論文 / 著書情報
Article / Book Information

題目(和文)	話し言葉音声認識のための統計的モデル化の研究
Title(English)	Statistical modeling for spontaneous speech recognition
著者(和文)	篠崎隆宏
Author(English)	
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第5848号, 授与年月日:2004年3月26日, 学位の種別:課程博士, 審査員:
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第5848号, Conferred date:2004/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Statistical Modeling for Spontaneous Speech Recognition

Takahiro Shinozaki

Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology

February, 2004

Contents

List of Figures

List of Tables

Chapter 1

Abstract

The characteristics of spontaneous speech are very different from read speech and recognition accuracy of conventional recognition systems drastically decreases for spontaneous speech. Since most of our speech is spontaneous, it is strongly desirable to improve recognition technique for spontaneous speech.

This study began with building a recognition system that was based on the Japanese spontaneous speech corpus CSJ. Experimental results show that acoustic and language modeling based on an actual spontaneous speech corpus is far more effective than conventional modeling based on read speech. However, recognition accuracy for spontaneous speech is still low, and a large number of research issues remain unresolved.

To understand problems of spontaneous speech recognition, various analyses were conducted. These analyses include correlation and regression analyses for individual differences in recognition performance, data-mining using decision trees, and comparison with performances by human listeners. As a result, a restricted set of attributes that are closely related to the recognition errors were identified.

One conclusion of the analyses was that difficulty of recognizing a word largely depends on the length and frequency of the word. Based on this observation, a new lexicon optimization method has been proposed. The proposed method optimizes the lexicon by making compound words or phrases step by step based on a word correctness probability model so as to improve the estimated recognition rate of the system. Experimental

results showed that the language model using the optimized lexicon improved the recognition rate. To cope with the degradation of recognition accuracy due to speaking rate fluctuation, a new acoustic model has been proposed. 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. In the experiments using the Bayesian network framework, the proposed model indicated consistently higher performance than conventional HMMs. To deal with utterances with various characteristics, the Massively Parallel Decoder (MPD) has been proposed. The MPD works with a cluster speech model that covers diversity of spontaneous utterances. In the experiment, MPDs with up to 400 decoding units were constructed on a GRID system. It was confirmed that the MPD were effective in improving recognition accuracy.

Chapter 2

Introduction

Read speech and similar types of speech, e.g. that from reading newspapers or from news broadcasts, can be recognized with accuracy higher than 90% using state-of-the-art speech recognition technology. However, recognition accuracy drastically decreases for spontaneous speech. This decrease is due to the fact that acoustic and linguistic models used have generally been built using written language or speech from written language whereas spontaneous speech is very different both acoustically and linguistically. Spontaneous speech includes many phenomena such as vague pronunciations, repetitions, and filled pauses. These phenomena make spontaneous speech recognition inherently difficult.

Since spontaneous speech comprises the major vehicle of human communication, broadening the application of speech recognition crucially depends on raising the recognition performance for spontaneous speech. Currently, our knowledge about the structure of spontaneous speech is inadequate to achieve necessary breakthroughs. Modeling of speech disfluencies is only just the beginning.

2.1 Research purpose

The purpose of this study is to improve recognition performance for spontaneous speech. This is the first study that tackles spontaneous speech recognition using a large scale Japanese spontaneous speech corpus. Since there is little knowledge about spontaneous

speech recognition, analysis of spontaneous speech from an engineering point of view and proposal of new techniques based on the analysis are two inseparable parts of this study.

2.2 Overview of this thesis

This thesis is organized as follows. Chapter 3 describes the conventional automatic speech recognition system and Chapter 4 describes human auditory system.

Analyses on spontaneous sounds are conducted in the following four chapters. In Chapter 5, various initial investigations on recognizing spontaneous presentation speech in connection with the "Spontaneous Speech" national project started in 1999 are reported. Chapter 6 reports an analysis of individual differences in spontaneous presentation speech recognition performances. Chapter 7 proposes the use of decision trees for analyzing errors in spontaneous presentation speech recognition. In Chapter 8, an automatic speech recognizer is evaluated in comparison with performances by human listeners to investigate problems of spontaneous speech recognition using N-grams and HMMs and estimates the room for improvement in the recognition rate.

Based on the analysis results in the previous chapters, three new method are proposed in the following three chapters. One observation obtained through the analyses is that difficulty of recognizing a word largely depends on the length and frequency of the word. A lexicon optimization method to improve recognition rate of large scale spontaneous speech recognition is proposed in Chapter 9. To cope with the degradation of recognition accuracy due to speaking rate fluctuation within an utterance, a new acoustic model for adjusting mixture weights and transition probabilities of the HMM for each frame according to the local speaking rate is proposed in Chapter 10. To achieve high recognition performance using cluster speech models, Massively Parallel Decoder (MPD) is proposed in Chapter 11. The MPD consists of a large number of decoding units and an integrator. It runs on a parallel computer and can process speech utterance with almost the same turnaround time as conventional decoders. Finally in Chapter 12,

conclusion and future works are presented.

Chapter 3

Automatic Speech Recognition

3.1 Mechanism

Speech recognition is a technique that transforms speech signals to a textual message. The most popular approach is to view speech as a signal that can be modeled as a stochastic process. In this approach, speech recognition problems can be mathematically described by equation (3.1).

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|X), \tag{3.1}$$

where W is a word sequence and X is an acoustic signal. Because directly modeling the conditional probability of W given X is difficult, the equation is transformed as shown in equation (3.2) using Bayes' theorem. In the argmax operator, $P(X)$ is constant for W and the term is simplified as shown in equation (3.3).

$$\hat{W} = \underset{W}{\operatorname{argmax}} \frac{P(X|W)P(W)}{P(X)} \tag{3.2}$$

$$= \underset{W}{\operatorname{argmax}} P(X|W)P(W). \tag{3.3}$$

The likelihood of $P(X|W)$ is estimated using an acoustic model and $P(W)$ is obtained using a language model. The role of a decoder is to search for the $\hat{W}$ that gives the highest total likelihood using the acoustic and language model. For large vocabulary continuous speech recognition, HMMs and Ngrams are commonly used as acoustic and language models, respectively.

3.2 Acoustic analysis

The input analogue speech signal is first sampled for digital processing. For speech recognition, 16000Hz is high enough for the sampling frequency. The digitized signal is then transformed into feature vectors that can be analyzed by the decoder. One of the most commonly used methods for analyzing digital speech signals is Mel-frequency cepstral coefficient (MFCC) analysis. In the MFCC analysis, the speech data is transformed using a Fourier transform and the magnitude is taken. The magnitude coefficients are then binned by accumulating its their values multiplied by the corresponding triangular filter gain. Thus, each bin holds the weighted sum representing the spectral magnitude in that filter-bank channel. The filters are equally spaced along the mel-scale which is defined by

$$Mel\,(f) = 2595 \log\left(1 + \frac{f}{700}\right). \tag{3.4}$$

MFCCs are calculated from the log filterbank amplitudes m_j using the Discrete Cosine Transform

$$c_i = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} m_j cos\left(\frac{\pi \times i}{N}\,(j - 0.5)\right), \tag{3.5}$$

where N is the number of filterbank channels. MFCCs are the parameterization of choice for many speech recognition applications. They give good discrimination and lend themselves to a number of manipulations.

In human speech perception, it was shown that dynamic features of the spectrum plays an important role. To use these features, the delta cepstrum was developed [1]. The delta cepstrum is defined as shown in equation (3.6),

$$d_t = \frac{\sum_{i=1}^{\theta} i\,(c_{t+i} - c_{t-i})}{2\sum_{i=1}^{\theta} i^2} \tag{3.6}$$

where d_t is a delta coefficient at time t computed in terms of the corresponding static coefficients c_{t+i} to c_{t-i}.

3.3 Hidden Markov models

An HMM can be characterized by a set of parameters

$$\lambda = \{S, O, A, B, \Pi, F\}. \tag{3.7}$$

S: A set of states, $S = \{s_1 \cdots s_n\}$.

O: A set of output symbols.

A: A set of states transition probabilities. $a_{i,j}$ represents transition probability from state s_i to state s_j.

B: A set of output distribution functions.

Π: A set of initial state probabilities.

F: A set of final states.

The output distribution function of a state can be either discrete or continuous. In the discrete case, the distribution function is a set of symbols in which each has an associated probability specifying the likelihood that the symbol will be observed in that state. In the continuous case, a mixture of Gaussian functions are widely used as the output distribution.

3.4 Ngram models

The language model $P(W)$ for word sequences W

$$W = w_1 w_2 \cdots w_k, \tag{3.8}$$

can be decomposed as follows by using the chain rule.

$$P(W) = \prod_{i=1}^{k} P(w_i | w_1 w_2 \cdots w_{i-1}). \tag{3.9}$$

Usually an N-gram approximation is applied to make it feasible to estimate the parameters from limited amounts of training data. An N-gram model approximates the word

sequence probability $P(W)$ as follows:

$$P_N(W) = \prod_{i=1}^{k} P(w_i | w_{i-N+1} \cdots w_{i-2} w_{i-1}). \tag{3.10}$$

The conditional probabilities $P(w_i | w_1 w_2 \cdots w_{i-N+1})$ can be estimated by using the relative frequencies of the word sequences.

$$P(w_i | w_1 w_2 \cdots w_{i-N+1}) = \frac{F(w_i w_{i-1} \cdots w_{i-N+1})}{F(w_{i-1} w_{i-2} \cdots w_{i-N+1})}, \tag{3.11}$$

where F is the number of occurrences of word sequences in the training corpus. Usually, a bigram in which N = 2 or a trigram in which N = 3 is used. To compensate for probabilities of word sequences that do not occur in the training corpus, back-off smoothing is widely used.

Chapter 4

Human Auditory System

Sound waves are captured by the auricle and lead into the external auditory meatus. The waves are then sent to oeil-de-boeuf of the cochlea via impedance matching by the auditory ossicle in the middle ear. Transformation from the sound waves to nerve impulses is conducted by the cochlea. The cochlea is a snail-shaped fistulous organ. The tube inside the cochlea is filled with liquid and separated into three rooms by the basilar membrane and the Reissner's membrane along its long direction.

On the basilar membrane, there is one line of inner hair cells and three lines of outer hair cells. Because of the mechanical structure of the cochlea, waves of different frequencies cause different segments of the membrane to vibrate. Each inner hair cell detects the vibration according to its position, which corresponds to a specific frequency, and outer hair cell adjusts its sensitivity. There are about 15000 hair cells in a human cochlea and they are connected to about 30000 spiral ganglion nerves by synapses. More than 90% of the synapses are for the inner hair cells. One inner hair cell is connected to about 20 nerves while approximately 20 outer hair cells are connected to one neuron. The hair cells have efferent synapses in addition to afferent synapses. The axons of the efferent synapses come from cells in the olivary complex.

The nerve fibers from the spiral ganglion go into the brainstem as cochlear nerves and terminate in dorsal and ventral cochlear nuclei. The nerve fibers from the cochlear nuclei go to nuclei of inferior colliculus by way of the trapezoid nucleus, superior olivary

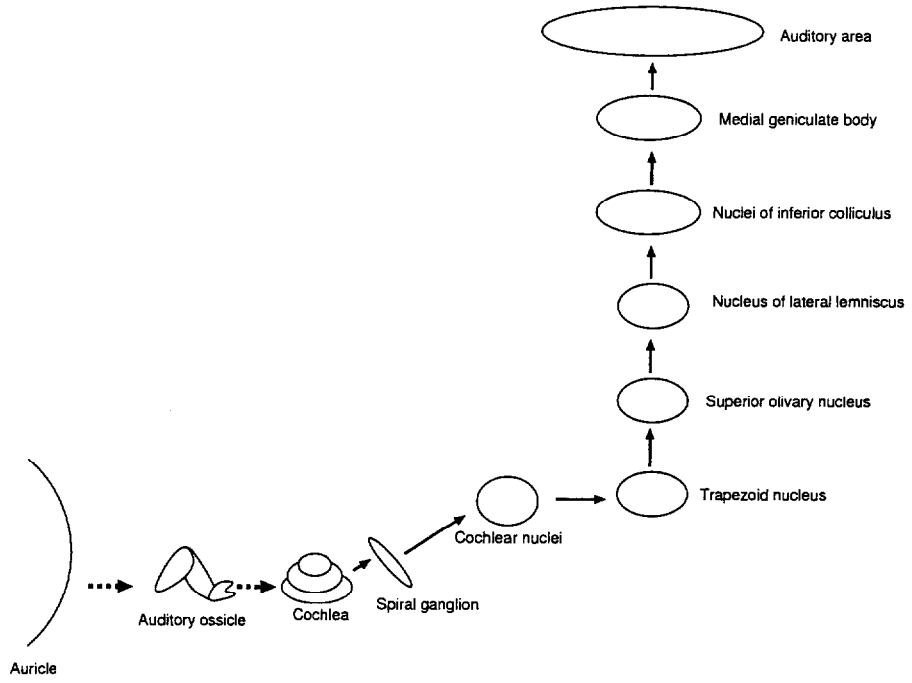

Figure 4.1: Human auditory system.

nucleus, and nucleus of lateral lemniscus. The nerve fibers from the inferior colliculus terminate in medial geniculate body. Right and Left relay nuclei of auditory pathway are connected by commissural neurofibers in every level. Gradually complex processes are applied by the relay nuclei such as depressing and accommodation.

The nerve fibers from the medial geniculate body terminate at the auditory area (Broadmann's Area 41, 42) in the temporal lobe of the cerebral cortex. Primary auditory area is a center of the sense of hearing and is situated at the top (Area 41) of the superior temporal gyrus. Secondary auditory area is around the primary auditory area and recognizes the meaning of the sound. Lesion in secondary auditory area causes sensory aphasia. Recently, fMRI has been used to observe activity of the cerebral cortex. It has been observed that noisy sound activates more areas than clean sounds [2].

Human cerebral cortex is made up of about 14 billion nerve cells. The whole central nervous system is estimated to consist of 100 to 200 billion nerve cells. The computa-

tional performance of a brain is roughly estimated using this equation

$$p = f \times s \times n, \tag{4.1}$$

where f is a frequency of nerve impulse, s is the number of synapses per nerve cell, and n is the number of nerve cells in a brain. By assuming that $f = 100$, $s = 10000$, and $n = 100000000000$, the performance of a human brain is estimated to $1000000000000000000 = 100$ Peta Flops. Supposing that the nerve cells related to auditory processing is 1% of the cells in the central nerve system, one Peta Flops machine is required to emulate the human auditory system which is 25 times faster than the earth simulator [3].

Chapter 5

Initial Results Using Corpus of Spontaneous Japanese

5.1 Introduction

Applying acoustic and language models based on written language to spontaneous speech results in poor recognition accuracy due to acoustic and linguistic mismatch. To improve technologies for spontaneous speech, a large scale spontaneous speech corpus is indispensable. However, until recently there was not such a Japanese corpus.

To build models and technology for spontaneous speech recognition, the Science and Technology Agency Priority Program (Organized Research Combination System) entitled "Spontaneous Speech: Corpus and Processing Technology" was started in 1999 under the supervision of Furui [4]. The project is being conducted over a 5-year period in pursuit of the following three major goals:

1. Building a large-scale spontaneous speech corpus consisting of approximately 7M words with a total speech length of 700 hours. The majority of the recordings will be monologues such as lectures, presentations, and news commentaries. They will be manually given orthographic and phonetic transcription. Since there is no clear definition of words in Japanese and no spacing between words in written Japanese sentences, a morphological analysis program will be used to split transcribed sentences into morphemes.

2. Acoustic and linguistic modeling for spontaneous speech understanding and summarization using linguistic as well as para-linguistic information.

3. Constructing a prototype of a spontaneous speech summarization system.

This chapter reports results of preliminary recognition experiments utilizing the corpus. Section 5.2 describes the task and experimental conditions. Section 5.3 describes acoustic and language models for recognition. Recognition results are presented in Section 5.4, and Section 5.5 gives some analysis on individual variations of recognition results. Section 5.6 reports the improvement by unsupervised speaker adaptation. Finally, some conclusions are given in Section 5.7.

5.2 Recognition task and experimental conditions

5.2.1 Recognition task

Presentation speech uttered by 10 male speakers was used as a test set of speech recognition. Table 5.1 shows an outline of the test set. The top four presentations in the table were on the subject of speech.

Morphemes (which will be called "words" hereafter in this chapter) were used as units for statistical language modeling. For all the following recognition performances, word-based performance is measured. Fillers are counted as words and taken into account in calculating the accuracy.

5.2.2 Experimental conditions

Sounds were digitized with 16kHz sampling and 16bit quantization. They were segmented into utterances using silence periods longer than 500ms. Feature vectors had 25 elements consisting of 12 MFCC, their delta, and delta log energy. CMS (cepstral mean subtraction) was applied to each utterance. HTK v2.2 [5] was utilized for acoustic modeling and speaker adaptation. Language models were made by the use of CMU SLM Tool Kit v2.05. The Julius v3.1 decoder [6] was used for speech recognition.

Table 5.1: Recognition test set of presentations

ID	Conference name	Length [min]
A22	Acoust. Soc. Jap.	28
A23	Acoust. Soc. Jap.	30
A97	Acoust. Soc. Jap.	12
P25	Phonetics Soc. Jap.	27
J01	Soc. Jap. Linguistics	57
K05	National Lang. Res. Inst.	42
N07	Assoc. Natural Lang. Proc.	15
S05	Assoc. Socioling. Sciences	23
Y01	Spont. Speech Corpus Meeting	14
Y05	Spont. Speech Corpus Meeting	15

5.3 Language and acoustic modeling

5.3.1 Corpora

The following two corpora were used for training.

- Spontaneous Speech Corpus (CSJ): A part of the corpus completed by the end of December 2000, consisting of approximately 1.5M words of transcriptions, was used. The training set consisted of 610 presentations; 274 academic conference presentations and 336 simulated presentations. The simulated presentations were specially recorded for the project and consisted of a wide variety of topics including subjects' talking about experiences in their daily lives.

- Web corpus: Transcribed presentations having roughly 76k sentences with 2M words were collected from the World Wide Web. Spontaneous speech usually includes various filled pauses but they were not included in this presentation corpus. An effort was thus made to add filled pauses to the presentation corpus based on statistical characteristics of the filled pauses. Their topics covered wide domains including social issues and memoirs.

5.3.2 Language modeling

The following three language models were built. Each model consisted of bigrams and reverse trigrams with backing-off. Their vocabulary sizes were all 30k.

SpnL: Made using 610 presentations in the CSJ. The speakers had no overlap with those of the test set. Since there were no punctuation marks in the transcription, commas were inserted at silences of 200ms or longer duration.

WebL: Made using the text of our Web corpus.

WebSpL: Made by adding whole the text of a textbook on speech processing authored by Furui to the Web corpus with equal weighting for task adaptation. The textbook contains about 63k words.

Table 5.2 shows an outline of the language models.

Table 5.2: Corpus size for training each language model

Language model	Corpus size [words]
SpnL	1.5 M
WebL	2 M
WSpL	2+0.06 M

5.3.3 Acoustic modeling

The following two tied-state triphone HMMs were made. Both models have 2k states and 16 Gaussian mixtures in each state.

SpnA: Using 338 presentations in the CSJ uttered by male speakers (approximately 59 hours). The speakers had no overlap with those in the test set.

RdA: The acoustic model made by the Information-technology Promotion Agency (IPA) and contained in the CD-ROM "Japanese Dictation Toolkit 99". Approximately 40-hours of read speech uttered by many speakers was used.

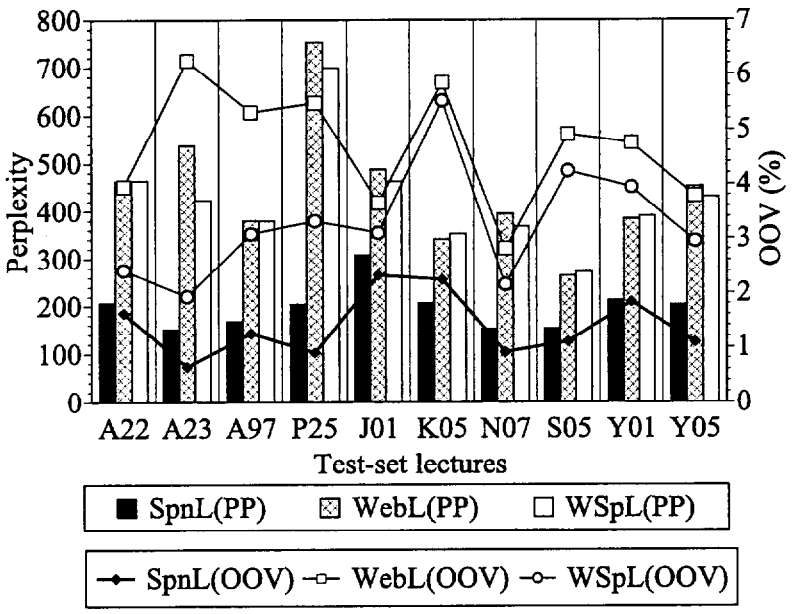

Figure 5.1: Test-set perplexity and OOV rate for the three language models.

5.4 Experimental results

5.4.1 Test-set perplexity and OOV rate

Figure 5.1 presents test-set perplexity of tri-grams and out-of- vocabulary (OOV) rate for each language model. The perplexity of SpnL made from the CSJ is clearly better than that of other models. WebL indicates high perplexity and OOV rate. This is because WebL is edited as a text and the topics are general. The OOV rate of WSpL is smaller than that of WebL for the four left-hand-side speeches. This shows that task adaptation by adding the textbook worked well. SpnL is superior to WSpL also in terms of the OOV rate.

5.4.2 Effects of language modeling

Figure 5.2 shows recognition results for the three language models when SpnA is used as the acoustic model. SpnL achieves the best results. WSpL achieves better results than WebL, especially for test sets A22, A23, A97 and P25, reflecting the test-set perplexity and OOV rate reduction. Mean accuracies are 64.3%, 54.9% and 57.1% for SpnL, WebL

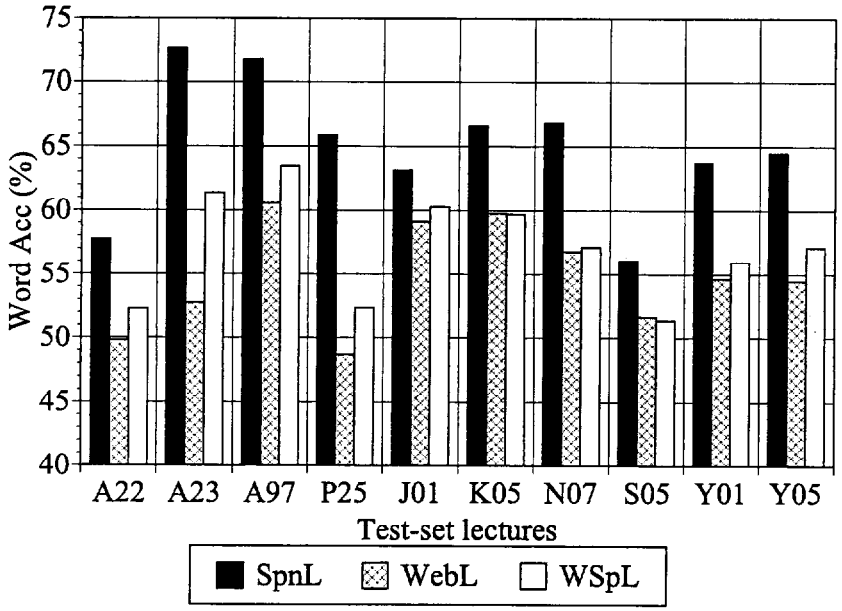

Figure 5.2: Word accuracy for the three language models.

and WSpL, respectively. A supplementary experiment was performed to analyze the effects of OOV rate and test set perplexity to the accuracy. In this experiment, OOV words were added to the language models as "unknown" class words; 489 words and 710 words were added to SpnL and WSpL, respectively. Resulting mean word accuracies using SpnL and WSpL were 65.8% and 59.9%, respectively. These results indicate that OOV is an equally important problem as an aspect of test-set perplexity in these models.

5.4.3 Effects of acoustic modeling

The recognition results for SpnA and RdA when SpnL is used as the language model are shown in Fig. 5.3. Mean accuracies are 64.3% and 53.0% for SpnA and RdA, respectively. SpnA made from the CSJ achieves much better results than RdA made from read speech. This is probably because SpnA has better coverage of triphones and better matching of acoustic characteristics corresponding to the speaking style.

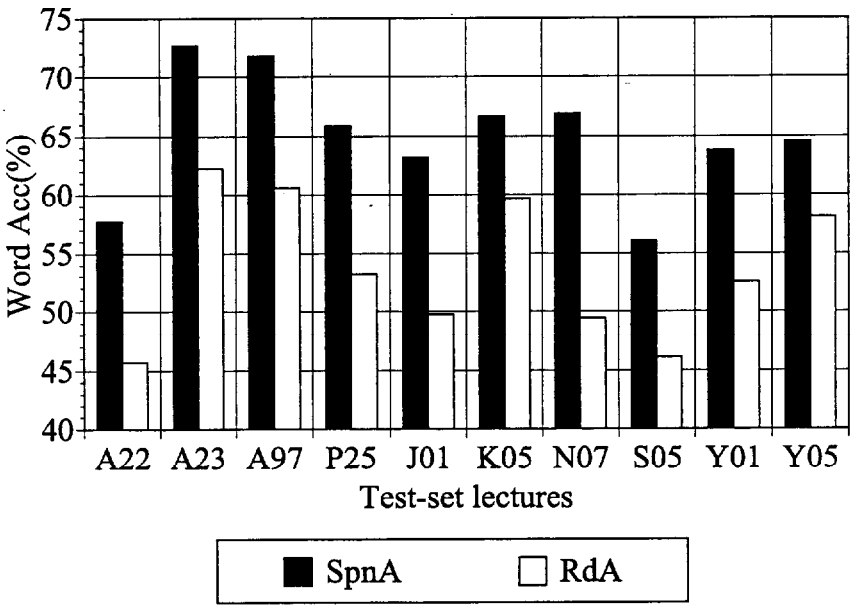

Figure 5.3: Word accuracy for the two acoustic models.

5.5 Individual differences

As shown in Figures 5.2 and 5.3, word accuracy varies largely from speaker to speaker. There exist many factors that affect the accuracy of spontaneous speech recognition. These factors include individual voice characteristics and speaking manner, including noises like coughing. Although all utterances were recorded using the same close-talking microphones, acoustic conditions still varied according to the recording environment.

Figure 5.4 presents relationship between speaking rate and word accuracy when SpnL and SpnA were used as language and acoustic models. The speaking rate was calculated using actual speech periods after removing pauses. 10 dots in the figure correspond to individual speakers. A MMSE line fitted to those dots is also shown in the figure. The correlation coefficient is -0.58. Faster speech generally produces more errors.

Figures 5.5 and 5.6 respectively show the effects of frequencies of fillers and repairs on word accuracy. The recognition conditions were the same as those for Fig. 5.4. There is a general tendency that the more frequently the filler and/or the repair occurs, the

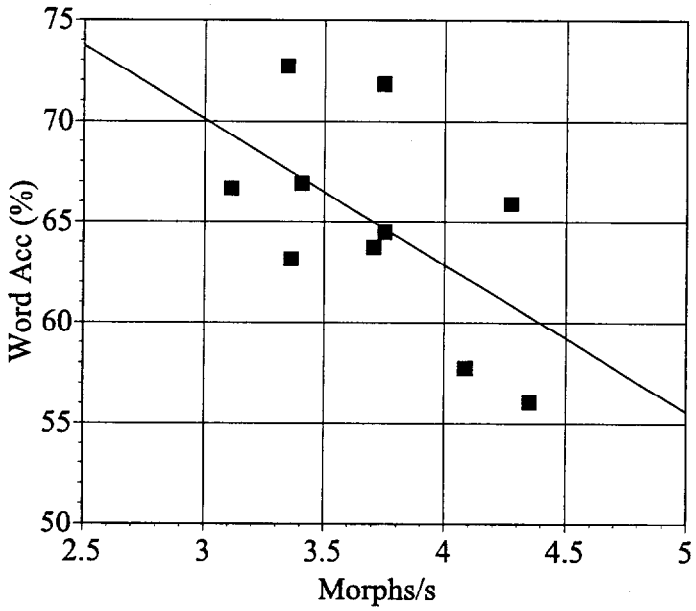

Figure 5.4: Speaking rate vs. word accuracy.

more recognition errors occur.

5.6 Unsupervised adaptation

A batch-type unsupervised adaptation method was incorporated to cope with speech variation due to speakers and recording environment. The MLLR method using a binary regression class tree to transform Gaussian mean vectors was applied to the HMM. The regression class tree was made using a centroid-splitting algorithm. The actual classes used for transformation were determined on run time according to the amount of data assigned to each class.

The following steps were carried out. The adaptation was performed based on recog-

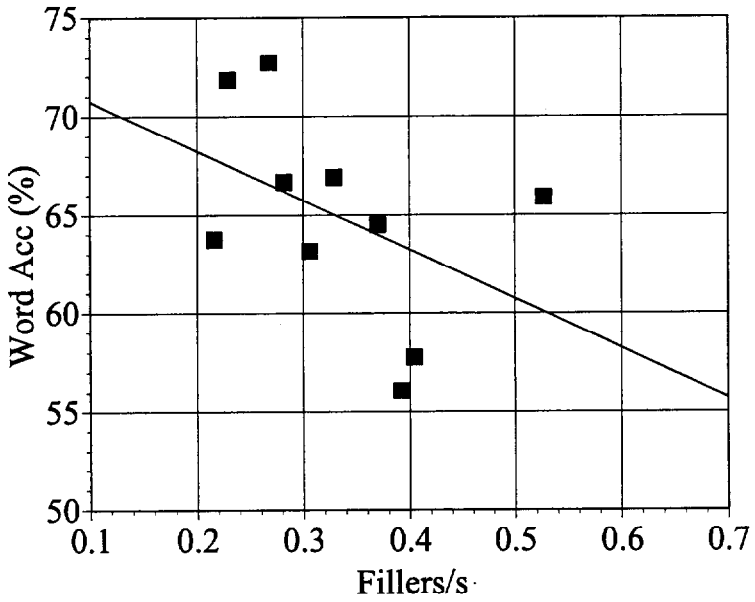

Figure 5.5: Filler frequency vs. word accuracy.

nition results and no confidence measure was applied.

1. Making a regression class tree having 64 leaf nodes for the SpnA phone model.

2. Recognizing the test-set utterances using the SpnA as a speaker independent model.

3. Applying the MLLR adaptation based on the recognition result for each utterance to make a speaker adaptive model.

4. Re-recognizing the test-set utterances using the speaker adaptive model.

5. Iterating the adaptation process using the resulting transcription.

Figure 5.7 presents the effect of the adaptation when SpnL was used as the language model. "SpnA" indicates the baseline condition. "mllr" indicates the result without iterations and "mllr-i" indicates the results after one iteration of adaptation. The single step of MLLR improved word accuracy by 2 to 6 %, and the second adaptation step

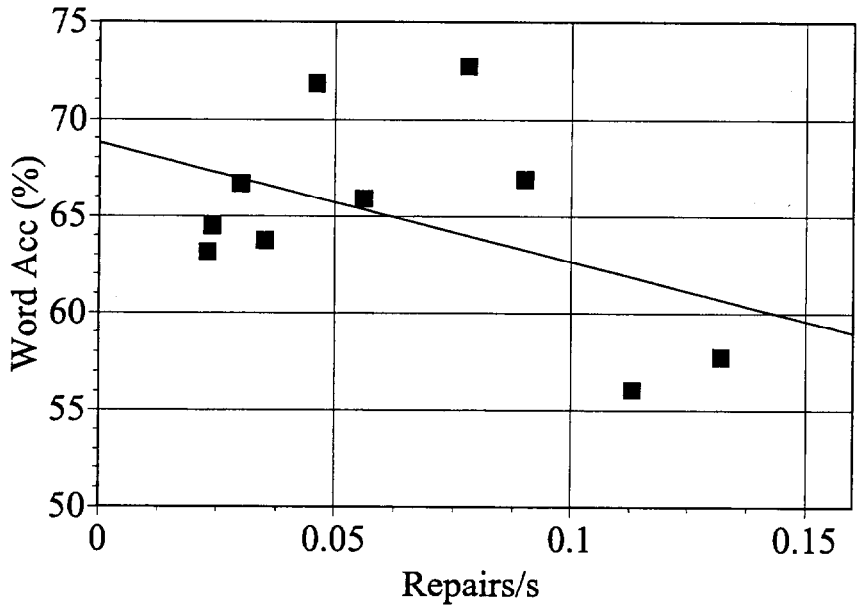

Figure 5.6: Repair frequency vs. word accuracy.

further improved accuracy by 1% on average. By applying the two steps of MLLR, the error rate was reduced by 15% relative to the speaker independent case.

5.7 Conclusions

This chapter reported experimental results for recognizing spontaneous presentation speech. Language models based on a spontaneous speech corpus and Web corpus were compared in terms of test-set perplexity, OOV rate, and word (morpheme) accuracy. Two acoustic models made by spontaneous speech and read speech were also compared. Both comparisons showed that models made from spontaneous speech were much superior to models based on read speech. It was revealed that recognition accuracy had a wide speaker-to-speaker variability. Correlation between word accuracy and speaking rate, filler and repair frequency was observed. When linguistic and acoustic models made form spontaneous speech were used, an average word recognition accuracy of 64.3% was achieved. This performance improved to 69.8% with the help of unsupervised MLLR adaptation for the acoustic model.

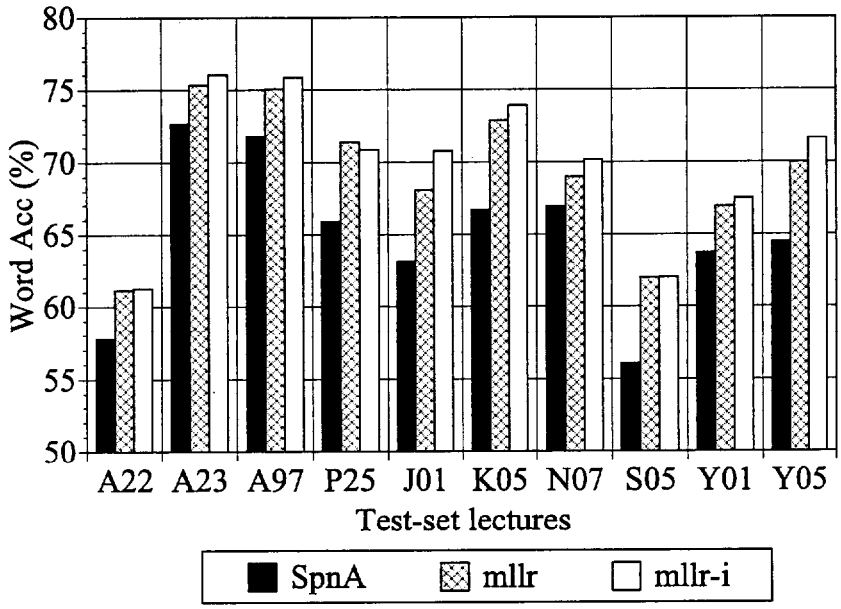

Figure 5.7: Results of unsupervised adaptation.

Although word accuracy was improved by using the spontaneous speech corpus, it is still not sufficient for building application systems. To understand problems in spontaneous speech recognition and improve recognition performance, further analysis of spontaneous speech is necessary.

Chapter 6

Analysis on Individual Differences

6.1 Introduction

In the previous chapter, it was shown that acoustic and language models made using the CSJ were significantly superior to conventional read-speech-based models when applied to spontaneous speech recognition [7]. However, recognition accuracy is still rather low, and there presumably exist many factors that affect recognition performance acoustically as well as linguistically.

It is presumable that variation of speaking style is larger in spontaneous speech than in read speech according to the degree of speaker's freedom. And so does the word accuracy of recognition systems. Knowing the structure of speaking style differences among individuals and the influence on word accuracy it exerts is very important to improve spontaneous speech recognition systems. This chapter reveals the structure of individual differences in word accuracy based on recognition results in presentation speech uttered by 50 male speakers processed by a state-of-the-art recognition system.

Section 6.2 describes the task and experimental set up. Experimental results and analyses are presented in Section 6.3. Finally, some conclusions are given in Section 6.7.

6.2 Recognition task and experimental set up

6.2.1 Recognition task

For the analysis of speaker variation, monologue presentation speech uttered by 50 different male speakers was used as a test set. Speakers in the test set had no overlap with those in the training set. The first 10 minutes of each presentation were used for analysis. Table 6.1 shows the details of the test set.

Table 6.1: Test set

Conference	No. presentations
Jap. Soc. AI	32
Acoust. Soc. Jap.	12
Others	6

6.2.2 Speaker attributes

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 which was defined as the number of phonemes per second and the averaged acoustic frame likelihood were calculated using the result of forced alignment of the reference tri-phone label after removing pause periods. Word perplexity was calculated using tri-grams, in which prediction of out of vocabulary words was not included. The filled pause rate and the repair rate were the percentage of filled pauses and repairs in total words, respectively. Tag information included in CSJ transcription was used to determine whether a word was a filled pause/repair or not. In CSJ, repairs are defined only for word fragments, and a whole word which is rephrased is not marked as a repair. The calculations of word accuracy, out of vocabulary rate and word perplexity were based on the reference sentence after excluding repairs.

6.2.3 Experimental conditions

Speech signals were digitized with 16kHz sampling and 16bit quantization. Feature vectors have 25 elements consisting of 12 MFCC, their delta and the delta log energy. The CMS (cepstral mean subtraction) was applied to each utterance. HTK v2.2 was used for acoustic modeling and adaptation. Language models were made by using the CMU SLM Tool Kit v2.05. Morphemes (which will be called "words" hereafter in this chapter) were used as units for statistical language modeling. The Julius v3.1 decoder [6] was used for speech recognition.

6.2.4 Language and acoustic modeling

A part of the CSJ, having approximately 1.5M words, was used as a training set. The training set consisted of 610 presentations; 274 academic conference presentations and 336 simulated presentations.

The language model used in the recognition consisted of bi-grams and reverse tri-grams with backing-off. It was made using the whole training set. The vocabulary size was 30k. Filled pauses were treated as words in modeling. Repairs were deleted from the training text and were not modeled. This is because modeling repairs effectively by N-gram is difficult due to a large amount of variations and few occurrences of each fragment.

A speaker independent (SI) acoustic model was made using 338 presentations uttered by male speakers (approximately 59 hours). It was a tied-state tri-phone HMM having 2k states and 16 Gaussian mixtures in each state. Each tri-phone HMM had three states with the left-to-right structure.

In addition, A batch-type unsupervised speaker adaptation was incorporated to see the effect on the individual differences. The MLLR method was applied to the speaker independent HMM in which a regression class tree having 64 leaves was made using a centroid-splitting algorithm. The resulting set of speaker adaptive HMMs for the 50 test set speakers is denoted as SA HMMs.

Table 6.2: Mean and standard deviation for each attribute

	Acc(SI)	Acc(SA)	AL(SI)	AL(SA)	SR	PP	OR	FR	RR
Mean	64.2	68.6	-55.4	-53.1	15.0	224	2.09	8.59	1.56
Standard deviation	7.4	7.5	2.3	2.2	1.2	61	1.18	3.67	0.72

Table 6.3: Correlation coefficient matrix: the lower triangular matrix shows the correlation coefficients and the upper triangular matrix shows the p-value, that is, the significance level. Bold face indicates a significant value with the significant level of 5%

	Acc(SI)	Acc(SA)	AL(SI)	AL(SA)	SR	PP	OR	FR	RR
Acc(SI)		–	5.4%	–	0.1%	0.5%	0.0%	0.6%	2.2%
Acc(SA)	–		–	2.4%	0.0%	1.6%	0.0%	0.6%	2.4%
AL(SI)	0.27	–		–	0.0%	65.1%	12.5%	6.9%	46.9%
AL(SA)	–	**0.32**	–		0.0%	52.3%	8.6%	7.0%	34.7%
SR	**-0.47**	**-0.49**	**-0.59**	**-0.64**		65.1%	1.2%	0.0%	34.0%
PP	**-0.39**	**-0.34**	-0.07	-0.09	0.07		0.0%	18.0%	44.8%
OR	**-0.54**	**-0.51**	-0.22	-0.25	**0.35**	**0.53**		0.3%	67.9%
FR	**0.38**	**0.38**	0.26	0.26	**-0.51**	-0.19	**-0.41**		32.9%
RR	**-0.32**	**-0.32**	-0.10	-0.14	0.14	0.11	-0.06	0.14	

6.3 Basic characteristics of the speaker attributes

Table 6.2 shows the mean and standard deviation over the 50 speakers for the word accuracy and other six kinds of speaker attributes. The calculation of the speaking rate is based on the SI HMM. The mean word accuracy of the 50 speakers is 64.2% and 68.6% for the SI and SA conditions, respectively. The standard deviation is 7.4% for the SI and 7.5% for the SA condition. As shown by the standard deviation, recognition accuracy largely varies from speaker to speaker. Correlation and regression analysis are discussed in 6.4 and in 6.5, respectively.

6.4 Correlation analysis

Table 6.3 shows the correlation matrix of speaker attributes. In the table, the lower triangular matrix shows the correlation coefficients and the upper triangular matrix shows the observed significance levels (p-values). The correlation coefficients written in

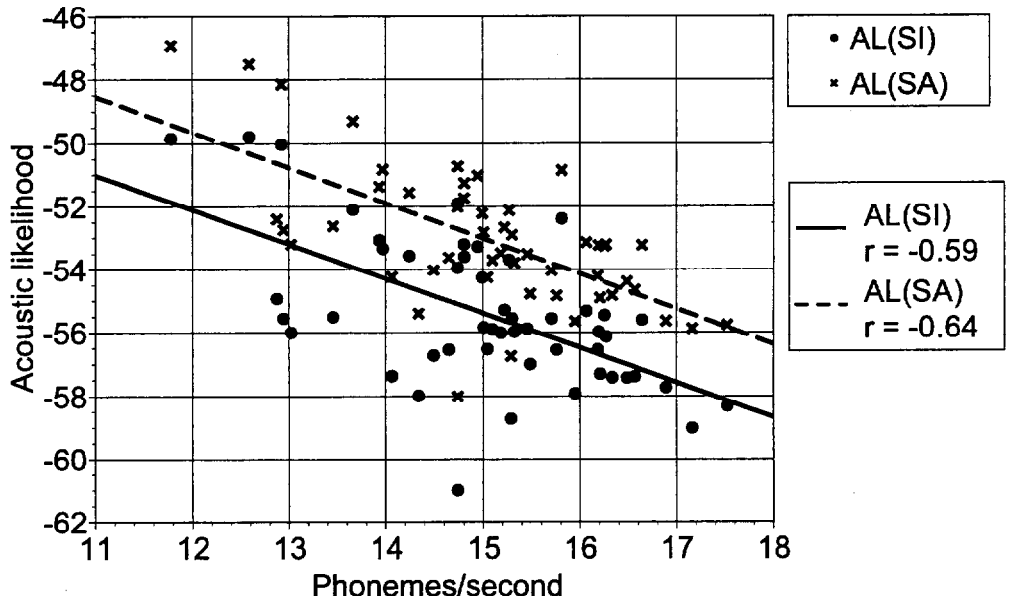

Figure 6.1: Speaking rate vs. acoustic likelihood.

bold face indicate significant values at 5% significance level (p-value < 0.05).

6.4.1 Correlation between acoustic likelihood and speaking rate

The correlation coefficient between acoustic likelihood and speaking rate is -0.59 for the SI acoustic model. Figure 6.1 shows the relationship between the speaking rate and the averaged frame likelihood. There is a tendency that the higher the speaking rate is, the lower the acoustic likelihood becomes. On the other hand, even very slow speaking rate does not cause a decrease of the acoustic likelihood. The Akaike Information Criterion (AIC) [9] also indicates that the first order regression model is better than the second order model for regressing the acoustic likelihood on the speaking rate. This indicate that there is a linear relationship between the speaking rate and the acoustic likelihood averaged over presentations. A stronger articulation effect in faster speakers is probably a cause of the decrease of likelihood.

The unsupervised adaptation increases the acoustic likelihood but leave the relationship between the speaking rate and the acoustic likelihood with only a slight increase

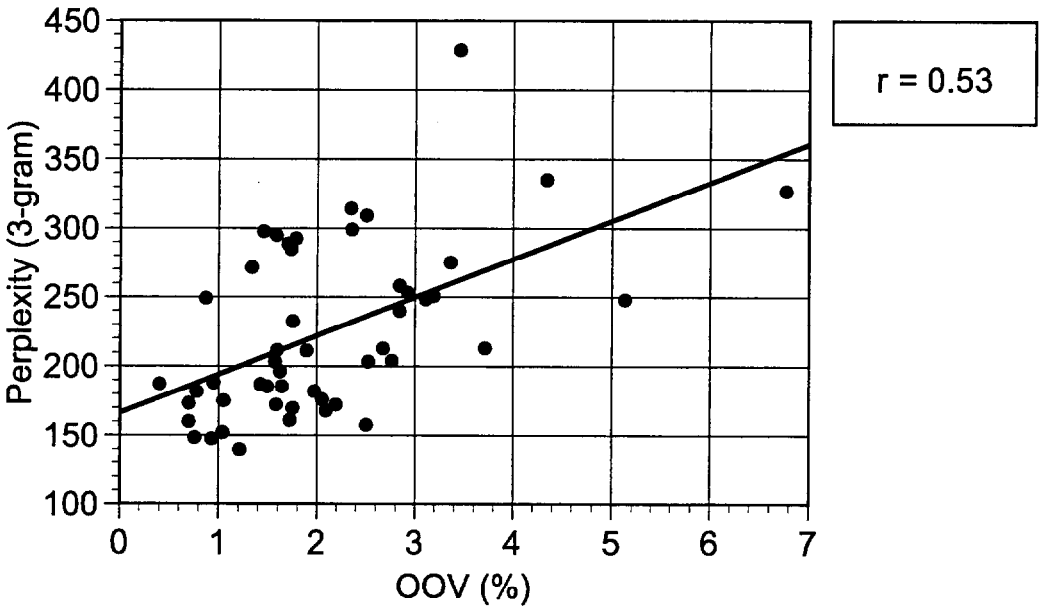

Figure 6.2: OOV vs. word perplexity.

in the correlation coefficient.

6.4.2 Correlation between word perplexity and several linguistic attributes

There exists significant correlation between word perplexity and out of vocabulary rate with a correlation coefficient of 0.53. Figure 6.2 shows the relationship between word perplexity and out of vocabulary rate. There is a tendency that presentations having a higher out of vocabulary rate show a higher perplexity.

The correlation coefficient of the filled pause frequency and the perplexity is -0.19 indicating that they are almost uncorrelated. The repair frequency and the perplexity have a correlation coefficient of 0.11. Since the perplexity was calculated after removing repairs, this result shows that the linguistic difficulty excluding repairs has almost no correlation with the repair rate.

6.4.3 Correlation between word accuracy and several attributes

The correlation coefficient between the word accuracy (SI) and the speaking rate is -0.47. Figure 6.3 shows the relationship between the word accuracy and the speaking rate. The relationship seems monotonic and even very slow speaking rate does not decrease the accuracy, which is similar to the result for the acoustic likelihood shown in Figure 6.1. The AIC also indicates that the first order model is superior to the second order model for regressing the word accuracy on the speaking rate.

Correlation between the word accuracy and the acoustic likelihood is not statistically significant, when the SI acoustic model is used. Their partial correlation coefficient adjusted for the speaking rate is -0.005. A partial correlation coefficient between the word accuracy and the speaking rate adjusted for the acoustic likelihood is -0.40, which is significant at a 1% significance level, and the partial correlation coefficient between the acoustic likelihood and the speaking rate adjusted for the word accuracy is -0.54, which is significant at a 1% significance level. This means that the correlation between the word accuracy and the acoustic likelihood is spurious. In other words, a fast speaking rate decreases the word accuracy and the acoustic likelihood independently. Similar results are obtained for SA conditions.

The correlation coefficient between the word accuracy and the repair frequency is -0.32. Figure 6.4 shows the scattergram of the word accuracy and the repair rate when the SI acoustic model is used.

There is a weak positive correlation of 0.38 between the word accuracy and the filled pause frequency, but this is also a spurious correlation, since the partial correlation coefficient adjusted for the speaking rate is 0.18.

Figure 6.5 shows the scattergram for word accuracy (SI) and out of vocabulary rate. The correlation coefficient between the word accuracy and the out of vocabulary rate is -0.54.

There is a weak negative correlation of -0.39 between the word accuracy (SI) and the perplexity, but this is also spurious; the partial correlation between the word accuracy

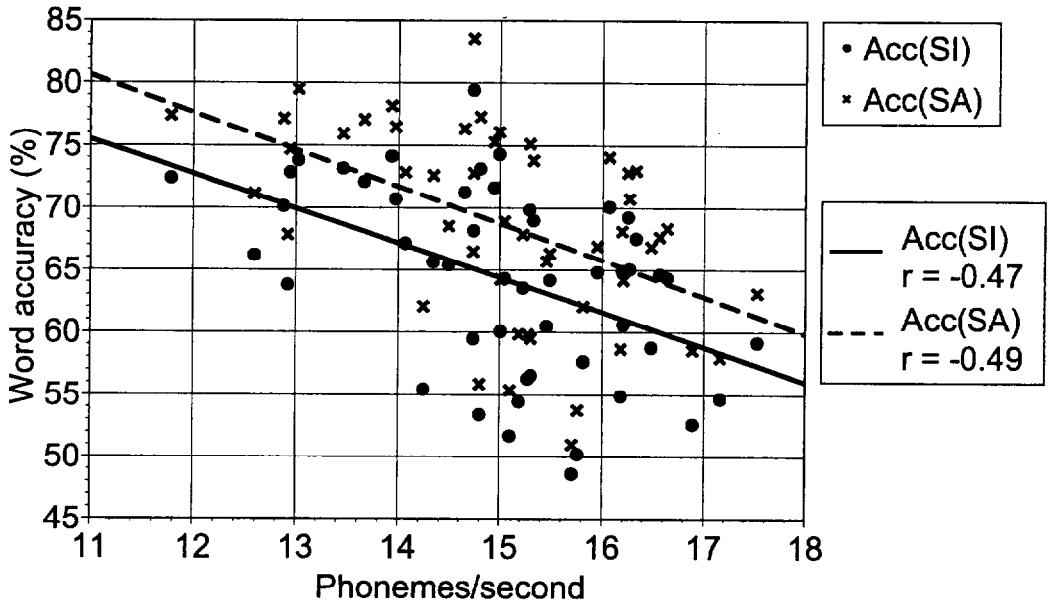

Figure 6.3: Speaking rate vs. word accuracy.

and the perplexity adjusted for the out of vocabulary rate is -0.14.

6.5 Regression analysis

The following equations (6.1) and (6.2) show linear regression models of the word accuracy with the six presentation attributes when the SI and SA acoustic model are respectively used for speech recognition.

$$\begin{aligned} Acc_{SI} &= -0.061AL_{SI} - 1.4SR_{SI} - 0.014PP \\ &\quad -2.3OR + 0.28FR - 3.3RR + 92 \end{aligned} \tag{6.1}$$

$$\begin{aligned} Acc_{SA} &= -0.061AL_{SA} - 1.6SR_{SI} - 0.010PP \\ &\quad -2.1OR + 0.30FR - 3.3RR + 98 \end{aligned} \tag{6.2}$$

In equation (6.1), the regression coefficient for the repair rate is -3.3 and the coefficient for the out of vocabulary rate is -2.3. This means that a 1% increase of the repair

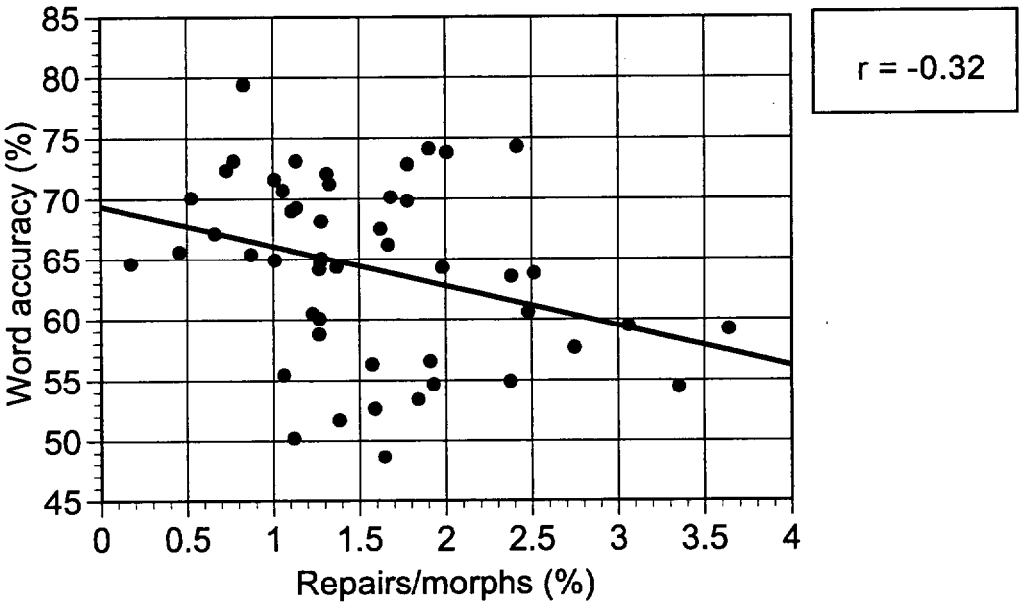

Figure 6.4: Repair frequency vs. word accuracy (SI).

rate or the out of vocabulary rate, respectively, corresponds to 3.3% or 2.3% 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. Regression coefficients before and after speaker adaptation are almost the same excepting the constant term. The coefficient of determination for the multiple linear regression (6.1) is 0.50 and that for (6.2) is 0.47, both are significant at a 1% level. This means that about a half of the variance of the word accuracy is explained by the model.

Table 6.4: Standardized regression analysis results, showing standardized regression coefficient (Coeff), p-value and 95% confidence interval (95% CI).

	Coeff(SI)	P	95% CI		Coeff(SA)	P	95% CI
AL(SI)	-0.02	0.885	(-0.29, 0.25)	AL(SA)	-0.02	0.904	(-0.31, 0.28)
SR(SI)	-0.23	0.149	(-0.55, 0.09)	SR(SI)	-0.26	0.135	(-0.60, 0.08)
PP	-0.12	0.374	(-0.38, 0.15)	PP	-0.08	0.549	(-0.36, 0.19)
OR	-0.36	0.015	(-0.65,-0.07)	OR	-0.33	0.028	(-0.63,-0.04)
FR	0.14	0.305	(-0.13, 0.41)	FR	0.15	0.301	(-0.14, 0.43)
RR	-0.32	0.008	(-0.55,-0.09)	RR	-0.32	0.010	(-0.55,-0.08)

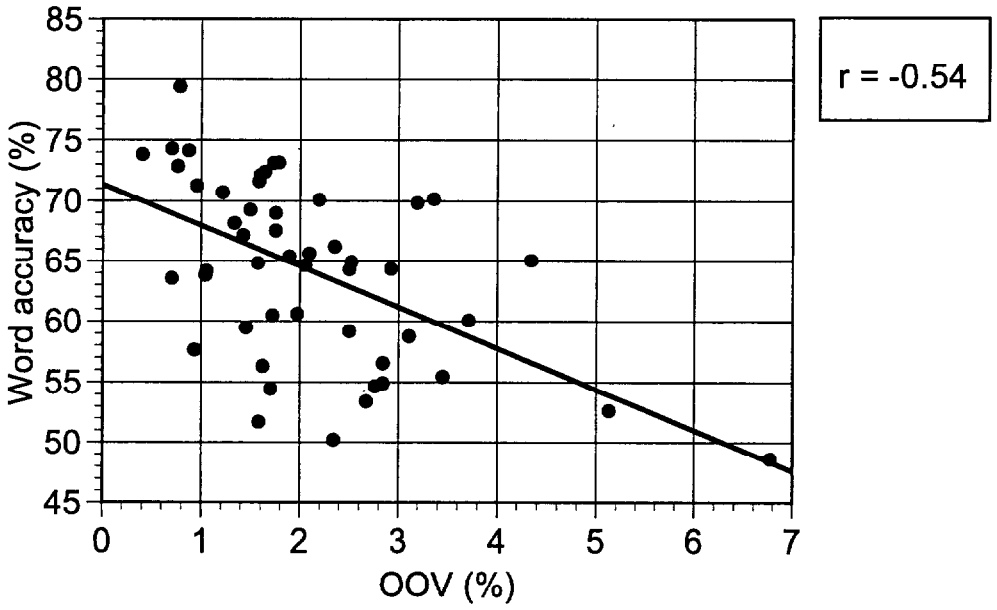

Figure 6.5: OOV rate vs. word accuracy (SI).

Table 6.4 shows standardized representation of the regression analysis with the equations (6.1) and (6.2), in which the variables are standardized before the analysis in order to show the effects of explaining variables on word accuracy. The table shows the standardized regression coefficient, the p-value and the 95% confidence interval. The standardized regression coefficients of the acoustic likelihood, the perplexity and the filled pause rate are relatively small for both the SI and SA regression models. Although most of these variables have statistically significant correlation with the word accuracy, these correlations are spurious as indicated in Section 6.4.

6.6 Discussion

As a supplementary experiment, a backward elimination procedure was employed to identify relatively important predictors of the word accuracy. The backward elimination process started with all of the six predictors in the model, and the model was refitted to the data after removing a variable with the largest p-value. The refitting process was iterated removing the least significant variable in the model until all remaining variables

had p-values smaller than 0.05. The important predictors identified were the speaking rate, the out of vocabulary rate and the repair rate, which correspond to the attributes showing relatively large coefficients in Table 6.4. Coefficients of determination of the regression models on these three attributes are 0.48 and 0.46 for speaker independent and adaptive cases, which are almost the same as that of the models for all attributes. It can be concluded that the main factors of individual differences in word accuracy are the speaking rate, the out of vocabulary rate and the repair rate.

6.7 Conclusion

In this chapter, the individual differences in spontaneous presentation speech recognition have been investigated. It was shown that the speaking rate, the out of vocabulary rate and the repair rate have relatively large effects on the individual differences of the word accuracy among a set of presentation/speaker attributes. It have found that the averaged acoustic likelihood of reference phoneme sequences and the test set perplexity are relatively minor factors in individual differences in word accuracy for the 50 male speakers in the test set.

Unsupervised MLLR speaker adaptation works well for improving the word accuracy but does not change the structure of the individual differences including the effects of the speaking rate. A special method for addressing speaking rate is crucial.

Approximately half of the variance of the word accuracy is explained by the regression model using the six explaining variables. The regression model with the three most important attributes also displays a similar prediction power.

To improve recognition performance, investigation into efficient methods for reducing the effects of the major attributes on the recognition accuracy is important.

Chapter 7

Analysis on Recognition Errors

7.1 Introduction

This chapter proposes an application of decision trees to analyze recognition errors. Words/phonemes contained in speech have many attributes, and the choice of a given word/phoneme by the speech recognition is either true (correct) or false (incorrect). To map the attributes to a true/false class, decision trees can be employed. It is expected the prediction capacity of a tree to be related to the explanation capability of the set of attributes used in this tree. In addition, it is investigated how these attributes cause recognition errors by analyzing the trees.

A "case" is defined as a set of attributes and a class. A decision tree is trained by using a set of cases. The performance of the tree is measured by applying the tree to a set of test cases and calculating what percentage of the classes are correctly predicted.

This chapter is organized as follows. Speech recognition task and an experimental set up of presentation speech are shown in section 7.2. In section 7.3, the principle of constructing decision trees is first reviewed, and then the construction and evaluation set up of the trees are shown. In sections 7.4 and 7.5, recognition performance and the experimental results of decision trees are shown. Finally in section 7.6, some conclusions are given.

7.2 Recognition task and experimental set up

7.2.1 Recognition task

Presentation speech uttered by 10 male speakers was used as a test set for speech recognition. Table 7.1 shows the contents of the test set.

Table 7.1: Recognition test set of presentations

ID	Conference name	Length [min]
A22	Acoust. Soc. Jap.	28
A23	Acoust. Soc. Jap.	30
A97	Acoust. Soc. Jap.	12
P25	Phonetics Soc. Jap.	27
J01	Soc. Jap. Linguistics	57
K05	National Lang. Res. Inst.	42
N07	Assoc. Natural Lang. Proc.	15
S05	Assoc. Socioling. Sciences	23
Y01	Spont. Speech Corpus Meeting	14
Y05	Spont. Speech Corpus Meeting	15

7.2.2 Experimental conditions

Speech signals were digitized with 16kHz sampling and 16bit quantization. Feature vectors had 25 elements consisting of 12 MFCC, their delta and the delta log energy. The CMS (cepstral mean subtraction) was applied to each utterance. HTK v2.2 was used for acoustic modeling. The language models were made by using the CMU SLM Tool Kit v2.05. Morphemes (which will be called "words" hereafter) were used as units for statistical language modeling. The Julius v3.1 decoder [6] was used for speech recognition. Filled pauses and repairs were taken into account as words in calculating the recognition accuracy.

7.2.3 Language and acoustic modeling

A part of the CSJ, having approximately 1.5M words, was used as a training set. Speakers had no overlap with those of the test set. The training set consisted of 610 presentations; 274 academic conference presentations and 336 simulated presentations. The

simulated presentations were specially recorded for the project and covered a wide variety of topics including the subjects talking about experiences in their daily lives.

The language model used in the recognition consisted of bigrams and reverse trigrams with backing-off. It was made using the whole training set. The vocabulary size was 30k. The acoustic model was made using 338 presentations uttered by male speakers (approximately 59 hours). It was a tied-state tri-phone HMM having 2k states and 16 Gaussian mixtures in each state.

7.3 Training and testing decision trees

7.3.1 Tree construction

The decision trees were made using a data-mining tool C4.5R8 [8]. In C4.5, trees are derived by a two-path strategy. First, questions about attributes are chosen step by step under a predefined criterion. Training cases are split by the question accordingly. This partitioning continues to subdivide the set of training cases until each subset in the partition contains cases of a single class, or until no question yields any improvement. Next, to correct over-training and make the tree robust against unseen data, the tree is pruned.

In this experiment, gain-ratio was employed for the question choosing criteria. Questions that maximize the gain-ratio were selected. Equation (7.1) shows the definition of the gain-ratio.

$$gainratio = \frac{H(Y) - H(Y|X)}{H(X)}, \tag{7.1}$$

where X is a random variable defined for each question, whose value is its answer. $H(X)$ denotes the entropy for the distribution of X. $H(Y)$ denotes the entropy for the distribution of a class. $H(Y|X)$ is the conditional entropy of the distribution of a class given an answer to the question. Entropy is calculated based on the distribution of the training cases for each tree node.

7.3.2 Decision trees for words

Decision trees for words were constructed by defining a case as a set of attributes of a reference word and the correctness of its recognition hypothesis. The correctness was determined by matching the reference word sequence and recognition hypothesis. Only substitution and deletion errors were analyzed; insertion errors were not considered in this study since inserted words do not have corresponding reference words. Compound words were not considered in the matching process, and errors included the cases where only word segmentation boundaries were different. Decision trees were pruned by error-based pruning. The threshold was set to 10 based on a preliminary study.

Table 7.2 shows the attributes in consideration. They are either discrete or continuous. In the table, "D" or "C" indicates that the attribute is treated in C4.5 as discrete or continuous, respectively. The JTAG3.03 morphological analysis program was used to obtain part of speech information. For the judgment of filled pauses and repairs, annotated information in the CSJ transcription was used. The speaking rate and frame likelihood attributes were calculated by using the result of phoneme alignment to the reference sentence. The first 2320 cases were used for each presentation in order to unify the condition in terms of the amount of data. Trees were created and tested using a cross validation method; the data set made of all selected cases was divided into 10 subsets and one of them was used for testing.

7.3.3 Decision trees for phonemes

Decision trees for phonemes are built in the same way using phonemes as units instead of words. Like for the word analysis, Only substitution and deletion errors were considered, and insertion errors were neglected. The pruning threshold was set to 10 based on a preliminary experiments.

Table 7.3 shows the phoneme attributes used in the experiments. Frame-by-frame information such as likelihood and power is averaged over the period of each reference phoneme obtained by the phoneme alignment. The likelihood value for each HMM state

Table 7.2: Word attributes

Number of phonemes in the word	C
Word duration (number of frames)	C
Speaking rate (number of phonemes/number of frames)	C
Averaged acoustic frame likelihood	C
Ratio of a certain phoneme class such as vowel or nasal	C
Part of speech (noun, verb, etc.)	D
Filled pause or not	D
Repair or not	D
Quotation or not	D
Loanword or not	D
Word frequency in the training set	C
Bigram score	C
Trigram score	C
Back off class	D
Word order in the sentence from either beginning or end	C
Part of speech of the left/right context word	D
Left/Right context word is filled pause or not	D
Left/Right context word is repair or not	D
Left/Right context word is quotation or not	D
Left/Right context word is loanword or not	D

does not include transitional probability. Whether or not a phoneme is uttered in a filled pause or repair is determined according to the annotation of the CSJ. Trees are created and tested using a cross validation method, dividing the data into 5 subsets. The first 8600 cases per presentation were used to equalize the amount of data.

7.4 Recognition results of the task

Figure 7.1 presents test-set perplexity and out-of-vocabulary (OOV) rate of the task using the trigram language model. Figure 7.2 shows word and phoneme recognition accuracies. In the phoneme recognition, no linguistic constraint was used. The results show that the accuracies vary greatly from speaker to speaker.

Table 7.3: Phoneme attributes

Kind of phoneme (a, u:, sh, etc.)	D
Left/Right phoneme kind context	D
Phoneme class (voiced, nasal, etc)	D
Left/Right phoneme class context	D
Filled pause or not	D
Repair or not	D
Left/Right context is filled pause or not	D
Left/Right context is repair or not	D
Max frame likelihood over all states	C
Minimum frame likelihood over all states	C
Average frame likelihood over all states	C
Number of states whose frame likelihood is greater than frame max minus delta	C
Frame likelihood variance over all states	C
Phoneme duration	C
Frame energy	C
Delta frame energy	C
Mono-phone frequency in the corpus	C
Tri-phone frequency in the corpus	C

7.5 Error analysis using decision trees

7.5.1 Decision trees for words

A set of decision trees for words was made using all the attributes listed in Table 7.2. Figure 7.3 shows prediction correctness of the trees. For comparison, word (recognition) correctness (WCorr) is also shown in the figure. TSpk denotes prediction correctness when trees are built for each speaker. TAll is also prediction correctness but when trees are built using the training data by all the 10 speakers.

The word correctness corresponds to the prediction correctness of a tree having only the root node. As can be seen, prediction correctness is higher than word correctness. This difference is believed to result from recognition errors caused by the attributes found in the tree.

Questions assigned near the root of the trees are the repair, the word occurrence frequency, the ratio of voiced phonemes, the ratio of long (double) consonants, etc.

TAll indicates better prediction correctness than TSpk. This means that the amount

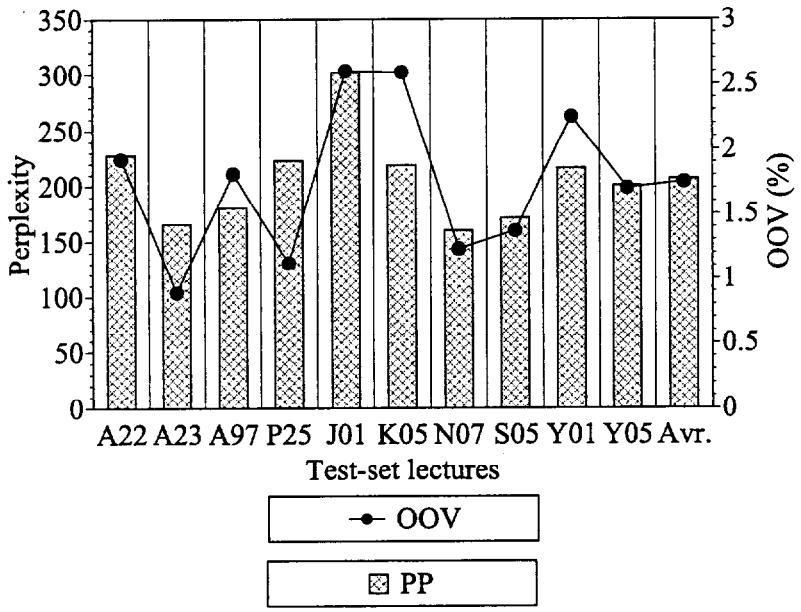

Figure 7.1: Test-set perplexity and OOV rate of the task.

of data is more significant than speaker-specific variations in this analysis.

7.5.2 Error factor analysis of word recognition

To analyze what attributes have strong correlation with recognition errors, various subsets of attributes were selected and the performance of trees were measured. As a result, it turned out that just considering three attributes produced almost the same performance as considering all the attributes in Table 7.2. The three attributes are the number of phonemes in a word, the speaking rate, and the frequency of word occurrence. Word recognition error tends to be higher if the word has a relatively small number of phonemes, is spoken fast, and is observed less frequently in the language-model training corpus. But strictly speaking, the relationships are not monotonic. For example, a very slow speaking rate also tends to increase errors. The other attributes are either less informative about word error or the information they provide is already included in the three major attributes.

Figure 7.4 shows the prediction correctness of the trees for subsets of attributes. The

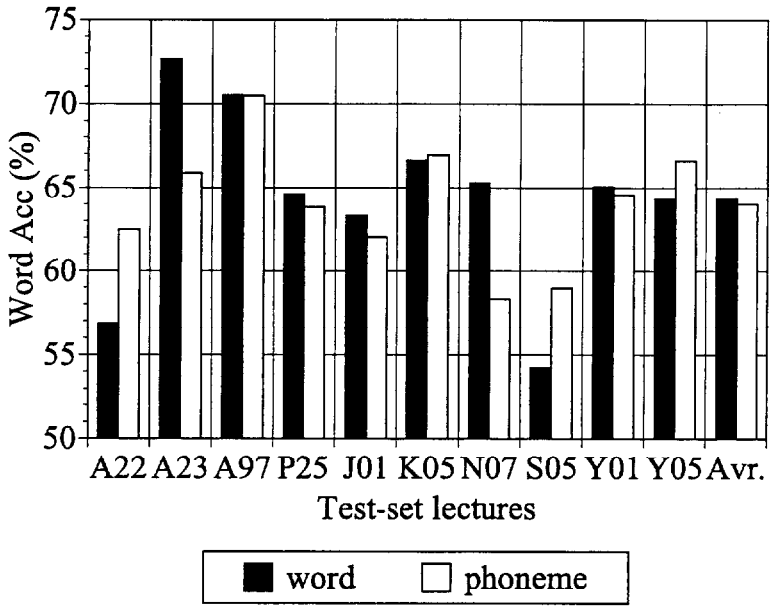

Figure 7.2: Word/phoneme recognition accuracy.

trees were built using training data by 10 speakers. AllAtt indicates the correctness of trees using all attributes. P, R and W indicate the number of phonemes in the word, the speaking rate and the word frequency, respectively. It can be seen that omitting any one of them degrades the prediction correctness.

In order to analyze the sources of word recognition accuracy variation among speakers, the success rate of recognition was estimated using the decision tree with the three most significant attributes. predicted success rate (PSR) was defined for each utterance as follows.

$$PSR = \frac{T}{T+F}, \tag{7.2}$$

where T indicates the number of test cases in the utterance that are predicted to be true (correctly recognized) by the tree, and F indicates those predicted to be false (incorrectly recognized). Figure 7.5 shows the relationship between PSR and the actual recognition correctness for the 10 speakers in the test set. The correlation coefficient for this result is 0.87, meaning that differences in the three attributes are highly related to variation in recognition accuracy.

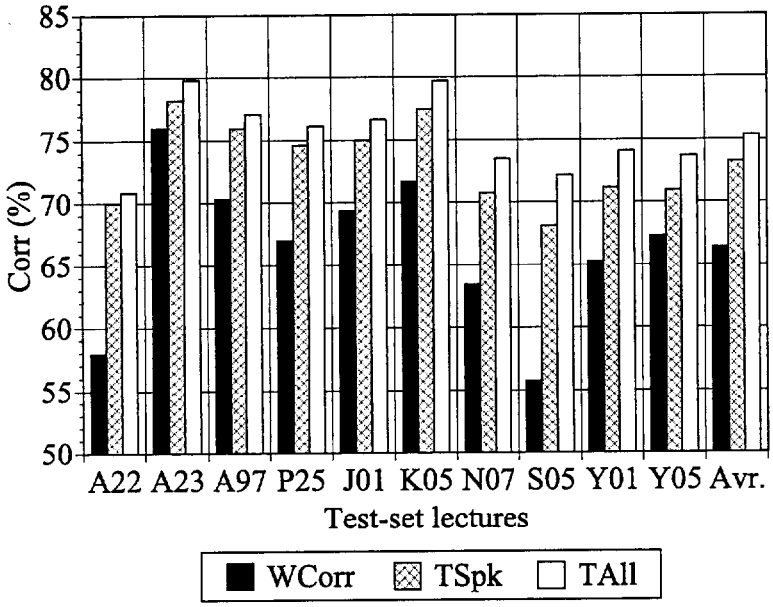

Figure 7.3: Recognition and prediction correctness.

7.5.3 Decision trees for phonemes

Figure 7.6 shows the prediction correctness of the trees for phonemes that are made using all the attributes in Table 7.3. TSpk denotes the tree made for each speaker. TAll denotes the tree made by using all the training data from the 10 presentations. For comparison, results of phoneme correctness (PCorr) are also shown.

The prediction correctness of TAll is higher than that of TSpk. This suggests that the factors contributing to recognition errors are similar among speakers.

7.5.4 Error factor analysis of phoneme recognition

Various subsets of attributes were selected and the performances of the trees were compared. It was found that a subset of attributes that indicates almost the same prediction correctness as all attributes in the Table 3 consisted of the frame-max and frame variance (F), the phoneme class and phoneme class context (P), the phoneme duration (D), and the mono-phone frequency in the training data (M). Figure 7.7 shows the prediction correctness for several attribute sets. Among these attributes, the phoneme duration

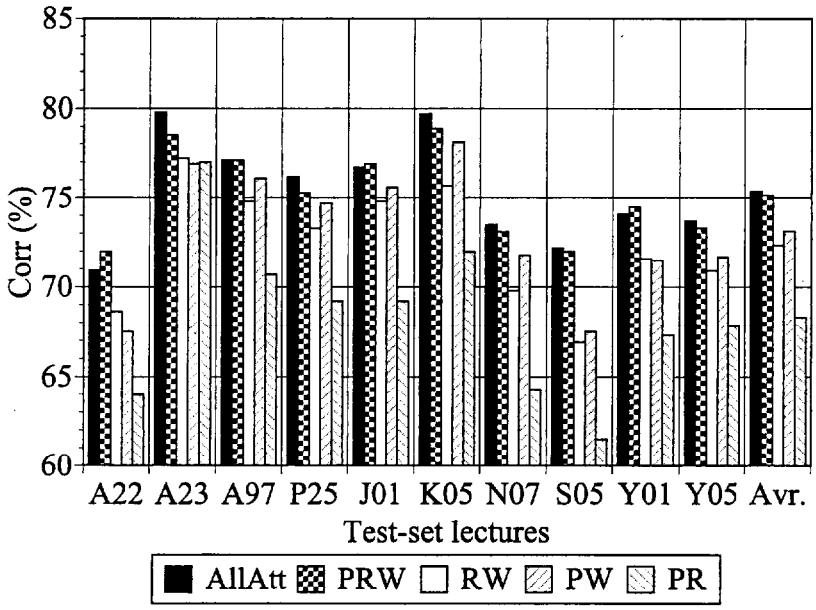

Figure 7.4: Analysis of word attributes.

seems to contribute the most to the correct recognition of phonemes.

7.6 Conclusion

In this chapter, the use of a decision tree for analyzing recognition errors was proposed. To what extent the recognition error can be explained by a set of attributes was quantitatively analyzed. In word recognition, it was found that the number of phonemes in the word, the speaking rate and the word frequency in the training data are highly related to the recognition rate. In phoneme recognition, a set of attributes consisting of the frame-max, the frame variance, the phoneme class, the phoneme class context, the phoneme duration and the mono-phone occurrence count has been found to have the same prediction power as all the attributes used in the experiment. To increase the recognition accuracy, the following issues are important: designing words considering the number of included phonemes; modeling the effects of speaking rate; and, properly increasing the training data. It might also be useful to use the decision-tree-based framework for estimating the confidence measure for recognition.

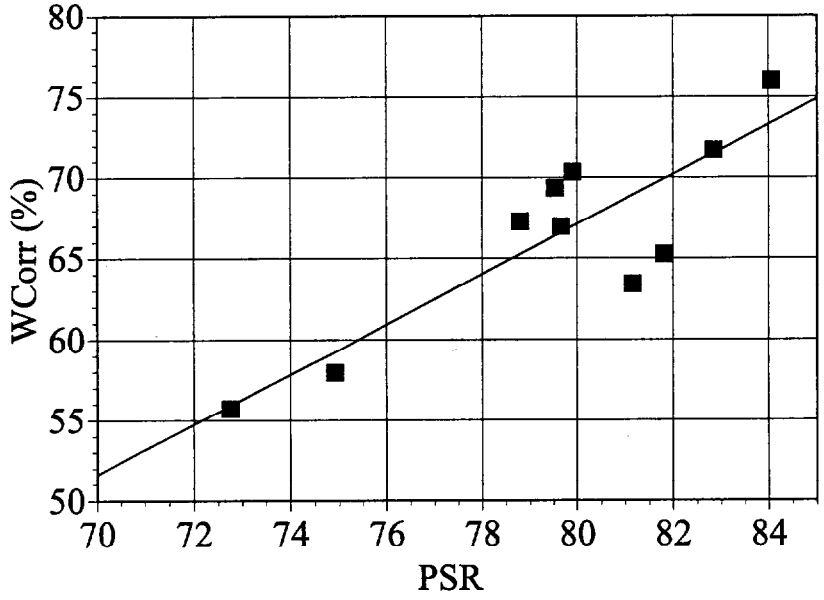

Figure 7.5: The predicted success rate (PSR) and the recognition correctness.

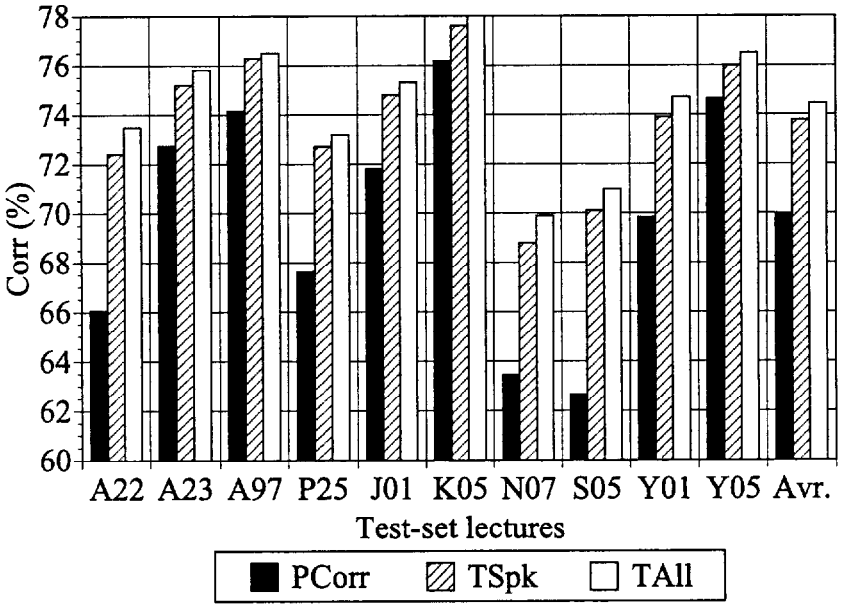

Figure 7.6: Prediction correctness.

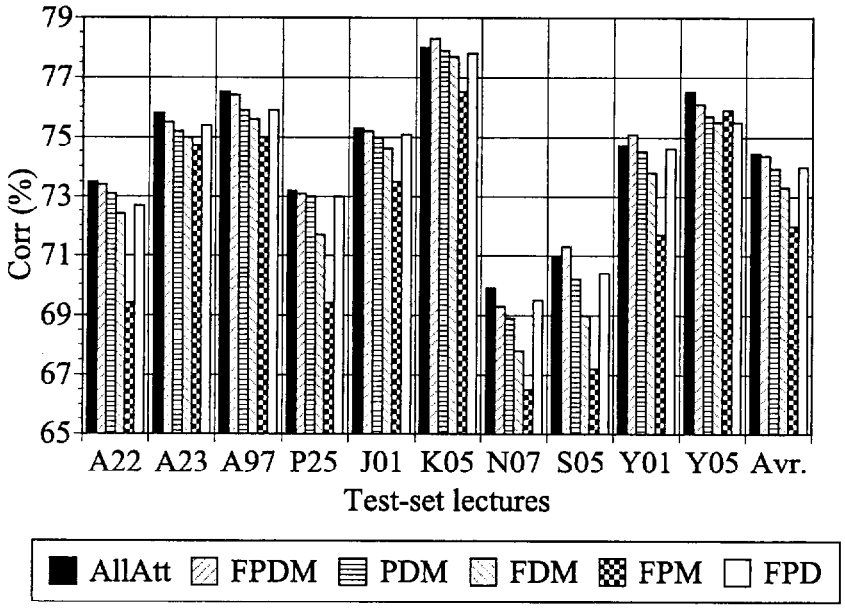

Figure 7.7: Analysis of phoneme attributes.

Chapter 8

Comparison of Human and Automatic Recognition Performance

8.1 Introduction

There is a large variation of difficulties for recognizing words in spontaneous speech; while some words are easy, others require specific knowledge or longer context. Some words are impossible to recognize even for humans. Recognition methods need to be improved in different ways for each class of words according to the variation of the difficulties. However, it still remains unclear as to what are the most important problems and to what extent they are significant. In order to investigate the possibility of improvement, recognition performance of an automatic speech recognizer is evaluated in comparison with human recognition performance.

Several comparisons have been conducted evaluating the difference in recognition performances between computers (decoders) and humans. For speech reading text, an order of magnitude higher word error rate was reported when comparing decoders with human listeners using sentences extracted from CSR'94 spoke 10 and CSR'95 Hub3 database under various SNR and microphone conditions [10]. Another experiment using sentences extracted from the Wall Street Journal database indicated roughly a five times higher error rate for a decoder [11]. For spontaneous speech, an order of magnitude

higher word error rate for a decoder was reported for the Switchboard task [12].

This chapter explores the possibility of improvement in spontaneous speech recognition by searching for conditions where the prescription would be relatively easy. For this purpose, decoder and human performances are compared in the same condition. Recognition results from the decoder and listeners are compared word by word and reasons for the errors by the decoder are analyzed.

8.2 Experimental set up

8.2.1 Recognition task

Recognition results were evaluated using the same recognition task performed by an automatic speech recognizer and human listeners. The task was to recognize a word in an excerpted period of utterance including a ± one word context. Both the decoder and human listeners chose the most likely words from the same vocabulary set, which consisted of the most frequent 25k words occurring in 455 academic presentations in the Corpus of Spontaneous Japanese (CSJ) [4]. The presentations in the corpus were given spontaneously and recorded using close-talking microphones.

Five hundred test utterances were randomly chosen from seven academic presentations in the CSJ given by different male speakers. Table 8.1 shows the contents of the test set presentations. Each test utterance was a three-word sequence excerpted from the presentations by forced alignment of an HMM sequence corresponding to the true word sequence. Since there is no spacing between words in Japanese sentences and even no clear definition of words, the JTAG morphological analysis program was used to define words. JTAG was also used to annotate pronunciations of the words. The resulting pronunciations were manually checked so that errors would not affect the segmentation accuracy. Utterances with severe errors were eliminated after the random selection of the test set. Approximately one percent of the center words of the 500 test utterances, target words to recognize, were not included in the vocabulary. In the evaluation process, recognition results were manually checked and simple transcription variations were

Table 8.1: Test set presentations

Presentation ID	Conference
A01M0035	Acoust. Soc. Jap
A01M0007	Acoust. Soc. Jap
A01M0074	Acoust. Soc. Jap
A02M0117	Soc. Jap. Linguistics
A03M0100	Assoc. Natural Lang. Proc.
A05M0031	Phonetics Soc. Jap.
A06M0134	Assoc. Socioling. Sciences

normalized. The major reason why the ± one word context was given, instead of the previous two words, was to avoid explicitly determining word boundaries of the center word in the wave form which is sometimes difficult to do due to coarticulation effects.

8.2.2 Recognition by decoder

For recognition by the decoder, a word network, as shown in Figure 8.1, was prepared for each test utterance. Finding the most likely path in the network corresponds to choosing the center word given the ± one word context. Note that in the decoding process, the optimum word boundary may be different from path to path. A language probability was assigned to each center word in the network as shown in equation (8.1).

$$P(w_c|w_f, w_b) \tag{8.1}$$

$$= \frac{P(w_f) \cdot P(w_c|w_f) \cdot P(w_b|w_f, w_c)}{\sum_w P(w_f) \cdot P(w|w_f) \cdot P(w_b|w_f, w)}. \tag{8.2}$$

Here w_c is a center word, and w_f and w_b are the front and back context words, respectively. The conditional probability of equation (8.1) was calculated using a trigram language model as shown in the equation (8.2).

The language model was trained using a corpus with 2.9M words consisting of 1289 academic and non-academic presentations given by both male and female speakers. Acoustic feature vectors had 25 elements consisting of 12 MFCC, their delta and the delta log energy. The CMS (cepstral mean subtraction) was applied to the sentence utterance including each three-word length test utterance. A tied state triphone model

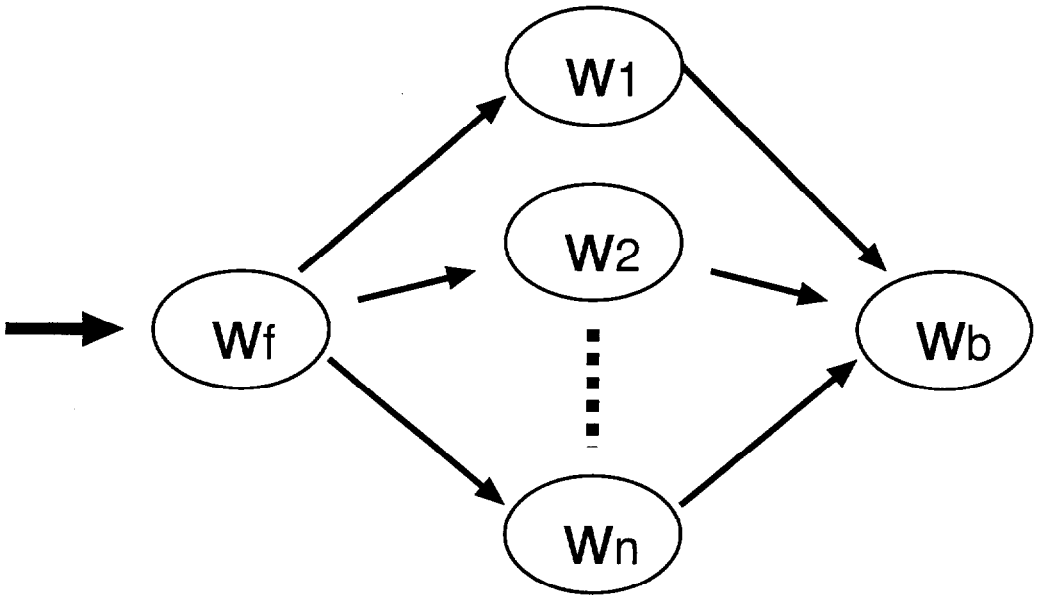

Figure 8.1: Word network for the decoder.

consisting of 2k states and 16 Gaussian mixtures in each state was used as a speaker independent (SI) acoustic model. The model was trained using 455 academic presentations in the CSJ given by male speakers, which had a total length of 94 hours. In addition to the SI model, speaker adaptive (SA) acoustic models were also constructed using an unsupervised adaptation method. The SI model was adapted for each speaker with the MLLR technique using the entire presentation. There was no overlap between the speakers in the training set and those in the test set.

The HTK was used for decoding. A language weight of 10 was determined by preliminary experiments. A relatively light pruning level that was also determined by preliminary experiments was used so as not to affect the recognition rate.

8.2.3 Recognition by humans

Human listeners were given the capability of playing back the test utterances and choosing words from the vocabulary using a GUI-based system. They could listen to the same utterance as many times as they liked to make a decision. However, once they made

a decision, they could not repeat the same task. Since the utterances were randomly selected from presentations, it was impossible for the human listeners to use a longer context beyond ± one word. They were informed that target words in the test utterances might not be in the vocabulary, and instructed to select the closest word even when they did not find the exact word. To facilitate finding the words from the large vocabulary, the GUI was equipped with a dictionary search using regular expressions.

Fifteen listeners, consisting of 14 male and one female, were divided into five groups, each having three listeners. They were students and staff of our laboratory. The 500 test utterances were partitioned into five blocks and each block was assigned to one of the groups. The same utterance was recognized by three different listeners in each group to mitigate the effects of careless mistakes and individual variations due to differences in familiarity with presented topics. An upper limit of human recognition ability was estimated by determining the selected word based on a majority rule among the three listeners. The estimated upper limit was used for comparison with results by the decoder. If there was no overlap between the words given by the three listeners, an answer by the listener having overall the best performance among the three listeners was adopted.

The listeners practiced the task using 10 examples before performing for the test utterances. Experiments were conducted in an office using a headphone. It took about one to two hours for the listeners to process the 100 test utterances.

8.3 Experimental results

8.3.1 Human recognition results

Figure 8.2 shows the recognition performance of individual listeners and that by the majority rule. Unknown words were not counted in the recognition rate. There were no insertion or deletion errors because of the experimental settings. The variation in score from listener to listener was mostly due to a difference in familiarity with the academic presentations. Averaged recognition rate of the majority-rule based result is 95.3%.

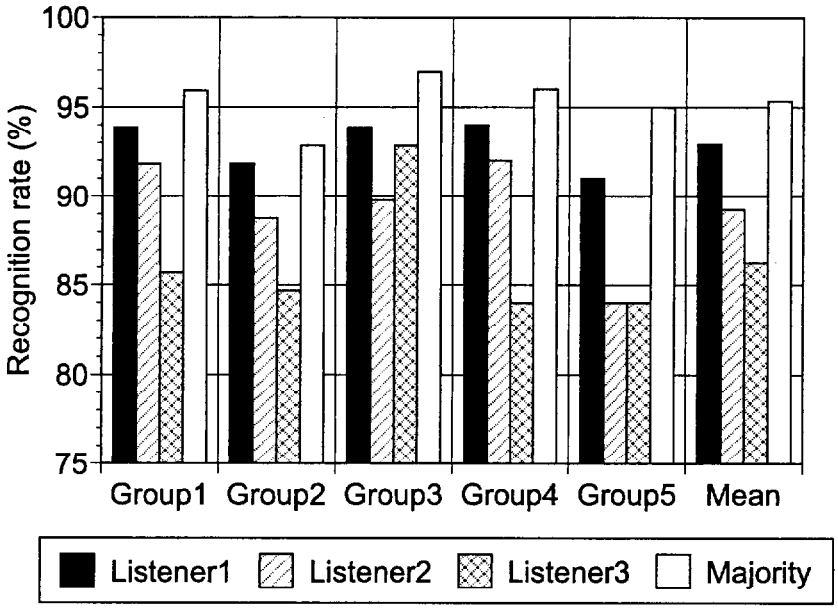

Figure 8.2: Human recognition rates.

8.3.2 Comparison between decoder and human

Comparison of the recognition rates by human listeners and the decoder is shown in Figure 8.3. The majority-rule based result is used as the human recognition rate. The averaged recognition rate of the decoder is 88.7% when the SI model is used, and 91.3% when the SA models are used. The human recognition rate is superior to that of the decoder under the same conditions defined for the context. The recognition error rate for human listeners is roughly half of that for the decoder. The differences of the recognition/error rate between humans and the decoder are significant at a 1% level for the SI results, and at a 5% level for the SA results.

8.3.3 Analysis of decoding errors

Table 8.2 shows the classification of the experimental results using the SI acoustic model. The results based on the majority rule were used in the case of human experiments. There exist twelve words that were successfully recognized by the decoder but not by human listeners. The reasons for the errors made by the humans include vague pronun-

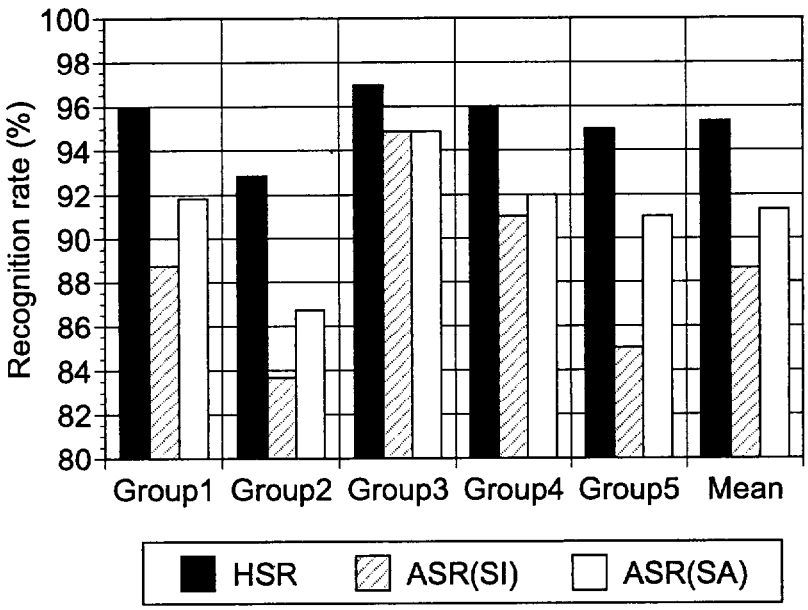

Figure 8.3: Human and decoder recognition rates.

ciations which made the recognition difficult and caused the same inattention errors by two or more listeners at the same time.

Table 8.2: Classification of recognition results

		ASR	
		Correct	False
HSR	Correct	426	45
	False	12	11

UNK: 6

There were 45 words that could be correctly recognized by humans but not by the decoder. Among these 45 words, 33 words were correctly recognized by all three listeners. If the decoder is improved so that these words can be correctly recognized, a 6% improvement in the accuracy can be expected. In order to investigate why the decoder failed to recognize these words, acoustic and linguistic likelihood values of the true words and the outputs of the decoder were compared. The result is shown in Figure 8.4. The acoustic likelihood was calculated including the ± one word context, and the

language weight was incorporated into the language likelihood.

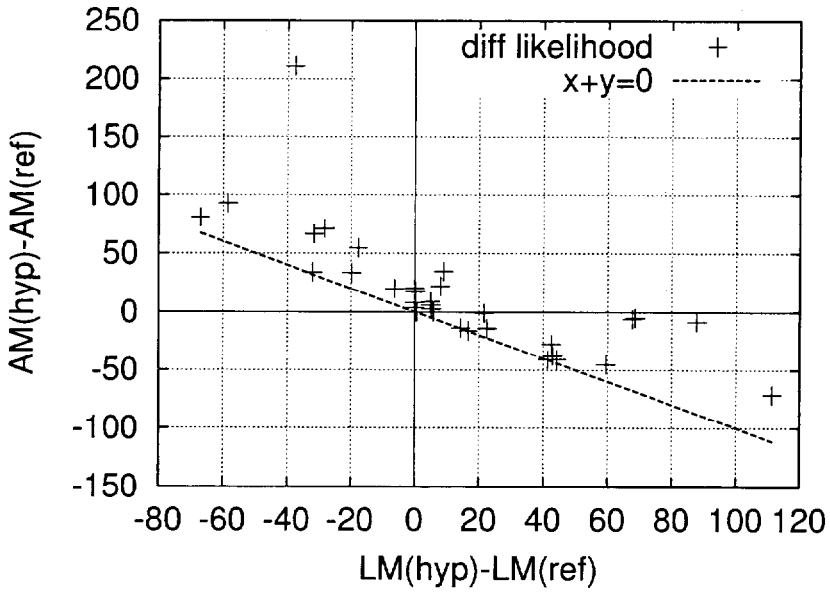

Figure 8.4: Comparison of likelihood values.

There are 13 samples in which the acoustic likelihood of the incorrect hypothesis word is lower but the language likelihood is higher than the true word. The inversion of the language likelihood is due to either an excessive likelihood assignment to the incorrect hypotheses or an unusual occurrence of the correct three word sequences; both are almost equally observed. The excessive likelihood seems to be caused by a backing off applied because of the sparsity of the training data. To recognize unusual true word sequences, improvement of the acoustic model is also required.

On the other hand, there are nine samples in which the language likelihood of the misrecognized word is lower but the acoustic likelihood is higher than the true word. Among these samples, one of them was totally unvoiced and another was contaminated with a low noise. The other seven samples have no problem as long as the three-word sequences are listened to. But when the center words are listened to in isolation, roughly half of the seven samples sound somewhat different from the correct word.

8.3.4 Relationship with continuous speech recognition

Word recognition rates of individual word recognition, given the ± one word context, conducted in the above experiments were compared to recognition rates using whole sentence continuous speech using various acoustic and language models which have different modeling accuracy. The results are shown in Figure 8.5. In this experiment, 3000 words and 280 sentences in the test set presentations listed in Table 8.1 are used. The recognition rate of this result is slightly lower than that of the task in subsection 8.3.2 even when the same model is used, since the results are not manually normalized.

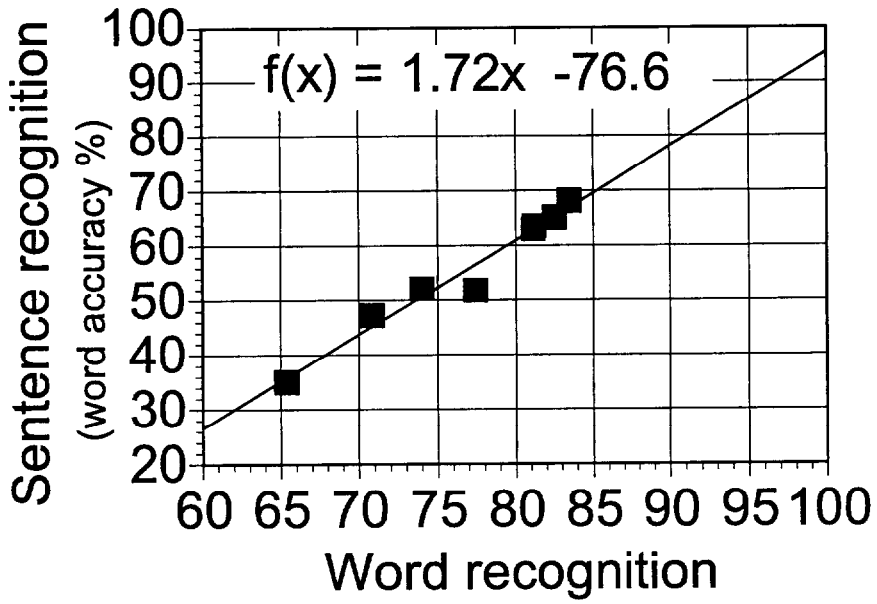

Figure 8.5: Word vs. sentence recognition rates.

The relationship between the recognition rates can be approximated by a straight line with a gradient of 1.72, that passes through the point of (100%, 96%). If the individual word recognition performance can improve by 6% as stated in the previous subsection, a 10% improvement in the continuous speech recognition is expected. However, to achieve word accuracy of 90% or more, the 10% improvement is not enough and wider context

information should be incorporated.

8.4 Conclusion

Recognition performance of an automatic speech recognizer has been evaluated in comparison with human recognition performance in spontaneous presentation recognition. The recognition error rate of human listeners was roughly half of that of the decoder. There existed roughly 6% of words that were easy for the humans to recognize but difficult for the decoder. Causes of the recognition errors by the decoder include problems of model accuracy and lack of robustness against vague and variable pronunciations. If the decoder could be improved to overcome these problems and the 6% of words could be correctly recognized, approximately a 10% improvement could be expected in continuous speech recognition without using contexts longer than trigrams. To achieve word accuracy of 90% or more, however, wider context information should be incorporated.

Chapter 9

Lexicon Optimization

9.1 Introduction

In this chapter, a new statistical lexicon optimization method for speech recognition based on both linguistic and acoustic features of words is proposed. This study was motivated by the observation described in Chapter 7 that less frequent and shorter words are generally difficult to recognize [21].

Language models based on words or morphemes are widely used for large-vocabulary continuous-speech recognition (LVCSR). For languages like English, words are well defined since they are separated by space symbols in the written text. Other languages like Japanese have no spacing between words and even no clear definition of words. Therefore, for these languages, it is common to preprocess text with a morphological analysis program to automatically split sentences into morphemes to make language models. In all languages including English and Japanese, it is not clear whether these conventional words or morphemes are optimal units for speech recognition.

From this point of view, several studies have been conducted to optimize the lexicon for improving the performance of language models[13, 14, 15, 16, 17, 18, 19, 20]. Some ideas presented in these works include a) automatically building a lexicon based on some criteria/rules for languages having no clear word definition, b) concatenating word pairs to model longer context by N-grams without increasing N so as not to increase the parameter dimension and data sparsity, and c) concatenating words to balance

the occurrence of all the word units. In these methods, basic units are concatenated based on evaluation functions such as unit pair frequency and mutual information. In [13, 14, 15, 16, 17, 19], performance was evaluated in terms of the recognition accuracy as well as the test-set perplexity. In [14], it was reported that certain word phrases were very frequent in dialogs of a limited domain. A phrase finding algorithm based on the mutual information criterion was found to improve the accuracy of a recognition system using a bigram language model. Paper [17] reported a mutual information based method using a large newspaper corpus with no improvement being achieved in the recognition accuracy. Paper [19] reported that a word pair frequency based method improved the recognition accuracy by 0.2% using a language model trained on WSJ.

One of the problems of these methods is that they are based only on a linguistic aspect and no acoustic characteristics have been considered. Although perplexity is a useful measure, decrease in perplexity does not necessarily guarantee improvement in the recognition rate. In the method proposed in this chapter, a word correctness probability model is estimated and used to directly estimate the recognition correctness of a system. A process of choosing and concatenating a word pair which maximizes the estimated word correctness is iterated.

This chapter is organized as follows. In Section 9.2, an experimental set up is described. In Section 9.3, the relationship between word frequency, word length and recognition correctness is analyzed and modeled. The proposed optimization method is described in Section 9.4. The proposed method is applied to a large scale Japanese spontaneous speech corpus in Section 9.5. It is shown that the method improves the recognition rate. Experimental results are discussed in Section 9.6. Finally in Section 9.7, the chapter is summarized and concluded.

9.2 Experimental set up

9.2.1 Baseline recognition system

Language models were trained using transcriptions of 610 academic and non-academic lectures from the large-scale Japanese spontaneous speech corpus (CSJ)[4]. JTAG morphological analysis program was used to convert Japanese text into morpheme sequences. The training set turned out to have approximately 1.5M morphemes (morpheme will be called "word" hereafter in this chapter). The most frequent 30k words were selected as the vocabulary for recognition and a trigram language model was made.

An acoustic model was made using 338 CSJ lectures presented by male speakers. The total length was approximately 59 hours. The Julius 3.1 decoder was used for speech recognition.

9.2.2 Recognition task

Two kinds of utterance sets were used, both of which consisted of academic lectures presented by male speakers in the CSJ.

Table 9.1: Development set

Conference name	No.lecture
Jap. Soc. Artif. Intell.	32
Acoust. Soc. Jap.	9
Soc. Jap. Linguistics	3

Table 9.2: Evaluation set

Lecture ID	Conference name	Length [min]
A22	Acoust. Soc. Jap.	28
A23	Acoust. Soc. Jap.	30
A97	Acoust. Soc. Jap.	12
J01	Soc. Jap. Linguistics	57
K05	National Lang. Res. Inst.	42
N07	Assoc. Natural Lang. Proc.	15
P25	Phonetics Soc. Jap.	27
S05	Assoc. Socioling. Sciences	23
Y01	Spont. Speech Corpus Meeting	14
Y05	Spont. Speech Corpus Meeting	15

A development set, consisting of 44 lectures, was used for analyzing and modeling the word correctness probability. Table 9.1 shows the content, in which the first 10 minutes of each lecture was used.

An evaluation set was used for evaluating the proposed method. This consisted of 10 lectures having no overlap with those in the development set. Table 9.2 shows its content, in which the entire length of each lecture was used. The OOV rate was 1.3% excluding word fragments.

All the speakers in these sets had no overlap with those in the training set for building acoustic and basic language models.

9.3 Relationship between word occurrence count, word length and recognition correctness

There exist many factors that affect the difficulty of recognizing each word. Among them, it has been shown in Chapter 7 that the number of occurrences of a word in the language model training set, as well as its length have a strong relationship with its correctness [21]. Generally speaking, less frequent and shorter words are harder to recognize. This is probably because the N-gram probability of a less frequent word is more difficult to model correctly and because shorter words are acoustically more easily confused in the decoding process.

To investigate these relationships, word attributes were defined and calculated as follows.

Cor: Word correctness (%). (0 or 100).

WF: Number of occurrences of a word in the language model training set.

LF: Logarithmic value of the *WF*; $\log_{10}(WF + 10^{-6})$.

NP: Number of phonemes in the word.

Cor is a binary attribute, taking 100 if the word was recognized correctly and 0 if it was

miss-recognized by the decoder. Insertion errors were not considered since they do not have corresponding reference words.

These attributes were computed for each word in the development set. The dotted lines in Figure 9.1 show the relationship between the number of occurrence of words and the recognition correctness where words were classified according to the number of phonemes they contain. The correctness was averaged over the group of words having the same range of *NP* and shown in the figure at each value of *LF*. Similarly, Figure 9.2 shows the relationship between the number of phonemes and the recognition correctness.

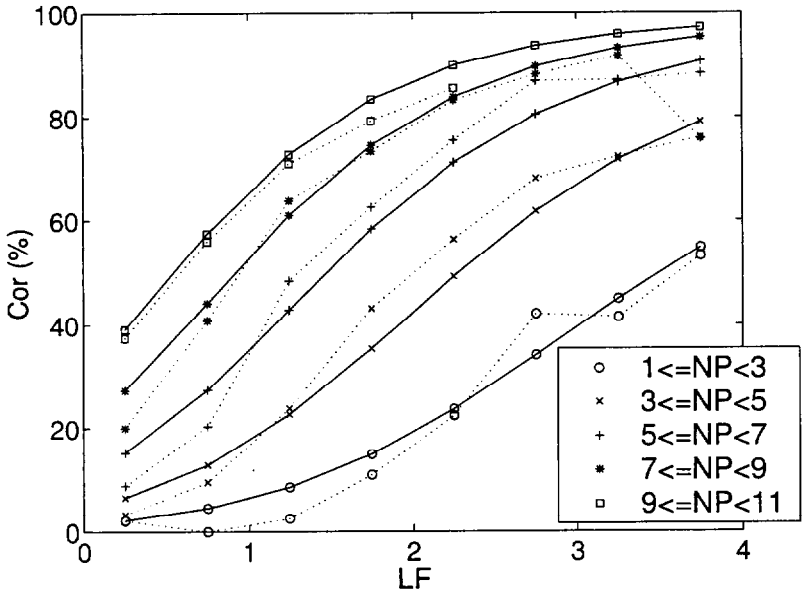

Figure 9.1: Log word occurrence count and recognition correctness.

From the Figures 9.1 and 9.2, it can be seen that the averaged word correctness changes largely according to the number of occurrences and the length of the word. The results are approximately continuous and monotonous.

The averaged word correctness can be regarded as an expected probability that a word is successfully recognized as a function of the attributes. The probability was modeled by a logit model having the logarithmic occurrence count (LF) and the number of phonemes (NP) as explanation variables. Second order terms of the attributes were

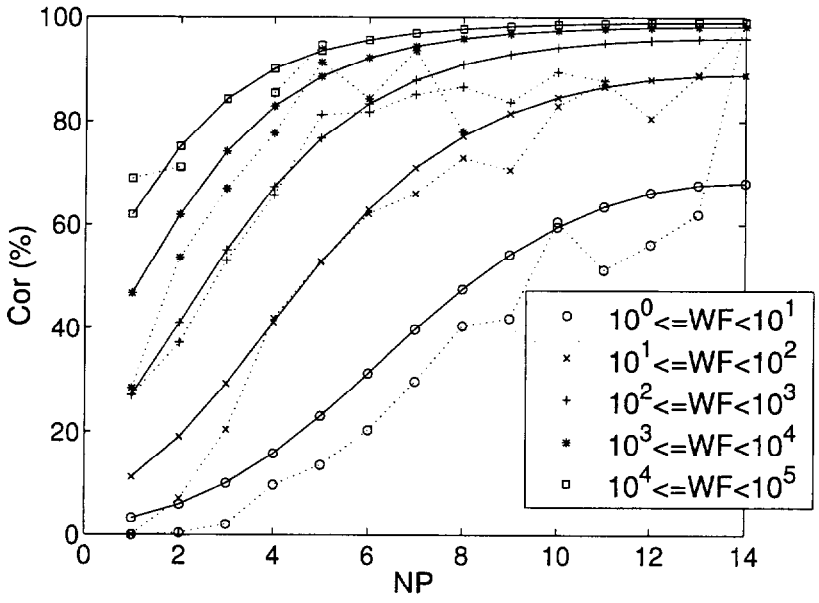

Figure 9.2: Word length and recognition correctness.

also incorporated to improve the modeling accuracy. Equation (9.1) shows the obtained model in which parameters were estimated by the maximum likelihood method.

$$\begin{aligned} &P\left(Cor = 100|NP, LF\right) \\ &= \Lambda\left(0.70NP - 0.03NP^2 + 1.60LF - 0.12LF^2 - 5.17\right) \end{aligned} \tag{9.1}$$

where Λ is a logistic function. Λ is expressed as follows.

$$\Lambda\left(x\right) = \frac{e^x}{1+e^x}. \tag{9.2}$$

Solid lines in the Figures 9.1 and 9.2 indicate the probability estimated by the logit model (9.1), which show that the logit model successfully indicates the global characteristics of the correctness.

9.4 Lexicon optimization method

In the previous section, it was shown that the word correctness probability was effectively modeled by the logit model as a function of the word occurrence count and the word

length. In this section, a new recognition lexicon optimization method based on the logit model is proposed, in which the lexicon is optimized by concatenating basic words iteratively.

The occurrence of words in the parent set of the recognition task is approximated by using the occurrence of words in the training set for building the language model. By using this assumption and the word correctness probability model, a recognition rate of the system can be estimated as an expected value of word correctness as formulated in equation(9.3), in which $\alpha(w)$ is a "recognition rate normalization factor".

$$E[Cor] = \frac{\sum_w P(Cor = 100|NP(w), LF(w)) \cdot WF(w) \cdot \alpha(w)}{\sum_w WF(w)\,\alpha(w)}. \quad (9.3)$$

The recognition rate normalization factor is introduced to impartially compare recognition systems having different recognition units. Suppose there are two recognition systems, one has "ice" and "cream" as separate words and the other has "ice+cream" as a single word for recognition. If "peach ice cream" is spoken and recognized as "beach ice cream" and "beach ice+cream" by the two systems, these two results are basically the same but their correctness values are $2/3$ and $1/2$, respectively.

This inconsistency can be alleviated using either recognition rate based on the initial word unit or the character recognition rate for the comparison. The character recognition rate has been sometimes used in speech recognition evaluation for the languages using Chinese characters, such as Japanese and Chinese. To approximate these recognition rates, the number of initial words before concatenation or the number of characters in the word can be assignment to $\alpha(w)$. The former assignment corresponds to using the word recognition rate based on the initial words and the latter assignment corresponds to using the character recognition rate.

Let's consider selecting a word pair $\langle w_1, w_2 \rangle$ and concatenating all the sequence of these words in this order to create a new word $w_{1,2}$ in the language model training set. New attributes after the operation can be expressed as follows.

- $NP(w_{1,2}) = NP(w_1) + NP(w_2)$

- $WF(w_{1,2})$ = The number of sequences "w_1 w_2" in the original training set.

The attributes of w_1 and w_2 also need to be updated as follows.

- $NP\left(w_i^{'}\right) = NP(w_i)$
- $WF\left(w_i^{'}\right) = WF(w_i) - WF(w_{1,2})$

where $w_i^{'}$ denotes the w_i (i = 1 or 2) after the operation. Note that by definition $\alpha(w_{1,2}) = \alpha(w_1) + \alpha(w_2)$.

A delta() evaluation function of a word pair concatenation is defined as the difference of the estimated correctness of the recognition system before and after the operation as shown in equation (9.4).

$$\Delta(w_1, w_2) \quad = \quad E_{after}[Cor] - E_{before}[Cor]. \tag{9.4}$$

A word pair is selected among all possible word combinations which maximizes the evaluation function and concatenate those word sequences in the training set. The process is iterated by choosing a new word pair step by step. The optimization process is summarized by optlexicon() as shown below.

```
procedure optlexicon() {
 for i = 1:maxiter {
   select word pair <w1,w2>
     which maximizes delta(w1,w2);
   break if delta(w1,w2) < 0;
   merge(w1,w2);
 }
}
```

9.5 Experimental results

9.5.1 Application to the training set

The optimization method was applied to the training set of the baseline language model. The initial vocabulary size, meaning the number of different words, in the training set was 34,895. $\alpha(w)$ was initialized as the number of characters in each word. After iterating the concatenation process for 500 times, the training set which had 1.5M words

at the beginning was reduced to 1.3M words. The estimated α-weighted correctness gain was 1.39%, which means that using the new language model trained using the optimized training set ideally improves the character correctness by 1.39% in the absolute value.

Figure 9.3 shows the concatenation evaluation score and its accumulated value. The evaluation score decreases exponentially as a function of the number of iterations.

Table 9.3 shows the mean and standard deviation of the attributes calculated for the training set before and after the optimization. The averaged number of phonemes increases and the averaged word frequency decreases as the result of the optimization.

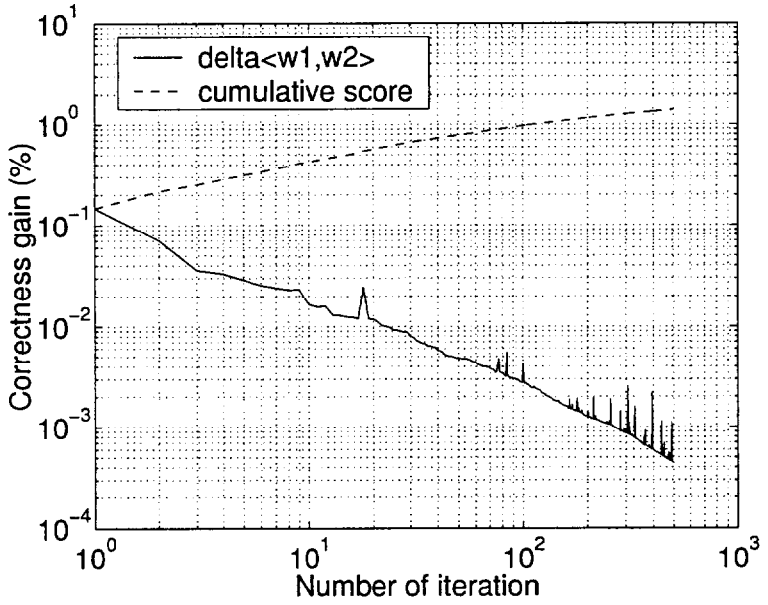

Figure 9.3: Changes of the evaluation and accumulated values in 500 iterations.

Table 9.3: Mean and standard deviation of word attributes before (*base*) and after (*opt*) the optimization

	NP	*WF*	*LF*
Mean(*base*)	3.79	13541	3.23
Standard deviation(*base*)	2.14	19510	1.20
Mean(*opt*)	4.39	7829	2.94
Standard deviation(*opt*)	2.51	13637	1.13

9.5.2 Recognition results

A trigram language model having 30k words was trained using the optimized training set and used for speech recognition. Other conditions were the same as the baseline system. Recognition performance for the evaluation set using the new language model was compared with the baseline system. The performance measured by the character correctness and accuracy is shown in Figure 9.4.

The figure shows that correctness and accuracy were improved for nine and eight lectures, respectively, out of 10 lectures. Averaged improvements were 0.48% and 0.33% in the absolute values for the correctness and accuracy, respectively. As a supplementary experiment, lexicon optimization based on word pair frequency criterion was also tried for comparison. The improvement in averaged accuracy was 0.11%, which was 1/3 of the improvement of the proposed method.

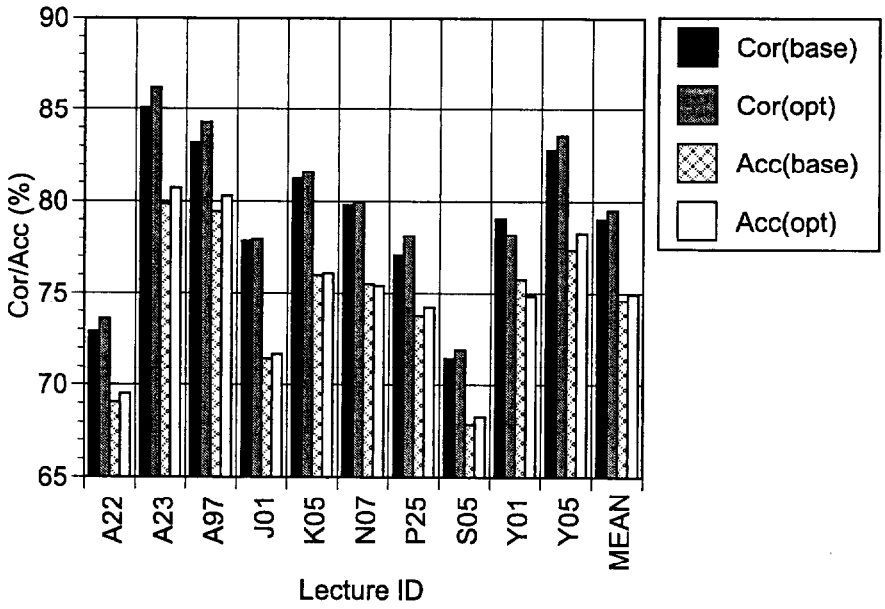

Figure 9.4: Character correctness and accuracy before (*base*) and after (*opt*) the optimization.

9.6 Discussion

The proposed method improved the recognition rate, but the improvement of 0.48% character correctness was much smaller than the estimated improvement of 1.39%.

This may be due to the estimation error by the word correctness probability model (9.1). It was assumed that the model parameters are fixed during the optimization steps. If the word correctness probability changes in the iteration, it may cause errors and the errors may increase as the iteration proceeds. The problem could be reduced by re-estimating the model parameters at some intervals. Another reason may be the fact that only two word attributes, the word frequency and the length, have been considered in the model. Other important attributes which could contribute to improve the performance may exist. Another possible reason is that insertion errors were ignored in the analysis. Considering the insertion error in the method could possibly also improve the recognition rate.

9.7 Conclusion

This chapter first investigated the relationship between the difficulty of word recognition and two word attributes, that is the number of occurrences of each word in the language model training set and the number of phonemes in the word, using a large scale Japanese spontaneous speech corpus. It was shown that the probability of successfully recognizing each word largely varies depending on the two attributes. The relationship was then modeled using the logit model, and the model was used to build a new lexicon optimization method. The proposed method is novel in the sense that it optimizes the lexicon considering both linguistic and acoustic features of words. Recognition results showed that the trigram language model using the optimized lexicon improved the recognition rate.

Chapter 10

Speaking Rate Modeling

10.1 Introduction

One of the main factors which creates difficulty in recognition of spontaneous utterances is the large variation of the speaking rate as indicated in Chapter 5,6,7 and in papers [7, 21, 22, 23]. This chapter explores several ways to extend the HMM to explicitly model the effects of the speaking rate variation. These models are realized by using the dynamic Bayesian network framework which has the ability to model complex probabilistic dependencies.

The reasons for the adverse effect of speaking rate fluctuation include spectral modification, and more directly, the deviation of the phone state duration which then causes a mismatch in transition probabilities modeled by the HMM.

A possible strategy to manage this problem is to first estimate the speaking rate and then adjust a recognizer based on the speaking rate. Sentence level acoustic model selection has been described in [24]. The fastest sentences are selected based on the speaking rate calculated by using the 1st pass recognition results, and re-recognized using an acoustic model adapted to those fastest sentences. In [25], frame level regulation using regression HMMs has been proposed. A way of modifying pronunciation and acoustic likelihood using a hidden mode variable was shown in [26]. This modification of pronunciations based on the hidden variable was implemented in [27]. Modification of the acoustic likelihood has been conducted in [28] where speaking rate information is

used for each frame.

Since the standard HMM is not powerful enough to model complex dependencies, several extensions have been made. However, such kind of extensions often require large efforts for their realization and many other possible extensions are then left untouched. For example, there are many possibilities for how to use the hidden mode variable. The Bayesian network is a flexible statistical framework on which such novel probabilistic models can be rapidly employed [29, 30, 31, 32]. In [29], the idea of using a Bayesian network for compensating for a changing speaking rate is also suggested, but experiments using the network were not conducted. This chapter explores possibilities of several Bayesian network based acoustic models that have a hidden mode variable to deal with speaking rate variation. These models extend a conventional HMM by modifying the parameters of Gaussian mixtures and/or transition probabilities according to the speaking rate frame by frame. These models are evaluated using utterances from meetings and lectures as test sets by rescoring N-best lists which are generated by a Bigram decoder with a 30k vocabulary size.

This chapter is organized as follows. In Section 10.2, the conventional and proposed models are formulated as a Bayesian network. In Section 10.3, several techniques for measuring speaking rate are reviewed. Experimental results are described and discussed in Section 10.4. It is shown that the proposed models using a hidden mode variable are more effective in improving the recognition rate than a regression HMM using the same speaking rate information. Especially, a hidden mode HMM that adjusts both mixture weights and transition probabilities depending on the speaking rate is the most effective. Finally, the chapter is concluded in Section 10.5.

10.2 DBN based acoustic modeling

In this section, a way of formulating the HMM as a Bayesian network is reviewed and a baseline network for encoding the HMM is defined. Then, several models that extend the HMM are described. Since model complexity and estimation accuracy of the parameters

from a training set always pose a trade-off, the number of parameters of the models are given special attention.

10.2.1 Bayesian network

Bayesian networks are directed graphs in which nodes represent random variables, and edges represent probabilistic dependency relations. A Bayesian network is defined by the graph structure and the Conditional Probability Distribution (CPD) at each node. There are several ways in which the CPDs can be defined. For example, if the variable of the node and those of its parents are both discrete, the CPD can be represented as a Conditional Probability Table (CPT), which lists the probability that the node takes on each of its different values for each combination of values of its parents. When the variable of the node is continuous and the parents are discrete-valued, a set of Gaussian mixtures can be used where each element corresponds to a combination of values of its parents [33].

Since speech recognition is a process of time series of feature vectors, Dynamic Bayesian Networks (DBN) [34] are ideally suited for this purpose. DBNs are Bayesian networks that have directed edges pointing in the direction of time. DBNs have a repeating topology of a common core structure, and its CPDs do not change with time.

10.2.2 Baseline model

Figure 10.1 shows an example of a phone HMM set modeling phones /a/ and /b/. Each phone model consists of three states with a left-to-right topology. Figure 10.2 shows the DBN structure that models the phone HMM sequence for model training and N-best rescoring [29], where the discrete variable **Phone-Counter** indicates position in the phone sequence and its value is incremented when binary random variable **Phone-Transition** shows it is phone transition. The node **End-of-utterance** is necessary to ensure that the process ends with a transition out of the last phone. In the figure, observed variables are indicated by shading their nodes. Also, continuous nodes are denoted by circles while discrete nodes are expressed by squares.

In the phone HMM set, a probability distribution for acoustic feature vectors is specified by a phone index and the state index of the phone. The Bayesian network has a node **Phone** that represents a phone index and **Phone-State** that represents the state index of the phone. As abbreviated in Figure 10.3, the node **Observation** which corresponds to acoustic observation, has incoming arrows from the nodes **Phone** and **Phone-State**. This means that the probability the value of **Observation** takes is dependent on these values since each node in a Bayesian network represents a random variable. Similarly, a phone state transition probability to the next HMM state is modeled by a node **Phone-State-Transition** that has incoming arrows from **Phone** and **Phone-State**, indicating probabilistic dependency on these variables. Equation (10.1) and Equation (10.2) show these dependencies of the acoustic observation and the transition probabilities, respectively.

$$P(O|P,S), \tag{10.1}$$

$$P(T|P,S). \tag{10.2}$$

In the equations, O is a single-letter abbreviation of the **Observation** variable for referential convenience, P is **Phone**, S is **Phone-State**, and T is **Phone-State-Transition**.

The node **Phone-State-Transition** represents a binary random variable that indicates either staying at the HMM state or moving to the next state, since the HMM has a left-to-right topology. In this example, cardinalities of the discrete random variables **Phone** and **Phone-State** are two and three, respectively, corresponding to the number of phones and the maximum number of states for each phone. The acoustic observation is a vector of real numbers and **Observation** is a continuous random variable.

A Bayesian network used as a baseline acoustic model has the same structure but a larger cardinality for **Phone**. The CPD of the observation node **Observation** is defined using a set of diagonal covariance Gaussian mixtures. Parameters of the network are trained using the EM/GEM algorithm on a Bayesian network. Decoding is performed by assigning values for all the hidden variables so as to maximize the joint probability

of the entire network. Hereafter, the baseline network is referred to as **BASE**.

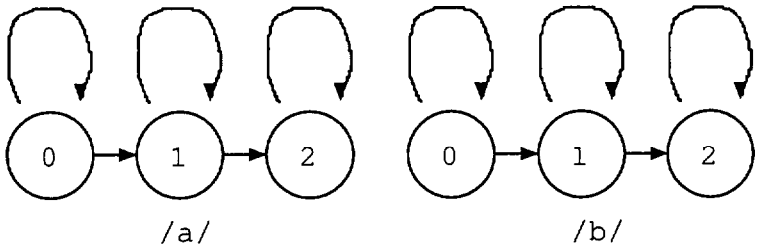

Figure 10.1: A phone HMM set consisting of two phones. Each phone is modeled by a three-state left-to-right HMM.

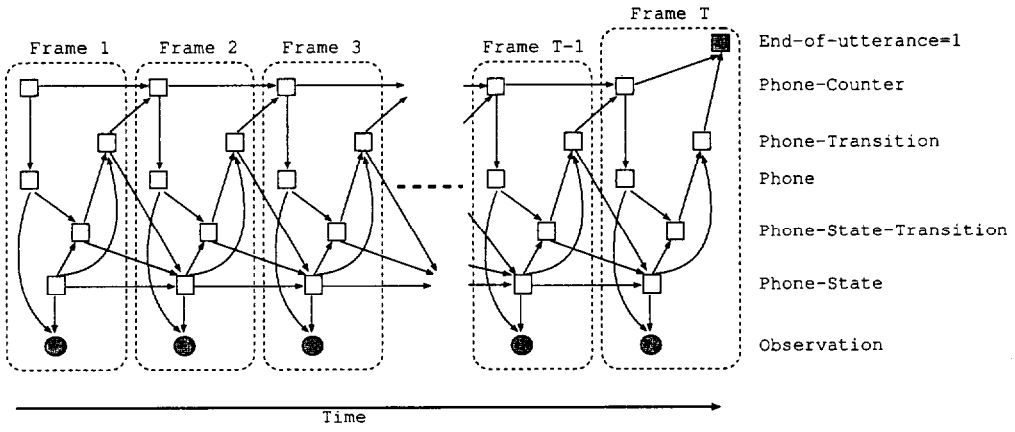

Figure 10.2: DBN representation of the phone HMM sequence. Circles denote continuous-value nodes, squares denote discrete nodes, clear means hidden, and shaded symbols indicate observed nodes.

10.2.3 Regression HMM

One possible way of controlling acoustic observation probability density is to use regression models, in which mean values of the Gaussian components are modeled by linear combination of explanation variables. A multiple-regression HMM has been proposed in [35] where F0 information was used as an auxiliary feature for the explanation variables. The mean vector μ of each Gaussian component is expressed as,

$$\mu = R \cdot \xi + \mu_0, \tag{10.3}$$

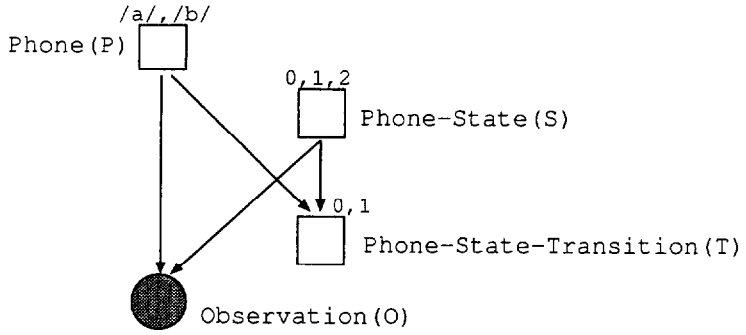

Figure 10.3: A portion of a time slice of the DBN in Figure 10.2 that encodes the conventional HMM. (BASE)

where R is the regression coefficient matrix, μ_0 is the constant term, and ξ is the auxiliary vector. Similar models have been proposed and implemented as DBNs in [36, 25] in which F0 and speaking rate are used as auxiliary information.

In this chapter, a DBN version of the multiple-regression HMM is evaluated using a speaking rate and the second and third order terms as explanation variables. The parameters added to the **BASE** model are regression coefficient matrix components that have the same row dimension as the mean vectors and a column dimension of three. The matrices are tied among Gaussian mixture components in each phone to reduce the number of parameters required to define the model. The Bayesian network representation of this model is shown in Figure 10.4 where there is an additional node **Speaking-Rate** that represents the speaking rate compared to **BASE**. An arrow directly connecting **Speaking-Rate** and **Observation** expresses the dependency between **Observation** and **Speaking-Rate**. The acoustic observation probability is expressed as shown in Equation (10.4), where O is the **Observation** variable, P is **Phone**, S is **Phone-State** and R is **Speaking-Rate**. This model is hereafter called **REG**.

$$P(O|P,S,R) \tag{10.4}$$

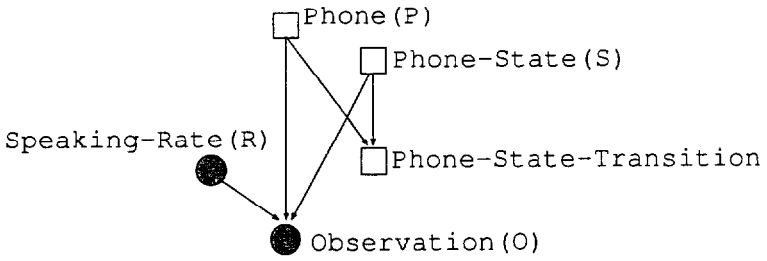

Figure 10.4: Regression model. (REG)

10.2.4 Hidden mode mixture weight model

Figure 10.5 shows a Bayesian network of a proposed model in which the acoustic observation node **Observation** has different probability density according to a "mode" of the speaking rate. In this network, two nodes are added to **BASE**; **Mode** and **Speaking-Rate**. **Mode** is a discrete hidden random variable that represents a "mode" of the speaking rate. As indicated by a dotted line in the figure, **Mode** depends on its counterpart of the previous time slice. This dependence is introduced based on an assumption that the speaking rate changes continuously. A CPT is used at this node. **Speaking-Rate** is a one-dimensional continuous random variable of the speaking rate and a set of Gaussian distributions are used for CPD at this node.

In this configuration, both the acoustic observation node **Observation** and the speaking rate observation node **Speaking-Rate** have the node **Mode** as their parent. According to this network, the joint observation probability of **Observation** and **Speaking-Rate** given **Phone**, **Phone-State**, and **Mode** is factorized, as shown in Equation (10.6), using the local Markov property that a node is independent of all its nondescendants given its parents. In the equation, O is the **Observation** variable, R is **Speaking-Rate**, P is **Phone**, S is **Phone-State**, and M is **Mode**. Equation (10.5) is obtained by applying the chain rule to the joint probability. Equation (10.6) is derived by using the conditional independence relationships of $P(O, R|P, S, M) = P(O|P, S, M)\, P(R|P, S, M)$ and $P(R, P, S|M) = P(R|M)\, P(PS|M)$.

$$
\begin{aligned}
& P(O, R|P, S, M) \\
= \; & P(O|P, S, M, R)\, P(R|P, S, M) && (10.5) \\
= \; & P(O|P, S, M)\, P(R|M). && (10.6)
\end{aligned}
$$

The CPD at node **Observation** has a different Gaussian mixture for each combination of the values of **Phone**, **Phone-State**, and **Mode**. This means that the CPD has $|Mode|$ times more Gaussian mixtures than **BASE**, where $|Mode|$ is the cardinality of the **Mode** variable. Usually, Gaussian mixtures dominate the number of parameters of an HMM. To reduce the number of parameters for accurate model estimation, the Gaussian components are tied for the different values of **Mode**. That is, different values of **Mode** specify different Gaussian mixture weights for the same Gaussian component.

Speaking-Rate has different distributions of the speaking rate depending on **Mode**, and this is used to detect a mode of the speaking rate. The Gaussian mixtures of **Observation** are modified based on a value of **Mode** by choosing different Gaussian mixture weights, and this is how compensation for spectral change is accomplished. Note that the speaking rate mode of each frame is not completely determined simply by the speaking rate but by considering the entire likelihood of the network using an inference algorithm on a Bayesian network. Hereafter, this model adjusting the mixture weights for each time frame by using the hidden mode variable is referred to as **HM-MW**.

Newly introduced parameters in addition to those used in **BASE** are: a CPT of size $|Mode| \times |Mode|$ at **Mode**, a one-dimensional Gaussian distribution for each value of **Mode** for CPD at **Speaking-Rate**, and $|Mode| - 1$ mixture weight vectors for each combination of the values of **Phone** and **Phone-State** at **Observation**. Note that this configuration is applicable not only to the speaking rate but also to any temporal fluctuation that affects speech features.

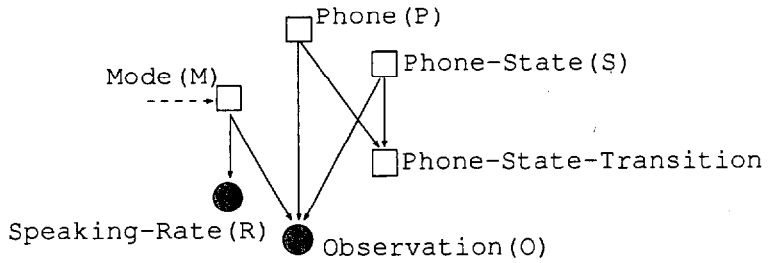

Figure 10.5: Hidden mode mixture weight model. The dotted link represents an edge from the previous time frame. (HM-MW)

10.2.5 Hidden mode transition probability model

In the model described in the previous subsection, observation probabilities of an underlying HMM are controlled by a hidden mode variable. It is also possible to control transition probabilities by using the hidden mode variable as shown in Figure 10.6. The parameterization for the variables **Mode** and **Speaking-Rate** are the same as **HM-MW**. **Mode** is a discrete hidden random variable used to represent the speaking rate mode and **Speaking-Rate** is a one-dimensional continuous random variable modeling the speaking rate. The joint probability of **Phone-State-Transition** and **Speaking-Rate** given **Phone**, **Phone-State**, and **Mode** is factorized as shown in Equation (10.7) based on conditional independence assumptions encoded in the network. In the equation, T is the **Phone-State-Transition** variable, R is **Speaking-Rate**, P is **Phone**, S is **Phone-State**, and M is **Mode**.

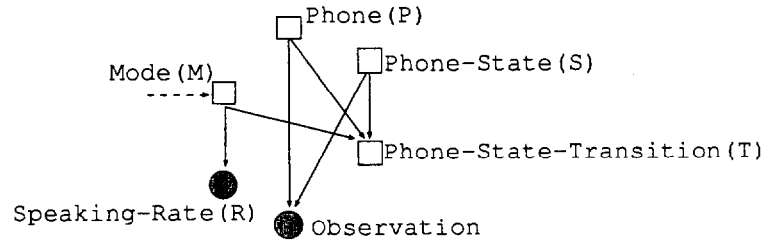

Figure 10.6: Hidden mode transition probability model. (HM-TRP)

Additional parameters to those used in **BASE** are: a CPT of size $|Mode| \times |Mode|$

$$P(T, R|P, S, M) = P(T|P, S, M)\, P(R|M). \tag{10.7}$$

at **Mode**, a one-dimensional Gaussian distribution for each value of **Mode** for CPD at **Speaking-Rate**, and $|SRmode| - 1$ transition probabilities for each combination of the values of **Phone** and **Phone-State** at **Observation**. Since the number of parameters required for modeling transition probabilities are fewer than for the Gaussian mixtures, they are separately modeled for each value of **Mode**. This model is hereafter called **HM-TRP**.

10.2.6 Hidden mode HMM

The controls of the mixture weights and the transition probabilities can be combined as shown in Figure 10.7. The variables introduced to control the underlying HMM parameters are **Mode** and **Speaking-Rate**. **Mode** is a discrete hidden random variable to represent the speaking rate mode and **Speaking-Rate** is a one-dimensional continuous random variable to model the speaking rate, as already explained in the previous subsections. The joint probability of **Observation**, **Phone-State-Transition**, and **Speaking-Rate** given **Phone**, **Phone-State**, and **Mode** is factorized as shown in Equation (10.8), based on conditional independence assumptions encoded in the network. In the equation, O is the **Observation** variable, T is **Phone-State-Transition**, R is **Speaking-Rate**, P is **Phone**, S is **Phone-State**, and M is **Mode**.

$$\begin{aligned} & P(O, T, R|P, S, M) \\ = \; & P(O|P, S, M)\, P(T|P, S, M)\, P(R|M). \end{aligned} \tag{10.8}$$

Additional parameters to those used in **BASE** are a union of the additional parameters of **HM-MW** and **HM-TRP**, that is, a CPT of size $|Mode| \times |Mode|$ at **Mode**, a one-dimensional Gaussian distribution for each value of **Mode** for CPD at **Speaking-Rate**, and $|Mode| - 1$ mixture weight vectors and transition probabilities

for each combination of the values of **Phone** and **Phone-State** at **Observation**. This model is hereafter called **HM-HMM**.

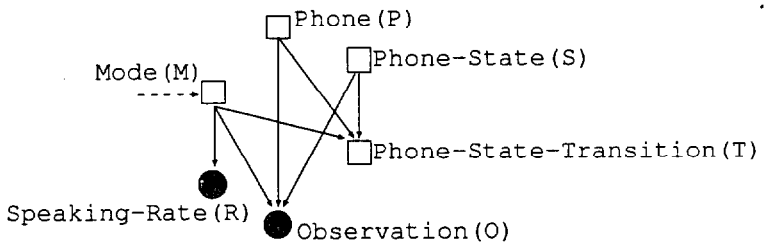

Figure 10.7: Hidden mode HMM. (HM-HMM)

10.3 Measurement of speaking rate

Many approaches have been reported for calculating/defining the speaking rate. They can be roughly divided into two categories, that is, lexical measures and signal based measures.

Lexical measures count units such as words or phones in a certain period. When correct transcription is available, these measures can be calculated by the forced alignment technique. When the correct transcription is not available, a recognition hypothesis can be used instead. A disadvantage of this method is that the hypothesis is not always correct and the errors degrade the reliability of the estimated speaking rate. Thus, when the estimated speaking rate is used to control the recognition system, it is possible that the estimate is less accurate for speech segments where the control by speaking rate is more important.

The signal based measures directly estimate speaking rate without relying on the transcription and thus can avoid the problem of the lexical measures. Enrate, proposed in [37], is one such measures. This is defined as the first spectral moment for the wideband energy envelope of the speech signal. The spectral range is approximately restricted between 1 and 16Hz. The concept of the Enrate is based on the fact that the energy envelope of speech rapidly changes when the speaking rate is high. The Enrate

can be considered as a conversion of the TEMAX-gram [38], which was developed to observe the speaking rate as a spectrogram, into a scalar value. Although the correlation between the Enrate and the phone or syllable rate is not high, it has been shown in [37] that the Enrate is a good predictor of recognition errors.

To improve the correlation with lexical measures, Mrate was proposed in [39]. This is a linear combination of the Enrate and peak-counting estimators. The correlation between the syllable rate and the Mrate is over 0.6, whereas correlation with the Enrate is approximately 0.4 for manually transcribed Switchboard data.

In [40], another way of estimating the speaking rate by detecting vowels has been shown. Modified loudness defined as a difference of higher frequency band loudness and lower frequency band loudness is calculated for every frame. The main part of the energy of a vowel concentrates on lower frequencies, whereas that for most consonants is located at higher frequencies. Therefore, vowels make peaks in the modified loudness and they can be thus detected by finding maxima of the modified loudness. Speaking rate is obtained by taking an inverse of the vowel frequency.

In the following experiments, lexical measures derived from correct and hypothesized transcriptions and the Enrate signal based measure are used. These measures are calculated for each frame of acoustic observation features using significantly overlapped analysis windows.

10.4 Experiments

10.4.1 Corpora and tasks

Two spontaneous speech corpora were used to train and evaluate the DBN based acoustic models. One was a corpus of the Meeting Recorder Project [41] and the other was the Corpus of Spontaneous Japanese (CSJ) [4]. Utterances gathered by the Meeting Recorder Project are recorded from meetings in natural settings, and contain background noises and speech overlaps by other speakers. CSJ consists of Japanese academic lecture speech and extemporaneous public speech. Speaker dependent experiments were

Table 10.1: Characteristic of the acoustic models of the tasks

Task	ICSI meetings	CSJ lectures
Language	English	Japanese
Model type	Speaker dependent	Speaker independent
Feature kind	MFCC_0_D_A	MFCC_E_D_N_Z
Feature dimension	39	25
Window width	25ms	25ms
Frame shift	10ms	10ms
# of phones	45	42
# of mixtures per state	64	28

conducted for the meeting data and speaker independent systems were evaluated using the lecture data. For both of the experiments, utterances recorded using close talking microphones were used.

Speaker dependent models were made using the utterances produced by one male speaker extracted from the meeting corpus. Utterances at nine meetings were used for training, and one meeting was used for testing. Lengths of the utterances for training and testing were 97 and 10 minutes, respectively. Experiments for the speaker independent condition was conducted using academic lectures given by male speakers from the CSJ. Ten lectures were selected as a training set and five lectures were used for testing. These lectures are part of the official test sets of the CSJ. The length of the training set was 116 minutes and the test set was 16 minutes. Table 10.1 shows these conditions.

10.4.2 Model training

First a monophone HMM set was made using the training set and HTK. The parameters of the DBN based acoustic models were initialized with the HMM. Then they were trained by the EM/GEM algorithms using GMTK [42] with 10 iterations.

Each phone of the monophone set was modeled by a three state HMM with a left-to-right topology. The number of Gaussian mixtures per monophone state was determined so as to maximize the recognition rate of the task by preliminary experiments; 64 for the meetings and 28 for the lectures. Table 10.1 shows the characteristic of the acoustic models.

Since the parameters of **Mode** and **Speaking-Rate** do not have corresponding values in the HMM, they were initialized with arbitrary values. For **HM-MW** and **HM-HMM**, the mixture weights were initialized by copying the mixture weights of the monophone HMM. Similarly, for **HM-TRP** and **HM-HMM**, the transition probabilities were initialized by copying those of the monophone HMM. Regression coefficient matrices for **REG** were initialized by giving zeros to all the elements.

After the initialization, most of the trainable parameters, including that of the **Mode**, **Speaking-Rate**, and the regression coefficient matrices, were trained. Only the variances of the Gaussian components in the acoustic observation nodes **Observation** of the networks used for the meeting task were kept constant. This is because the number of mixtures is large in contrast with the amount of the training data. For these DBN acoustic models other than **BASE**, speaking rate information was also used in addition to the normal acoustic features. For **REG**, the speaking rate was normalized so that the mean value became zero for the training set. This made it reasonable to initialize the Gaussian components of the model using those of the monophone HMM.

10.4.3 Experiments using oracle speaking rate

To investigate the effect and limit of the acoustic models, speaking rate information derived from forced alignment of correct phone state sequences with the utterances were used for both training and testing the acoustic models. The speaking rate was defined as an inverse value of the state holding time. The observed values were smoothed using Equation (10.9), where $SR_I(t)$ and $SR_S(t)$ indicate time series of the speaking rate before and after smoothing.

$$SR_S(t) = \sum_{s=-20}^{20} SR_I(t+s) \cdot (20 - |s|). \tag{10.9}$$

The DBN based acoustic models were evaluated by rescoring N-best lists using GMTK with a single pass of max-product inference. The N-best lists were generated using the monophone HMM that was used to initialize the DBN models and a Bigram language model. The Bigram model used for the meeting task was trained on

the HUB5E and the one used for the lecture task was trained on 6.7 million words of transcriptions from the CSJ. Their vocabulary sizes were both 30k. The number of hypotheses generated for each utterance was 50 and 100 for the meeting and the lecture tasks, respectively. The cardinality of the hidden discrete variable **Mode** was set to four.

Figure 10.8 shows the recognition results of the meeting task. The word accuracy of the baseline model **BASE** was 52.7%, and the absolute improvement of the word accuracy by **REG** and **HM-MW** compared to **BASE** was 0.4% and 1.7%, respectively. By controlling the transition probabilities, **HM-TRP** improved the accuracy by 1.7%. The most effective model was **HM-HMM** combining **HM-MW** and **HM-TRP**. This model improved the accuracy by 3.2% for the absolute value by controlling both the mixture weights and the transition probabilities. Similar results were obtained for the lecture task as shown in Figure 10.9. The improvement by **HM-HMM** was 2.1% in this case.

Although both **REG** and **HM-MW** models modify Gaussian mixtures based on the speaking rate, **HM-MW** achieved higher improvement than **REG**. One disadvantage of **REG** might be that it deterministically changes the mean values of the Gaussian components according to the speaking rate. Even if the true speaking rate information is used, it is possible that at some time frame a given speaking rate does not match the local effects of the speaking rate in terms of the changes of the acoustic characteristics, since it has been smoothed as mentioned above. Moreover, it is possible that the relationship between the speaking rate and the change of speech spectra is essentially probabilistic. **HM-MW**, on the other hand, probabilistically chooses a speaking rate mode considering the entire likelihood of the network and therefore it has the capability to select a mode that does not directly match the speaking rate. This feature was obtained by introducing the hidden variable **Mode** for representing the mode of speaking rate.

Mean deletion and substitution error rates with **BASE** and **HM-HMM** for different

speaking rates are shown in Figure 10.10 for the meeting task. The speaking rate was classified into four classes; SR0 is the slowest and SR3 is the fastest. The speaking rate was calculated for each correct word by averaging phone rates using correct transcription. Therefore, insertion errors were not counted. As can be seen in the figure, both the deletion and substitution errors increase for **BASE** as the speaking rate increases. Reduction of the deletion errors by **HM-HMM** is higher at faster speaking rates. For substitution errors, **HM-HMM** has a relatively uniform effect across different speaking rates. Similar error tendencies are observed for the lecture task, though the result is not shown in the figure.

The proposed models, **HM-MW**, **HM-TRP**, and **HM-HMM** have a discrete hidden variable **Mode** that represents a speaking rate mode as explained in Section 10.2. Although the cardinality of the variable is specified beforehand, the correspondence between the value of the variable and the speaking rate is obtained through a training process using a set of Gaussian distributions at **Speaking-Rate**. The distributions are estimated so as to maximize the entire likelihood of the network taking the dependencies on mixture weight and/or transition probability into account. Figure 10.11 shows the four one-dimensional Gaussian distributions of **HM-HMM** corresponding to each value of the **Mode** estimated using the ICSI meetings. As can be seen in the figure, different values of the **Mode** have different features of the speaking rate.

10.4.4 Experiments without using oracle speaking rate

Rescoring experiments without relying on the true transcription were conducted using two different speaking rate measures for **REG** and **HM-HMM**. One measure was **HYP**, which was similar to the one used in the oracle experiments with the exception that **HYP** uses the one-best hypothesis in the N-best list as an approximation of the true transcription. For the rescoring, the same acoustic models as the previous experiments were used. The other was **ENRATE** which was the Enrate measure. Window width for the Enrate calculation was set at 400ms based on preliminary experiments. When

rescoring, acoustic models trained with Enrate were used.

Tables 10.2 and 10.3 show the results for the meeting and lecture tasks, respectively. In the table, the results by the baseline model without using the speaking rate information indicated by **BASE** and those by using the speaking rate calculated from true transcription indicated by **ORACLE** are also shown. The cardinality of **Mode** was set to three and four.

As can be seen in Table 10.2, no improvement was obtained by the regression model **REG** for the meeting task regardless of using **HYP** or **ENRATE** measures. This is probably because the regression model is vulnerable to the decrease of the quality of the speaking rate. Because the one-best hypothesis includes recognition errors, **HYP** is not an accurate approximation of the oracle speaking rate. Although **ENRATE** is free from the recognition errors, it seems to be less effective in explaining the change of acoustic features compared to the oracle speaking rate. **HM-HMM** succeeded in exploiting the speaking rate information to improve the word accuracy. When the cardinality of **Mode** was set to three, an absolute improvement of 0.7% and 0.8% was obtained for **HYP** and **ENRATE**, respectively. For the lecture task, as Table 10.3 indicates, the highest improvement of 1.3% was found for **HM-HMM** with a **HYP** measure, where the cardinality of **Mode** was set to four. The optimal cardinality of **Mode** probably depends on the underlying HMM complexity such as number of mixtures, amount of training data, and estimation accuracy of the speaking rate.

10.5 Conclusions

This chapter explored several dynamic Bayesian network based acoustic models for improving recognition accuracy of spontaneous speech using an explicitly modeled effect of the speaking rate. Although the DBN based recognition system is slower than conventional systems that are highly tuned for the speech recognition domain, it provides a flexible framework and is well suited for analyzing underlying principles and prototyping.

When speaking rate information obtained from the true transcription was given,

Table 10.2: Word accuracy of the meeting task

	REG	HM-HMM |Mode|=3	HM-HMM |Mode|=4
BASE		52.7	
HYP	52.4	53.4	53.0
ENRATE	52.5	53.5	53.1
ORACLE	53.1	55.3	55.9

Table 10.3: Word accuracy of the lecture task

	REG	HM-HMM |Mode|=3	HM-HMM |Mode|=4
BASE		48.5	
HYP	49.0	49.3	49.7
ENRATE	48.6	48.8	48.7
ORACLE	49.3	50.0	50.5

the proposed models, **HM-MW**, **HM-TRP**, and **HM-HMM** indicated higher performances than **BASE** which encodes conventional HMM, and **REG** which encodes regression HMM using the same speaking rate information. The absolute improvement achieved by using **HM-HMM** was 3.2% and 2.1% for the meeting and lecture tasks, respectively. These DBN based acoustic models were also evaluated using speaking rate measures without using true transcriptions. Two measures were used for this purpose, best hypothesis-based speaking rate and Enrate. Although the regression model **REG** sometimes failed in making use of these speaking rates, **HM-HMM** showed improvement over the conventional models for both tasks. In the best condition, **HM-HMM** improved word accuracy by 0.8% for a meeting task and 1.3% for a lecture task. For both of the experiments with and without oracle speaking rate, the proposed models indicated consistently higher performance than conventional HMMs and regression HMMs using the same speaking rate information.

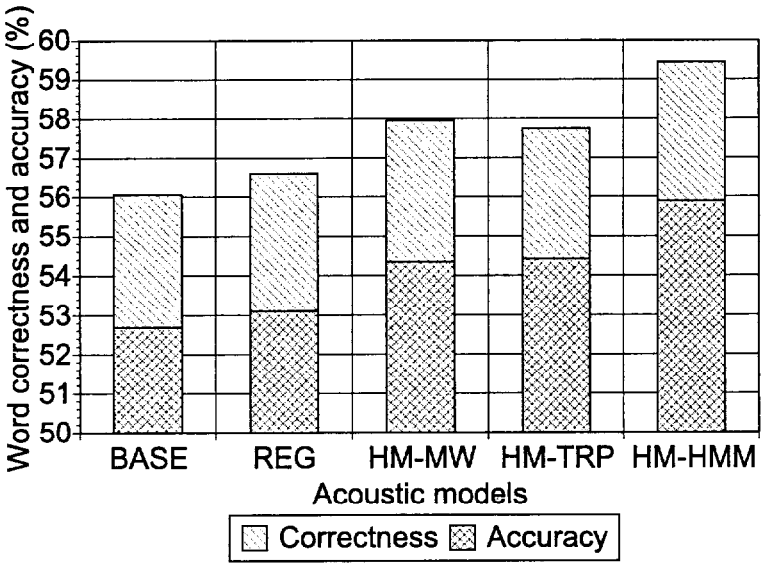

Figure 10.8: Word correctness and accuracy of the meeting task given speaking rate measured using true transcript.

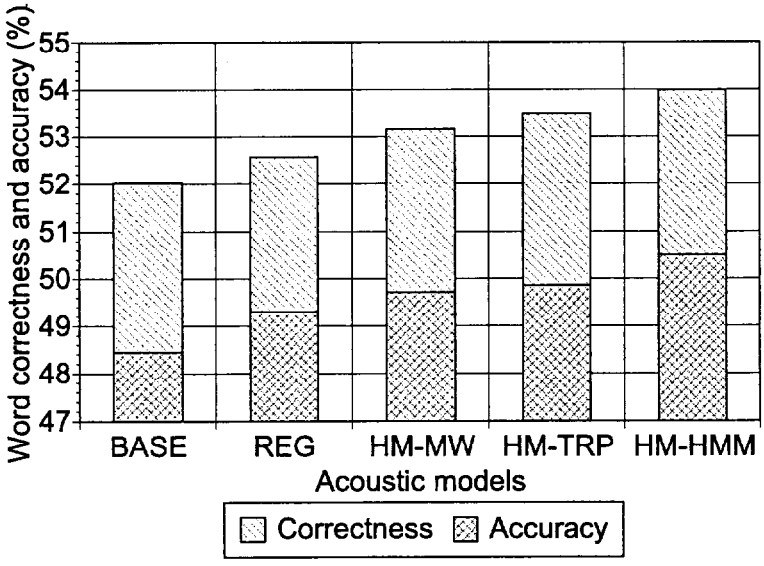

Figure 10.9: Word correctness and accuracy of the lecture task given speaking rate measured using true transcript.

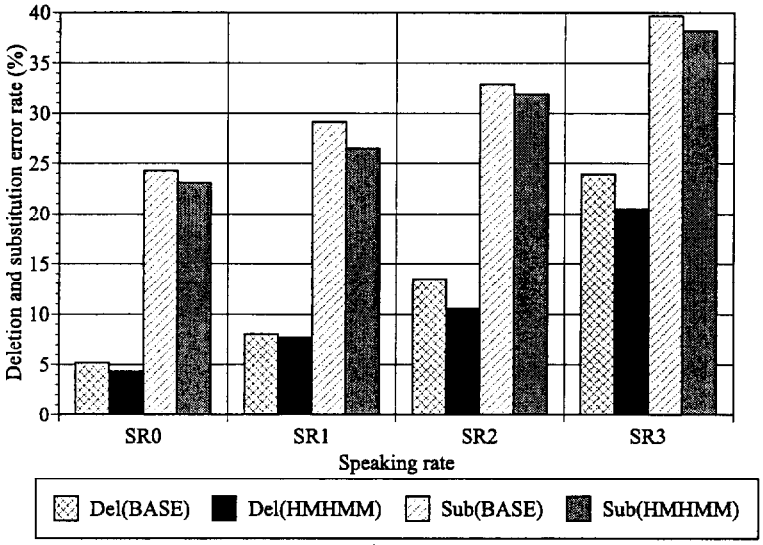

Figure 10.10: Error distribution for speaking rate.

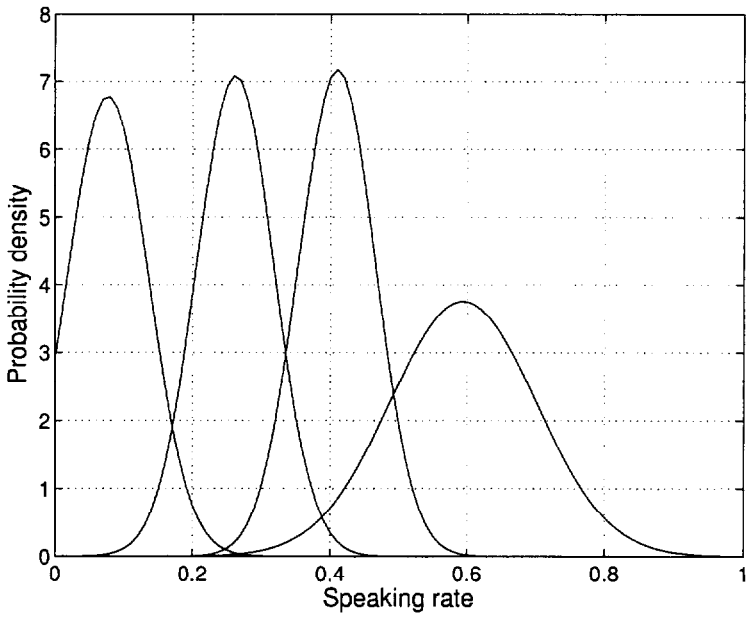

Figure 10.11: Gaussian distributions for the variation of the speaking rate mode trained on the ICSI meetings.

Chapter 11

Massively Parallel Decoder

11.1 Introduction

Recently, several large scale spontaneous speech corpora have become available and recognition performance for spontaneous speech has been greatly improved by using a large amount of spontaneous data to make speech models. However, recognition rates are still not adequate to satisfy the demands of most applications. This is because spontaneous speech has many variations between speakers and can vary greatly even in different utterances of one speaker. These speaker and utterance specific characteristics can not be modeled by a speaker independent or general model since they are averaged in the process. One solution is to model speech sounds as a set of speech models that will include a suitable model for every specific input utterance.

In [43], speaker cluster based HMMs were used to process broadcast news speeches where the speakers change frequently. The recognition system selects one of the HMMs for each utterance. In [44], sentence level mixture Ngrams were investigated to capture topic related dependencies. The component Ngrams are based on article clusters and they are mixed at sentence level instead of choosing the component. Since it is computationally expensive to search for hypothesis calculating the maximum or sum of the likelihood of all the component models, some approximation has been required such as using GMMs to choose a component HMM and rescoring N-best lists generated by using a speaker independent model. However, to choose an HMM by using GMMs,

GMM likelihood must be calculated before the selection and the recognition process can not be started until a certain period passes after an utterance begins. Moreover, the selection might not be optimal since GMM is not an accurate model. When a cluster model is used to rescore N-best lists, the improvement in recognition accuracy is limited by the restricted search space.

In order to take advantage of the modeling strategy to its fullest extent, this chapter proposes a Massively Parallel Decoder (MPD). MPD consists of a large number of decoding units and an integrator. It runs on a parallel computer and can process speech utterances with almost the same turnaround time as conventional decoders.

This chapter is organized as follows. Architecture and processing time of the MPD are described in Section 11.2. Experimental conditions are described in Section 11.3 and the results are presented in Section 11.4. Finally, some conclusions are given in Section 11.5.

11.2 Massively Parallel Decoder

The Massively Parallel Decoder (MPD) is a decoder that runs using an cluster speech model. Figure 11.1 shows its architecture. It consists of a set of decoding units (DUs) and an integrator. Each DU is just a conventional decoder that uses one of the element speech models in the cluster model. An input speech utterance is sent to all the DUs and each DU processes the signal independently based on its speech model. The recognition hypotheses of the DUs are gathered to the integrator and a final output is synthesized. While there are many ways to integrate the hypotheses from the DUs [45], a maximum likelihood criteria is used in the following experiments. The integrator selects the hypothesis with the highest likelihood.

MPD can work efficiently on parallel computers such as Grid [46], MPP and SCM [47]. Grid and MPP connects many computers or processors to form a parallel computer. The SCM integrates many processing units in a single chip. Although Grid and MPP currently require much power and space, they will be realized as a System on Package

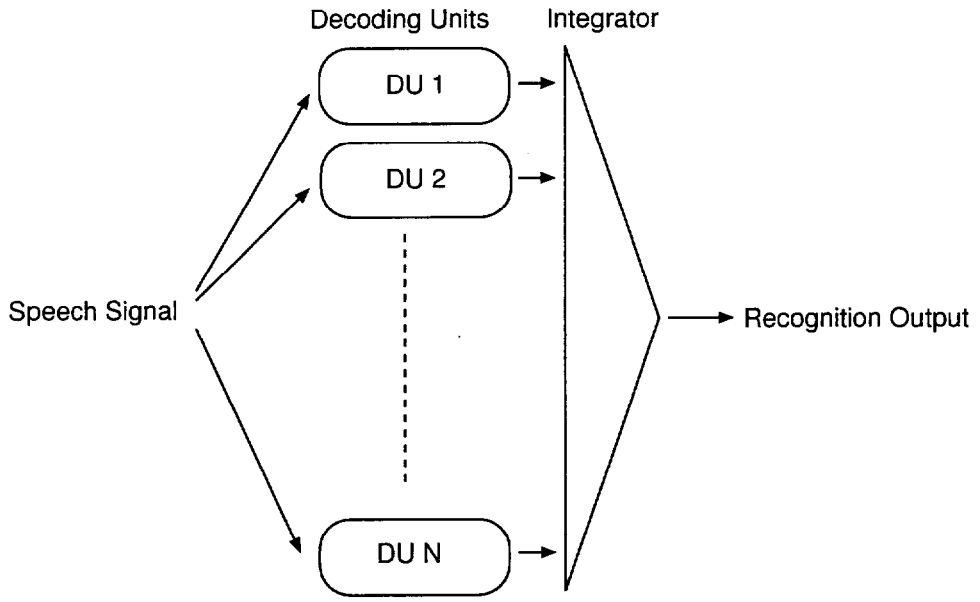

Figure 11.1: Architecture of the Massively Parallel Decoder.

(SOP) and will become easy to use.

In any case, parallel computers will become popular in the near feature because they solve critical problems of single processor systems such as line delay. To take advantage of parallel computers, parallel algorithms are crucial. In this aspect, MPD is well suited to parallel computers since it has highly parallelized structure.

Figure 11.2 shows the processing time of a MPD running on a parallel computer. By assigning each DU to a different processing unit (PU), the turnaround time T of the MPD becomes constant to the number of DUs as shown in equation (11.1). In the equation, t and β represent processing time of the DU and the integrator, respectively. Since the processing time of the integrator is negligible compared to that of the decoding unit, equation (11.1) can be approximated as equation (11.2). Thus, the turnaround time of MPD is about the same as conventional decoders using single acoustic and language models.

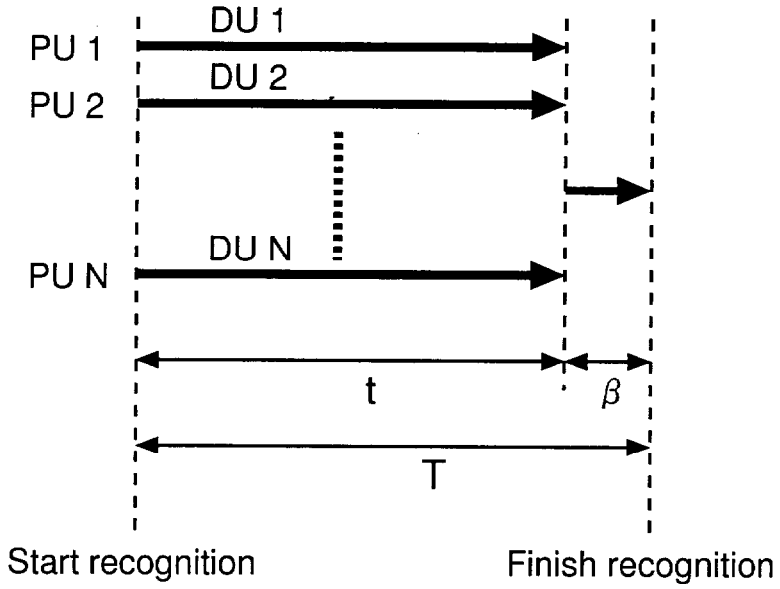

Figure 11.2: Processing time of MPD

$$T = t + \beta \tag{11.1}$$

$$\approx t \tag{11.2}$$

11.3 Experimental conditions

11.3.1 Recognition task

The recognition task was the test-set 1 of the Corpus of Spontaneous Japanese (CSJ). The test set consisted of ten academic lectures given by different male speakers. In the experiment, utterances were extracted based on silence periods longer than 500ms and five minutes of utterances were excerpted from each lecture. Figure 11.1 shows the lecture IDs and number of utterances in the five minute sets.

11.3.2 Acoustic models

The training set for acoustic models was the CSJ academic presentations given by male speakers. It includes 787 lectures and amounts to 186 hours. Feature vectors had 38

Table 11.1: Test-set

Conference name	# of utterances used
A01M0097	58
A04M0051	77
A04M0121	73
A03M0156	88
A03M0112	43
A01M0110	65
A05M0011	31
A03M0106	27
A01M0137	45
A04M0123	23

elements consisting of 12 MFCC, their delta, delta delta, delta log energy and delta delta log energy. The CMS (cepstral mean subtraction) was applied to each utterance. HTK [5] was utilized for model training and adaptation.

For a baseline system, a speaker independent triphone HMM was made that had 3k states and 16 Gaussian mixtures in each state. A regression class tree with 64 leaves was associated with the HMM that classified Gaussian mixtures for MLLR adaptation. This baseline HMM is hereafter denoted as **SIAM** (Speaker Independent Acoustic Model).

Two types of cluster acoustic models were made. One was made by adapting the speaker independent triphone HMM to 400 different male speakers in the training set using the MAP adaptation method. This model is denoted as **SCAM** (Speaker Cluster Acoustic Model). The other was based on utterance clustering. The clustering was conducted as follows using the whole training set for acoustic models.

1. Randomly distributes utterances to N clusters so that all the clusters have the same number of utterances. N is the number of clusters desired. Makes HMMs for each cluster by adapting the speaker independent HMM.

2. Calculates the likelihood for all the utterances for all the clusters.

3. For each cluster, selects an utterance with the highest likelihood by rotation until all the utterances are classified. This ensures that all the clusters have the same number of utterances.

4. Makes HMMs for each cluster by adapting the HMMs made in the previous stage.

5. Goes to step 2 or ends after sufficient iteration. The iteration number of ten was chosen in the following experiments.

Based on the obtained definitions of the utterance clusters, cluster acoustic models were made by adapting the speaker independent HMM using the MAP adaptation method. These models are denoted as **UCAM(N)** (Utterance Cluster Acoustic Model with N clusters), where N is the number of element models.

11.3.3 Language models

The training set used for language models included academic and extemporaneous lectures. It consists of 2485 CSJ lectures and contains 6.1 million words. The baseline language model was a word trigram interpolated with a word class trigram which was based on 100 word classes. The vocabulary size of the baseline model was 30k. Interpolation weights of 0.7 and 0.3 were used for word and class models, respectively. The word class definition was trained using the incremental greedy merging algorithm [48]. This model is denoted as **SILM** (Speaker Independent Language Model).

Similar to the cluster acoustic models, two types of cluster language models were made. One was made by training a set of speaker dependent models. The speaker dependent models were made by weighting the transcriptions of the speaker. The target speakers were the same as those selected for the cluster acoustic models. For the weighting, the transcription was multiplied so that the cumulative number is about 5% for the entire training set. This model is hereafter denoted as **SCLM** (Speaker Cluster Language Model).

The other cluster model was based on utterance clusters. The clusters were trained using the same algorithm as for the utterance cluster acoustic models but using perplexity instead of acoustic likelihood. The whole training set for language models was used for the training. In the iteration process of the algorithm, the cluster based models were made by adapting models in the previous stage by duplicating the utterances belonging

to the cluster. Based on the obtained definition of the utterance clusters, word trigrams were made from the transcriptions in which corresponding utterances were duplicated. The final utterance cluster models were made by interpolating the word trigrams with the word class trigram using the fixed interpolation weights of 0.7 and 0.3. These models are denoted as **UCLM(N)** (Utterance Cluster Language Model with N clusters), where N is the number of element models.

11.3.4 Recognition Systems

The Julius decoder [6] was used both as a baseline recognition system and for decoding units of MPDs. A Grid system was used for the MPDs. The baseline decoding system used the speaker independent acoustic model (SIAM) and the language model (SILM). The MPDs use the cluster acoustic model and/or cluster language model. To specify which models are used, the recognition systems using MPD are denoted as **MPD(AM, LM)**, where AM is the acoustic model and LM is the language model. For example, a MPD based recognition system using UCAM(40) in combination with SILM is denoted as MPD(UCAM(40), SILM).

11.3.5 Unsupervised adaptation

For some applications, response time is not an issue and recognition can be performed off-line. In such cases, batch-type unsupervised adaptation is a useful way to improve the recognition rate. Unsupervised acoustic and language model adaptations were applied to the speaker independent model and the cluster models. The adaptations were conducted for each lecture using the recognition results of the baseline or the MPD based recognition systems.

For the baseline system, unsupervised MLLR acoustic model adaptation and language model adaptation using word class [49, 50] are applied at the same time based on the recognition results using the speaker independent models. The word class based language model adaptation method updates word probability given word class by maximum likelihood criteria using the recognition hypotheses. The adaptation for the cluster

models are conducted in a similar way by adapting all the element models using recognition results given by the MPD.

11.4 Experimental results

Table 11.2 shows the results using MPDs with the cluster acoustic models and the speaker independent language model. As can be seen, all the results using MPDs indicated lower word error rate than the baseline system using the speaker independent acoustic and language models. Among the utterance cluster based cluster models with different number of elements, UCAM(10) showed the lowest word error rate. The error rate of UCAM(10) was even lower than the speaker based cluster model SCAM(400) that has a 40 times larger number of clusters. Table 11.3 shows the results using the MPDs with the cluster language models. UCLM(20) indicated the lowest word error rate among them.

Table 11.2: Recognition results of the MPDs using the cluster acoustic models.

Recognition system	Word error rate
BASE	24.9
MPD(UCAM(6),SILM)	23.6
MPD(UCAM(10),SILM)	23.1
MPD(UCAM(20),SILM)	23.3
MPD(SCAM(400),SILM)	23.5

Table 11.3: Recognition results of the MPDs using the cluster language models.

Recognition system	Word error rate
BASE	24.9
MPD(SIAM, UCLM(6))	23.7
MPD(SIAM, UCLM(10))	23.6
MPD(SIAM, UCLM(20))	23.3
MPD(SIAM, UCLM(40))	23.7
MPD(SIAM, SCLM(400))	23.9

Figure 11.3 shows the results of the MPD(UCAM(10), UCLM(10)) that combines the utterance cluster acoustic model UCAM(10) and the utterance cluster language model UCLM(10). The MPD has 100 decoding units according to the number of

combinations of the element acoustic and language models. The word error rate using MPD(UCAM(10), UCLM(10)) is 22.3% and the relative reduction compared to the baseline system is 10.5%. By combining the cluster models, MPD(UCAM(10), UCLM(10)) indicated lower error rate than MPD(UCAM(10), SILM) and MPD(SIAM, UCLM(10)). Figure 11.4 compares lecture word error rate using the baseline system and MPD(UCAM(10), UCLM(10)). As can be seen, error rates were reduced for all the test set lectures.

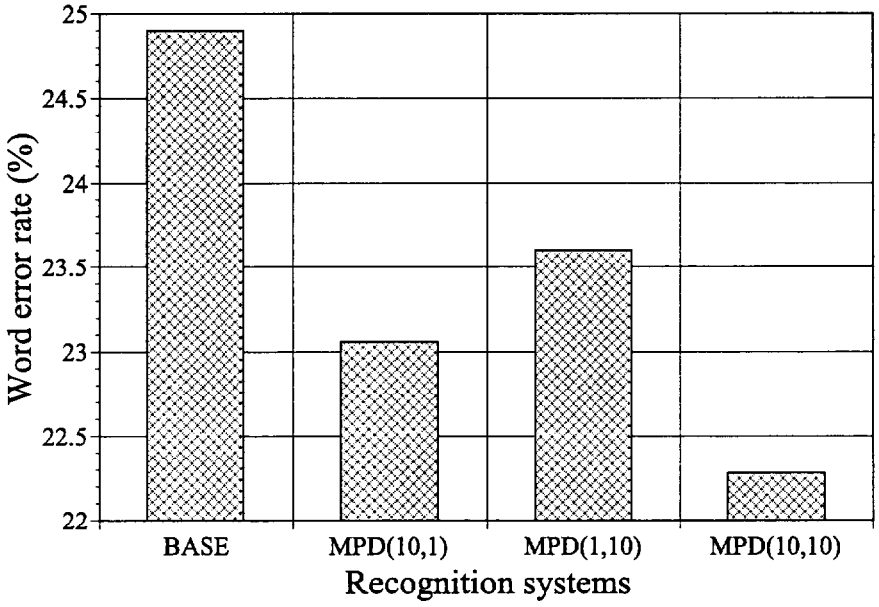

Figure 11.3: Word error rate. BASE denotes results of the baseline system. MPD(10,1), MPD(1,10), and MPD(10,10) denote results using the massively parallel decoder of MPD(UCAM(10),SILM), MPD(SIAM,UCLM(10)), and MPD(UCAM(10),UCLM(10)),respectively.

Figure 11.5 shows results when unsupervised acoustic model and language model adaptation were applied for the baseline system and MPD based system. The MPD based system was MPD(UCAM(10), UCLM(10)). By combining the MPD and the unsupervised adaptation, a word error rate of 20.4% was obtained.

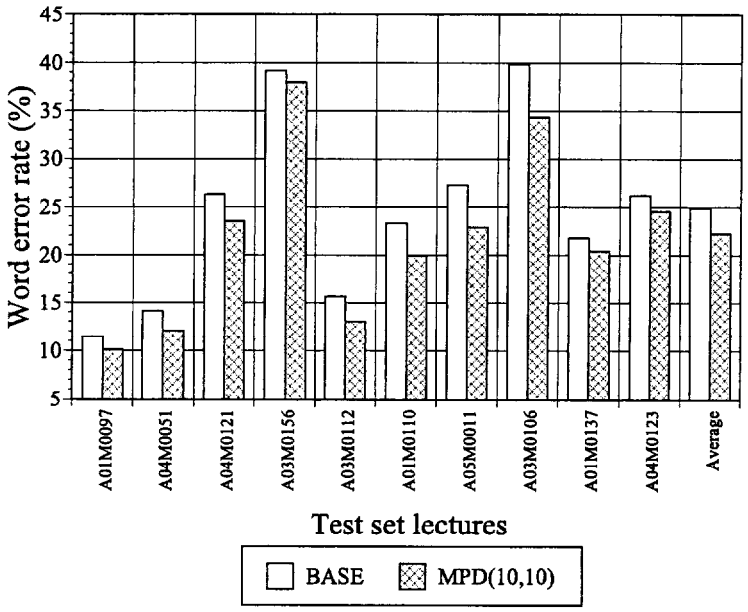

Figure 11.4: Word error rate of each test set lecture.

11.5 Conclusion

This chapter has proposed the Massively Parallel Decoder which consists of a large number of decoding units and an integrator. The MPD runs using cluster acoustic and/or language models. By using parallel computers, there is almost no increase of the turnaround time compared to conventional decoders which use single acoustic and language models. A relative error rate reduction of 10.5% was obtained by using MPD compared to the baseline system using speaker independent speech models. It was also confirmed that MPD is effective for off-line decoding with unsupervised acoustic and language model adaptation.

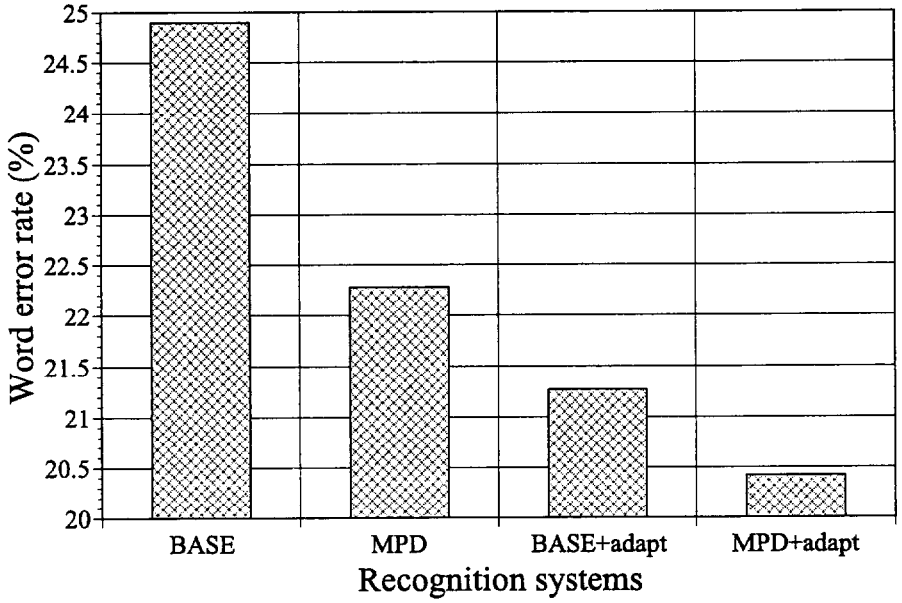

Figure 11.5: Word error rate when combined with unsupervised acoustic model and language model adaptation. BASE+adapt and MPD+adapt denotes results when the adaptation is conducted.

Chapter 12

Conclusion

12.1 Summary of accomplishments

The characteristics of spontaneous speech are very different from read speech and recognition accuracy of technological systems drastically decreases for spontaneous speech. It is now crucial to improve recognition techniques for spontaneous speech.

The study began with building a recognition system that was based on the Japanese spontaneous speech corpus CSJ. It is the first system that was based on the CSJ and various initial investigations on recognizing spontaneous presentation speech have been reported in Chapter 5. Presentation speech uttered by 10 male speakers of approximately 4.5 hours duration was recognized. Experimental results show that acoustic and language modeling based on an actual spontaneous speech corpus is far more effective than conventional modeling based on read speech. Recognition accuracy has a wide speaker-to-speaker variability according to the speaking rate, the number of fillers, the number of repairs, etc. It was confirmed that unsupervised speaker adaptation of acoustic models was effective for improving the recognition accuracy. It was also shown that the recognition accuracy for spontaneous speech is, however, still rather low, and there remain a large number of research issues.

To understand problems of spontaneous speech recognition, various analyses were conducted in Chapter 6, 7, and 8. In Chapter 6, an analysis of individual differences in spontaneous presentation speech recognition performances was presented. Ten minutes

from each presentation given by 50 male speakers, for a total of 500 minutes, was automatically recognized for the analysis. Correlation and regression analyses were applied to the word recognition accuracy and various speaker attributes. A restricted set of speaker attributes comprising the speaking rate, the out of vocabulary rate and the repair rate was found to be most significant in yielding individual differences in the word accuracy. Unsupervised MLLR speaker adaptation worked well for improving the word accuracy but did not change the structure of the individual differences. Approximately half of the variance in the word accuracy was explained by a regression model using a limited set of the three attributes.

Chapter 7 proposed the use of decision trees for analyzing errors in spontaneous presentation speech recognition. The trees were designed to predict whether a word or a phoneme can be correctly recognized or not, using word or phoneme attributes as inputs. The trees were constructed using training "cases" by choosing questions about attributes step by step according to the gain ratio criterion. The errors in recognizing spontaneous presentations given by 10 male speakers were analyzed, and the explanation capability of attributes for recognition errors was quantitatively evaluated. A restricted set of attributes closely related to the recognition errors were identified for both words and phonemes.

In Chapter 8, an automatic speech recognizer was evaluated in comparison with performances by human listeners to investigate problems of spontaneous speech recognition using N-grams and HMMs and to estimate potential gains for improvement in recognition rate. The evaluation task was to recognize spontaneous speech presentations from the Corpus of Spontaneous Japanese. Both the automatic recognizer and human listeners were requested to choose the most likely word from a dictionary, given a speech signal of three words in length including $\pm$ one word context extracted from a presentation. Recognition performances were compared using the same criteria for both experiments. The results showed that the recognition error rate by human listeners is roughly half of that by the recognizer. By examining words that were easy for humans

but difficult for the recognizer, it was found that causes of the recognition errors by the decoder included insufficiency of model accuracy and lack of robustness against vague and variable pronunciations. While there is room for improvement even in conditions that do not use contexts longer than trigrams, wider context information should be incorporated to achieve high word accuracy.

One conclusion of the analyses was that the difficulty of recognizing a word largely depends on the length and frequency of the word. Based on this observation, a new lexicon optimization method to improve recognition rate of large scale spontaneous speech recognition has been proposed in Chapter 9. First, a word correctness probability model was made to model correlation between the difficulty of recognizing a word and the attributes of the word. The proposed method optimizes the lexicon by making compound words or phrases step by step based on the word correctness probability model so as to improve the estimated recognition rate of the system. The optimization method was applied to a large scale Japanese spontaneous speech corpus. Experimental results showed that the language model using the optimized lexicon improved the recognition rate.

To cope with degradation of recognition accuracy due to speaking rate fluctuation within an utterance, which is one of the most significant problems in spontaneous speech recognition, a new acoustic model for adjusting mixture weights and transition probabilities of the HMM for each frame according to the local speaking rate has been proposed in Chapter 10. The proposed model was implemented along with variants and conventional models using the 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 the Bayesian networks. Utterances from meetings and lectures were used for evaluation where the Bayesian network-based acoustic models were used to rescore the likelihood of the N-best lists. In the experiments, the proposed model indicated

consistently higher performance than conventional HMMs and regression HMMs using the same speaking rate information.

To deal with spontaneous utterances that include many variations, the Massively Parallel Decoder (MPD) has been proposed in Chapter 11. MPD consists of a large number of decoding units and an integrator. MPD recognizes utterances using a speech cluster model. Each decoding unit of MPD uses one of the element models in the cluster. By using a parallel computer, the processing time of MPD is almost the same as conventional decoders which use single model and processor. Experiments were conducted using MPDs that have up to 400 decoding units. By running a MPD with an utterance cluster HMM and an utterance cluster Trigram, 11% reduction in word error rate was obtained. By combining MPD with an unsupervised MLLR adaptation and a class model based Ngram adaptation, an averaged word accuracy of 80% was obtained for the CSJ test-set lectures.

12.2 Future work

Three new techniques were proposed in this study: the lexicon optimization method, the hidden mode HMM, and the Massively Parallel Decoder (MPD). For the lexicon optimization method, further progress is expected by improving the word correctness probability model and the concatenation algorithm. Currently, the word correctness probability model does not take account insertion errors. Also, it uses only two kinds of word attributes. The problem of the concatenation algorithm is that it were the hypothesis that the relationship between the word attributes and the recognition difficulty is fixed during the optimization steps.

Future works for the hidden mode HMM include investigating more efficient ways of utilizing speaking rate information, finding better methods for speaking rate estimation, incorporating other spontaneous speech features to further improve the recognition accuracy, and implementing computationally efficient systems that can work with more general LVCSR conditions for promising probabilistic models which can be found by

using flexible DBN toolkits.

For the MPD, some prior probabilities can be introduced for selecting an element speech model given a history of the previous choices. Also, some improvement is expected by investigating integration methods for the integrator and clustering techniques for cluster speech models.

Apart from these proposed methods, problems which have been left untouched are the out of vocabulary and repair words. Also, a modeling method for inter and intra acoustic fluctuation should be further improved. Since in spontaneous speech recognition, local acoustic and linguistic clues are not always enough to recognize words, utilizing wider context, domain knowledge, and semantic factoring is also important.

Chapter 13

Acknowledgment

I am deeply grateful to Prof. Sadaoki Furui for his invaluable discussion and exact guidance. I wish to express my thanks to the members of the "Spontaneous Speech" national project for constructing the corpus and for fruitful discussions. I would also like to thank all members of Furui Laboratory for their help and friendship.

Bibliography

[1] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans .Acoust. Speech and Signal Processing, ASSP-34 No.1, pp.52–59, 1986.

[2] T. Nakai, K. Matsuo, C. Kato, S. Tanaka, G.H. Glover, T. Moriya and T. Okada, "The auditory attention system during dual listening Task performance," ISMRM, 894, 2000.

[3] http://www.es.jamstec.go.jp/esc/en

[4] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki, and T. Ohdaira, "Toward the realization of spontaneous speech recognition — Introduction of a Japanese Priority Program and preliminary results," Proc. ICSLP, Beijing, pp. 518-521, 2000.

[5] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, "The HTK Book, Version 2.2," Entropic Ltd, 1999.

[6] A. Lee, T. Kawahara, and S. Doshita, "An efficient two-pass search algorithm using word trellis index," Proc. ICSLP, Australia, pp.1831-1834, 1998.

[7] T. Shinozaki, C. Hori, and S. Furui, "Towards automatic transcription of spontaneous presentations," Proc. EUROSPEECH, Denmark, Vol. 1, pp. 491-494, 2001.

[8] J.R.Quinlan, "C4.5: programs for machine learning," Morgan Kaufmann, 1996.

[9] H. Akaike, "Information theory and an extension of the maximum likelihood principle," Proc. ISIT, (B. N. Petrov and F. Csaki eds.) Akademiai Kiado, Budapest, pp.267-281, 1973.

[10] N. Deshmukh, R. J. Duncan, A. Ganapathiraju, and J picone, "Benchmarking human performance for continuous speech recognition," Proc. ICSLP, vol. 4, pp. 2486–2489, 1996.

[11] D. A. van Leeuwen, L. G. Van den Berg, and H. J. M. Steeneken, "Human benchmarks for speaker independent large vocabulary recognition performance," Proc. Eurospeech, vol. 2, pp. 1461–1464, 1995.

[12] R. P. Lippmann, "Speech recognition by machines and humans," Speech Communication, vol. 22, pp. 1–15, 1997.

[13] E. Giachin, P. Baggia, and G. Micca, "Language models for spontaneous speech recognition: a bootstrap method for learning phrase bigrams," Proc. ICSLP, vol. 2, pp. 843–846, Sept. 1994.

[14] S. Suhm and A. Waibel, "Towards better language models for spontaneous speech," Proc. ICSLP, vol. 2, pp. 831–834, Sept. 1994.

[15] E. Giachin, "Phrase bigrams for continuous speech recognition," Proc. ICASSP, vol. 1, pp. 225–228, May 1995.

[16] K. Hwang, "Vocabulary optimization based on perplexity," Proc. ICASSP, vol. 2, pp. 1419–1422, Apr. 1997.

[17] H. Inaba, T. Kawahara, and S. Doshita, "Reconstruction of vocabulary for large vocabulary continuous speech recognition," Proc. JSAI, pp. 495–498, June 1998, (in Japanese).

[18] D. Klakow, "Language-model optimization by mapping of corpora," Proc. ICASSP, vol. 2, pp. 701–704, May 1998.

[19] H. Kuo and W. Reichl, "Phrase-based language models for speech recognition," Proc. EUROSPEECH, vol. 4, pp. 1595–1598, Sept. 1999.

[20] Z. Jun, "Lexicon optimization for Chinese language modeling," Proc. ISCSLP, Oct. 2000.

[21] T. Shinozaki and S. Furui, "Error analysis using decision trees in spontaneous presentation speech recognition," Proc. ASRU, Dec. 2001.

[22] T. Shinozaki and Sadaoki Furui, "Analysis on individual differences in automatic transcription of spontaneous presentations," Proc. ICASSP2002, Orlando, FL, vol.1, pp. 729–732, May 2002.

[23] H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," in Proc. ICASSP, Orlando, FL, vol.1, pp. 725–728, May 2002.

[24] N. Mirghafori, E. Fosler, and N. Morgan, "Towards robustness to fast speech in ASR," in Proc. ICASSP, Atlanta, GA, Vol.1, pp. 335–338, May 1996.

[25] T. Stephenson, M. Magimai-Doss, and H. Bourlard, "Speech recognition of spontaneous, noisy speech using auxiliary information in Bayesian networks," in Proc. ICASSP, Hong-Kong, May 2003.

[26] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld, "Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode," in Proc. ICSLP, Philadelphia, Vol. supplement, PA, Oct. 1996.

[27] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in Proc. Eurospeech, Rhodes, Greece, vol.5, pp. 2379–2382, Sept. 1997.

[28] A. Tuerk and S. J. Young, "Indicator variable dependent output probability modelling via continuous posterior functions," in Proc. ICASSP, Salt Lake City, UT, vol.1, pp. 473–476, May, 2001.

[29] G. Zweig, "Speech recognition with dynamic Bayesian networks," Ph.D. thesis, University of California, Berkeley, 1998.

[30] G. Zweig and S. Russell, "Speech recognition with dynamic Bayesian networks," AAAI. pp. 173–180, 1998.

[31] G. Zweig and M. Padmanabhan, "Dependency modeling with Bayesian networks in a voicemail transcription system," in Proc. Eurospeech, Budapest, Hungary, pp. 1135–1138, Sept. 1999.

[32] J. Bilmes, "Graphical models and automatic speech recognition," Technical Report UWEETR-2001-005, University of Washington, Dept. of EE, Seattle, WA, Nov. 2001.

[33] K. Murphy, "A brief introduction to graphical models and Bayesian networks," http://www.ai.mit.edu/ murphyk/Bayes/bnintro.html [Online], 1998.

[34] T. Dean and K. Kanazawa, "Probabilistic temporal reasoning," AAAI. pp. 524–528, 1988.

[35] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden markov model," in Proc. ICASSP, Salt Lake City, UT, pp. 513–516, May 2001.

[36] T. Stephenson, J. Escofet, M. Magimai-Doss, and H. Bourlard, "Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables," in Proc. NNSP, Martigny, Switzerland, pp. 637–646, Sept. 2002.

[37] N. Morgan, E. Fosler, and N. Mirghafori, "Speech Recognition using on-line estimation of speaking rate," in Proc. Eurospeech, Rhodes, Greece, vol.4, pp. 2079–2082, Sept. 1997.

[38] S. Kitazawa, H. Ichikawa, S. Kobayashi, and Y. Nishinuma, "Extraction and representation of rhythmic components of spontaneous speech," in Proc. Eurospeech, Rhodes, Greece, vol.2, pp. 641–644, Sept. 1997.

[39] N. Morgan and E. Fosler, "Combining multiple estimators of speaking rate," in Proc. ICASSP, Seattle, WA, Vol.2, pp. 729–732, May 1998.

[40] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in Proc. ICASSP, Seattle, WA, Vol.2, pp. 945–948, May 1998.

[41] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in Proc. Human Language Technology Conference, San Diego, CA, pp. 246–252, Mar. 2001.

[42] J. Bilmes and G. Zweig, "The graphical models toolkit: an open source software system for speech and time-series processing," in Proc. ICASSP, Orlando, FL, vol.4, pp. 3916–3919, May 2002.

[43] Z. Zhang, S. Furui and K. Ohtsuki, "On-line incremental speaker adaptation with automatic speaker change detection," Proc. ICASSP, vol.2, pp.961–964, 2000.

[44] R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic Mixtures vs. Dynamic Cache Models," Proc. ICSLP, Vol.1, pp.236–239, 1996.

[45] J. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser Output Voting Error Reduction," Proc. ASRU, 347-352, 1997.

[46] S. Itoh, "Technical trends on Grid computing," Trans of IPSJ, Vol.44, No.6, pp. 576–580, 2003.

[47] R. Tummala and V. Madisetti, "System on Chip or System on Package?," IEEE Design & Test of Computers, Vol.16, Issue 2, pp 48–56, 1999.

[48] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, R. L.Mercer, "Class-based n-gram models of natural language," Computational Linguistics, vol.18, no.4, pp.467–479, 1992.

[49] T. Yokoyama, T. Shinozaki, K. Iwano, and S. Furui, "Unsupervised class-based language model adaptation for spontaneous speech recognition," Proc. ICASSP, vol.1, pp. 236–239, 2003.

[50] L. Lussier, E. Whittaker, and S. Furui, "Combinations of language model adaptation methods applied to spontaneous speech," Proc. SSST, pp. 73–78, 2004.

Publications

1. Takahiro Shinozaki and Sadaoki Furui, "Presentation Transcription Using a Japanese Spontaneous Speech Corpus," Trans of IPSJ, Vol.43, No.7, pp.2098-2107 (2002-7) (in Japanese).

2. Takahiro Shinozaki, Sadaoki Furui, "Dynamic Bayesian Network-based Acoustic Models Incorporating Speaking Rate Effects," Trans of the IEICE (Conditional Acceptance).

3. Takahiro Shinozaki, Chiori Hori and Sadaoki Furui, "Towards Automatic Transcription of Spontaneous Presentations," Proc. EUROSPEECH2001. 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, vol.1, pp.491-494 (2001-9).

4. Takahiro Shinozaki and Sadaoki Furui, "Error Analysis Using Decision Trees in Spontaneous Presentation Speech Recognition," Proc. ASRU2001. Automatic Speech Recognition and Understanding workshop, Madonna di Campiglio, Trento, Italy (2001-12).

5. Takahiro Shinozaki and Sadaoki Furui, "Analysis on Individual Differences in Automatic Transcription of Spontaneous Presentations," IEEE International Conference on Acoustics, Speech, and Signal Processing, Proc. ICASSP2002, Orlando, U.S.A., vol.1, pp.729-732 (2002-5).

6. Takahiro Shinozaki and Sadaoki Furui, "A New Lexicon Optimization Method for LVCSR based on Linguistic and Acoustic Characteristics of Words," International

Conference on Spoken Language Processing, Proc. ICSLP2002, Denver, U.S.A., vol.1, pp.717-720 (2002-9).

7. Takahiro Shinozaki and Sadaoki Furui, "An Assessment of Automatic Recognition Techniques for Spontaneous Speech in Comparison With Human Performance," ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, pp.95-98, (2003-4).

8. Takahiro Shinozaki and Sadaoki Furui, "Time Adjustable Mixture Weights for Speaking Rate Fluctuation," EUROSPEECH 2003, Vol.2, pp973-976 (2003-9).

9. Takahiro Shinozaki and Sadaoki Furui, "Hidden Mode HMM using Bayesian Network for Modeling Speaking Rate Fluctuation," ASRU 2003.

10. Sadaoki Furui, Kikuo Maekawa, Hitoshi Isahara, Takahiro Shinozaki and Takashi Ohdaira, "Toward the Realization of Spontaneous Speech Recognition —Introduction of a Japanese Priority Program and Preliminary Results," Proc. ICSLP2000. 6th International Conference on Spoken Language Processing, Beijing, Vol.3, pp.518-521 (2000-10).

11. Sadaoki Furui, Koji Iwano, Chiori Hori, Takahiro Shinozaki, Yohei Saito and Satoshi Tamura, "Ubiquitous Speech Processing," Proc. ICASSP 2001, IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, U.S.A., vol.1, pp.13-16 (2001-5).

12. Tadasuke Yokoyama, Takahiro Shinozaki, Koji Iwano, Sadaoki Furui, "Unsupervised class-based language model adaptation for spontaneous speech recognition," Proc ICASSP 2003, IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, (The conference was canceled due to SARS (Severe Acute Respiratory Syndrome))

13. Tatsuya Kawahara, Hiroaki Nanjo, Takahiro Shinozaki and Sadaoki Furui, "Benchmark Test for Speech Recognition Using the Corpus of Spontaneous Japanese,"

Proc. SSPR2003, Tokyo, Japan, pp.135-138 (2003-4).

14. Tadasuke Yokoyama, Takahiro Shinozaki, Koji Iwano and Sadaoki Furui, "Unsupervised Language Model Adaptation Using Word Classes for Spontaneous Speech Recognition," Proc. SSPR2003, Tokyo, Japan, pp.71-74 (2003-4).