoff algorithm [26], which works well when there is a large difference in the n-gram entries between the models. Finally, both adaptation methods are combined. The proposed methods successfully reduce the perplexity and word error rate.

6.4 Massively Parallel Decoder-Based Recognition

Shinozaki et al. have proposed using a combination of cluster-based language models and acoustic models in the framework of a Massively Parallel Decoder (MPD) to cope with the problem of acoustic as well as linguistic variations of presentation utterances [27]. MPD is a parallel decoder that has a large number of decoding units, in which each unit is assigned to each combination of element models. Likelihood values produced by all the decoding units are compared, and the hypothesis having the largest likelihood is selected as the recognition result. The system runs efficiently on a parallel computer, and thus the turnaround time is comparable to the conventional decoder using a single model and processor. In experiments conducted using presentation speeches from the CSJ, two types of cluster models have been investigated: presentation-based cluster models and utterance-based cluster models. It has been confirmed that utterance-based cluster models give significantly lower recognition error rate than presentation-based cluster models in both language and acoustic modeling. It has also been shown that roughly 100 decoding units are sufficient in terms of recognition rate; and, in the best setting, 12% reduction in word error rate was obtained in comparison with the conventional decoder.

7. Indexing and Summarization

7.1 Automatic Speech Summarization

Spontaneous speech is ill-formed and very different from written text. Spontaneous speech usually includes redundant information such as disfluencies, fillers, repetitions, repairs and word fragments. In addition, irrelevant information caused by recognition errors is usually inevitably included when spontaneous speech is transcribed. Therefore, an approach in which all words are simply transcribed is not an effective one for spontaneous speech. Instead, speech summarization which extracts important information and removes redundant and incorrect information is ideal for recognizing spontaneous speech. Speech summarization is also expected to reduce time needed for reviewing speech documents and improve the efficiency of document retrieval.

Speech summarization has a number of significant challenges that distinguish it from general text summarization. Applying text-based technologies [28] to speech is not always workable and often they are not equipped to capture speech specific phenomena [29], [30]. One fundamental problem with the speech summarization is that they contain speech recognition errors and disfluencies. We have proposed a two-stage summarization method consisting of important sentence extraction and word-based sentence compaction, as shown in Fig. 4 [31]–[33]. After removing all the fillers based on speech recognition results, a set of relatively important sentences is extracted, and sentence compaction is applied to the set of extracted sentences. The ratio of sentence extraction and compaction is controlled according to a summarization ratio initially determined by the user. Sentence and word units are extracted from the speech recognition results and concatenated for producing summaries so that they maximize the weighted sum of linguistic likelihood, amount of information, confidence measure,

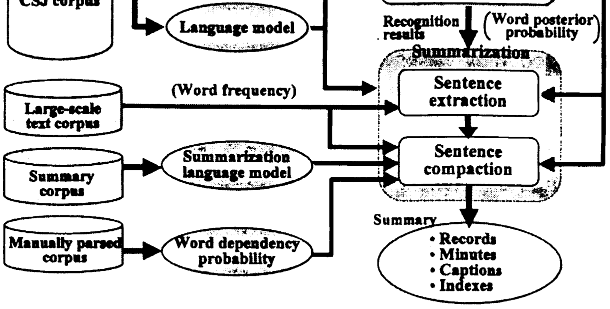

Fig. 4 A two-stage automatic speech summarization system.

and grammatical likelihood of concatenated units. The proposed method has been applied to summarization of broadcast news utterances as well as unrestricted-domain spontaneous presentations and has been evaluated by objective and subjective measures. It has been confirmed that the proposed method is effective in both English and Japanese speech summarization.

Speech summarization technology can be applied to any kind of speech document and is expected to play an important role in building various speech archives including broadcast news, lectures, presentations, and interviews. Summarization and question answering (QA) perform similar tasks, in that they both map an abundance of information to a (much) smaller piece which is then presented to the user. Therefore, speech summarization research will help the advancement of QA systems using speech documents. By condensing important points of long presentations and lectures and presenting them in a summary speech, the system can provide the listener with a valuable means for absorbing more information in a much shorter time.

7.2 Automatic Indexing of Presentations

Kawahara et al. have proposed a method for automatic detection of section boundaries and extraction of key sentences that can index the section topics from presentation audio archives [18]. The method makes use of pauses, language model information and "discourse markers", which are characteristic expressions used in initial utterances of sections. The discourse markers are derived in an unsupervised manner based on word statistics. The statistics of the presumed discourse markers are used to define the importance of sentences. This measure is also combined with the tf-idf measure based on content words. An experimental evaluation using the CSJ demonstrates that the proposed method provides better indexing of section boundaries compared with a simple baseline method using pause information only; and, that it is robust against speech recognition errors.

7.3 Voice QA Systems

Voice QA systems that respond to queries given by voice to retrieve exact answers or written articles from a wide range of domains are important and useful applications of speech recognition technology. In these systems, how to reduce speech recognition errors in the query utterances and their effect on the answers is one of the most important research issues from the usability point of view. We have investigated a method of generating effective domain-dependent language models for voice query recognition and a new dialogue strategy [34]. In the proposed interactive dialogue strategy using multimodal user interfaces, users are requested to indicate correct keywords, and incorrect keywords are automatically replaced by most probable keywords in the N-best list based on domain-dependent word co-occurrence scores. The word co-occurrence scores are used as a long-distance language model to augment the trigrams used in the voice query recognition. A preliminary QA system using voice input and graphical user interface has been implemented using NTT's SAIQA open-domain QA system.

8. Conclusions and Future Research

Recent progress of large-volume storage devices and high-speed networks has enabled digital archiving and streaming of audio and video materials. At many universities, multimedia archives of lectures are now being constructed. They are intended to help students audit lectures at convenient times and places at their own pace. In these audio archives, appropriate indices are necessary for efficient browsing and searching for portions of specific topics or speakers. Speech recognition needs to be used for automating the indexing process which would cost a lot if manually done.

Speech recognition technology is expected to be applicable not only to indexing of speech data (lectures, broadcast news, etc.) for information extraction and retrieval, but also to preparing minutes of meetings, closed captioning, and aids for the handicapped. Broadening these applications depends crucially on raising the recognition performance of spontaneous speech.

Various spontaneous speech corpora and processing technologies have recently been created under several recent projects including the "Spontaneous Speech" project in Japan. However, how to incorporate filled pauses, repairs, hesitations, repetitions, partial words, and disfluencies still poses a big challenge in spontaneous speech recognition.

The large-scale spontaneous speech corpus, CSJ (Corpus of Spontaneous Japanese) will be stored with XML format in a large-scale database system developed by the COE (Center of Excellence) program at Tokyo Institute of Technology so that the general population can easily access and use it for research purposes [35]. Since the recognition accuracy for spontaneous speech is still rather low, the collection of the corpus will be continued in the COE program in order to increase coverage of variations in spontaneous speech.

Human speech recognition is a matching process whereby heard speech is matched to existing various knowledge as shown in Fig. 5. To make significant progress in speech recognition, it is necessary to create a new paradigm

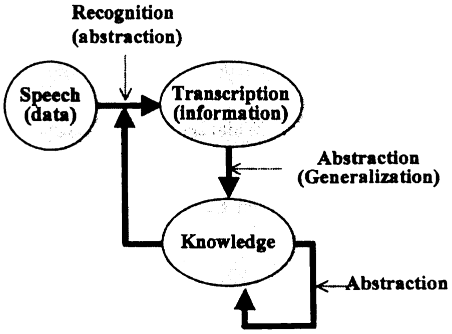

Fig. 5 Knowledge resources for speech recognition.

and rely much more on use of various knowledge sources in the recognition process. In speech summarization, future research includes investigation of other information/features for important unit extraction. Therefore, how to systematize and use various knowledge, such as domain and topic-related knowledge, context and speaker identity, are key issues in research on spontaneous speech recognition and summarization.

Although it is quite obvious that human beings effectively use prosodic features in speech recognition, how to use them in automatic speech recognition is still difficult. This is mainly because prosodic features are difficult to extract automatically and correctly from speech signal and difficult to model due to their dynamic natures. This is especially true for spontaneous speech.

This paper has focused on corpus-based spontaneous speech recognition issues mainly from the viewpoint of human-to-human monologue speech processing (Category II). Most of the issues discussed in this paper, however, are expected to be applicable to another important category, human-computer dialogue interaction (Category III).

Acknowledgments

The author would like to thank all members of the "Spontaneous Speech" project as well as students at Tokyo Institute of Technology and Kyoto University who contributed to building the large-scale corpus and creating the new technologies described in this paper.

References

[1] S. Furui, "Recent advances in spontaneous speech recognition and understanding," Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.1–6, Tokyo, 2003.

[2] S. Furui, "Toward spontaneous speech recognition and understanding," in Pattern Recognition in Speech and Language Processing, ed. W. Chou and B.-H. Juang, pp.191–227, CRC Press, New York, 2003.

[3] "The ICSI Meeting Recorder Project," http://www.icsi.berkeley.edu/Speech/mr/

[4] A. Waibel and I. Rogina, "Advances on ISL's lecture and meeting trackers," Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.127–130, Tokyo, 2003.

[5] Y. Akita, M. Nishida, and T. Kawahara, "Automatic transcription of discussions using unsupervised speaker indexing," Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.79–82, Tokyo, 2003.

[6] D.W. Oard, "Transforming access to the spoken word," Proc. International Symposium on Large-scale Knowledge Resources, pp.57–59, Tokyo, 2004.

[7] J. Hirschberg, M. Bacchiani, D. Hindle, P. Isenhour, A. Rosenberg, L. Stark, L. Stead, S. Whittaker, and G. Zamchick, "SCANMail: Browsing and searching speech data by content," Proc. Eurospeech2001, pp.2377–2380, Aalborg, 2001.

[8] G. Rigoll, "An overview on European projects related to spontaneous speech recognition," Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.131–134, Tokyo, 2003.

[9] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.7–12, Tokyo, 2003.

[10] K. Maekawa, H. Kikuchi, and W. Tsukahara, "Corpus of spontaneous Japanese: Design, annotation and XML representation," Proc. International Symposium on Large-scale Knowledge Resources, pp.19–24, Tokyo, 2004.

[11] K. Uchimoto, C. Nobata, A. Yamada, S. Sekine, and H. Isahara, "Morphological analysis of the Corpus of Spontaneous Japanese," Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.159–162, Tokyo, 2003.

[12] J. Venditti, "Japanese ToBI labeling guidelines," OSU Working Papers in Linguistics, 50, pp.127–162, 1997.

[13] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti, "X-JtoBI: An extended J_ToBI for spontaneous speech," Proc. ICSLP 2002, pp.1545–1548, Denver, 2002.

[14] L. Boves and N. Oostdijk, "Spontaneous speech in the spoken Dutch corpus," Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.171–174, Tokyo, 2003.

[15] T. Shinozaki, C. Hori, and S. Furui, "Towards automatic transcription of spontaneous presentations," Proc. Eurospeech2001, pp.491–494, Aalborg, Denmark, 2001.

[16] T. Ichiba, K. Iwano, and S. Furui, "Relationships between training data size and recognition accuracy in spontaneous speech recognition," Proc. Acoustical Society of Japan Fall Meeting, 2-1-9, 2004. (in Japanese)

[17] J. Ueberla, "Analysing a simple language model – some general conclusion for language models for speech recognition," Comput. Speech Lang., vol.8, no.2, pp.153–176, 1994.

[18] T. Kawahara, T. Kitade, K. Shitaoka, and H. Nanjo, "Efficient access to lecture audio archives through spoken language processing," Proc. Special Workshop in Maui (SWIM), 3.10, 2004.

[19] T. Kawahara, H. Nanjo, and S. Furui, "Automatic transcription of spontaneous lecture speech," Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, a01tk096, Madonna di Campiglio, Italy, 2001.

[20] T. Shinozaki and S. Furui, "Analysis on individual differences in automatic transcription of spontaneous presentations," Proc. IEEE ICASSP, pp.I-729–732, Orlando, 2002.

[21] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the corpus of spontaneous Japanese," Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.135–138, Tokyo, 2003.

[22] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Bayesian acoustic modeling for spontaneous speech recognition," Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.47–50, Tokyo, 2003.

[23] T. Shinozaki and S. Furui, "Hidden mode HMM using Bayesian network for modeling speaking rate fluctuation," Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.417–422, St. Thomas, 2003.

[24] L. Lussier, E.W.D. Whittaker, and S. Furui, "Combinations of language model adaptation methods applied to spontaneous speech," Proc. Third Spontaneous Speech Science & Technology Workshop, pp.73–78, Tokyo, 2004.

[25] H. Nanjo and T. Kawahara, "Unsupervised language model adaptation for lecture speech recognition," Proc. IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.75–78, Tokyo, 2003.

[26] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yamada, A. Ito, K. Itou, and K. Shikano, "Continuous speech recognition consortium – an open repository for CSR tools and models –," Proc. IEEE Int. Conf. on Language Resources and Evaluation, pp.1438–1441, Las Palmas de Gran Canaria, Spain, 2002.

[27] T. Shinozaki and S. Furui, "Spontaneous speech recognition using a massively parallel decoder," Proc. Interspeech-ICSLP, pp.1705–1708, Jeju, Korea, 2004.

[28] I. Mani and M.T. Maybury Ed., Advances in Automatic Text Summarization, MIT Press, Cambridge, MA, 1999.

[29] B. Kolluru, H. Christensen, Y. Gotoh, and S. Renals, "Exploring the style-technique interaction in extractive summarization of broadcast

news," Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.495–500, St. Thomas, 2003.

[30] H. Christensen, Y. Gotoh, B. Kolluru, and S. Renals, "Are extractive text summarization techniques portable to broadcast news," Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.489–494, St. Thomas, 2003.

[31] C. Hori and S. Furui, "A new approach to automatic speech summarization," IEEE Trans. Multimedia, vol.5, no.3, pp.368–378, 2003.

[32] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel, "A statistical approach to automatic speech summarization," EURASIP Journal on Applied Signal Processing, pp.128–139, 2003.

[33] T. Kikuchi, S. Furui, and C. Hori, "Two-stage automatic speech summarization by sentence extraction and compaction," Proc. ISCA-IEEE Workshop on Spontaneous Speech Processing and Recognition, TAP10, Tokyo, 2003.

[34] D. Kim, S. Furui, and H. Isozaki, "Language models and dialogue strategy for a voice QA system," Proc. ICA2004, pp.V-3705–3708, Kyoto, 2004.

[35] S. Furui, "Overview of the 21st century COE program "Framework for Systematization and Application of Large-scale Knowledge Resources,"" Proc. International Symposium on Large-scale Knowledge Resources, pp.1–8, Tokyo, 2004.

Sadaoki Furui is currently a Professor at Tokyo Institute of Technology, Department of Computer Science. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction and has authored or coauthored over 400 published articles. From 1978 to 1979, he served on the staff of the Acoustics Research Department of Bell Laboratories, Murray Hill, New Jersey, as a visiting researcher working on speaker verification. He is a Fellow of the IEEE, the Acoustical Society of America and the IEICE. He was President of the Acoustical Society of Japan (ASJ) from 2001 to 2003, and is currently President of the International Speech Communication Association (ISCA) and the Permanent Council for International Conferences on Spoken Language Processing (PC-ICSLP). He has served as Editor-in-Chief of both the Journal of Speech Communication and the Transaction of the IEICE. He is an Editorial Board member of Speech Communication, the Journal of Computer Speech and Language, and the Journal of Digital Signal Processing. He has received the Yonezawa Prize and the Paper Awards from the IEICE (1975, 88, 93, 2003), and the Sato Paper Award from the ASJ (1985, 87). He has received the Senior Award from the IEEE ASSP Society (1989) and the Achievement Award from the Minister of Science and Technology, Japan (1989). He has received the Technical Achievement Award and the Book Award from the IEICE (2003, 1990). In 1993, he served as an IEEE SPS Distinguished Lecturer.